PSYCHOMETRIKA—VOL. 89, NO. 1, 4–41
MARCH 2024
https://doi.org/10.1007/s11336-024-09965-6

EXAMINING DIFFERENTIAL ITEM FUNCTIONING FROM A MULTIDIMENSIONAL IRT PERSPECTIVE

TERRY A. ACKERMAN

THE UNIVERSITY OF IOWA

YE MA

AMAZON WEB SERVICES

Differential item functioning (DIF) is a standard analysis for every testing company. Research has demonstrated that DIF can result when test items measure different ability composites, and the groups being examined for DIF exhibit distinct underlying ability distributions on those composite abilities. In this article, we examine DIF from a two-dimensional multidimensional item response theory (MIRT) perspective. We begin by delving into the compensatory MIRT model, illustrating and how items and the composites they measure can be graphically represented. Additionally, we discuss how estimated item parameters can vary based on the underlying latent ability distributions of the examinees. Analytical research highlighting the consequences of ignoring dimensionally and applying unidimensional IRT models, where the two-dimensional latent space is mapped onto a unidimensional, is reviewed. Next, we investigate three different approaches to understanding DIF from a MIRT standpoint: 1. Analytically Uniform and Nonuniform DIF: When two groups of interest have different two-dimensional ability distributions, a unidimensional model is estimated. 2. Accounting for complete latent ability space: We emphasize the importance of considering the entire latent ability space when using DIF conditional approaches, which leads to the mitigation of DIF effects. 3. Scenario-Based DIF: Even when underlying two-dimensional distributions are identical for two groups, differing problem-solving approaches can still lead to DIF. Modern software programs facilitate routine DIF procedures for comparing response data from two identified groups of interest. The real challenge is to identify why DIF could occur with flagged items. Thus, as a closing challenge, we present four items (Appendix A) from a standardized test and invite readers to identify which group was favored by a DIF analysis.

Key words: multidimensional IRT, differential item functioning, compensatory and noncompensatory MIRT models.

1. Introduction

Dimensionality has long posed challenges for testing practitioners attempting to model test response data. Most tests inherently measure different composites of requisites skills outlined in their test specifications. It is important to understand that response data represent an interaction between examinees and the test items. While the resulting response data may appear unidimensional for one group of examinees, it could manifest as multidimensional for another group. For example, consider a math test that includes story problems that require both reading and math skills to answer correctly. If the test is written at a 4th-grade reading level and administered to fourth graders—some of whom may not read at the expected level—their responses may reflect deficits in either reading or math skills or both. However, when the same test is given to fifth graders who read at or above the fourth-grade level, the items should primarily differentiate based on math skills rather than reading abilities. Thus, due diligence demands that test practitioners

Correspondence should be made to Terry A. Ackerman, The University of Iowa, 8 North Shore Drive, Edwardsville, IL62025, USA. Email: tackerman@uiowa.edu

TABLE 1.
A compilation of seminal DIF methodology.

Method	Seminal research
Chi-square using IRT parameter estimates	Lord (1980), Cohen et al. (1993)
Hierarchical general linear modeling	Williams & Beretvas (2006)
Invariance alignment	Muthen & Asparouhov (2018)
IRT ICC area difference	Raju (1988), Raju et al. (1995), Flowers et al. (1999)
Likelihood ratio test	Thissen et al. (1988)
Logistic Regression	Clauser and Mazzor (1998), Swaminathan & Rogers (1990)
Mantel–Haenszel (raw score)	Holland et al. (1988)
Propensity Scoring	Liu et al. (2016)
Random DIF	De Boeck (2008)
Regularization techniques	Bauer et al. (2020), Huang (2018)
Residual-based DIF	Lim et al. (2022)
SEM using MIMIC modeling	Fleishman and Lawrence (2003)
SIBTEST (raw score)	Shealy & Stout (1993a,b)
Variance Estimators	Camilli & Penfield (1997), Penfield & Algina (2006)

thoroughly examine the dimensionality of the response data for individual subgroups as well as the entire test-taking population.

If the data are multidimensional, practitioners need to consider how the skills or subsequent scores may be misrepresented if the data are modeled as unidimensional. Fitting a two-dimensional model can help the practitioner understand substantively what composites are being measured and if the potential for differential item functioning (DIF) exists

DIF is a standard post-administration subgroup analysis conducted to ensure that test items do not favor one identifiable subgroup (e.g., males, females, whites, blacks, or Hispanics) when compared conditionally to another. The goal is to confirm test fairness. Over the years, many different approaches have been developed to detect DIF (Table 1). However, the challenge lies not merely in statistically identifying when items significantly favor one group over another, but rather in understanding the underlying reasons for why the DIF occurs.

Kok (1988), Ackerman (1992), Camilli (1992), and Shealy and Stout (1993a; 1993b have hypothesized that DIF can occur when items inadvertently measure invalid skills, and the two groups being examined have different ability distributions related to these invalid skills. These researchers described DIF from a two-dimensional perspective as

$$\varepsilon_{\theta_2}[P_{i,Ref}(u=1|\theta_1,\theta_2)|\theta_1] \neq \varepsilon_{\theta_2}[P_{i,Foc}(u=1|\theta_1,\theta_2)|\theta_1] \quad (1)$$

where

- $P_{i,Ref}$ and $P_{i,Foc}$ represent the probability of correct response for the Reference and Focal groups to item *i,*
- $\theta_1$represents the valid skill that is intended to be measured by the test publisher, and
- θ_2 represents an invalid or unintended-to-be-measured-skill (e.g., speededness, test-wiseness, reading ability on a test designed to measure mathematics ability) that affects the correctness of an examinee's response.

Even though the Reference and Focal group examinees have the same θ_1 -level of proficiency, DIF occurs because the θ_2-latent ability distributions for the two groups are different. Equation

(1) can only hold if

$$G_{Ref}\left(\theta_2|\theta_1\right) = G_{Foc}\left(\theta_2|\theta_1\right), \tag{2}$$

where G_{Ref} and G_{Foc} denote the conditional distribution of θ_2 given fixed values of θ_1. That is, for DIF to occur items must measure invalid skills **and** the two groups of interest being examined must differ in their ability distributions on the invalid skill. It must be a "perfect storm," both situations must occur. If no invalid skills are being measured, then ability differences on any invalid skill are moot and DIF should not occur. Likewise, if a test contains items that measure invalid skills, but the two groups of interest have identical underlying distributions on these invalid skills, no DIF should occur. Specifically, DIF manifests itself as a function of differences in underlying ability distributions.

While most DIF researchers simulate DIF by changing the parameters for one group versus another (and never stating why), our approach focuses on maintaining identical generating item parameters for each group., we manipulate the underlying ability distributions (i.e., G_{Ref} and G_{Foc}), resulting in distinct parameter estimates for each group. This perspective emphasizes the interaction between the skill composites being measured and the underlying ability distributions of the examinees. Consequently, it will be essential to focus initially on the two-dimensional IRT model and review relevant notation and characteristics that do not occur in the unidimensional IRT model.

In this address, we will build upon the five research studies cited above, and illustrate how the multidimensional nature of a test can be used to comprehend and explore the underlying mechanisms of DIF using multidimensional item response theory (MIRT). Our analyses assume that testing practitioners have already conducted dimensionality assessments of their test response data (e.g., using scree plots (Cattell, 1966) or specialized software such as DETECT (Zhang & Stout, 1999) or DIMTEST (Stout, 1987)) and in concert with test specifications, determined that their response data exhibit a two-dimensional structure. Subsequently, two-dimensional item response theory item parameters have been estimated.

The format for this article is as follows. First, a comprehensive review of the two-dimensional compensatory MIRT model is provided. This review includes an examination of how items can be graphically represented in a two-dimensional latent ability plane by detailing response surfaces, contour plots, item vector plots, and conditional centroid plots. Item vector plots provide insight into the range of composites being measured and assist testing practitioners in providing validity evidence regarding the test's intended-to-be-measured skills. Items that measure the intended skills identified in a test's specifications typically lie in an identifiable "validity sector."

Following this, the work of Wang (1985) and Camilli (1992) is explained, focusing on the analytical derivation of the reference composite (RC) resulting from a unidimensional calibration of two-dimensional data. Using the RC direction, a centipede plot is created to illustrate how examinees' latent abilities (θ_1, θ_2) are mapped onto the unidimensional IRT scale.

Next, Camilli's (1992) work is examined, concentrating on the analytical derivation of unidimensional two-parameter logistic (2PL) IRT item parameter estimates when the underlying generating model is the two-dimensional compensatory MIRT model. Example results using these derivations are demonstrated for a hypothetical 19-item test specifically designed to showcase how changes in an examinee group's underlying ability distribution affect the estimation of $\hat{a}$ and $\hat{b}$.

These explanations provide the terminology and graphical conceptualization as background for three studies that explore how dimensionality and disparate underlying latent ability distributions can influence DIF results. The first study adopts a strictly analytical approach, while the final two studies utilize two DIF statistics that condition on the number correct scores: Mantel –Haenszel (Holland et al., 1988) and Sibtest (Shealy & Stout, 1993a,b) . Simulated datasets are created for illustrative purposes to emphasize how DIF can occur. The article concludes with

a challenge based on DIF analyses conducted on standardized test data. Readers are encouraged to inspect four items and identify the group indicated as significantly favored.

2. The Compensatory Two-Dimensional IRT Model

Before exploring MIRT models, it is best to examine the 2PL IRT model, which is widely used in measurement and standardized testing. This model describes the probability of a correct response for an examinee j, with latent ability θ_j, responding to an item i, with difficulty and discrimination values denoted by the parameters b_i and $a_{i,}$ respectively. The model is written as

$$P\left(u_{ij}=1|\theta_j,a_i,b_i\right)=\frac{1.0}{1.0+e^{-1.7a_i(\theta_{jk}-b_i)}} \tag{3}$$

It should be noted that θ and b are on the same metric, (usually ranging from -3 to +3) and b equals the θ value for which the$p = .5$. Graphically the model is represented as an item characteristic curve (ICC), where the b is the θ value corresponding to the point of inflection and a corresponds to 2.35 times the slope of the ICC at this point.

McKinley and Reckase (1982) extended the unidimensional 2PL to the multidimensional case, M2PL, which can be written as

$$P\left(u_{ij}=1|\boldsymbol{\theta_j},\boldsymbol{a_i},d_i\right)=\frac{1.0}{1.0+e^{-1.7(\sum_{k=1}^{m}a_{ik}\theta_{jk}+d_i)}}$$

where $\mathbf{a}_\mathrm{i}$ is a vector of discrimination parameters, d_i is a scalar difficulty parameter for item i, and θ_j is a vector of ability parameters for person j. It is important to note that d_i is added in the logit, so unlike in the 2PL model, negative values represent difficult items. For each dimension, there is a discrimination parameter and a latent ability. However, regardless of the number of dimensions, there is only one difficulty parameter because in this model a difficulty parameter for each dimension is indeterminate, i.e., there is an unlimited number of a_i and d_i values that yield the same probability of correct response.

To find equivalent counterparts to the unidimensional discrimination and difficulty they made the following substitution: a point in the ability space is redefined as:

$$\theta_j=\zeta_j\cos\alpha_j,$$

where ζ_j is the distance from the origin to the point, and α_j is the angle created from the point to the j^{th} axis. Using this trigonometric substitution, the model can then be written as

$$P\left(u_{ij}=1|\zeta_j,\boldsymbol{a_i},\boldsymbol{\alpha_j},d_i\right)=\frac{1.0}{1.0+e^{-\left(\zeta_j\sum_{k=1}^{m}a_{ik}\cos\alpha_{jk}+d_i\right)}}.$$

To find the location of the steepest slope in the $\boldsymbol{\alpha}$-vector direction, it is necessary to compute the second derivative with respect to ζ_j and set it equal to zero. Like the unidimensional 2PL model, the maximum slope in $\boldsymbol{\alpha}_{\mathrm{j\,direction}}$ occurs when P_{ij} =.5. McKinley and Reckase (1982) defined the multidimensional discrimination analog to the unidimensional a-parameter for item i as

$$\mathrm{MDISC}\ =a_i=\sqrt{\sum_{k=1}^{m}a_{ik}^2}, \tag{4}$$

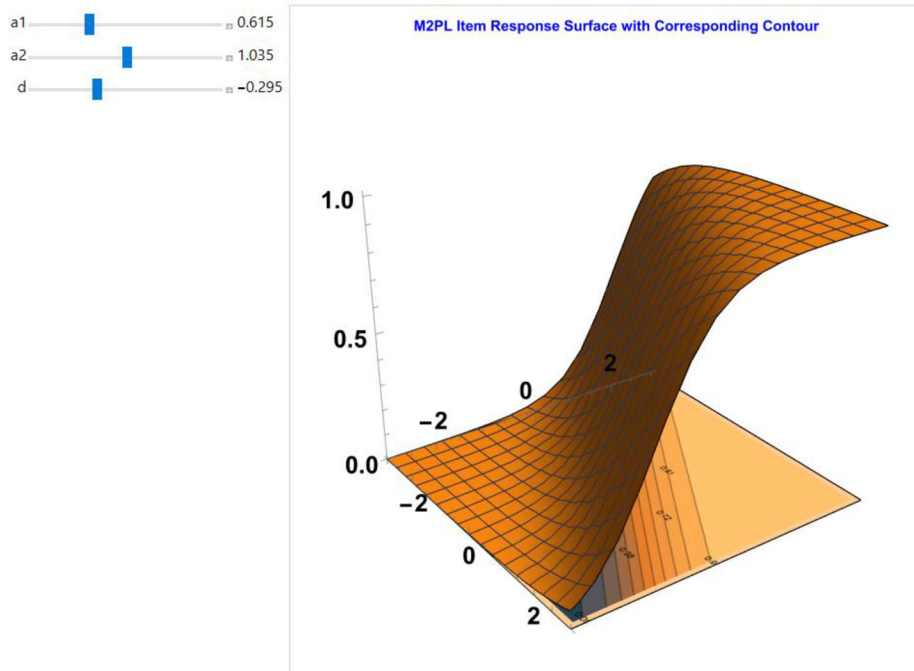

FIGURE 1.
Graphic representation of the response surface for the compensatory model and its corresponding contour.

and the M2PL parameter corresponding to the unidimensional difficulty parameter, b_i, can be written as:

$$\text{MDIFF} = b_i = \frac{-d_i}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}} = \frac{-d_i}{a_i}. \tag{5}$$

For this presentation, the focus will be on the two-dimensional case and the probability of correct response to item i is expressed as

$$P\left(u_{ij} = 1 | \theta_{1j}, \theta_{2j}, a_{1i}, a_{2i}, d_i\right) = \frac{1.0}{1.0 + e^{-1.7(a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i)}}. \tag{6}$$

Graphically, this function represents an item response surface. For an item i, we can inspect the surface plot or the contour plot to gain further insight. Using the software program, Mathematica (Wolfram, 2020) allows one to manipulate the item parameters and observe changes in the response surface. Such a Mathematica plot is shown in Fig. 1. The plot is configured to enable the user to change the item parameters using the parameter sliding bars on the left of the plot. The surface plot is also rotatable for viewing from different perspectives.

