JOURNAL OF APPLIED MEASUREMENT, *25*(3/4), 190-202

A Graphical Framework Using Item Response Modeling to Detect Nonuniform Differential Item Functioning

Bunyong Dejanipont
University at Buffalo

Mark Wilson
University of California, Berkeley

Detecting differential item functioning (DIF) is fundamental to improving the fairness and validity of virtually any assessment. However, many commonly used DIF indices are suitable for uniform DIF, but not for nonuniform DIF. Furthermore, using graphs such as category characteristic curves (also known as category probability functions) to examine DIF in a polytomous item is relatively involved and convoluted. To address such limitations, we propose a graphical DIF detection framework that is intuitive and sensitive to both nonuniform DIF and uniform DIF. Using the partial credit model as an illustration, we compare a DIF graphical-based approach that leverages the proposed framework against a DIF parametric test approach that utilizes a DIF index and its accompanying DIF test statistic. For uniform DIF detection, there is a substantial agreement between the two approaches. But for nonuniform DIF detection, the discrepancy is stark: the graphical-based approach flags many more items having nonuniform DIF than the parametric test approach does. Altogether, the results suggest that a DIF investigation that only relies on a uniform-DIF-oriented index and its DIF test statistic is likely to provide specious conclusions about item bias, since such an index is generally much less sensitive to nonuniform DIF.

Keywords: nonuniform differential item functioning, item bias, category characteristic curve, item response modeling, partial credit model

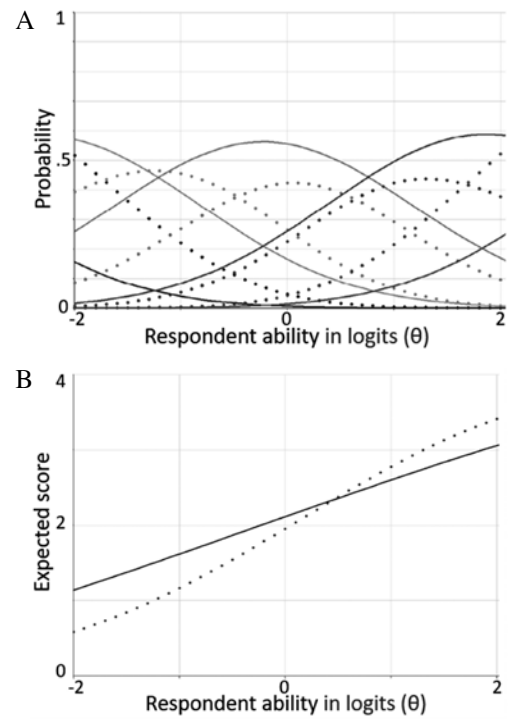
Bunyong Dejanipont https://orcid.org/0000-0002-3274-910X
Bunyong Dejanipont is now at University of California, Berkeley.
Correspondence concerning this article should be addressed to Bunyong Dejanipont, Berkeley School of Education, University of California, Berkeley, 4415 Berkeley Way West Building, Berkeley, CA 94720, United States; email: bunyong.dejanipont@berkeley.edu

Differential item functioning (DIF) is the extent to which an item is differentially difficult for respondents with the same ability from different groups (Holland & Thayer, 1988). And DIF detection procedures are a crucial component of ensuring the fairness of virtually any measure. However, our review suggests that many psychometric studies that investigate DIF either examine only uniform DIF (while leaving out nonuniform DIF) or use an approach that is insufficient for nonuniform DIF detection. In either case, conclusions about item bias drawn from such studies provide only partial, if not misleading, insights. And naturally, relying on such insights for measurement development could adversely affect the measurement validity. Moreover, studies examining DIF often report only estimated values of a DIF indicator along with its test statistic while omitting to report the graphs of items under examination. As this paper later demonstrates, a reliance on a DIF index that is not sensitive to nonuniform DIF and its test statistic by themselves is hardly a sound DIF investigation because of the possibility of nonuniform DIF.

Furthermore, even when graphs are used to examine polytomous items for DIF, they are often in the form of (or akin to) category characteristic curves (CCCs)[1], which are relatively involved and convoluted, thereby providing difficulty to many substantive researchers and practitioners. As an illustration of how CCCs are relatively complicated to interpret and overloaded with information for the purpose of examining DIF, consider a 5-point scale with ten items. In this example, using CCCs to investigate DIF between two groups would involve dealing with 100 different CCCs (ten CCCs for each item, as shown in Figure 1 Panel A). Alternatively, one can use expected score curves (ESCs) to examine DIF for the same scale, which involves dealing with only 20 ESCs (two ESCs for each item, as shown in Figure 1 Panel B). As such, instead of having to work with 100 CCCs, one can strategically use 20 ESCs to efficiently investigate DIF in the ten 5-point items, thereby substantially reducing the number of plots that one needs to inspect by 80%. And logically, using CCCs to examine DIF not only involves many comparisons but also entails that each comparison is in terms of probability difference, which is less intuitive than score difference. Thus, when using graphs to examine DIF, ESCs are not only more optimal but also more readily interpretable than CCCs and the like.

