Looking at DIF From a New Perspective: A Structure-Based Approach Acknowledging Inherent Indefinability

Anna Doeblér¹

Abstract

Differential item functioning (DIF), although highly relevant for psychometric assessment in various fields of psychology, is mathematically not well-defined. Especially the impact, the difference between the means of the person parameters in the focal and the reference group, is a parameter that is not identified without further assumptions. Common DIF detection methods necessarily impose such assumptions, however, in most cases the specific constraints remain quite vague and are implicit in the mathematical algorithms. As an alternative, a structure-based approach is proposed that is independent of the impact and, therefore, not affected by the identification problem. The approach allows to (a) reveal all DIF-relevant information from the data, (b) define indices that quantify the amount of DIF in a test as a whole, and (c) perform item-level DIF analyses. For the last application, it is unavoidable to assume additional identification constraints. However, compared to other existing methods (which also rely on such constraints), the approach suggested here provides advantages as it is not only easier to perform but also much more transparent concerning its underlying assumptions.

Keywords

differential item functioning (DIF), identification problem, item response theory (IRT), DIF structure, effect size

Introduction

According to the definition of differential item functioning (DIF), an item of a test is a DIF item if and only if test participants from different groups have different probabilities of solving this item correctly (if it is a dichotomous item) or answering in a specific category (if the item is polytomous), even though they have the same level of ability (e.g., Magis, Béland, Tuerlinckx, & De Boeck, 2010). Although this definition seems reasonable and is indeed widely used in practical testing, it has been shown that it cannot be sufficiently formalized due to an identification problem in the corresponding mathematical model (Bechger & Maris, 2015; Tutz &

¹University of Mannheim, Germany

Corresponding Author:

Anna Doeblér, University of Mannheim, L 13, 15 D-68161 Mannheim, Germany.
Email: adoebler@mail.uni-mannheim.de

Schauberger, 2015). In particular, this identification problem makes it impossible to view DIF as an absolute item property: If an item shows DIF or not does not depend on model fit, but can only be determined relative to an identification constraint, which in turn cannot be inferred from the data. These mathematical identification issues of DIF are discussed in more detail in the following section.

Identification problems are generally solved by imposing additional constraints. However, so far there is no general consensus about how additional constraints identifying DIF are to be made. Because DIF is a problem of high practical relevance, this leads to the unsatisfactory situation that on one hand there is a high demand for statistical methods to detect it, yet, on the other hand, these methods necessarily need to be insufficient. To make it worse, it seems that many practitioners are not even aware of the fundamental theoretical problem they are facing. The idea of DIF is easy to understand and to communicate, but the identification issue and the resulting problems are much less obvious.

Even though DIF is not well-defined, a huge amount of different DIF detection methods, covering item response theory (IRT)-based approaches (classic approaches include Lord, 1980; Raju, 1990; Thissen, Steinberg, & Wainer, 1988) as well as logistic regression (Swaminathan & Rogers, 1990) or Mantel-Haenszel strategies (Holland & Thayer, 1988) have been developed. More recent approaches include anchor item selection procedures (e.g., Kopf, Zeileis, & Strobl, 2015a, 2015b), item purification (e.g., Magis et al., 2010), mixture modeling (Frick, Strobl, & Zeileis, 2015), and model-based recursive partitioning (Strobl, Kopf, & Zeileis, 2015). Another class of methods uses nonlinear multiple indicators multiple cause (MIMIC) models (Jöreskog & Goldberger, 1975) for DIF detection (Müthen, 1985, 1989). However, application of these models also relies on anchor item assumptions or purification methods (e.g., Finch, 2005; Shih & Wang, 2009; Wang, Shih, & Yang, 2009; Woods & Grimm, 2011).

The reason why most of these approaches can indeed provide a distinction between DIF and non-DIF items is that the corresponding algorithms implicitly impose additional constraints, which identify the DIF concept. However, there are three major problems here: First, because different approaches generally use different constraints, the final assignments of DIF and non-DIF items often vary between different methods. Second, in most cases the DIF assignments do also depend on the specific items that are included in the analysis. Third, imposing additional constraints that are necessary for identification is equivalent of making (direct or indirect) assumptions about the so-called impact μ , which is the difference between the mean value of abilities in both groups. Because the value of the impact is of high interest, statistically and as a substantial measure of group-level differences, the exact consequences a certain DIF detection method has on the impact should be communicated to practitioners. However, because the constraints used in common approaches are often defined through iterative algorithms, this rarely is the case. Most often it remains unclear on which specific assumptions about the impact the final DIF assignments of single items are actually based.