Some researchers refer to the M2PL as "partially compensatory," because the abilities are additive in the logit, allowing for compensation (i.e., being high on one ability can "compensate" for being low on the second ability) can occur. As shown in Fig. 2, the equiprobability contour plot for an item with M2PL parameters $a_1 = a_2 = 1.0$ and $d = .0$, two examinees, A and B, having exact opposite ability profiles such as high on θ_1 and low on θ_2 versus low on θ_1 and high on θ_2

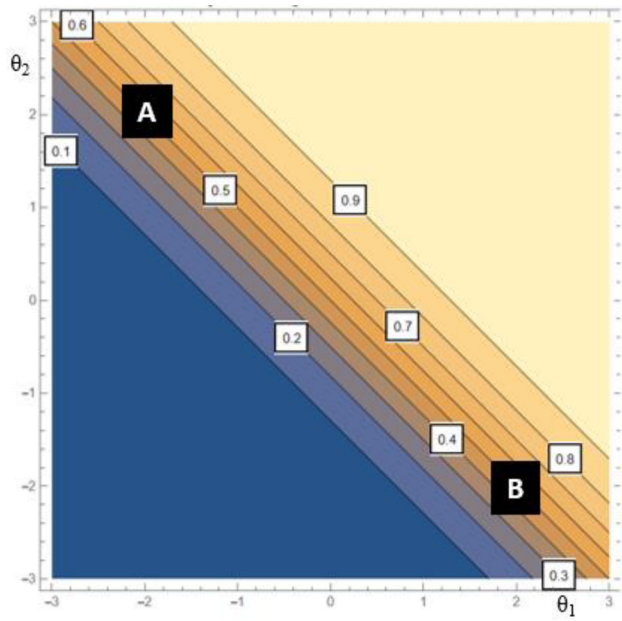

FIGURE 2.
Contour plot of compensatory model item with $a_1 = a_2 = 1.0$ and d=.0.

can have the same probability of correct response. This occurs because of the additive nature of the logit of the M2PL model. When $a_1 = a_2$, $p(\theta_{1i}, \theta_{2i}) = p(\theta_{1j}, \theta_{2j})$ for all examinees i and j where the sum $\theta_1 + \theta_2$ equals the same value (e.g., $p(\theta_1 = 2, \theta_2 = -2) = p(\theta_1 = 0, \theta_2 = 0) = p(\theta_1 = -2, \theta_2 = 2)$. Note, that the slopes of the contours are all equal to $-a_1/a_2$, and all (θ_1, θ_2) combinations such that $\theta_2 = (-a_1/a_2)(\theta_1)$ will have the same probability of correct response or lie on the same equiprobability contour. Furthermore, when $a_1 = 0$ or $a_2 = 0$ there is no compensation and the model is equivalent to the 2PL unidimensional model. That is, when $a_2 = 0$, $p(\theta_1, \theta_2) = p(\theta_1)$ and when $a_1 = 0$, $p(\theta_1, \theta_2) = p(\theta_2)$.

2.1. Item Vector Representation

The drawback that occurs in the representation of M2PL two-dimensional items is that only one surface or contour can be examined at a time. This problem can be solved by representing items in the two-dimensional latent ability plane as a vector. This is accomplished using the following guidelines (Reckase, 2009) :

- All vectors lie on lines that pass through the origin.
- Vectors can lie only in the first and third quadrants because the a-parameters are constrained to be positive.
- Vectors representing easy items lie in the third quadrant; those representing difficult items lie in the first quadrant. (Note that if a_1 is negative the vector will lie in the second quadrant and if a_2 is negative the vector will lie in the fourth quadrant.)
- The tail of the vector lies on the $p = .5$ equiprobability contour and the vector is always orthogonal to this contour.

Using the derived information for multidimensional discrimination and difficulty, these vectors are created where the length of the vector indicates how discriminating the item is equal to MDISC (4). The tail of the vector lies on the $p = .5$ equiprobability contour. The signed distance

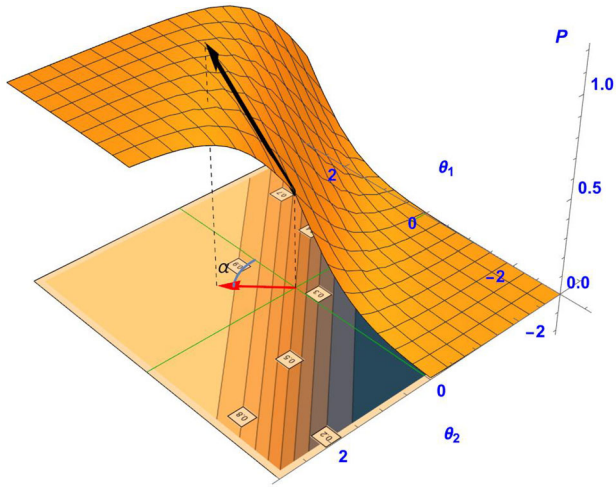

FIGURE 3.
Illustration of the direction of maximum slope for a compensatory item projected onto the latent ability plane to form its item vector.

from the origin perpendicular to this contour is the unidimensional analog of difficulty, MDIFF (5). The angular direction from the θ_1-axis indicates the (θ_1, θ_2)-composite of ability that the item i is best measuring:

$$\alpha_i = \cos^{-1}\left(\frac{a_{1i}}{\text{MDISC}}\right).$$

As illustrated in Fig. 3, vectors are projections, indicating the composite direction of maximum discrimination or maximum slope onto the latent ability plane. As is shown, the corresponding contour with the response surface over the third quadrant is removed so that the projected item vector is illustrated in relationship to the underlying contour surface. The greater the discrimination of the item, the steeper the response surface, causing the corresponding contours to become closer together and the greater the length of the vector. Vectors of easy items appear in the third quadrant and vectors representing difficult items are in the first quadrant.

Angle item vectors differ by only a few degrees these nuances cannot be attributed solely to phrasing or vocabulary. Item writers and psychometricians need to examine the content sectors that contain different item contents or test specifications. These sectors help determine which (θ_1, θ_2)-composites an item measures best (Ackerman, 1991) . Ultimately, these content sectors will help in understanding or defining the θ_1 and θ_2 latent abilities and defining an imposed unidimensional score scale.

Item vectors are often color-coded based on their content classification. Ideally, vectors with the same content should cluster in a narrow *content sector*, indicating that they are measuring similar (θ_1, θ_2)-composites. For example, consider a standardized graduate admissions test with 101 items. In Fig. 4, observe how different content areas occupy unique sectors. The one vector in quadrant two had a negative a_2 value.

2.2. *The Validity Sector*

Ackerman (1992) defined the ***validity sector*** as the sector containing vectors of items measuring practitioner-determined valid composites. Unlike the 101-item test shown above, most

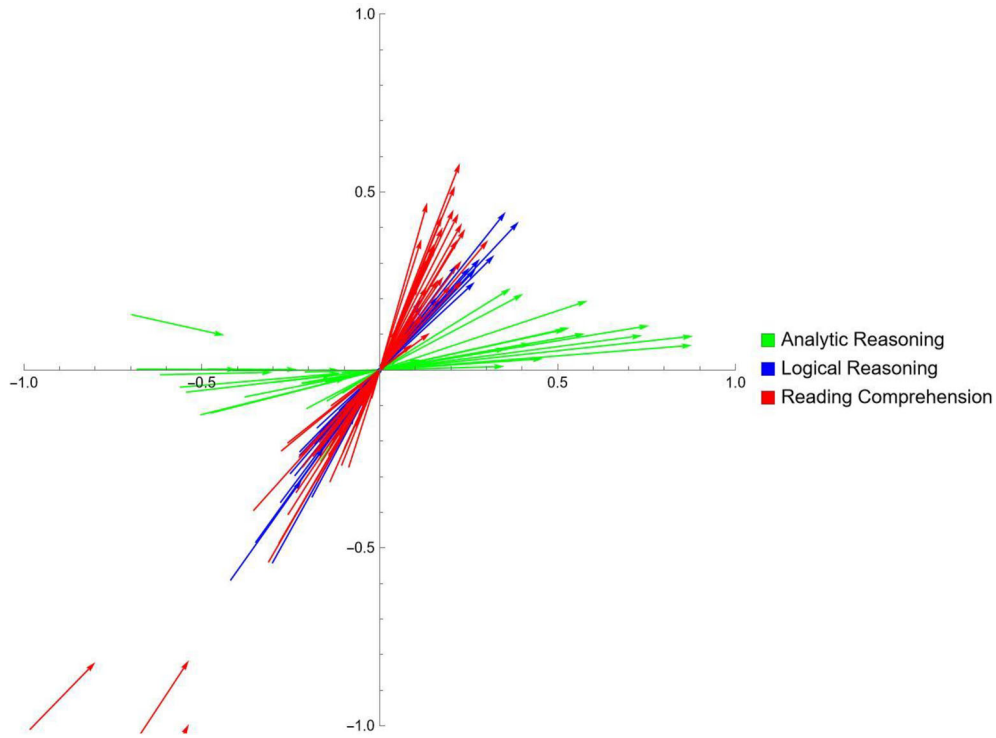

FIGURE 4.
Item vectors for a standardized 101-item test with three content areas.

standardized tests yield vectors that can be enclosed by a 30°–45° degree sector as illustrated in Fig. 5. In this figure, the green item vectors (and red dotted RC) are enclosed in a 45° validity sector and are believed to be vectors of items measuring the valid (θ_1, θ_2)-composites as described in the test's specifications. Items whose vectors (red) fall outside the validity sector should be examined for DIF because they are likely measuring invalid or nuisance dimensions. DIF has the potential to occur if the groups of interest differ in their distributions of underlying abilities on these invalid skill composites. Ma et al. (2023) demonstrated that differences in multidimensional latent ability distribution on the invalid dimension can result in DIF especially when items measure primarily the invalid dimension. The insight gained from item vectors and the validity sector can guide psychometricians in refining assessments, ensuring validity, and understanding the intricate interplay of latent abilities.

2.3. Score Scale Consistency Using Conditional Centroids

Another way to understand how DIF can occur from a two-dimensional perspective is to examine whether an imposed unidimensional score scale consistently represents the same (θ_1, θ_2)-composite as one progresses across the observable score scale. When different parts of the unidimensional score scale represent different skill composites, score scale consistency breaks down.

For example, consider the 101-item test represents a dimensional scenario in which there is a lack of score scale consistency. That is, the same (θ_1, θ_2)-composite is not being measured equally well throughout the observable score range. To examine scale consistency, "centroid" plots can be created to show $(\bar{\theta}_1, \bar{\theta}_2)$ for each number correct observed score, x, (i.e., $(\bar{\theta}_1, \bar{\theta}_2)|X = x$). This

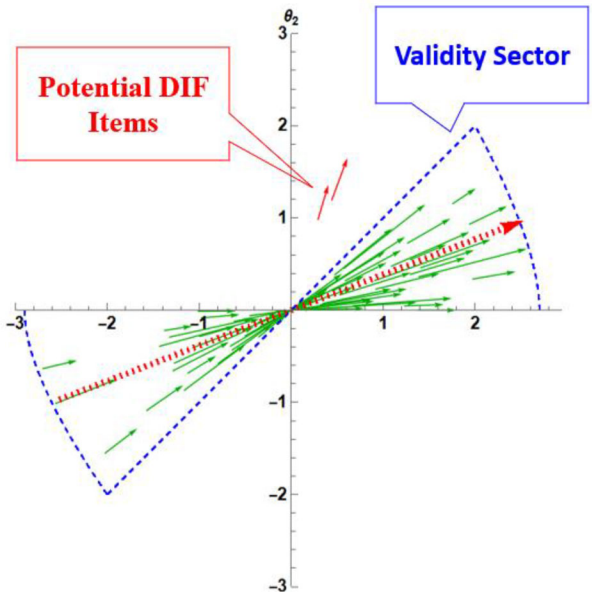

FIGURE 5.
A validity sector enclosing item vectors (green) for a 60-item standardized test. Vectors outside the sector (red) are measuring composites that could result in DIF. (Color figure online)

plot should be linear across the latent ability plane, indicating consistent measurement of (θ_1, θ_2) across the scale. In the top illustration of Fig. 6, centroid plots for each content category are graphed across the score range. The Analytic Reasoning tends to be linear representing primarily differences in θ_1. Logical Reasoning and Reading Comprehension represent differences in θ_2, but not consistently. However, the results of this test are reported as a single score. The lower figure shows the centroid plot for the total test score scale. Scores in the range from 0 to 35 represent differences in the θ_1-ability. Scores in the range from 35 to 80 indicate differences in the θ_2-ability. Scores from 85 to 100 represent proficiency differences in the upper θ_1-ability. Work by Strachan et al. (2022) found that if there is a confounding of difficulty and dimensionality (e.g., easy items measure one dimension and difficult items measure a second dimension) the composite may not be linear.

Such inconsistency makes score interpretation and certain psychometric procedures such as equating and computer adaptive testing pool development incredibly challenging. This variation also affects DIF procedures that group examinees according to their number correct scores for conditional analyses because different scores reflect different skills.

2.4. Reference Composite: Mapping a 2PL Scale in a Ttwo-Dimensional Latent Space

Wang (1985) demonstrated that when calibrating multidimensional data, the estimated unidimensional 2PL model essentially combines latent abilities into a weighted composite known as the *reference composite (RC)*. The *RC* is key because it indicates the (θ_1, θ_2)-composite being best measured by the unidimensional IRT θ-scale and the number correct score scale. It is important to note that for, say a two-dimensional two-item test, the RC (θ_1, θ_2)-composite does not directly correspond with the composite skills being measured by either of the two items. This has important implications for how to substantiate or define the unidimensional scale for a test containing two-dimensional items. The *RC* is a useful tool for demonstrating the parallelism of multiple two-dimensional test forms. That is, after calibration and rescaling, the *RC*s of parallel forms should

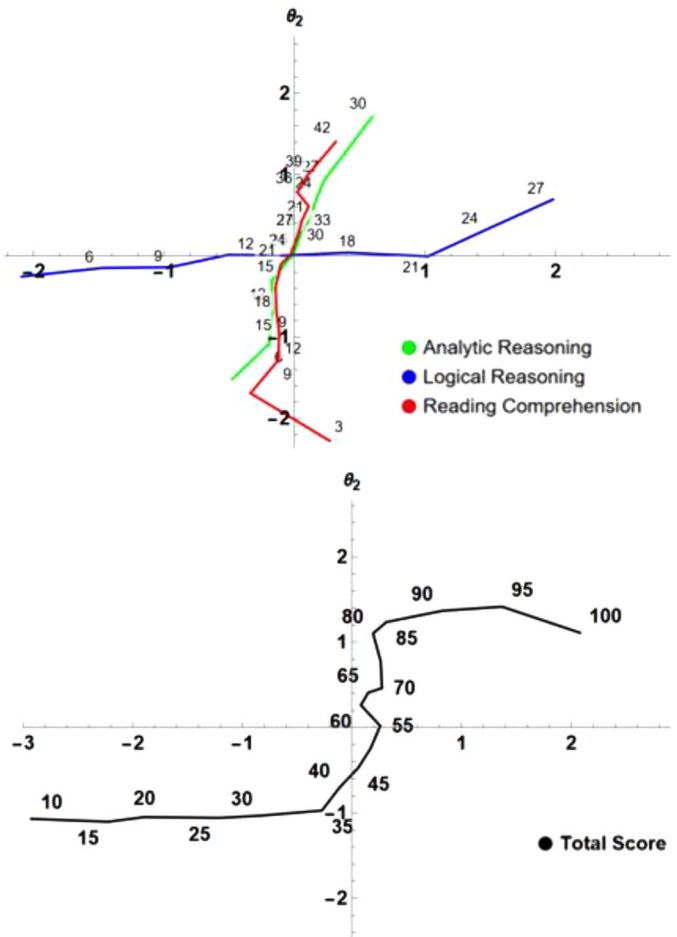

FIGURE 6.
Conditional centroid plots for each content category (top) and total test score (bottom).

align closely within measurement error and other constraints based on the test specification (e.g., cut score measurement precision, content constraints).