Figure 1

Plots of Functions of Respondent Ability for a 5-Point Item



Note. Panel A: Category characteristic curves (CCCs) of a 5-point item. CCCs have the vertical axis that represents the probability, ranging from 0 to 1.

Panel B: Expected score curves (ESCs) of the same item. ESCs have the vertical axis that represents the score (as opposed to the probability).

The dotted curves (···) are one group of respondents, and the solid curves (—) are another group of respondents.

1 For a polytomous item, category characteristic curves are also referred by some as categorical response curves or category probability functions. (Category characteristic curves are akin to cumulative category response functions, which also have an axis that represents the probability.)

As such, we propose a graphical DIF detection framework that leverages expected score differences to detect both nonuniform DIF and uniform DIF, thereby taking advantage of ESCs (e.g., efficient, intuitive). For illustration, we use the partial credit model (Masters, 1982) to demonstrate the potential of a graphical-based approach that applies the proposed framework. Notably, we illustrate that while the graphical approach and the parametric test approach perform similarly for uniform DIF detection, the graphical approach is more effective for detecting nonuniform DIF. Therefore, using a DIF index and its test statistic by themselves is not advisable in a DIF investigation because such a DIF index is generally less sensitive to nonuniform DIF and is therefore likely to provide specious conclusions about item bias.

The remainder of this article is organized as follows: we briefly review the concept of DIF and discuss the partial credit model and its parametric test approach in detecting DIF. We then discuss a methodological issue commonly observed when testing for statistically significant DIF. Next, we introduce our graphical DIF detection framework that incorporates expected score differences in ESCs. We then empirically compare the graphical approach with the parametric approach in detecting DIF and conclude with a discussion of implications.

Differential Item Functioning

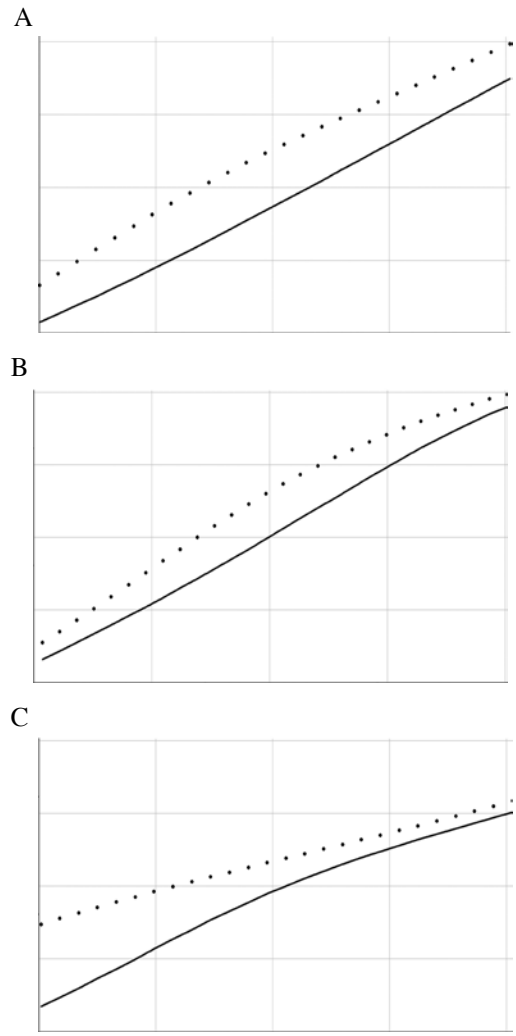
Differential item functioning (DIF) is the degree to which an item shows differential difficulties for persons with the same ability from different groups (Holland & Thayer, 1988). That is, DIF means that an item favors persons in one group over persons in another group even though they have the same level of ability (or whatever the construct) being measured. Thus, DIF for an item can threaten the validity and fairness of the instrument. Typically, DIF is thought of as the difference in the item parameters between groups. The grouping is usually based on a demographic variable (e.g., gender, race, nationality), although it can be based on any variable of interest. In practice, a DIF analysis can provide evidence, for example, of differences in linguistic interpretations or social considerations of items among different groups of people.

Uniform DIF

Uniform DIF occurs when an item exclusively favors one group over another group

Figure 2

Examples of Uniform DIF



Note. A dotted curve (···) represents an expected score curve (ESC) plot of a focal group; a solid curve (—) represents an ESC plot of a reference group; the horizontal axis presents respondent ability in logits, ranging from –2 to 2; the vertical axis represents expected score, ranging from 0 to 4.

across the ability spectrum—though the extent of such favoritism can vary. Visually, uniform DIF can be seen as a noncrossing of ESCs. Figure 2 illustrates examples of uniform DIF, such that across the whole ability spectrum for a given item, the item favors one group and only that group over another group. Figure 2 Panel C, for instance, depicts the focal group finding the particular item easier than the reference group across the estimated ability spectrum and to varying extents. Notably, as Figure 2 shows, ESCs of an item do not have to be parallel to each other to be considered uniform DIF so long as they do not cross each other (within the estimated ability range).