Bechger and Maris (2015) recently proposed an approach for DIF analyses that stands out very positively in the DIF literature in that it explicitly takes the identification problem into account. These authors suggest to test for differential item pair functioning in the Rasch model, meaning that DIF is not evaluated for single items but rather for pairs of items. Yet, as Bechger and Maris (2015) stated in their discussion, this test is still inconsistent with the notion that DIF is the property of an item and, therefore, might be of little value to practitioners who are looking for DIF items.

The approach outlined in this article does not primarily focus on evaluating DIF as an item property, but rather as a concept referring to the test as a whole. It is based on a *DIF structure*, which will be defined below. This DIF structure is invariant of the chosen scale and, therefore, not affected by the identification problem. It can be used to address three main practical

applications: First, providing all DIF-relevant information that can possibly be inferred from the data in a very intuitive and easily comprehensible way. Second, quantifying the amount of DIF in a test as a whole. Third, allowing for traditional single-item DIF analyses (if these are still wanted) that are optimal regarding transparency of how the impact is defined and easiness of calculations.

Although DIF has also been analyzed in classical test theory (CTT), an IRT perspective will be adopted, as this allows for precise mathematical descriptions on both the test and the item level. Specifically, as interest is in uniform DIF, that is, DIF that affects only the difficulty parameters, the framework of the Rasch model is employed. Note, however, that the identification problem is independent of the specific measurement theory.

The remainder of this article is organized as follows: First, identification problem in a Rasch model accounting for DIF is shortly explained, thereby introducing the notation, which will be used throughout this work. Then the idea, estimation and a simple graphical representation of the DIF structure are described. Afterward, a detailed section covering the three applications of the DIF structure is explained. Finally, the author closes with a short discussion.

The Identification Problem

The Rasch DIF Model (R-DIF Model)

The typical situation underlying a uniform DIF analysis within binary IRT is described by the *Rasch DIF model* (R-DIF model), which is formally close to a nonlinear MIMIC model with factor loadings constrained to equality. For J persons, an indicator variable $x_j, j = 1, \dots, J$, is set to zero if the person is in the reference group, and to one if the person is in the focal group. Likewise, the answers to a test with I items are recorded by binary variables $Y_{ji}, j = 1, \dots, J, i = 1, \dots, I$.

The R-DIF model is defined by two distributional assumptions and a model equation. The distributional assumptions describe the marginal distributions of the person parameters in both the reference and the focal group on an assumed common scale:

$$\theta_R \sim D_R(\eta_R) \text{ and } \theta_F \sim D_F(\eta_F). \quad (1)$$

Here θ_R is the person parameter variable in the reference group, θ_F the person parameter variable in the focal group and D_R, D_F are the corresponding distributions with parameters η_R and η_F . The approach is not limited to a specific class of distributions, yet, for simplicity of exposition, normal marginal distributions are assumed. Without loss of generality, the expectation in the reference group is set to zero:

$$\theta_R \sim N(0, \sigma_R^2) \text{ and } \theta_F \sim N(\mu, \sigma_F^2). \quad (2)$$

The model equation defines the probability of a positively scored response:

$$P(Y_{ji} = 1) = \frac{\exp(\theta_j - (\beta_i + x_j \Delta \beta_i))}{1 + \exp(\theta_j - (\beta_i + x_j \Delta \beta_i))}. \quad (3)$$

Local independence is assumed, that is, statistical independence of item answers conditional on a given person parameter. The expectation in the focal group, μ , equals the difference between the expectations in both groups, which is commonly referred to as the impact. The DIF-free items are exactly those items with $\Delta \beta_i = 0$. The quantity $\exp(-\Delta \beta_i)$ has a natural interpretation as an odds ratio: For any θ_j ,

$$\frac{P(Y_{ji}=1|x_i=1)}{1 - P(Y_{ji}=1|x_i=1)} \frac{1 - P(Y_{ji}=1|x_i=0)}{P(Y_{ji}=1|x_i=0)} = \exp(-\Delta\beta_i). \quad (4)$$