For the two-dimensional case, Wang (1985) determined that the *RC* is a function of the L'A'AL matrix, where **A** is the n x 2 matrix of discrimination parameters for a given n-item test and **L** is the Cholesky decomposition of the underlying $\theta_1 - \theta_2$ variance–covariance matrix, Ω. The angle between the positive θ_1-axis and the *RC* can be calculated as the arccosine of the first element of the eigenvector associated with the larger of the two eigenvalues of the **L'A'AL** matrix.

A simple example will help to clarify. Assume a two-item case where $a_1 = 1.3$, $a_2 = .4$ and $d = -1.2$ for Item 1 and $a_1 = .4$, $a_2 = 1.3$ and $d = -1.2$ are the parameters for Item 2. The item vectors for Item 1 and Item 2 are, respectively 17.10° and 72.90° from the positive θ_1-axis, respectively. Assume further a group of examinees, Group A, whose two-dimensional underlying ability is $N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1.0 & .0\\ .0 & .5\end{pmatrix}\right]$ and Group B's underlying distribution is $N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}.5 & .0\\ .0 & 1.0\end{pmatrix}\right]$. The Cholesky decomposition of Ω_A is $\begin{bmatrix}1.0 & .0\\ .0 & .7071\end{bmatrix}$ for both groups and the **L'A'AL** matrix is

equal to $\begin{bmatrix} 1.8500 & .7353 \\ .7353 & .9250 \end{bmatrix}$. This matrix is negative definite. The eigenvalues of this matrix are 2.2562 and.5187 and the eigenvector that corresponds to the larger eigenvalue is $\begin{bmatrix} -.8753 \\ -.4835 \end{bmatrix}$. The squared elements sum to 1.0, so this eigenvector can be considered as direction cosines. The angle associated with the *RC* can be computed by taking the arccosine of the absolute value of the elements since the **L'A'AL** matrix is negative definite. These calculations indicate that the RC for the Reference group lies 28.91° from the positive θ_1-axis. Similarly, the RC for the Focal group lies 61.09° from the positive θ_1-axis. Notice the angular difference between the *RC*s is 32.18°, which is the angular between the red dashed *RC* and green dashed *RC* figure in Fig. 7. The underlying joint and marginal distributions, *RC*s, and item vectors are also displayed in the figure.

Trying to compare scores from different *RC*s can be problematic because it could result in examinees being ordered differently on the two *RC*s. Ramsay (1990) and Junker and Stout (1991) examined the effects of differential ordering in a DIF context. Even though DIF procedures that condition upon the number correct score may not be sensitive to the dissimilar substantive interpretation of the two *RC*s, they are sensitive to differential ordering. This situation is graphically illustrated on the right in Fig. 7. In this diagram, two examinees, X and Y, would have one ordering if they are in the Focal group and orthogonally mapped onto the RC_{FOC}, but a reverse ordering if they belong to the Reference group and are mapped onto RC_{REF}.

In an attempt to overcome scaling problems, DIF methodologies use different approaches. DIF analyses that use IRT methodology (Raju 1988) require that before comparing item characteristic curves for Reference and Focal groups, the item parameters for one group must be rescaled and placed on the other group's scale. This can be accomplished using either a mean–mean or mean–sigma rescaling (Kolen & Brennan, 2014) . It should be noted that these linking/rescaling procedures apply optimally when the reference composites have identical composite directions. They are designed to account for mean and standard deviation scale differences that are established in calibrating response data to fit a unidimensional model.

DIF Methodologies are not designed to compensate for scale differences that would occur when scales represent different ability composites. The greater the angular separation between the *RCs,* the less effective rescaling becomes. For example, consider an extremely unrealistic case. If the *RC* for the Reference group is 10°, the *RC* for the Focal group is 70°, placing the Focal group parameters on the Reference group's scale would adjust primarily θ_1 differences, but the Focal group's scale would primarily measure θ_2. This problem is discussed further in Study 1 below. Several researchers, including Li and Lissitz (2000) and Oshima et al. (2000), examined linking from a multidimensional perspective which would align the RC's for two distinct groups, before adjusting for scale differences.

It should be further noted that the Mantel–Haenszel accomplishes rescaling by including the studied item in the calculation of the total score. Sibtest calculations utilize a regression correction. It is important to recognize that these rescaling techniques work optimally when the *RC's* angular directions are similar. Because the *RC* direction can vary for different content subsets of items, Shealy and Stout (1993a,b) recommend that practitioners identify valid test items to ensure the conditioning score (i.e., *RC*) represents a valid composite. As the number of items on a test increases, (say > 40), the influence of any one item decreases. Usually, the RC lies within the validity sector.

Additional research by Ackerman and Xie (2019) compared Camilli's approach two other two unidimensional approaches to explore how well they capture the representation of two-dimensional latent ability space. Carlson (2017) and Strachan et al. (2020) conducted research in which the RC is nonlinear when there is a confounding of difficulty and dimensionality. Addition-

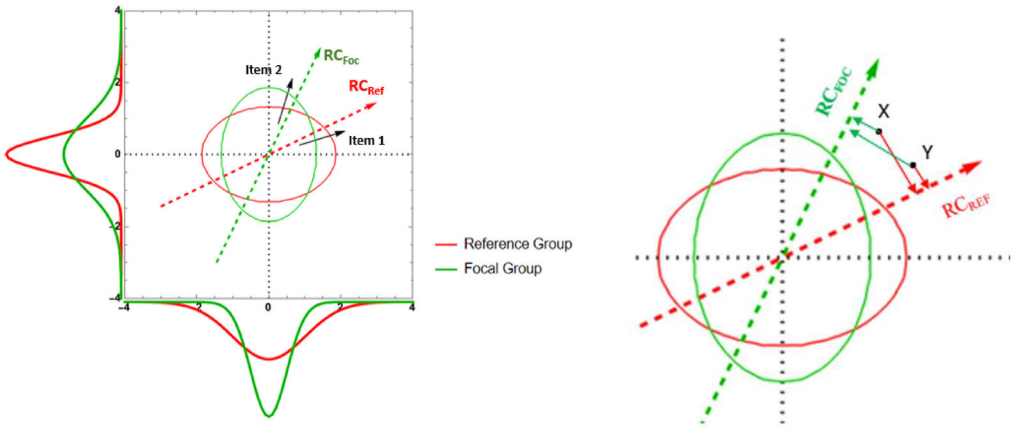

FIGURE 7.
RCs for two groups having different underlying ability distributions based on a two-item test (left) and orthogonal mappings upon the composites for two examinees, X and Y, (right).

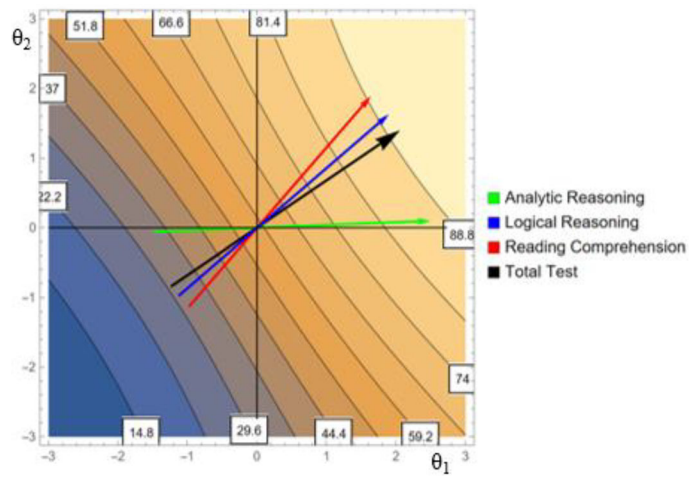

FIGURE 8.
RCs for the 101-item test for the three subsections and the total test.

ally, Ma et al. (2023) evaluated the efficiency of DIF detection using both Camilli's approach and the projective IRT approach (Ip, 2010) when potential DIF items fell outside the validity sector.

The contour plot of the standardized 101-item test is displayed in Fig. 8. Within this plot are the *RC*s for the three individual content areas, as well as the total test. The *RC*s align with the direction of the sector containing content vectors. Interestingly, the wide angular range of the content *RC*s results in curved contours for higher score categories, resembling patterns seen in the noncompensatory model (10).

2.5. *Centipede Plot: Graphically Mapping the Two-Dimensional IRT Latent Ability Space Onto the Unidimensional IRT Scale*

This *RC* enables one to create an interesting visualization of how the two-dimensional latent ability plane gets mapped onto the unidimensional ability scale. This mapping can be illustrated by a "centipede" plot. In this type of plot, the compensatory model (6) test characteristic curve

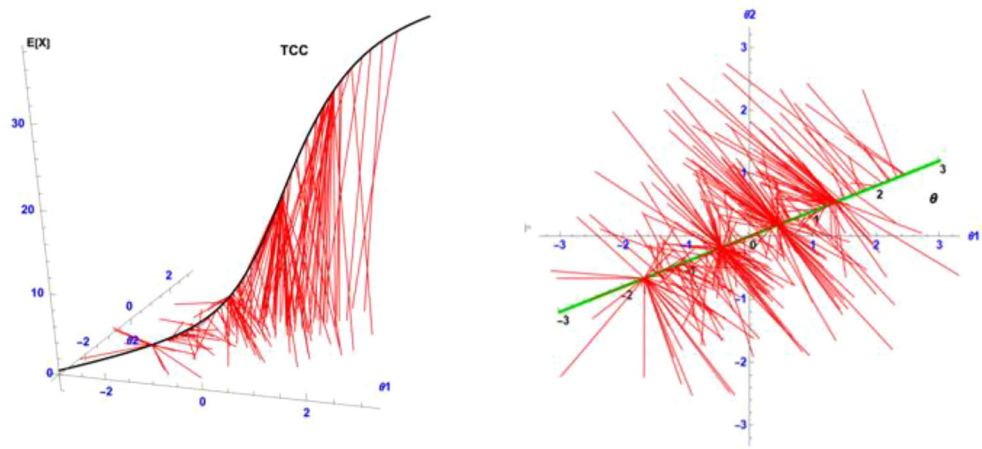

FIGURE 9.
Two perspectives illustrating the mapping of the two-dimensional latent abilities onto the expected number correct score scale.

is first drawn in the *RC*-direction. Then, vectors are drawn from the generated (θ_1, θ_2) to their expected score, obtained using the estimated $\hat{\theta}$. Figure 9 displays two perspectives of a centipede plot for a 40-item test: one from a side view and another from an overhead view. The vertical axis represents the proportion correct true score. Vectors are displayed for a sample of 200 examinees. It is informative to observe which (θ_1, θ_2)-combinations map onto the same proportion correct true scores. This information helps psychometricians and test developers explore how regions of the latent ability space with opposite (θ_1, θ_2)-profiles, (high, low) vs (low, high), get mapped onto the same conditioning number correct score and thus will be in the same 2x2 contingency table often used in DIF approaches such as the Mantel–Haenszel.

3. Analytically Estimating 2PL Item Parameters from a Two-Dimensional Latent Space

Although most tests are multidimensional, practitioners often fit unidimensional IRT models to the response data. Camilli (1992) analytically determined how the unidimensional 2PL model can be extracted from data where the true model is a two-dimensional model (6). The estimated unidimensional IRT model can be expressed in terms of the two underlying factor scores as,

$$\varepsilon_{\upsilon_2}[P(u=1\,|\upsilon_1,\upsilon_2)|\,\upsilon_1] = \int_{-\infty}^{+\infty} P\,(u=1|\upsilon_1,\upsilon_2)G(\upsilon_2|\upsilon_1)\,\mathrm{d}\upsilon_2, \tag{7}$$

where ε_{υ_2} is the expected value of the unidimensional item response function anchored over the factor score, υ_1, the RC; υ_2 is the second factor score which is orthogonal to υ_1, the first factor score, and G($\upsilon_2|\upsilon_1$) is the underlying conditional distribution (Camilli, 1992, p.133). Using this formulation, Camilli derived the formulas to calculate the unidimensional $\hat{a}$ and $\hat{b}$ as

$$\hat{a}_j = \frac{\boldsymbol{a}_j^{'}\boldsymbol{W}_1}{\sqrt{2.89 + \boldsymbol{a}_j^{'}\boldsymbol{W}_2\boldsymbol{W}_2^{'}\boldsymbol{a}_j}} \tag{8}$$

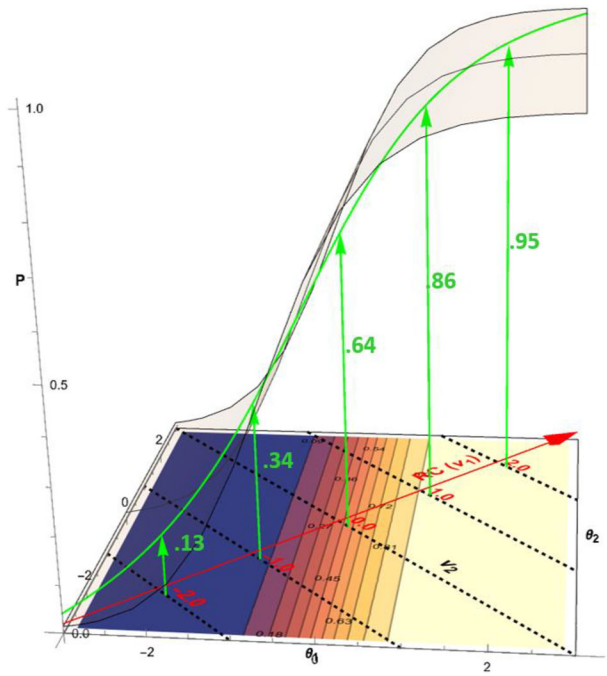

FIGURE 10.
Projected unidimensional ICC for a M2PL item with a1 = 1.5, a2 = 0, and d =.5, and a reference composite (RC) angle of 45° yielding 2PL $\hat{a} = .73$ and $\hat{b} = -.47$.

and

$$\hat{b}_j = \frac{d_j - a_j' \boldsymbol{\mu}}{\boldsymbol{a}_j \boldsymbol{W}_1}. \tag{9}$$

where $\boldsymbol{a}_j$ is the two-dimensional discrimination vector for the M2PL model (4), d_j is the difficulty parameter for the M2PL model, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the first and second standardized eigenvalues of the matrix $\boldsymbol{L}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{L}$, where $\boldsymbol{A}$ is the matrix of discrimination parameters for all the items on the test and $\boldsymbol{L}'\boldsymbol{L}$ **is** the Cholesky decomposition of the two-dimensional latent ability variance–covariance matrix $\boldsymbol{\Omega}$. A visual representation using (7), (4), and (9) is illustrated in Fig. 10. For all items, the unidimensional ICCs would lie in the RC plane.

A more detailed graphical example for a projected ICC for M2PL item with $a_1 = 1.5, a_2 = 0$, and $d = .0$, and a *RC*- angle of 45° and an underlying a bivariate normal with a mean vector of {0,0}and the covariance matrix, Ω, as [{1,.4}, {.4,1}] yielding $\hat{a} = .73$ and $\hat{b} = -.47$ is illustrated in Fig. 10. In this figure, the components of Camilli's formulation (7) are illustrated including the translucent M2PL response surface, the contour of this surface, the *RC* (ν_1), the second principal component, ν_2, and the estimated unidimensional 2PL ICC with calculated p values (see values in green) for $\upsilon_1 = \{-2, -1, 0, 1, 2\}$. The complete calculations to derive these values are provided in Appendix B.