Nonuniform DIF

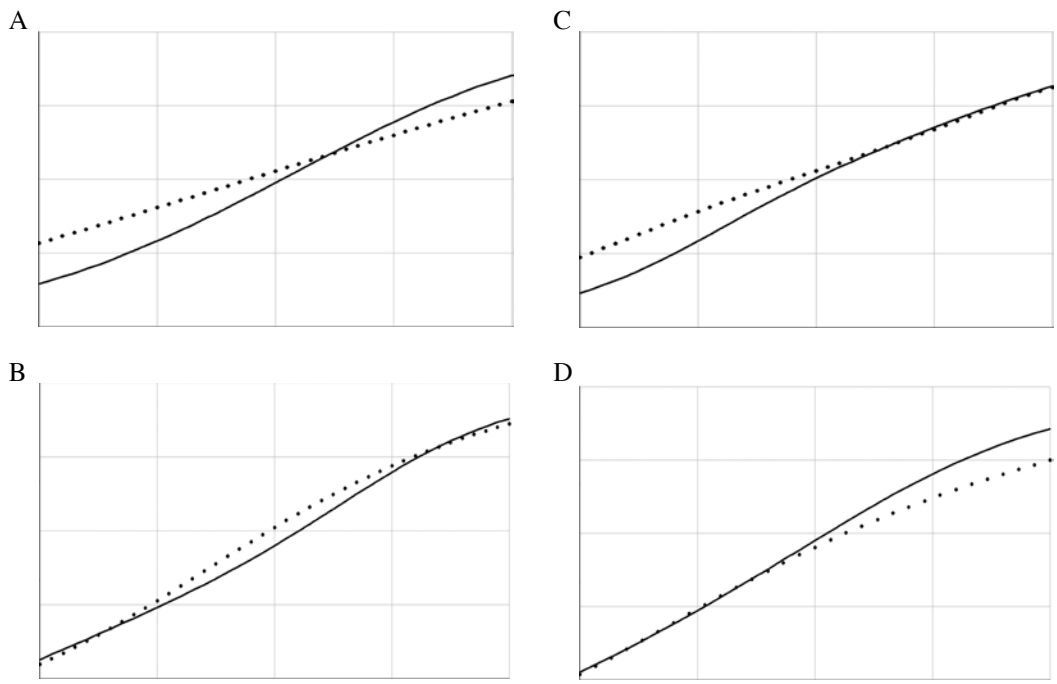
Nonuniform DIF occurs when an item favors one group over the other group for a part of the ability spectrum while favoring the latter group over the former group for another part of the ability spectrum. Visually, nonuniform DIF can be seen when ESCs of an item cross each other. Figure 3 illustrates examples of nonuniform DIF. Figure 3 Panel A shows that on the left of the ability spectrum, the focal group (dotted curve) finds the item easier than the reference group (solid curve), whereas on the right of the ability spectrum, the focal group finds the item harder than the reference group. Because of the crossing, nonuniform DIF is sometimes referred to as "crossing DIF."

Partial Credit Model

The partial credit model (PCM; Masters, 1982) is a Rasch-family item response model for polytomous items. Consider a polytomous item i having the following response categories: $0, 1, 2, \ldots, m$ such that there are $m + 1$ response categories, and there are m steps. In other

Figure 3

Examples of Nonuniform DIF



Note. A dotted curve (···) represents an expected score curve (ESC) plot of a focal group; a solid curve (—) represents an ESC plot of a reference group; the horizontal axis presents respondent ability in logits, ranging from –2 to 2; the vertical axis represents expected score, ranging from 0 to 4.

words, the polytomous item has $m + 1$ response categories (which are category 0, category 1, category 2, …, and, lastly, category m) and has m steps (which are 1st step, 2nd step, …, and, lastly, m^{th} step). Denoting X_{pi} as the response variable for a person p to the item i, the PCM models the probability of $X_{pi} = k$ (in the sense of "the probability of selecting response category k") given the person (or respondent) ability θ_p, item difficulty δ_i, and step deviation τ_{ij} (Penfield et al., 2008). (Item difficulty plus step deviation equals step difficulty δ_{ij}: $\delta_i + \tau_{ij} = \delta_{ij}$.) Formally, the expression for the PCM is the following (which also defines the category characteristic curves):

$$\mathrm{P}\left(X_{pi} = k \mid \theta_p, \delta_i, \tau_i\right) = \frac{exp\sum_{j=0}^{k}\left(\theta_p - \delta_i - \tau_{ij}\right)}{\sum_{t=0}^{m}\left[exp\sum_{j=0}^{t}\left(\theta_p - \delta_i - \tau_{ij}\right)\right]}, \quad (1)$$

where $\sum_{j=0}^{0}\left(\theta_p - \delta_i - \tau_{ij}\right) \equiv \sum_{j=0}^{0}(0) = 0$; and thus $\exp\sum_{j=0}^{0}\left(\theta_p - \delta_i - \tau_{ij}\right) \equiv \exp(0) = 1$. t is a dummy subscript running through the set of response categories for item i from 0 to m. The step deviation parameter τ_{ij} represents the difficulty associated with the transition from category k–1 to category k in relation to the item difficulty δ_i (Penfield et al., 2008). τ_i is the vector of step deviation parameters for item i.