An alternative parametrization of the R-DIF model is given by,

$$P(Y_{ji}=1) = \begin{cases} \frac{\exp(\theta_j - \beta_i^R)}{1 + \exp(\theta_j - \beta_i^R)} & \text{if } x_j = 0, \\ \frac{\exp(\theta_j - \beta_i^F)}{1 + \exp(\theta_j - \beta_i^F)} & \text{if } x_j = 1. \end{cases} \quad (5)$$

Here β_i^R and β_i^F are the difficulties of the i th item in the reference and focal group, respectively, and the original parametrization is obtained via $\Delta\beta_i = \beta_i^F - \beta_i^R$ and $\beta_i = \beta_i^R$.

The marginal likelihood for the data from person j is given by

$$L_j = L_j(\beta, \Delta\beta, \mu, \sigma_R^2, \sigma_F^2) = \int \prod_{i=1}^I P(Y_{ji}=y_{ji}) \varphi(\theta_j, x_j) d\theta_j. \quad (6)$$

Here φ is the appropriate marginal distribution, that is, the density of D_R or D_F . For the complete data matrix, the marginal likelihood is then $\prod_{j=1}^J L_j$. Using the alternative parametrization, it can be seen that the likelihood factors in a part for the reference and a part for the focal group if a μ is fixed before estimation:

$$L^R = \prod_{j \in \{j^* | x_{j^*} = 0\}} L_j \text{ and } L^F = \prod_{j \in \{j^* | x_{j^*} = 1\}} L_j. \quad (7)$$

Thus, the marginal likelihood can be maximized by maximizing each of the factors L^R and L^F , conditional on some μ .

Identification of the R-DIF Model

Detecting DIF in practical applications is challenging, and there is an enormous amount of literature dealing with the best strategies for corresponding analyses (Osterlind & Everson, 2009). The main reason for the trouble is that the R-DIF model is not identified. Inspecting Equations 2 and 3, it becomes clear that adding any arbitrary number c to the impact μ (i.e., the difference between the expectations in the reference and the focal group) as well as to all the $\Delta\beta_i$, does not change the probability of a positively scored response: For a person of the reference group, this is obvious, as neither μ nor $\Delta\beta_i$ are needed for calculations. For a person of the focal group, it follows because the term $(\theta_j - (\beta_i + x_j \Delta\beta_i))$ stays the same when θ_j as well as $\Delta\beta_i$ are simultaneously increased by c . Therefore, the likelihood of the R-DIF model always stays the same if μ is shifted by the same amount as all the $\Delta\beta_i$.

For the detection of DIF items one cannot simply fix μ : Any given μ leads to a specific set of estimated $\Delta\beta_i$. These estimates of $\Delta\beta_i$ allow for a division into DIF and non-DIF items, where the former are the items with $\Delta\beta_i \neq 0$ and the latter those with $\Delta\beta_i = 0$. However, using a different value for the impact, for example, one that differs by a certain amount c from the previous one, leads to estimates of the $\Delta\beta_i$, which differ exactly by this amount c from those obtained before. Therefore, items formerly seen as DIF-free (because corresponding estimates of $\Delta\beta_i$ were zero) are now DIF items, as their estimated $\Delta\beta_i$ is now c . Likewise, items first identified as DIF items might now be DIF-free (if their former estimate of $\Delta\beta_i$ was $-c$) or have at least a different amount of DIF, that is, a different $\Delta\beta_i$. Thus, determining the value μ is

most crucial for any DIF analysis, because it directly determines the distinction between DIF and DIF-free items (cf. Tutz & Schauberger, 2015).