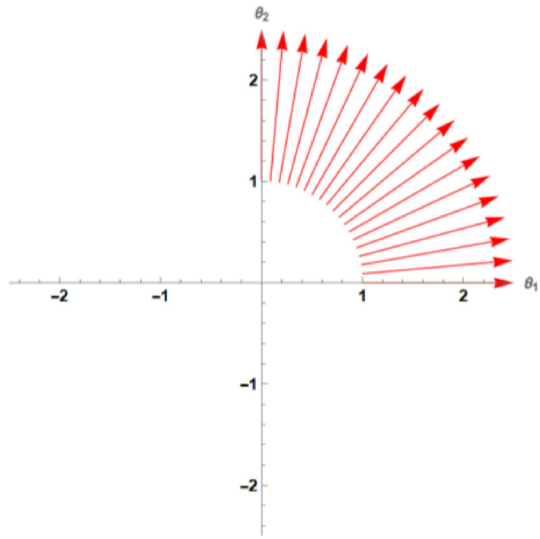

FIGURE 11.
Item vectors for a hypothetical 19-item test.

3.1. *Illustration of Changes in a Two-Dimensional Latent Ability Distribution can Affect $\hat{a}$ and $\hat{b}$ Values Using Camilli's Formulation*

Using (6) and (7), one can observe how changes in the underlying two-dimensional distribution of an examinee population impact the direction of the *RC* and consequently affect estimated a-parameters. As an item's vector angle approaches the angle of the *RC*, the estimated unidimensional $\hat{a}$-parameter increases. Conversely, when an item's vector angle deviates from the RC's angular direction, the estimated $\hat{a}$ becomes smaller. For illustrative purposes, consider a generated test of 19 items. The vectors of these items span angles in 5° increments from 0° to 90°. Assume all items have an MDISC value of 1.5 and an MDIFF value of −1.0. (Note in (6) the d-parameter is added in the logit, thus these would be considered to all be difficult items.) The vectors for this test are shown in Fig. 11. These parameters were selected for illustration purposes only and would never reflect an actual test. In most cases, real tests have more items and item vectors typically lie within a relatively narrow validity sector (e.g., 30°-45°). Only in extreme distributional cases would the *RC* be pulled out of the validity sector.

Using (6), $\hat{a}$ values were calculated for all items under four different two-dimensional latent ability distributional conditions: $\mathrm{N}\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & .0\\ .0 & 1\end{pmatrix}\right]$, $\mathrm{N}\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & .5\\ .5 & 1\end{pmatrix}\right]$, $\mathrm{N}\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & .0\\ .0 & .5\end{pmatrix}\right]$, $\mathrm{N}\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}.5 & .0\\ .0 & 1\end{pmatrix}\right]$. The *RC*-angles for these cases are 45°, 52.15°, 60.22°; and 22.78°, respectively. As the correlation increased, the *RC* began to shift in the 45° direction. When the variances of the two groups are unequal, the *RC* shifts toward the axis of the ability with the greater variance. The results are displayed in Fig. 12 (top). For each of the four distributional conditions, the $\hat{a}$'s tend to approach the MDISC value (4) as an item's angular direction aligns more closely with that of the *RC*. These trends are depicted in the plot by the respective colored vertical lines. The range of $\hat{a}$ was (.52 (0°) to.66(45°)), (.44(0°) to.67(60°)), (.34(0°) to.72(75°)), and (.34(90°) to.72(15°)) for the four respective conditions.

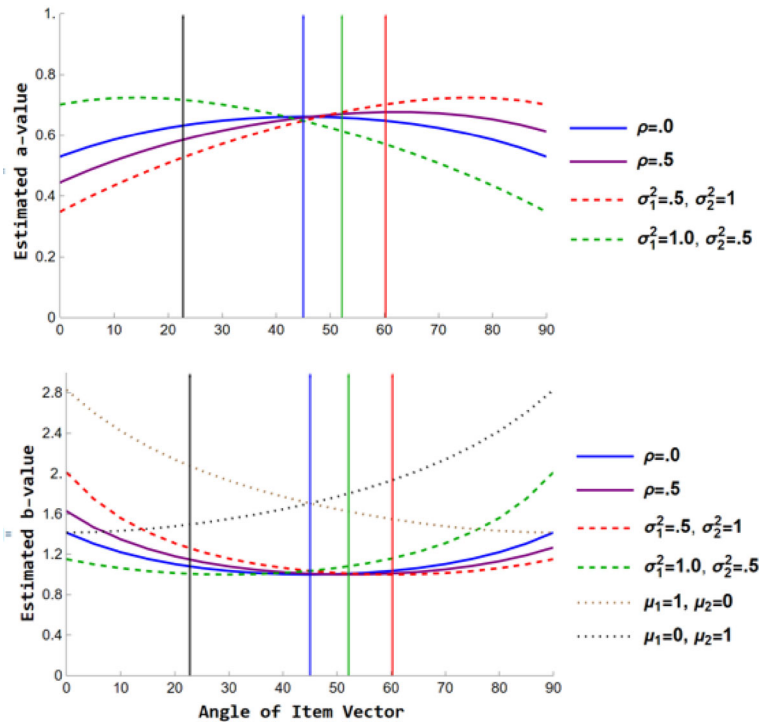

FIGURE 12.
Estimated a-(top) and b values (bottom) for different underlying ability distributions.

Additionally, two more conditions were considered in the calculation of $\hat{b}$: $N\left[\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}1 & .0\\ .0 & 1\end{pmatrix}\right]$ and $N\left[\begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}1 & .0\\ .0 & 1\end{pmatrix}\right]$. Although mean differences do not affect the RC or $\hat{a}$, they do affect $\hat{b}$, along with changes in variances and correlations. Figure 12 (bottom) shows how the estimated difficulties change as the items' measurement angles vary in reference to the RC (denoted by the colored vertical lines). The range of $\hat{b}$ was 1.41(0°) to.99(45°)), 1.62 (0°) to 1.00(50°), 2.01(0°) to 1.00 (60°), −2.01(90°) to −1.00 (30°), −2.82(0°) to −1.41(90°) and −2.82(90°) to −1.41(0°) for the six respective conditions. As an item's angular direction approaches the angle of the *RC*, the closer the $\hat{b}$ value will be to the item's MDIFF value (5).

Figure 13 presents a composite graph containing the 19 M2PL response surfaces and the *RC* plane (depicted in green, top left panel). The *RC plane* intersects the latent probability space at a 45° angle from the θ_1 axis. Additionally, the 19 ICCs are graphed in the *RC plane* (right panel). The top right panel shows the test characteristic surface and the corresponding contours, marked by the reference composite (indicated by the red arrow). Drawn in blue is the test characteristic curve (sum of the estimated 19 unidimensional ICCs using (4) and (9). This illustrates how closely the analytical estimates of *a* and *b* align with the two-dimensional model. Notably, the unidimensional TCC is not as steep as the TCS, indicating that the discrimination parameters may be underestimated. Wang (1986) previously compared 2PL estimates derived from generated and real data with estimates using the two-dimensional compensatory model. However, more research needs to be done in this area.

At the bottom of Fig. 13 are the 19 ICCs. The red ICCs represent vectors furthest from the 45° and have the lowest $\hat{a}$ values or flattest ICCs. The green ICCs correspond to item vectors with angles closest to the *RC* angle that have the largest $\hat{a}$ values or steepest ICCs.

In the next sections, three studies are examined. Each study provides a different perspective on how DIF can occur when response data are two-dimensional. The goal is to provide further insights for testing practitioners enabling them to conduct more informed DIF analyses and better understand the underlying causes when DIF occurs.

4. Study 1: DIF Illustrated Analytically Using Unidimensional IRT Item Calibration of a Two-Dimensional Latent Space

For this example, two pieces of research are foundational. The first builds upon the work of Shealy and Stout (1993a,b) and Ackerman (1992). This research identifies one of the causes of DIF. Imagine a scenario where valid items measure the intended to-be-measured ability, θ_1, while the test also contains items that inadvertently measure invalid skills, denoted as θ_2. By adopting a multidimensional perspective, we can estimate the potential for DIF by calculating the difference between the conditional expectations of the Reference and Focal groups, $E\left[\theta_{2R}|\theta_1\right] - E\left[\theta_{2F}|\theta_1\right]$. Assuming the regression of θ_2 on θ_1 is linear and homoscedastic, we can express the expected conditional difference, ECD, as follows:

$$\begin{aligned} \text{ECD} &= E\left[\theta_{2R}|\theta_1\right] - E\left[\theta_{2F}|\theta_1\right] \\ &= \left(\mu_{\theta_2 R} - \mu_{\theta_2 F}\right) + \left(\rho_R \frac{\sigma_{\theta_2 R}}{\sigma_{\theta_1 R}}\right)\left(\theta_1 - \mu_{\theta_1 R}\right) - \left(\rho_F \frac{\sigma_{\theta_2 F}}{\sigma_{\theta_1 F}}\right)\left(\theta_1 - \mu_{\theta_1 F}\right), \end{aligned} \tag{10}$$

where for the Focal group, $\theta_{1F,}$ and θ_{2F} are the two latent variables that are bivariate normally distributed with mean vector components $\mu_{\theta_1 F}$ and $\mu_{\theta_2 F}$ and variance–covariance is given as

$$\begin{pmatrix} \theta_{1F} \\ \theta_{2F} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{\theta_1 F} \\ \mu_{\theta_2 F} \end{pmatrix}, \begin{pmatrix} \sigma_{\theta_1 F} & \rho_F \sigma_{\theta_1 F} \sigma_{\theta_2 F} \\ \rho_F \sigma_{\theta_1 F} \sigma_{\theta_2 F} & \sigma_{2F} \end{pmatrix}\right],$$

and for the Reference group, θ_{1R} and θ_{2R} are the two latent variables that are bivariate normally distributed with mean vector components $\mu_{\theta_1 R}$ and $\mu_{\theta_2 R}$ and variance–covariance is given as

$$\begin{pmatrix} \theta_{1R} \\ \theta_{2R} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_{\theta_1 R} \\ \mu_{\theta_2 R} \end{pmatrix}, \begin{pmatrix} \sigma_{\theta_1 R} & \rho_F \sigma_{\theta_1 R} \sigma_{\theta_2 R} \\ \rho_R \sigma_{\theta_1 R} \sigma_{\theta_2 R} & \sigma_{2R} \end{pmatrix}\right].$$

Equation 10 serves as a "Rosetta stone" for understanding how the potential for DIF could occur. If we have estimates of the characteristics of the underlying ability for the two groups of interest, we gain insight into how different parameters of the Reference and Focal distributions contribute to conditional differences with the potential for DIF. The conditional difference serves as a weighting of the differences between ICCs that have been calculated separately for each group and then rescaled. While the ECD is not a DIF analysis, it becomes valuable when subgroup distributional differences are estimated. It helps identify the potential for DIF to be significant using more traditional DIF approaches (e.g., Sibtest, Mantel–Haenszel). When the ECD is close to zero no DIF should be expected. If the ECD results in a constant value, uniform DIF could occur. Uniform DIF occurs when the rescaled ICCs for two groups differ only in their difficulty or $\hat{\boldsymbol{b}}$ values (i.e., one group consistently has a lower probability of correct response across all

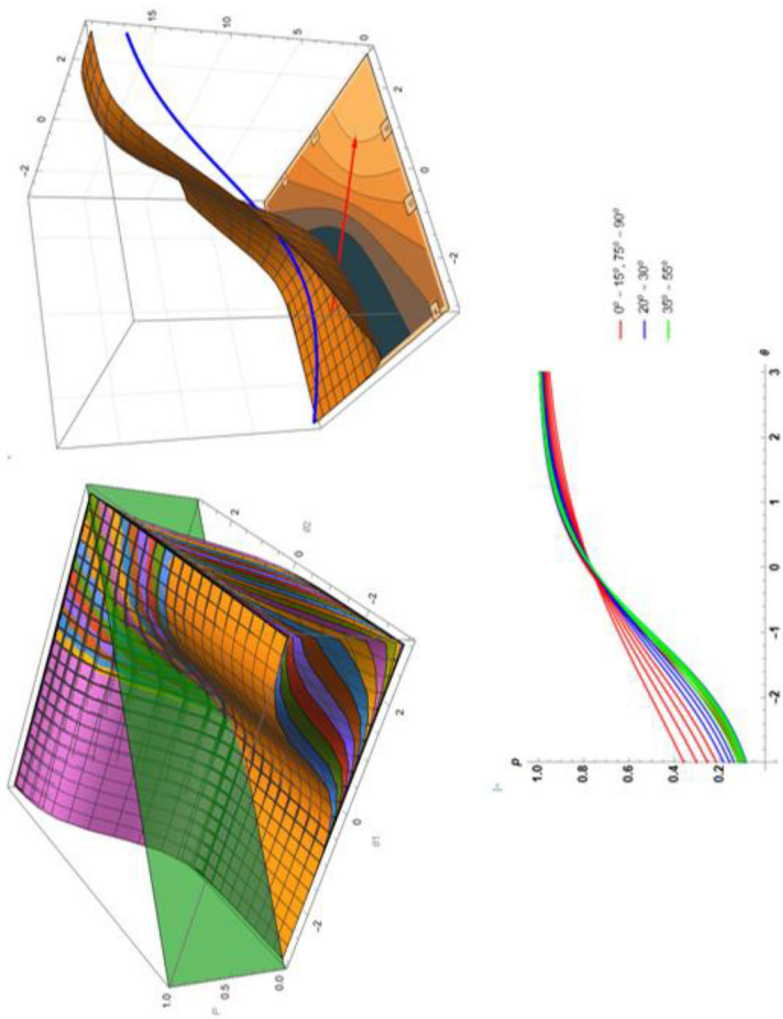

FIGURE 13.
Composite graph illustrating the 19 M2PL item response surfaces and green RC plane (top left); the test characteristic surface, contour, reference composite (red arrow), and unidimensional TCC (top right); and, the 19 estimated unidimensional ICCs colored by item vector angle (bottom). (Color figure online)

TABLE 2.
Generating compensatory model parameters for a 10-item test.

Item	a1	a2	d
1	.614	.000	−.579
2	1.270	.013	.422
3	1.081	.021	−.109
4	.698	.020	−.533
5	1.331	.052	.233
6	1.001	.049	.124
7	.961	.056	−.726
8	.764	.052	.415
9	.932	.072	.074
10[a]	.050	1.250	−.200

[a] This item is considered to be a biased item.

levels of θ). When ECD is a function of θ (e.g., $.4\theta$), it suggests that nonuniform DIF could occur. Nonuniform DIF occurs when the rescaled ICCs for two groups differ only in their discrimination or $\hat{\boldsymbol{a}}$ values (i.e., for some levels of θ the Reference group has a higher probability of correct response, while for other levels of θ the Focal group has a higher probability of correct response*)*.