Alternatively, the PCM can be expressed as the log ratio of the probability of $X_{pi} = k$ to the probability of $X_{pi} = k$–1:

$$\ln\frac{\mathrm{P}\left(X_{pi} = k \mid \theta_p, \delta_i, \tau_i\right)}{\mathrm{P}\left(X_{pi} = k - 1 \mid \theta_p, \delta_i, \tau_i\right)} = \theta_p - \delta_i - \tau_{ij} \text{ for } k = 0, 1, \cdots, m, \quad (2)$$

where $\theta_p - \delta_i - \tau_{ij}$ equals $\theta_p - (\delta_i + \tau_{ij})$.

Modeling DIF by Using the PCM

In the case of binary grouping variables, the PCM model can be extended to model DIF by incorporating a DIF parameter α_{ig}. The α_{ig} (i.e., item deviation step DIF) represents the effect of group g on the difficulty of item i. Using the PCM as the foundation and letting there be one binary grouping variable (e.g., smoking status: 0 = nonsmoker, 1 = smoker) with two group levels (indexed as $g = 0, 1$), the PCM DIF model is

$$\ln\left[\frac{\mathrm{P}\left(X_{pi} = k \mid \theta_p\right)}{\mathrm{P}\left(X_{pi} = k - 1 \mid \theta_p\right)}\right]_g = \theta_p - (\delta_i + \tau_{ij}) - \alpha_{ig} \text{ for } k = 0, 1, \cdots, m, \quad (3)$$

where $\mathrm{P}(X_{pi} = k \mid \theta_p)$ is the probability of selecting response category k on item i for person p; $\mathrm{P}(X_{pi} = k–1 \mid \theta_p)$ is that of selecting response category k–1; and θ_{pg} is the ability of person p in group g. The multidimensional PCM DIF model follows a similar logic, with the subscript d on parameters to represent a particular subdimension.

Parametric Test Approach in Detecting DIF

Under the Rasch-family modeling, which the PCM falls under, there are two common statistical tests for DIF: the z test and the likelihood ratio test. The former involves the ratio of α_{ig} to its standard error, and the latter involves the ratio of the likelihood of the model with DIF parameters to the likelihood of the model without DIF parameters. While the two DIF tests use estimates of DIF parameters differently to compute their DIF test statistics, both are summary DIF test statistics in the sense that they, essentially, generalize across the range of group differences in a given item. Or put differently, the two DIF tests examine the very same type of DIF (Paek & Wilson, 2011). Considering this central commonality, this paper uses the z test as an example to illustrate differences between a parametric test approach and a graphical approach in detecting not only uniform DIF but also nonuniform DIF. Formally, the z test for DIF under the Rasch-family modeling is defined as follows:

$$z - \text{test statistic} = \frac{\alpha_{ig}}{SE_{\alpha_{ig}}}, \quad (4)$$

where a DIF parameter estimate is statistically significant if the absolute value of its z-test statistic is more than 1.96 (i.e., significance level at .05).

Moreover, "With two groups, only one 'free' DIF parameter in item *i* will be estimated. In such a case, the difference in difficulties between the reference and focal groups of [the] item *i* is equal to $2\alpha_{ig}$" (Wang, 2004, p. 232). Under the Rasch-family modeling framework, the DIF index can be conceptualized as the absolute value of the DIF parameter estimate times two:

$$\text{DIF index} = 2\left|\alpha_{ig}\right| . \quad (5)$$

Notes on Testing for Statistically Significant DIF

Given the widespread confusion about the testing for the statistical significance of DIF, it is important to note that, generally, for a DIF investigation that examines DIF between two groups of respondents, decreasing the critical value for statistical significance based on the number of items or the number of DIF indices is not appropriate. And this is because doing so unduly deflates the power of a DIF test statistic, thereby allowing the rate of Type II error (i.e., the error of failing to reject a false null result or, more specifically, the error of failing to detect DIF that is present and relevant) to become nonsensically high. Put simply, for such a DIF investigation, a critical value that is considerably less than .05 greatly decreases the power to detect DIF, thereby providing an inaptly indiscriminate DIF benchmark. Indeed, through simulations, Penfield (2001) finds that items containing small-to-medium (uniform) DIF will often go undetected once the critical value is less than .05.

As a case in point of overcorrection, purporting to use "conservative rules for identifying DIF" between meditators and nonmeditators, Baer et al. (2011) "adopted a *p* value of .00043 (.05 divided by 117) as the criterion for statistical significance" and, conveniently, found only one item that "showed significant DIF using this strict criterion" (p. 5). However, note that for such a DIF investigation, which examines DIF between two groups of respondents, regardless of how many items an instrument has (or how many DIF indices, along with their DIF test statistics, are computed), its underlying core focus is nonetheless on *a single pair of scores within an individual item* and not on *multiple pairs of scores between various items*. Thus, for such a DIF investigation, a more appropriate standard for statistical significance level is simply the *p* value of .05, which would reasonably control for the error rate of failing to identify DIF that materially exists.

Altogether, the widespread misapplication of the statistical significance criterion for DIF indices and the uniform-DIF-oriented nature of many DIF indices further necessitate incorporating graphical information to examine DIF and having a framework to do so.

Proposed Graphical Framework for Detecting DIF

DIF parametric test approaches usually use a DIF index that, essentially, generalizes a wide range, if not an entire range, of differences between respondent groups across ability (or construct) levels for an item, thereby reducing the magnitude of DIF when an item has response curves that cross each other, as in Figure 3 Panel A. Departing from the reliance on such a DIF index, we propose a graphical framework that leverages the expected score differences in the expected score curves to detect DIF.

Moreover, as shown in Figure 1, unlike graphs such as category characteristic curves (in Panel A) that have a vertical axis that represents the *probability*, expected score curves (in Panel B) have a vertical axis that represents the *score*, thereby allowing for a more intuitive comparison because the comparison is in terms of score difference rather than probability difference. Therefore, using the maximum of expected score differences between expected score curves is not only intuitive but also more effective at detecting nonuniform DIF than a parametric approach.

In the context of DIF detection, the expected score curve (ESC) can be conceptualized as a function of the expected

score for person p responding to item i:

$$\mathrm{E}\left[X_{pi} = k \mid \theta_p, \delta_{ij}, \boldsymbol{\delta}_i\right] = \sum_{k=0}^{m_i}\left[k \bullet \mathrm{P}\left(k \mid \theta_p, \delta_{ij}, \boldsymbol{\delta}_i\right)\right], \quad (6)$$

where $\mathrm{P}(X_{pi} = k \mid \theta_p, \delta_{ij}, \boldsymbol{\delta}_i)$ is the probability of selecting response category k that a person p with ability θ_p will score a response category k in item i. The item i has step difficulty δ_{ij} such that $\boldsymbol{\delta}_i$ represents the vector of step parameters. That is, ESC is the sum of the probability of scoring in response category k where the probability is weighted by the scoring category k (i.e., the sum of score-weighted probabilities). Conventionally, as Figure 1 Panel B shows, ESCs have the vertical axis that represents expected scores. As such, for a given item, a maximum vertical distance between ESCs represents the maximum of differences in expected score between respondent groups, which can be more effective at detecting nonuniform DIF than a compensatory summary DIF index—as we empirically demonstrate next.

Method

The multidimensional PCM DIF model is used to model dimensions simultaneously and analyze the empirical data. We first begin with the DIF parametric test approach discussed earlier that uses the (summary) DIF index. We then move to the DIF graphical approach, applying our proposed framework that leverages the expected score curves. Next, we compare and contrast the two approaches in detecting uniform and nonuniform DIF. As an example, we then provide an interpretation of the results of the graphical approach. We used ACER ConQuest 5 (Adams et al., 2020) for all data analyses and used marginal maximum likelihood to estimate parameters.

Data

Instrument

The Five Facet Mindfulness Questionnaire (FFMQ; Baer et al., 2006) is a 39-item Likert response format instrument with five response categories ranging from *very rarely or never true* (recoded as 0) to *very often or always true* (recoded as 4). The FFMQ is used to measure five dimensions of mindfulness: (a) observing, (b) describing, (c) acting with awareness, (d) nonjudging of inner experience, and (e) nonreactivity to inner experience—each of the dimensions shares no common items with other dimensions. All 19 negatively worded items were reverse-scored before data analysis. Baer et al. (2006) provide a detailed description of the instrument.

For the DIF parametric test approach using the aforementioned DIF index, we estimate the statistical significance using the z test (Equation 4), calculate the effect size of the DIF index, and use the criteria proposed by Paek and Wilson (2011), such that any FFMQ item that has a DIF index of at least .426 is flagged as having DIF. As an illustration for a graphical DIF detection approach using the proposed framework, any FFMQ item that has a maximum vertical distance between the expected score curves of at least .5 is flagged as having DIF.[2]

Participants

In this paper, the empirical examples are based on women and men (N = 201) in two different age cohorts: (1) the youth cohort is comprised of 56 women and 45 men (n = 101, M_{age} = 20.2, SD_{age} = 1.9) and (2) the older adult cohort is comprised of 50 older women and 50 men (n = 100, M_{age} = 59.0, SD_{age} = 6.5). Moreover, the samples of women and men were

2 To illustrate a graphical approach that applies the graphical DIF detection framework, the score of .5 is arbitrarily selected for the FFMQ, whose items have the possible score range of 4. Even so, this cutoff point is unlikely to make the graphical approach substantially and systematically less conservative in its uniform DIF detection than the parametric approach's DIF index and its criteria. Indeed, our analyses reveal that for FFMQ items with noncrossing expected score curves, the parametric approach flags three items, while the graphical approach flags two items, such that the one item flagged only by the parametric approach is merely .1 short of being flagged by the graphical approach. However, unlike the parametric approach using the DIF index, this graphical approach has a constant sensitivity to DIF regardless of whether it is uniform or nonuniform.