The second approach involves a systematic analytic approach to explore DIF and builds upon the research by Camilli (1992). In this study, we investigate scenarios where unidimensional 2PL item parameters are analytically estimated even when the true model is the M2PL model. This situation characterizes practitioners who ignore the dimensionality of their test data, as described by Equations (4) and (9). For illustration purposes, we created a ten-item two-dimensional set of items. The M2PL item parameters are shown in Table 2. The first nine items fall in a narrow validity sector and primarily measure θ_1. However, Item 10 serves as a potential DIF item, primarily measuring θ_2 with a vector angle of 87.72°. A vector plot of the ten items is shown in the upper left panel of Fig. 14. Although the same parameters for all ten items were used for both the Reference and Focal groups, their underlying ability distributions differ. Consequently, their *RC*s will also differ, leading to distinct unidimensional item parameter estimates. Using Camilli's formulation, Equations (4) and (9) were used to examine three different cases with dissimilar underlying distributions for the Reference and Focal groups. After calculating the $\hat{a}$- and $\hat{b}$ values, the item parameters for the Reference group were placed on the Focal group's scale, using a mean–mean transformation (Kolen & Brennan, 2014) .

In Table 3, results are presented for two cases in which the Focal group and Reference group underlying ability distributions are identical and three cases where the distributions are different. In this table, the underlying bivariate normal distributional parameters and the angular direction of the RC for each group are listed in the first two columns. The third column provides the ECD based on each group's distributional parameters. In the fourth and fifth columns are the 2PL $\hat{a}$ and $\hat{b}$ values for *Item 10* using Camilli's formulation (4) and (9) for each group. Note that once the Reference parameters were calculated, they were rescaled using a mean–mean transformation. In the final column, the ICC differences are designated as displaying no DIF, uniform DIF, or nonuniform DIF. It is essential to reemphasize that DIF is caused by the underlying ability distributional differences.

As can be seen in the first two rows of Table 3, when the underlying distributions are identical, the ECD =.0, and the $\hat{a}$ and $\hat{b}$ values for *Item 10* are identical. Thus, whenever the underlying distributions are identical, regardless of the different composites being measured by the items, there should be no DIF.

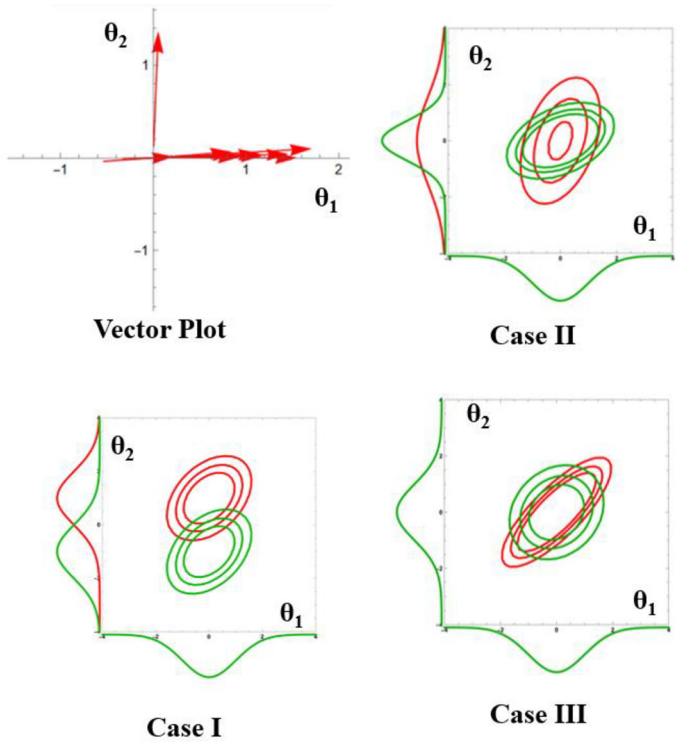

FIGURE 14.
Graphical displays of the item vectors, and the underlying Reference (red) and Focal (green) distributions for each of the three cases outlined in Table 3. (Color figure online)

4.1. Case 1: Unequal $\mu_{\theta 1}$- and Unequal $\mu_{\theta 1}$ Values: Uniform Bias

This case has two parts, one in which the Reference and Focal groups have mean differences in θ_1 and mean differences in θ_2. ECD differences were 2 and -2, respectively. There were no $\hat{a}$ differences. However, when the Reference group had the greater θ_1-mean, Item 10 was much easier for the Reference group with $\hat{b} = 1.78$ compared to $\hat{b} = -.24$ for the Focal group. Conversely, the Reference group had the greater mean for θ_2, the Reference group's $\hat{b} = -2.30$ versus $\hat{b} = 1.73$ for the Focal group. Notably, the *RC*-angle was 26.68° for both groups. Since these examples resulted in only $\hat{b}$ differences, they were an indication of uniform DIF.

4.2. Case 2: Unequal η Variance: Nonuniform Bias

In Case 2a, the Reference group has $\sigma^2_{\theta_1} = 2.5$ and $\sigma^2_{\theta_2} = 1.$ and $\rho = .4$, whereas the Focal group has $\sigma^2_{\theta_1} = .5$ and $\sigma^2_{\theta_2} = 1.0$ and $\rho = .4$. Note that the $\sigma^2_{\theta_1}$ is five times larger for the Reference group. In Case 2b, the variance differences were on θ_2: $\sigma^2_{\theta_1} = 2.5$ for the Reference group and $\sigma^2_{\theta_2} = .5$ for the Focal group. The ECD was -.31θ_1 and the *RC* angles were 16.40° and 42.05° for the Focal group and 37.78° and 18.23° for the Reference group for Case 2a and Case 2b, respectively. In both of these cases, both $\hat{a}$ and $\hat{b}$ differed between the two groups resulting in nonuniform DIF.

TABLE 3.
Analytical results of estimated 2PL item parameters for *Item 10* for the Reference and Focal group based on their underlying different distributions.

Distributions		Estimated 2PL parameters for Item 10					
Focal	Reference	ECD	Focal		Reference[1]		DIF?[2]
			$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	
Identical Distributions: $\rho = 0$							
$N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & .0\\.0 & 1\end{pmatrix}\right]$ 3.12°	$N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & .0\\.0 & 1\end{pmatrix}\right]$ 3.12°	0	.05	1.69	.05	1.69	No
Identical Distributions: $\rho = .4$							
$N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & .4\\.4 & 1\end{pmatrix}\right]$ 26.68°	$N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & .4\\.4 & 1\end{pmatrix}\right]$ 26.68°	0	.29	−.33	.29	−33	No
Case 1a: θ_1-Mean Difference: $\rho = .4$							
$N\left[\begin{pmatrix}-1\\0\end{pmatrix},\begin{pmatrix}1 & .4\\.4 & 1\end{pmatrix}\right]$ 26.68°	$N\left[\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}1 & .4\\.4 & 1\end{pmatrix}\right]$ 26.68°	−2	.29	−24	.29	1.58	U
Case 1b: θ_2Mean Differences: $\rho = .4$							
$N\left[\begin{pmatrix}0\\-1\end{pmatrix},\begin{pmatrix}1 & .4\\.4 & 1\end{pmatrix}\right]$ 26.68°	$N\left[\begin{pmatrix}\mathbf{0}\\\mathbf{1}\end{pmatrix},\begin{pmatrix}1 & .4\\.4 & 1\end{pmatrix}\right]$ 26.68°	2	.29	1.73	.29	−1.90	U
Case 2a: θ_1-Variance Differences: $\rho = .4$							
$N\left[\begin{pmatrix}\mathbf{0}\\\mathbf{0}\end{pmatrix},\begin{pmatrix}2.5 & .63\\.63 & 1.0\end{pmatrix}\right]$ 16.40°	$N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}.5 & .28\\.28 & 1.0\end{pmatrix}\right]$ 37.78°	$-.31\theta_1$	.40	−.24	.16	−.55	NU
Case 2b: θ_2-Variance Differences: $\rho = .4$							
$N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1.0 & .63\\.63 & 2.5\end{pmatrix}\right]$ 42.05	$N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1.0 & .28\\.28 & .5\end{pmatrix}\right]$ 18.23	$-.35\theta_1$	.21	−.45	.55	−.20	NU
Case 3: Correlational Differences							
$N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & .2\\.2 & 1\end{pmatrix}\right]$ 16.17°	$N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & .8\\.8 & 1\end{pmatrix}\right]$ 40.67°	$.6\theta_1$	.18	−.50	.53	−.21	NU

[1]Reference group estimated parameters were rescaled to the Focal group's scale using the mean–mean transformation
[2]U = Uniform DIF; NU = Nonuniform DIF.

4.3. Case 3: Unequal Correlations: Nonuniform Bias

In this case, the θ_1- and θ_2 means and variances are equal but the correlations differ: Reference $\rho_{\theta 1,\theta 2} = .8$ and Focal $\rho_{\theta 1,\theta 2} = .2$. The expected difference is $E[\eta_R|\theta] - E[\eta_F|\theta] = .6\theta$. The *RC*-angles were 16.17° for the Focal group and 40.67° for the Reference group. Again, both $\hat{a}$ and $\hat{b}$ differed between the groups indicating nonuniform DIF.

Using the software package Mathematica (Wolfram, 2020) , Ackerman and Xie (2019) created a DIF Graphical Simulator. This simulator allows researchers to change the underlying two-dimensional latent distributions for the Reference and Focal groups as well as the M2PL item parameters for a given suspect item. A graphical example is given in Appendix C.

5. Study 2: Examining the Effect of Different Conditioning Scores on DIF Analyses

Ackerman and Evans (1994) employed two DIF approaches that are standard DIF analysis by testing practitioners, Mantel–Haenszel and Sibtest. We will first provide a brief background of each approach and then examine how the approaches were used to assess the effect of changing the conditioning scores and attempt to account for the complete latent ability used by examinees to respond to a hypothetical two-dimensional test. The biggest cause of DIF when conditioning on raw scores (e.g., Mantel–Haenszel) is that the raw score may not always account for the complete latent ability space that examinees used to respond to the items. In this study, two DIF statistics

TABLE 4.
A 2 x 2 contingency used in the MH computation.

Group	1 (Correct)	0 (Incorrect)	Total
Reference (R)	A_j	B_j	N_{Rj}
Focal (F)	C_j	D_j	N_{Fj}
Total	$N_{1.j}$	$N_{0.j}$	$N_{..j}$

were used: the Mantel–Haenszel (*MH*) (Holland et al., 1988) and Sibtest β_u (Shealy & Stout, 1993a,b) . Both statistics are conditional analyses grouping subjects by their number correct score.

5.1. Background: Mantel–Haenszel and Sibtest DIF Detection Methods

When calculating the *MH*-statistic for item i, we consider two groups of examinees: the Reference group and the Focal group. The Focal (F) group typically represents a minority group (e.g., Hispanic examinees). The Reference (R) group is frequently a nonminority group (e.g., White examinees). Examinees from each group are matched based on their number correct score.

For each score category j, a 2 x 2 contingency table (Table 4) is created. This table notes the frequency of correct and incorrect answers for each group, along with the marginal and total frequencies. It is tacitly assumed that examinees in the same contingency table are matched on their latent abilities, that they used to respond to the item being examined. The centipede plot above (Fig. 9) illustrates how even though examinees may have the same number correct score, they could have quite different two-dimensional latent ability profiles.

Summing over the contingency tables for item i and using a continuity correction, the MH statistic is calculated as

$$\mathrm{MH}_i = \frac{\left[\left|\sum_j A_j - \sum_j E\left(A_j\right)\right| - \frac{1}{2}\right]^2}{\sum_j Var\left(A_j\right)},$$

where the expected value of Cell A frequency is given as

$$E\left(A_j\right) = \frac{N_R N_{1.j}}{N_{..j}}$$

and the variance of cell A frequencies equals

$$Var\left(A_j\right) = \frac{N_{Rj} N_{Fj} N_{1.j} N_{0.j}}{\left(N_{..j}\right)^2 \left(N_{..j} - 1\right)}$$

Typically, the MH statistic is used to test the null hypothesis that for each raw score category j the odds of a Reference group examinee answering the item correctly equals the odds that a Focal group examinee will answer the item correctly (Holland et al., 1988) . That is, if p_{Rj} and p_{Fj}; are the probabilities of a Reference and Focal group examinee answering the item correctly,

respectively, and $q_{Rj}$and q_{Fj} are the probabilities of a Reference and Focal group examinee answering the item incorrectly, respectively,

$$H_o: \frac{p_{Rj}}{q_{Rj}} = \frac{p_{Fj}}{q_{Fj}} \text{ j} = 1, \ldots, \text{K}$$

is tested against the alternative of uniform DIF,

$$H_1: \frac{p_{Rj}}{q_{Rj}} = \alpha \frac{p_{Fj}}{q_{Fj}} \quad \alpha \neq 1, j = 1, \ldots, K$$

where H_o is the null hypothesis, H_1 is the alternative hypothesis, and α is the common odds ratio in the K 2 x 2 tables. Uniform DIF occurs when the rescaled, unidimensional item response functions differ only in difficulty. When H_o is true, MH is distributed as χ^2 with 1 degree of freedom. It should be noted that for a given examinee, the score (i.e., 0 vs 1) on the item being examined is part of the conditioning score.

DIF according to (Shealy & Stout, 1993a,b) , should be conceptualized by examining the difference in certain marginal item characteristic curves for the two groups of interest:

$$P\left(X_i = 1 | \Theta = \theta\right) \int P_i\left[\theta, \eta\right] f\left(\eta | \theta\right) \mathrm{d}\eta,$$

where $P_i\left[\theta, \eta\right]$ is the M2PL model (Eq 1) and $f(\eta|\theta)$ is a specified group's conditional distribution of the nuisance dimension, η, given a fixed value of θ, the target ability. That is, for a fixed value of θ, $P_i(X_i = 1|\Theta = \theta)$ is obtained by averaging $P_i[\theta, \eta]$ over η. That is, $P_i(X_i = 1|\Theta = \theta)$ is the ICC if the differences in the nuisance direction are integrated out. Y*

An estimate of the SIBTEST test statistic is given as

$$\hat{\beta}_U = \sum_{h=0}^{n} \hat{p}_h \left(\bar{\text{Y}}_{Rh} - \bar{Y}_{Fh}\right)$$

where

$$\hat{p}_h = \frac{\left(G_{Rh} - G_{Fh}\right)}{\sum_{j=0}^{n}\left(G_{Rh} - G_{Fh}\right)},$$

and G_{Rh} and G_{Fh} are the number of examinees in the Reference and Focal groups at the valid score X = h. The Sibtest test statistic is computed as

$$\beta_U = \frac{\hat{\beta}_U}{\hat{\sigma}\left(\hat{\beta}_U\right)}.$$

where the standard deviation in the denominator is calculated as

$$\hat{\sigma}\left(\hat{\beta}_U\right) = \left\{\sum_{h=0}^{n} \hat{p}_k^2 \left[\frac{1}{G_{Rh}} \hat{\sigma}^2\left(Y|h, R\right) + \frac{1}{G_{Fh}} \hat{\sigma}^2\left(Y|h, \right)\right]\right\}^2,$$

The test statistic has an approximate $N(0,1)$ distribution when no DIF is present. Unlike the MH statistic, an examinee's score on the studied item is not part of the conditioning score. Sibtest resolves rescaling issues by means of a regression correction (Shealy & Stout, 1993a,b) .

5.2. *DIF Detection with Different Conditioning Scores*

Using the Mantel–Haenszel and Sibtest statistics, Ackerman and Evans (1994) examined the impact of different conditioning scores on DIF results. Specifically, they looked at generated two-dimensional compensatory data where θ was the valid skill and η represented the invalid skill. The testing scenario involved a 30-item test measuring (θ, η)-composites spanning measurement angles from 0° to 90°, in 3° increments. A vector of these items is displayed in Fig. 15. All items had a difficulty parameter (d_i) value of zero and an MDISC value of 1.5. The Reference and Focal groups had different latent ability distributions. The bivariate normal distributions for the Reference and Focal were $N\left[\begin{pmatrix}1\\-1\end{pmatrix},\begin{pmatrix}1 & .4\\ .4 & 1\end{pmatrix}\right]$ and $N\left[\begin{pmatrix}-1\\1\end{pmatrix},\begin{pmatrix}1 & .4\\ .4 & 1\end{pmatrix}\right]$, respectively. Three different sample size pairings were used but results were similar for each pairing. The results shown here are for the pairing $N_{Ref} = 1000$, and $N_{Foc} = 500$.

The purpose of comparing the four different conditioning scores is to illustrate that DIF can occur when one has not accounted for the complete latent ability space. Here are the details for each conditioning variable:

- To condition on θ, the transformation used is $X_\theta = 10(\theta) + 25$. In this case, the ability η is not accounted for and DIF should increase the more the ability η is required (i.e., as the angle of the item vector increases toward 90°).
- To condition on η, the transformation used is $X_\eta = 10(\eta) + 25$ was used. When conditioning on η the ability θ is not accounted for, and DIF should increase the more the ability η is required (i.e., as the angle of the item vector decreases toward 0°).
- The number correct score is equivalent to the case where $\theta = \eta$, (i.e., the RC angle is at *45°).* Items that require an equal weighting of both skills, (i.e., $a_1 = a_2$), should show no DIF. However, DIF should increase as items require more of θ-skill (item vectors approach 0°) or more of η (items approach 90°).
- Finally, to condition on (θ, η) the latent ability plane was divided into 64-square regions using an 8 x 8 grid (Fig. 16). All examinees in the same square of the grid were assigned the same conditioning score, (i.e., examinees in the square $-3 \leq \theta_1 < -2.25$ and $2.25 < \eta \leq 3$ would be assigned a conditioning score of 1). Note that the conditioning score does not enter the calculation of either DIF statistic, but rather ensures that examinees with the same abilities are placed into the same 2 x 2 conditioning table. When the conditioning score is a function of (θ, η), the complete latent ability is accounted for and there should be no DIF.

Results of the DIF analyses using the four different conditioning scores are illustrated in Fig. 17. When the conditioning score was the number correct, both DIF procedures consistently identified items 1–9 and 22–30 as showing DIF 100% of the time. When the conditioning score was a linear transformation of the generating θ, items 9–30 were consistently rejected 100% of the time. Slight differences were observed between *MH* and Sibtest β_U results. β_U appeared to be more sensitive to DIF. Its rejection rate increased faster than MH as the angular composite of the item deviated from 0°. This is shown by the fact that the rejection rate reached or exceeded .9 by Item 7 for β_U, whereas it did not occur until Item 9 for *MH*. Note that rather than using the Sibtest regression correction for β_U, the conditioning variable was based on a latent trait parameter.

As hypothesized, the opposite results occurred when the valid test direction was along the η-axis. That is, items measuring θ (i.e., Items 1–23) consistently exhibited DIF in favor of the Reference group. For the final analysis, the 64 score categories matching examinees on both θ and η were used as the conditioning scores. The results, as shown in Fig. 19, showed no DIF for any of the 30 items, regardless of the DIF procedure. However, it is important to recognize that this purely hypothetical testing situation would not occur in practice. Its purpose was to

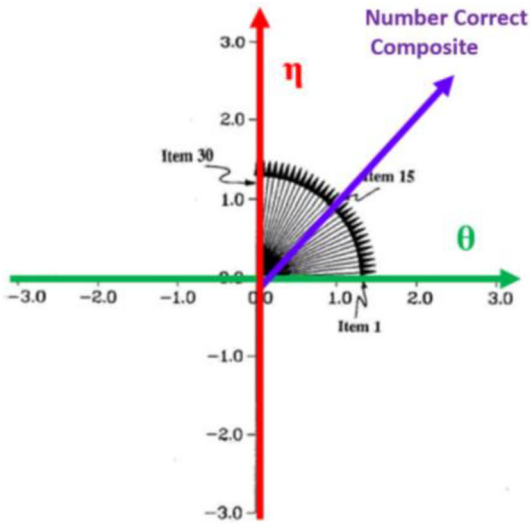

FIGURE 15.
Item vectors for the 30-item symmetric test and conditioning score composite directions.

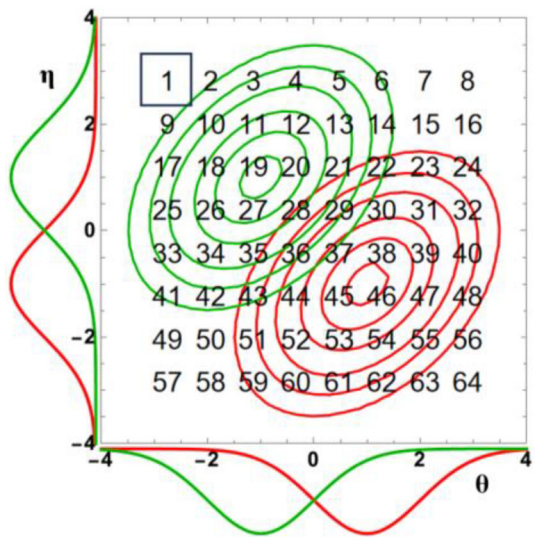

FIGURE 16.
Overlaying an 8 x 8 grid on the (θ, η) latent ability plane with Reference (red) and Focal (green) underlying and marginal distributions. (Color figure online)

illustrate the significance of underlying group distributional differences interacting with items that measure a spread of two-dimensional composite skills. Identifying the specific composite skills being measured remains a genuine challenge for testing practitioners. Several studies have examined using multiple conditioning scores, such as those by Clauser et al. (1996) and Mazor et al. (1998), in an attempt to condition on the complete latent ability space.

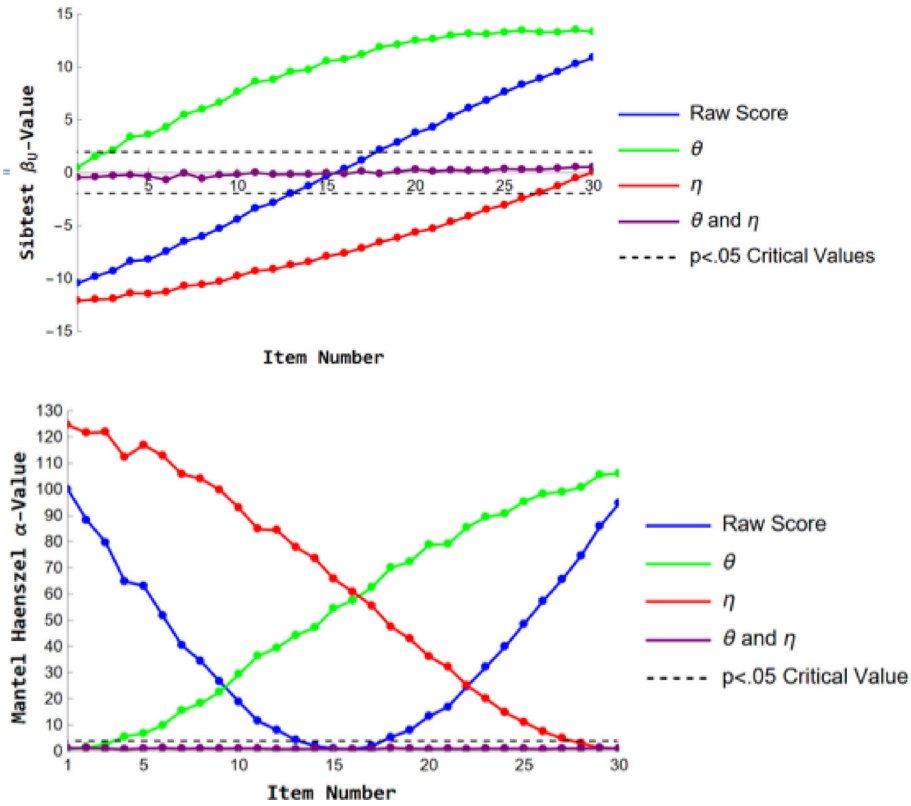

FIGURE 17.
MH and SIBTEST DIF results by item for each of the four condition scores.

6. DIF Even Though Reference and Focal Two-Dimensional Distributions are Identical

6.1. The Two-Dimensional Noncompensatory MIRT Model

Up to this point, the discussion has centered around how DIF can occur when items measure invalid skill composites and the two groups of interest have different ability distributions related to the invalid skill. Interestingly, DIF can also occur when the underlying two-dimensional ability distributions are identical and the vectors corresponding to the test items lie in a very narrow validity sector (e.g., a unidimensional test). This phenomenon was examined by Ackerman and Evans (1994), Bolt and Johnson (2009), and Ackerman et al. (2014). They found that DIF can occur when a two-dimensional test contains items for which different groups of students use distinct approaches to solve the same problem. These divergent solution strategies could occur due to pedagogical differences in how students were taught to information, particularly in items such as "story" problems. For instance, one group might have been explicitly taught to combine pieces of information, whereas another group was not instructed to integrate or combine these pieces.

In Ackerman and Evans (1994) and Ackerman et al. (2014), the different strategies were are modeled using two distinct MIRT models. The integration strategy was modeled using the compensatory model (6), and the nonintegration strategy was modeled using the MIRT noncompensatory model developed by Sympson (1978). This does not allow for compensation and can

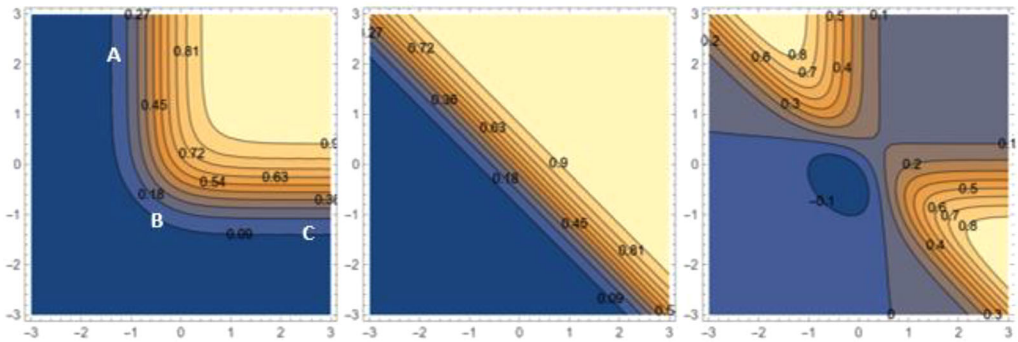

FIGURE 18.
Contour and difference plots for a matched noncompensatory and compensatory item.

be expressed as

$$P_{NC}\left(u_{ij}=1|\theta_{1j},\theta_{2j},a_{1i},a_{2i},b_{1i},b_{1i}\right)=\left[\frac{1.0}{1.0+e^{\left(a_{1i}(\theta_{1j}-b_{1i})\right.}}\right]\left[\frac{1.0}{1.0+e^{\left(a_{2}(\theta_{2j}-b_{2i})\right.}}\right]. \tag{11}$$

This model is essentially the product of two 2PL (1) models, with a discrimination and difficulty parameter for each dimension. Unlike the compensatory model (6) which assumes the abilities from different dimensions can compensate for each other, the noncompensatory model treats them independently. P_{NC}'s multiplicative nature ensures that P_{NC} can never be larger than the maximum value of either dimension's 2PL model.

The multiplicative nature of this model causes the response surface equiprobability contours to become curved. That is, unlike the compensatory model contours which are always parallel lines, the noncompensatory model contours are always parallel curves, the larger the a values, the more discriminating the item and the closer together the equiprobability curves. A contour plot of a noncompensatory response surface where $a_1 = a_2 = 1.6$ and $b_1 = b_2 = -.47$ is displayed in the left panel in Fig. 18. In the left panel, the letters A, B, and C denote three different (θ_1, θ_2)-profiles, (high, low), (low, low), and (low, high), respectively. Notice that all lie on the same equiprobability contour or have the same probability of correct response.

6.2. *DIF Study Simulation Using Matched Compensatory and Noncompensatory Items*

A 30-item test was created where the primary focus was on measuring θ_1 and θ_2 equally (i.e., the item vectors were enclosed in a narrow validity sector from 40° to 50°.) For the Reference group, all 30 items were modeled using the compensatory model (6). For the Focal group, for items 1–10 and 15–30 the probability of correct response followed the compensatory model, but for items 11–14, the probability of correct response was determined using the noncompensatory model (11) that matched their compensatory counterparts. To estimate noncompensatory item parameters that would match the compensatory for items 11–14, the approach proposed by Spray et al. (1990) was used. Specifically, PNC parameters were estimated by minimizing the function

$$\sum_{i=1}^{N}\left\{\left[P_C\left(\boldsymbol{\theta_i},\boldsymbol{a},d\right)-P_{NC}(\theta_i,\hat{\boldsymbol{a}},\hat{b})\right]\right\}^2$$

TABLE 5.
Compensatory and Noncompensatory item parameters matched on p value for a given underlying ability distribution.

Item	Compensatory				Noncompensatory				
	a_1	a_2	d	p	a_1	a_2	b_1	b_2	p
11	.4	.4	.0	.50	.64	.64	−.86	−.86	.50
12	.8	.8	.0	.50	1.13	1.13	−.51	−.51	.50
13	1.2	1.2	.0	.50	1.14	1.14	−.51	−.51	.50
14	1.6	1.6	.0	.50	1.25	1.25	−.47	−.47	.50

for 2000 randomly generated examinee abilities from the latent underlying bivariate normal distribution, $N\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & .4\\ .4 & 1\end{pmatrix}\right]$. The *Nminimize* function in Mathematica (2020) was used for this optimization. This process was repeated for 10 replications for each of the four items to ensure that the estimates obtained were not unduly influenced by the samples selected or the starting values. The matched set of parameters for items 11–14 are displayed in Table 5. In Fig. 18, the center panel contains the equiprobability contour plot for the matched Item 14 compensatory item. The right panel displays the compensatory—noncompensatory difference contour for item 14. From this plot, it appears that the compensatory item 14 and the noncompensatory item 14 would produce similar probabilities of correct response for examinees in the first and third quadrants, but noticeably different probabilities for examinees in the third and fourth quadrants.

Response data were generated for 1000 examinees for both the Reference and Focal groups using the same underlying bivariate distribution. A DIF analysis was conducted for each item using the Mantel–Haenszel and Sibtest procedures using the software program difR (Magis et al., 2010) . This process was replicated 100 times. For each item, we calculated the proportion of times it resulted in significant DIF using each method. The results are graphed in Fig. 19.

There was a clear demarcation between the items. Items 12–14, which were modeled discordantly, were flagged more frequently (65% to 85% of the replications) than the remaining items. All significant DIF results favored the Reference group, which is cross-validated by the right panel of Fig. 18 which shows the probability of correct response in quadrants 2 and 4 was greater for the compensatory model. These results parallel the findings of Ackerman and Evans (1994) which illustrated that DIF detection for two groups using different models was affected greatly by the discrimination power of the items. The MDISC value for item 11 was only.4, whereas for the remaining items it was.8, 1.2, and 1.6, respectively. It was also noted that for the remaining 24 matched items, some had Type I error rates as high as.12%, possibly affected by the conditioning total score for each group. The contour of the test characteristic surface for the Focal group exhibited a slight curve, resembling the contour of a noncompensatory item. Interestingly, even when conditioning on both θ_1 and θ_2, as illustrated in Study 2 above, the DIF for groups using two different MIRT models did not disappear as shown by Ackerman (2014).