Table 1

Demographic Characteristics of Women and Men

	Men (n = 106)	Women (n = 95)
Years of education	M = 14.8 (SD = 2.7)	M = 15.2 (SD = 2.7)
Aged 50 and older	47%	53%
Have seen a mental health professional	73%	67%
Currently employed	49%	58%
Religion		
Christian	50%	57%
No religion	33%	35%
White	27%	33%
Income		
30k to 49.9k	18%	19%
50k to 99.9k	22%	24%
Above 100k	20%	32%

matched on several demographic variables (e.g., ethnicity, education, religion, household income, and mental health professional service history); therefore, we examine DIF related to gender differences. Without matching such demographic variables, a gender difference in item function may be confounded by other demographic variables. The respondents were recruited from the West South-Central region of the United States. Table 1 shows the demographic characteristics of women and men in this sample.

Results

Parametric Test Approach in Detecting DIF

For the given empirical data, an omnibus test showed that DIF existed between women and men. We found a statistically significant interaction between gender ability and item difficulty ($\chi^2(34) = 132.02$, $p < .001$).[3] Table 2 shows item statistics of items flagged as having DIF by the parametric test approach. (See Table S1 in the supplemental materials for item statistics of all 39 items using the parametric test approach.) Item statistics for Item 20, for example, indicated statistically significant medium DIF, suggesting that even when controlling for overall mindfulness ability, women found the item harder than men by .43 logits.[4]

Table 2

Item Statistics of Items Flagged as Having DIF by the Parametric Test Approach

Item	Estimate of α	*SE* of α	*z* test	2|α|
20	.217	.058	3.741	.434
28	.329	.073	4.507	.658
30	.253	.064	3.953	.506
5	–.275	.069	–3.986	.550

Graphical Approach in Detecting DIF

Applying the proposed graphical DIF detection framework, we examined DIF by using the maximum of expected score differences between the expected score curves (ESCs) for each item. Eight items were flagged as having DIF by the graphical DIF detection approach: two for uniform DIF and six for nonuniform DIF. Figure 4 and Figure 5 together list items that were flagged as having DIF by the graphical approach or the parametric test approach or both.

Figure 4 and Figure 5 also show ESC plots of the flagged items, with the horizontal axis as the respondent mindfulness spectrum, ranging from –2 to 2, and the vertical axis as the expected score, ranging from 0 to 4. For example, ESC plots for Item 20 in Figure

3 When examining the term gender*items. From a modeling perspective, DIF is an interaction between group membership and item difficulty.

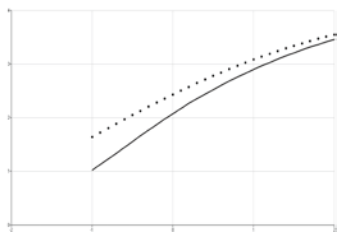
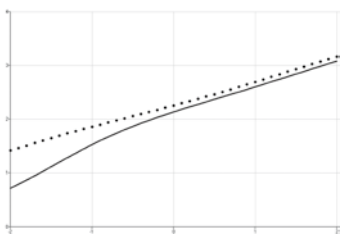
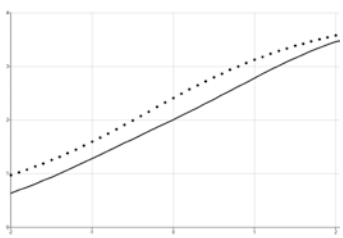
4 $2(.217) \approx .43$

4 show that on the less mindful side of the spectrum, women (solid curve) find the item harder than men (dotted curve); likewise, on the more mindful side of the spectrum, women (solid curve) find the item harder than men (dotted curve). Note that while the ESC plots for Item 20 show that the disparity in difficulty between the two groups decreases as respondent mindfulness increases along the spectrum, the item is always harder for one group and only that group, which is the hallmark of uniform DIF.

Importantly, as shown in Figure 4, for flagged items whose ESCs do not cross, the graphical approach and the parametric test approach led to very similar results. The only item that the graphical approach did not flag but was flagged by the parametric approach was Item 30, whose maximum expected score difference was .4, which is .1 short of being

Figure 4

Items With Noncrossing Expected Score Curves That Were Flagged as Having DIF by the Parametric Test Approach or the Graphical Approach or Both

Item	Expected score curve (ESC) plots	Parametric test approach	Graphical approach using the proposed framework
20		X	X
28		X	X
30 [a]		X	

Note. Dotted curves (···) represent ESC plots of men and solid curves (—) represent women, with the horizontal axis as respondent ability in logits, ranging from –2 to 2, and the vertical axis as expected score, ranging from 0 to 4.

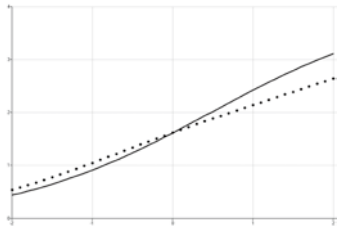
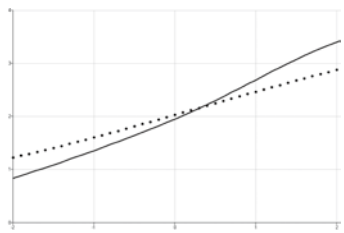
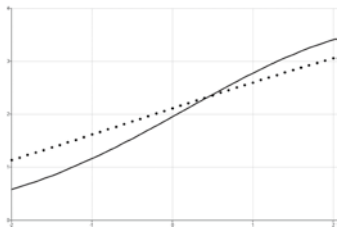
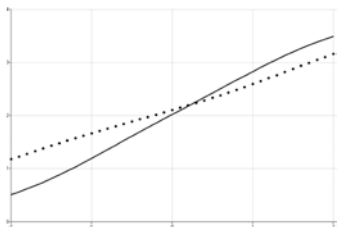
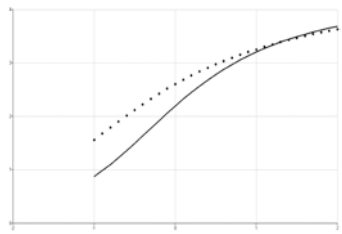
X indicates that the corresponding approach flagged the item as having DIF.

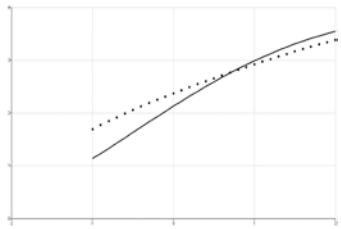
For Item 20, its estimated respondent ability range is from –1 to 2.

[a] For Item 30, the maximum expected score difference is .4, which is .1 short of being flagged by the graphical approach.

Figure 5

Items With Crossing Expected Score Curves That Were Flagged as Having DIF by the Parametric Test Approach or the Graphical Approach or Both

Item	ESC plots	Parametric test approach	Graphical approach using the proposed framework
5		X	X
9			X
12			X
18			X
26			X

36		X

Note. Dotted curves (···) represent ESC plots of men and solid curves (—) represent women, with the horizontal axis as respondent ability in logits, ranging from –2 to 2, and the vertical axis as expected score, ranging from 0 to 4.

X indicates that the corresponding approach flagged the item as having DIF.

For Items 26 and 36, their *estimated* respondent ability range is from –1 to 2.

flagged by the graphical approach. Taken together, there is a substantial agreement between the graphical approach and the parametric approach in detecting *uniform* DIF.

However, for *nonuniform* DIF detection, the discrepancy between the graphical approach and the parametric test approach is stark. For flagged items whose ESCs do cross (Figure 5), while both approaches flagged Item 5, five other items were flagged by the graphical approach but not the parametric approach. As discussed earlier, common DIF indices such as the one used by the parametric test approach are generally less sensitive to nonuniform DIF. Therefore, it is not surprising that the parametric approach did not flag these five items.

Moreover, another advantage of applying the graphical framework is that it can provide additional insights into the nature of DIF of the flagged items. For example, we identified several patterns of nonuniform DIF of the items in Figure 5. First, women (solid curves) with a low degree of mindfulness found these items easier than men (dotted curves) with the same degree of mindfulness, whereas women with a high degree of mindfulness found the items harder than men with the same degree of mindfulness. Second, ESC plots for the items in Figure 5 also show that the ESCs for women (solid curves) are steeper than those for men (dotted curves), suggesting that women were more likely than men to endorse extreme item responses for these items. Third, the ESCs of the first four items in Figure 5 show they have crossing points around an expected score of "2" or item response "Sometimes true."

Discussion

Detecting item response bias including DIF is crucial in improving the fairness and validity of virtually any assessment. Yet, many commonly used DIF indices are unsuitable for situations involving nonuniform DIF. Meanwhile, DIF detection approaches designed to identify nonuniform DIF in polytomous items are less accessible to those who are not specialized in psychometrics. To address such limitations, this article proposes a graphical DIF detection framework that is sensitive to both uniform DIF *and* nonuniform DIF—as opposed to only uniform DIF. Notably, the proposed framework takes the advantages of ESCs (e.g., efficient, intuitive, computable using common statistical software) and uses the maximum expected score difference to detect DIF. And through the empirical data, this article demonstrates that while a DIF graphical-based approach that leverages the proposed framework performs similarly to a DIF parametric test approach in detecting uniform DIF, the DIF graphical-based approach unequivocally outperforms the DIF parametric test approach in detecting nonuniform DIF.