7. The True Challenge: Substantively Identifying the Cause of the Manifested DIF

At a large national testing company, we routinely conducted DIF analyses after each administration using the Mantel–Haenszel procedure. Two DIF analyses were conducted, one looking for **gender DIF** (comparing Males vs Females) and a second analysis looking at **racial DIF** (comparing White Examinees vs Black Examinees). We would often share these results with the content editors to see if they could explain our DIF results. Sometimes we would do a "blind" test.

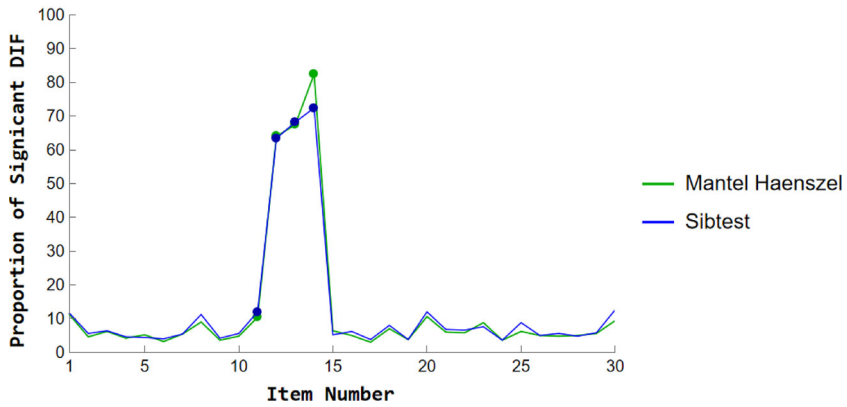

FIGURE 19.
Mantel–Haenszel and SIBTEST DIF results for each of the 30 items.

That is, we would assemble a set of items flagged for significant DIF. We also included a few items that did not show DIF. We asked the editors to determine which items exhibited significant DIF and which group was favored. Below are four actual items which item writers found very difficult to explain the DIF results. Using your psychometric knowledge and DIF expertise, determine if the items below favored Males, Females, White examinees, or Black examinees, or showed No DIF. While statistically detecting DIF is straightforward, understanding why it occurs is the true challenge! Answers are provided in Appendix A.

Item 1

The Cold War threatened to erupt into a "hot" war in October 1962 when President Kennedy demanded that the Soviet Union

A. dismantle naval bases located in Nicaragua.
B. withdraw all troops from South Korea.
C. remove all missiles and missile bases located in Cuba.
D. return captured American pilot Gary Powers to the USA.

Item 2

In comparison with normal males, those with Klinefelter's syndrome have:

A. 1 extra X-chromosome
B. 1 fewer Y-chromosome
C. 1 extra Y-chromosome
D. 2 extra X-chromosomes

Item 3

A bell was found fastened at a fork in a branch 15 feet from the ground in a 40-year-old tree. A person claimed that the bell was fastened to the tree about six feet from the ground when the tree was 10 years old. Of the following, the best evaluation of this story is that the claim is:

A. true, because the bell moved upward as the tree grew taller.
B. true, because the bell was fastened to a forked branch, which grew rapidly upward.
C. false, because trees do not grow taller that quickly.
D. false, because upward growth in trees occurs at the terminal buds, not within the trunk or branches.

Item 4

A customer at a service station asks the attendant to put 30 pounds of air in my right rear tire." Assuming that the tire is completely flat, air will be pumped into the tire until the:

A. tire's weight increases by 30 pounds.
B. air pressure inside the tire equals the atmospheric pressure.
C. air pressure inside the tire is 30 pounds per square inch greater than the atmospheric pressure.
D. air pressure inside the tire is 30 times greater than the atmospheric pressure.

8. Summary and Concluding Remarks

It is widely recognized that test response data often exhibit multidimensionality. Due diligence requires that testing practitioners should first examine the dimensionality of their data. By identifying the dominant dimensions and mapping them onto a test's specifications, each dimension can be well defined. Additionally, practitioners need to be vigilant and identify unintended skills that are being measured. Using this foundational analysis, appropriate calibration model(s) can be selected. These models play a crucial role in estimating item parameters, scaling examinee abilities, and understanding the potential for DIF to occur. Research by Kok (1988), Ackerman (1992), Camilli (1992), and Shealy and Stout (1993a,b) hypothesized that DIF can result when a test measures invalid or unintended skills and the groups of interest exhibit distinct conditional ability distributions on these skills for different levels of the valid skill.

Using this perspective as a starting point, this article provides a detailed examination of the two-dimensional compensatory MIRT model. Graphical representations of two-dimensional item response surfaces and their corresponding contours were examined. These plots provide a deeper understanding of how items perform across the latent ability space. Plots of items as vectors indicating the (θ_1, θ_2)-composite that each item is optimally measuring were illustrated and discussed. Vectors of valid items should lie within a sector, termed the validity sector. Further insight about (θ_1, θ_2)-composite consistency across the observable score scale was illustrated using plots of conditional centroids. Finally, centipede plots, detailing how examinees'(θ_1, θ_2)-abilities get mapped onto the unidimensional θ-scale were explained. Such graphical analytics, in concert with knowledge of underlying ability distributions for subgroups of interest, can provide detailed insight into the potential for DIF to occur.

The article then focused on the analytical work of Wang (1985) and Camilli (1992). They demonstrated how data generated using the two-dimensional compensatory model can be mapped onto a unidimensional 2PL IRT scale, referred to as the reference composite (RC). Additionally, they showed how estimated 2PL IRT item parameters can be derived given estimates of the underlying two-dimensional bivariate normal examinee ability distribution parameters $(\mu_{\theta_1}, \mu_{\theta_2}, \sigma_{\theta_1}, \sigma_{\theta_2}, \rho)$ and the compensatory model (6) item parameters (a_1, a_2, and d). Specific examples were provided to illustrate how 2PL IRT parameter estimates can change as the underlying two-dimensional ability distributions change.

We then reviewed three studies that illustrate how DIF can occur using a two-dimensional framework:

- The first study investigated how DIF can occur for an invalid item, one that deviates significantly from the validity sector. DIF results were examined across four different distributional scenarios. This study identified which distributional differences result in uniform DIF and which produce nonuniform DIF.
- The second study emphasized the importance of considering the complete latent ability space when using DIF. conditional approaches. Simulations revealed that DIF occurred when examinees were matched solely on their θ_1 value, or only their θ_2 value, or on their

number correct score where $\theta_1 = \theta_2$. However, when examinees were matched on their (θ_1, θ_2)-groupings, no DIF occurred.

- The third study detailed an educational scenario with identical underlying two-dimensional distributions for the two groups. Despite this similarity, certain items displayed DIF. The DIF occurred for items where one group's responses were generated using the compensatory model and the other group's responses were generated using the noncompensatory model. These models were chosen to simulate different response strategies that resulted from different instructional pedagogies.

The paper concludes with a test for readers to correctly identify the Mantel–Haenszel DIF analysis results for four given items from a nationally administered standardized test. This involves determining which examinee group was favored for each item or whether the results indicated no DIF. Conducting DIF analyses is relatively straightforward thanks to computer programs written for all the approaches listed in Table 1. Substantively explaining the results is where the real challenge lies. Psychometricians and item writers must collaborate to interpret DIF findings. One should also never discount the possibility of Type I error!

Generated sets of item parameters in this study were created specially to provide insight for the testing practitioner about how underlying ability distributions or different response styles can affect examinee performance and consequently unidimensional item parameter estimation to create DIF. They are not realistic. Real data are very messy, to say the least. We will never know the true item parameters, the true latent abilities, or whether the data are unidimensional or multidimensional. Only when we simulate data do we know the truth. It is paramount that psychometricians and testing practitioners always remember the words of wisdom by the noted British statistician George Box "*Essentially, all models are wrong, but some are useful*" (Box & Draper, 1987).

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Open science statement Data and the Mathematica code used in the illustration will be made available on the Open Science Framework upon publication.

Appendix A

Groups indicated as being favored in the Mantel–Haenszel analysis.

Item 1: Male examinees
Item 2: Black examinees
Item 3: No DIF
Item 4: Male examinees. This is the only item which has a possible explanation: that males, for the most part, know more about cars than females.

Appendix B

Example illustrating formulation of how the unidimensional 2PL model gets mapped into a two-dimensional latent ability space:

$$\varepsilon_{\upsilon_2}[P(u=1\,|\upsilon_1,\upsilon_2)|\,\upsilon_1] = \int_{-\infty}^{+\infty} P\,(u=1|\upsilon_1,\upsilon_2)G(\upsilon_2|\upsilon_1)\,\mathrm{d}\upsilon_2.$$

Assume you want to find the unidimensional 2PL $\hat{a}$ and $\hat{b}$ value for a two-item test where the two-dimensional compensatory parameters are given as $A = [\{1.5, 0\}, \{0, 1.5\}]$ and $D = \{.5, .5\}$ and the underlying model is given as

$$P\left(u_{ij}=1|\theta_{1j},\theta_{2j},a_{1i},a_{2i},d_i\right) = \frac{1.0}{1.0+e^{-1.7(a_{1i}\theta_{1j}+a_{2i}\theta_{2j}+d_i)}}.$$

It is also given that the underlying two-dimensional distribution is a bivariate normal with a mean vector of {0,0}and the covariance matrix, Ω, as [{1,.4}, {.4,1}]. Note these are chosen only for illustration purposes. Item 1 measures only θ_1 and item 2 measures only θ_2.

Following the work of Wang (1986) and Camilli (1992), we first determine the Cholesky decomposition, L, of Ω. L[{1., 0.4}, {0., 0.91651}]. To compute the reference composite, we first need to calculate the L'A'AL matrix which equals, {2.25, 0.9}, {0.9, 2.25}. The eigenvalues of this matrix are 3.15, 1.35} and the eigenvectors associated with the eigenvalues, w_{ij}, are {0.7071, 0.7071}, {−0.7071, 0.7071}.

The reference composite is then calculated as the arccosine of the first element of the eigenvector associated with the largest eigenvalue. The arccosine of.7071 corresponds to 45° which corresponds to the reference composite direction from the positive θ_1-axis. This is the composite that would represent the unidimensional θ-scale if the data were fit to the 2PL model.

It should also be noted that the first and second factor scores, υ_1 and υ_2 are then defined as:

$$\begin{pmatrix}\upsilon_1\\ \upsilon_2\end{pmatrix} = \begin{bmatrix} w_{11(\theta_1-\mu_{\theta_1})} + w_{12(\theta_1-\mu_{\theta_1})} \\ w_{21(\theta_1-\mu_{\theta_1})} + w_{22(\theta_2-\mu_{\theta_2})}\end{bmatrix} = \begin{bmatrix} .7071\theta_1 + .7071\theta_2 \\ -.7071\theta_1 + .7071\theta_2\end{bmatrix}$$

In Fig. 20, the left panel is a contour plot of Item 1 with the reference composite (υ_1) direction indicated with a solid red arrow and the perpendicular υ_2 direction indicated with a dotted red arrow. We then substitute υ_1 and υ_2 in for θ_1 and $\theta_{2\,\mathrm{in}}$ the compensatory model to get

$$p\left(u_{ij}=1|\upsilon_1,\upsilon_2\right) = \frac{1.0}{1.0+e^{-1.7(1.5\,\upsilon_1+.0\upsilon_2+.5)}}.$$

To determine G ($\upsilon 2|\upsilon 1$), we must first rotate the bivariate normal distribution 45o and then determine the conditional distribution. Assuming Σ is the original covariance $\Sigma = \begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$ and R_θ is the rotation matrix,

$$R_\theta = \begin{bmatrix}(\cos 45°) & -(\sin 45°) \\ (\sin 45°) & +(\cos 45°)\end{bmatrix} = \begin{bmatrix}\frac{\sqrt{2}}{2} & \frac{-\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}\end{bmatrix}$$

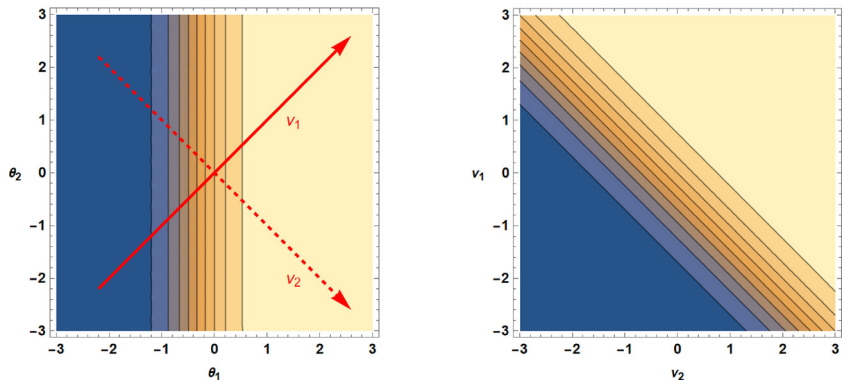

FIGURE 20.
The contour graph of the original item response surface with direction of first (ν_1) and second principal component (ν_2) (left) and contour surface rotated 45° (right).

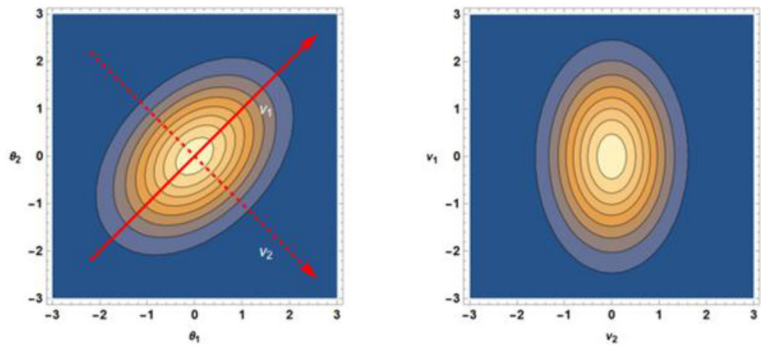

FIGURE 21.
A contour plot of the original bivariate normal distribution (left) and the contour plot of the rotated distribution (right).

then the rotated mean vector, μ', and rotated covariance matrix, Σ', are given by $\mu^{'} = R_\theta \mu = \frac{\sqrt{2}}{2}\begin{bmatrix}\mu_1 - \mu_2 \\ \mu_1 + \mu_2\end{bmatrix}$ and $\Sigma^{'} = R_\theta \Sigma R_\theta^T = \frac{1}{2}\begin{bmatrix}\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 & \sigma_1^2 - \sigma_2^2 \\ \sigma_1^2 - \sigma_2^2 & \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2\end{bmatrix}$, where σ_1^2, σ_2^2 and ρ are the original variances and correlation of the original random variables.
The rotated mean vector is [0,0] and the rotated covariance matrix, $\Sigma^{'}$ is [{.6,0}, {0,1.4}]. The formula for the conditional distribution of G(υ1|υ2) equals (Fig. 21)

$$G\left(\upsilon 1|\upsilon 2\right) =\sim N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}\left(x - \mu_X\right), \sigma_Y^2\left(1 - \rho^2\right)\right) =\sim N\left(0, 1.4\right).$$

In Fig. 22 on the left are conditional normal distributions, $G\left(\upsilon 1|\upsilon 2\right)$, for υ1 = -2, -1,0,1,2. On the right are the conditional ICCs, $p\left(u_{ij} = 1|\upsilon_1, \upsilon_2\right)$, for υ1 = -2, -1,0,1,2.
Using the formula

$$\left(u_{ij} = 1|\upsilon_1, \upsilon_2\right) = \int_{-6}^{+6} P(u = 1|\upsilon_1, \upsilon_2)G(\upsilon_2|\upsilon_1)\mathrm{d}\upsilon_2$$

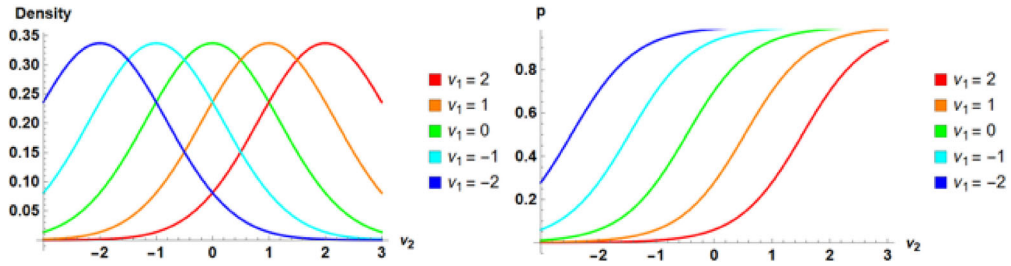

FIGURE 22.
Conditional normal distributions, $G(\upsilon 1|\upsilon 2)$, for $\upsilon 1 = -2, -1, 0, 1, 2$ (left) and conditional ICCs, $p(u_{ij} = 1|\upsilon_1, \upsilon_2)$, for $\upsilon 1 = -2, -1, 0, 1, 2$ (right).