Notably, the parametric approach especially failed to flag items whose DIF magnitudes somewhat evened out over the respondent

ability spectrum. This is likely because the DIF index of the parametric approach, essentially, cancels out the differences between two groups when an item has crossing ESCs with similar effect sizes for both groups, thereby giving a specious impression about the nature of DIF of the item. Thus, in contrast with the proposed graphical framework, the parametric method is much less sensitive to the nonuniform DIF because its DIF index represents the overall DIF that generalizes the whole range of group differences by balancing the differences between groups.

Research and Practical Implications

This paper offers a framework for using the graphical representation of expected score curves (ESCs) to detect both uniform DIF and nonuniform DIF, which promotes a more intuitive and comprehensive understanding of the nature of item bias, thereby enabling instrument developers and administrators to make a more informed decision.

Of course, mistakenly concluding that items contain uniform DIF (when they actually contain *nonuniform* DIF) or that items do not contain DIF altogether can erroneously impact the understanding about the items (and by extension the entire assessment), which in turn can adversely affect the decisions of not only assessment developers but also other relevant stakeholders. Many organizations routinely use assessment scores to make consequential decisions (e.g., determining eligibility for an opportunity, granting a promotion, deciding whether to start or stop an intervention, allocating resources, issuing a license). Accordingly, mistakenly thinking that items have uniform DIF when they in fact have nonuniform DIF would lead to a misunderstanding about the items' bias, which could result in a misguided remedy or a poor scoring system and unintended adverse consequences: for example, erroneously concluding that items are uniformly biased and thus implementing a uniform-DIF adjustment that, in effect, inaptly pushes many individuals in one group above a cutoff (i.e., making false positives) or pushes many individuals below it (i.e., making false negatives), thereby biasing decisions (e.g., who receives an opportunity, who earns a promotion, who passes an exam). Similarly, thinking that items contain merely small (or negligible) uniform DIF while they actually contain medium-to-large nonuniform DIF (as would be the case with the parametric test approach's results for FFMQ items 9, 12, 18, 26, and 36 shown in Figure 5) could lead to erroneously keeping or not addressing items that distort the measurement. To illustrate, consider FFMQ items 9, 12, and 18 (Figure 5): gender comparisons are likely to be biased for respondents with low mindfulness (as well as those with high mindfulness), even though for each of these items, the overall difference between men's average score and women's average score may be small. And not recognizing or addressing this could, for instance, lead instrument administrators and practitioners to adopt eligibility criteria for a mental health program partially based on FFMQ scores that seem psychometrically sound but actually are unfair. As such, a more thorough and intuitive understanding of the nature of DIF is essential to a more informed decision pertaining to an underlying measure.

To this end, the proposed graphical framework can provide reasonably comprehensive and intelligible insights into the nature of both uniform DIF and nonuniform DIF. Further, by plotting items' ESCs, researchers and practitioners can easily examine if there is a pattern of DIF, which may not be evident through DIF parametric indices alone. For example, one of the patterns discussed in the results section was that the ESCs for women were steeper than the ESCs for men, suggesting that women were more likely than men to endorse extreme item responses for the items flagged as having nonuniform DIF (Figure 5), whereas such patterns were not apparent through the parametric test approach (Table 2). Not realizing patterns of item bias, at best, hinders the effort to make assessments more equitable for all individuals, regardless of their group memberships. Likewise, the

framework can also highlight outliers that exist in an instrument. For instance, if most items show similar response patterns across groups but one item deviates conspicuously, the item can be marked for further examination. This facilitates pinpointing sources of bias and understanding the impact of such bias on the overall instrument.

Moreover, the framework enhances communication about DIF results. By leveraging the intuitiveness of ESCs to detect DIF, the proposed framework makes the relatively esoteric psychometric concepts more accessible and easier to understand. As such, researchers can communicate their findings on DIF more effectively to, for example, practitioners, policymakers, and other stakeholders, thereby fostering broader discussions and ensuring that the implications of DIF are clearly understood. And a better understanding of DIF, in turn, facilitates more informed decisions, for example, about whether to modify, remove, or retain items. Therefore, we believe that there are opportunities for researchers across various fields to incorporate the graphical framework into their work. We also encourage research on determining DIF effect size criteria for ESCs to promote a deeper understanding of the extent of item bias.

Note

Supplemental material can be found at: https://reurl.cc/9WeKWa

References

Adams, R. J., Wu, M. L., Cloney, D., Berezner, A., & Wilson, M. R. (2020). *ACER ConQuest: Generalised item response modelling software* (Version 5) [Computer software]. Australian Council for Educational Research. https://www.acer.org/au/conquest

Baer, R. A., Samuel, D. B., & Lykins, E. L. (2011). Differential item functioning on the Five Facet Mindfulness Questionnaire is minimal in demographically matched meditators and nonmeditators. *Assessment*, *18*(1), 3–10.

Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment*, *13*(1), 27–45.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel–Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement*, *71*(6), 1023–1046.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education*, *14*(3), 235–259.

Penfield, R. D., Myers, N. D., & Wolfe, E. W. (2008). Methods for assessing item, step, and threshold invariance in polytomous items following the partial credit model. *Educational and Psychological Measurement*, *68*(5), 717–733.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*(3), 221–261.