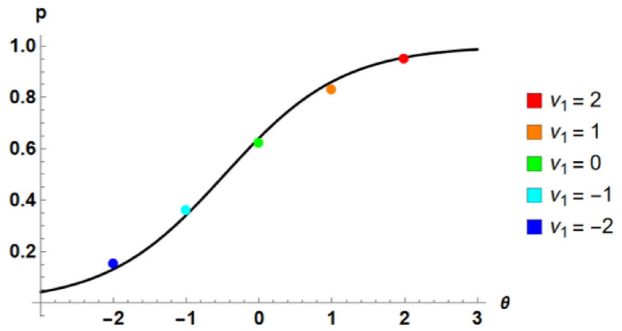

FIGURE 23.
Estimated unidimensional ICC with five estimated (υ_1, p) values plotted.

4where $d\upsilon_2 = .001$ we can estimate the unidimensional ICC for values of $\upsilon 1$ values of -2, -1,0,1,2. These values are.13,.34,.64,.86, and.95, respectively. Using Camilli's derivational formulas,

$$\hat{a}_j = \frac{\boldsymbol{a}_j' \boldsymbol{W}_1}{\sqrt{2.89 + \boldsymbol{a}_j' \boldsymbol{W}_2 \boldsymbol{W}_2' \boldsymbol{a}_j}}$$

and

$$\hat{b}_j = \frac{d_j - a_j' \boldsymbol{\mu}}{\boldsymbol{a}_j \boldsymbol{W}_1}.$$

we obtain the 2PL item parameter estimates: $\hat{a} = .73$ and $\hat{b} = -.47$. Fig. 23 shows the estimated ICC using the $\hat{a}$ and $\hat{b}$ and the five color-coded estimated (υ_1, p) values. Figure 24 illustrates three different perspectives of all the elements of Camilli's formulation, including the M2PL response surface and corresponding contour plot, the RC (v1) which represents the estimated unidimensional scale, v2 (the orthogonal second principal component, the RC plane, the underlying conditional latent ability distribution, $G(\upsilon 1|\upsilon 2)$, and the estimated unidimensional ICC.

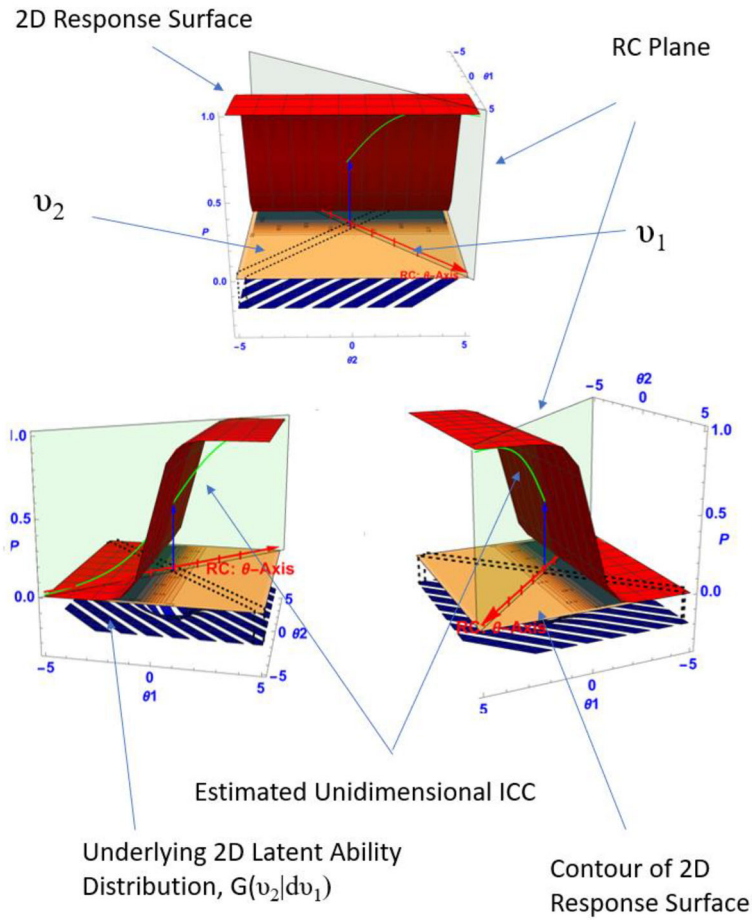

FIGURE 24.
Three different perspectives of different elements that were used in the mapping of the two-dimensional compensatory model onto a unidimensional ICC.

Appendix C

Ackerman and Xie (2019) created a DIF Graphical Simulator. This simulator enables researchers to modify the underlying two-dimensional latent distributions for the Reference and Focal groups and the M2PL item parameters for a given suspect item. Using the Camilli (1992) analytical derivations, the 2PL unidimensional discrimination (a) and difficulty (b) parameters are estimated and the resulting ICC is illustrated. A mean–mean transformation is used to place the Focal group's estimated parameters onto the scale of the Reference group. The transformed ICCs are then displayed, and the degree of misfit, defined as: $\sum_{\theta=-3}^{\theta=3} (P(\theta)_{Ref} - P(\theta)_{Foc})^2$, is calculated. The DIF Graphical Simulator is shown in Fig. 25

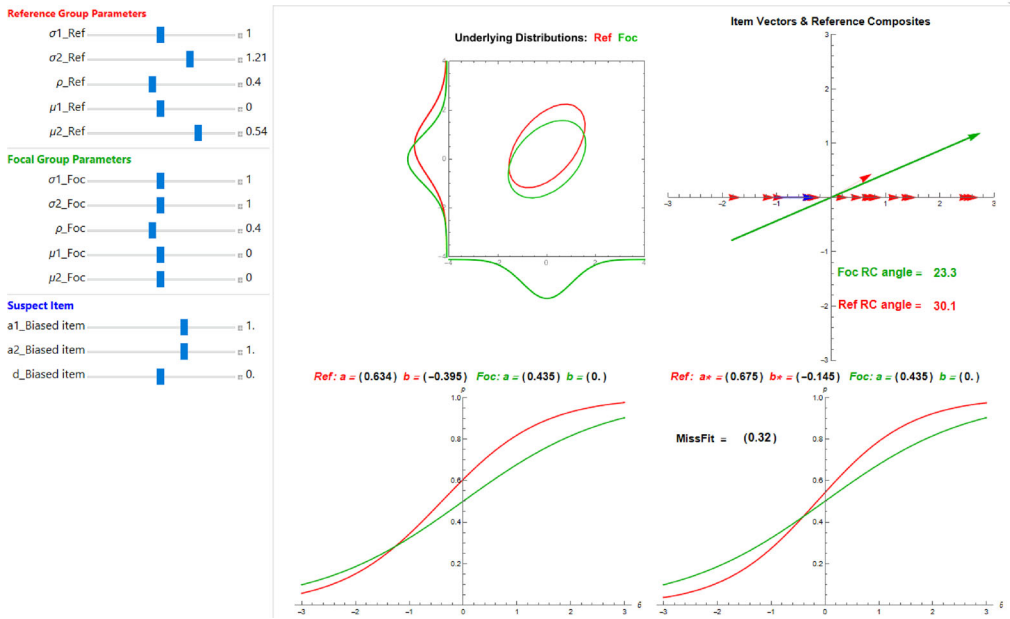

FIGURE 25.
The graphical display is shown by the DIF Graphical Simulator.

References

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement, 15*, 13–24. https://doi.org/10.1177/014662169101500103

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67–91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

Ackerman, T. A., & Evans, J. A. (1994). The influence of conditioning scores in performing DIF analyses. *Applied Psychological Measurement, 18*, 329–342. https://doi.org/10.1177/014662169401800404

Ackerman, T. A., McCallaum, B., & Ngerano, G. (2014). Differential item functioning from a compensatory-noncompensatory perspective. Invited address to the International Congress of Educational Research, Haceppette University, Ankara, Turkey.

Ackerman, T. A. & Xie, Q. (2019). DIF graphical simulator. *Educational Measurement: Issues and Practice, 38*(1), 5. https://doi.org/10.1111/emip.12171

Ackerman,T. A. & Xie, Q. (2019). DIF graphical simulator. *Educational Measurement: Issues and Practice*, 38(1), 5. https://doi.org/10.1111/emip.12171

Bauer, D. J., Belzak, W. C., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal, 27*, 43–55. https://doi.org/10.1080/10705511.2019.1642754

Bolt, D. M., & Johnson. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Assessment,33*(5), 335–352. https://doi.org/10.1177/0146621608329891

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*(2), 129–147. https://doi.org/10.1177/014662169201600203

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel–Haenszel statistics. *Journal of Educational Measurement, 34*(2), 123–139. https://doi.org/10.1111/j.1745-3984.1997.tb00510.x

Carlson, J. E. (2017). Unidimensional vertical scaling in multidimensional space. *ETS 11 Research Report Series, 2017*(1), 1–28. https://doi.org/10.1002/ets2.12157

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10. PMID 26828106.

Clauser, B. E. & Mazor, K. M. (1998). Using Statistical Procedures To Identify Differentially Functioning Test Items. An NCME Instructional Module. Educational Measurement: Issues and Practice, 17(1), 31–44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement, 33*(4), 454–464. https://doi.org/10.1111/j.1745-3984.1996.tb00501.x

Clauser, B. E. & Mazor, K. M. (1998). Using Statistical Procedures To Identify Differentially Functioning Test Items. An NCME Instructional Module. *Educational Measurement: Issues and Practice*, 17(1), 31–44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17*(4), 335–350. https://doi.org/10.1177/014662169301700402

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533–559. https://doi.org/10.1007/s11336-008-9092-x

Fleishman, J. A. & Lawrence, W. F. (2003). Demographic variation in SF-12 scores: true differences or differential item functioning. *Medical care, 41*(7), 75–86. https://doi.org/10.1097/01.MLR.0000076052.42628

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. British Journal of Mathematical and Statistical Psychology, 63, 395–416. https://doi.org/10.1348/000711009x466835

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. New York: Springer. https://doi.org/10.1007/978-1-4939-0317-7

Lim, H., Choe, E. M., & Han, K. (2022). A residual-based differential item functioning detection framework in item response theory. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12313

Liu, Y., Zumbo, B., Gustason, P., Huang, Y., Kroc, E., & Wu, A. (2016). Investigating causal DIF via propensity score methods. *Practical Assessment, Research and Evaluation, 21*(13), 1–24. https://doi.org/10.7275/ewqz-n96

Ma, Y., Ackerman, T., Ip, E., & Chung, J. (2023). The effect of the projective IRT model on DIF detection. IMPS 2023 Annual Meeting, College Park, Maryland, United States.

Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22*(4), 357–367. https://doi.org/10.1177/014662169802200404

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*(4), 309–326. https://doi.org/10.1177/01466219922031437

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning detection and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (pp.129–145). Hillsdale, NJ: Lawrence Erlbaum. http://www.books.google.co.ke/books?isbn=1109103204.

Huang, P. H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology, 71*, 499–522. https://doi.org/10.1111/bmsp. 121-130.

Junker, B., & Stout, W. F. (1991). Robustness of ability estimation when multiple traits are present with one trait dominant. Paper presented at the International Symposium on Modern Theories in Measurement: Problems and Issues. Montebello, Quebec.

Kok, F. (1988). Item bias and test multidimensionality. In R. Lange Heine & J. Rost (Eds.), Latent trait and latent class models (pp. 263–275). New York: Plenum Press. https://doi.org/10.1007/978-1-4757-5644-9_12

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement, 24*, 115–138. https://doi.org/10.1177/01466216000242002

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum https://eric.ed.gov/?id=ED312280

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

McKinley, R. L., & Reckase, M. D. (1982). The use of the general rasch model with multidimensional item response data.

Muthen, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research, 47*, 637–664. https://doi.org/10.1177/0049124117701488

Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. Journal of Educational Measurement *37*(4), 357–373. http://www.jstor.org/stable/1435246

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*(4), 295–312. https://doi.org/10.1111/j.1745-3984.2006.00018.x

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495–502. https://doi.org/10.1007/BF02294403

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353–368. https://doi.org/10.1177/014662169501900405

Ramsay, J. O. (1990). A kernel smoothing approach to IRT modeling. Talk presented at the Annual Meeting of the Psychometric Society at Princeton New Jersey.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer. https://doi.org/10.1007/978-0-387-89976-3

Shealy, R., & Stout, W. F. (1993). An item response theory model for test bias. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239). Hillsdale: Erlbaum.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159–19. https://doi.org/10.1007/BF02294572

Spray, J., Davey, T., Reckase, M., Ackerman, T. & Carlson, J. (1990). Comparison of two logistic multidimensional item response theory models. ACT Research Report ONR90-8.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589–617.

Strachan, T., Ip, E., Fu, Y., Ackerman, T., Chen, S. H., & Willse, J. (2020). Robustness of projective IRT to misspecification of the underlying multidimensional model. *Applied Psychological Measurement, 44*(5), 362–375. https://doi.org/10.1177/0146621620909894

Strachan, T., Cho, U. H., Ackerman, T., Chen, S.-H., de la Torre, J., & Ip, E. (2022). Evaluation of the linear composite conjecture for unidimensional IRT scale for multidimensional responses. *Applied Psychological Measurement, 46*(5), 347–360. https://doi.org/10.1177/01466216221084218

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Sympson, B. (1978) A model for testing with Multidimensional items. In Weiss, D. J. (ed) Proceedings of the 1977 Computerized Adaptive Testing Conference, University of Minnesota, Minneapolis.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale NJ: Erlbaum.

Wang, M. (1985). Fitting a unidimensional model multidimensional item response data: The effects of latent space misspecification on the application of IRT Unpublished manuscript, University of Iowa.

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement, 30*, 22–42. https://doi.org/10.1177/0146621605279867

Wolfram, 2020 Wolfram Research, Inc., (2020). Mathematica, (Version 12.2), [Computer Software]. Champaign, IL.

Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 213–249.

Manuscript Received: 28 NOV 2023
Published Online Date: 5 APR 2024