Common DIF analyses based on IRT often determine μ by choosing so-called *anchor items*, which are items that are assumed to be DIF-free. Indeed, anchor items solve the identification problem because they uniquely determine a value of μ : One can simply estimate μ from the R-DIF model by setting $\Delta\beta_i$ equal to zero for all anchor items. Once μ is estimated, assessing the DIF or no-DIF question for any other item reduces to applying an appropriate significance test. Theoretically, even a single anchor item would suffice. However, because of the identification problem, determining anchor items is not straightforward at all and, at least from a mathematical point of view, in no way easier than determining the DIF and non-DIF items directly. Again, it is only possible if the true μ could be known *before* the analysis. Yet, this resembles the attempt to drag oneself up by one's own bootstraps.

The DIF Structure

Preliminary Remarks

Basically, there are two steps to be carried out in any DIF analysis that aims to detect the DIF and non-DIF items of a test: First, defining the impact μ . Second, testing each item if it displays DIF or not by using an appropriate significance test.¹ Although the second step (once the first one has been completed) is more or less straightforward, this is not at all true for the first step, as described in the last section.

The new approach for DIF analyses that is described in the following focuses on the first step. However, it differs from common DIF analyses in that it does not attempt to determine the *true* μ (whose existence is indeed quite doubtful, not only because it is impossible to derive it from the data, but also because of epistemological considerations). Rather, it aims to uncover the underlying DIF *structure* in the data, which is invariant of the chosen value for μ .

Defining and Estimating the DIF Structure

The basic idea for the DIF structure is that, although the true μ is needed to estimate the true $\Delta\beta_i$, the structure of the $\Delta\beta_i$ is obtained, nevertheless, by estimating them with any arbitrary chosen value for μ . Estimating the $\Delta\beta_i$ under the assumption of two different values for μ results in estimates, which differ only in an additive constant that is given by the difference of these two values for μ . Hence, if the estimated values of the $\Delta\beta_i$ are displayed on the real line, then a shift in μ only causes a common corresponding shift of all the $\hat{\Delta\beta}_i$. Particularly, for each single item the assumed value of μ can be shifted in such a way that the item becomes DIF-free, that is, such that its $\hat{\Delta\beta}_i$ equals zero (in this case μ equals the expectation of the focal group).

Therefore, to estimate the DIF structure parameter estimation of the R-DIF model is carried out conditional on a fixed but arbitrary value for μ . To obtain a structure, which is especially intuitive to interpret, μ is set to zero. To identify the scale for the person and difficulty parameters, it is further necessary to fix a reference point, for example, by fixing the mean of the person parameters in the reference group. Following the notation in Equation 2, the mean of the person parameters in the reference group is set to zero.

With these assumptions, marginal maximum likelihood (MML) estimation can be used to obtain the difficulty parameters and variances of person parameters in the reference and the focal group. For any given μ this is possible, because the marginal likelihood is the product of two independent factors.

Estimates of the β_i^R and the β_i^F from Equation 5 result, and estimates of the $\Delta\beta_i$ are obtained by itemwise subtraction:

$$\hat{\Delta}\beta_i = \hat{\beta}_i^F - \hat{\beta}_i^R. \quad (8)$$

The standard errors are calculated from the standard errors of $\hat{\beta}_i^F$ and $\hat{\beta}_i^R$:

$$s_{\Delta\beta_i} = \sqrt{s_{\beta_i^F}^2 + s_{\beta_i^R}^2}. \quad (9)$$

The collection of the $\hat{\Delta}\beta_i$, together with their confidence intervals, provide an estimate of the DIF structure. An easy way to display this DIF structure graphically is shown in the upper part of Figure 1 for four hypothetical items. To illustrate that the DIF structure (the absolute differences between the $\hat{\Delta}\beta_i$ and the intersections of their confidence intervals) is indeed invariant under different identification constraints, the middle and lower part of Figure 1 show the estimated DIF structure for the same items using different values for μ . All three subfigures look very much alike, the only difference is the scale of the x axis, which is shifted to the right or to the left, according to the sign and absolute value of the chosen value for μ . However, this difference is quite important because it directly determines which of the items show DIF. These are exactly those, whose confidence intervals do not include zero, which applies to different items for different identification constraints.

Three Applications of the DIF Structure

In this section, three applications of the DIF structure are described that can be efficiently used to perform an overall DIF analysis. The first two, revealing all DIF-relevant information and quantifying DIF through appropriate indices, are invariant under the identification problem as they do not rely on the assumption that DIF is a property of single items (which it cannot be without further constraints). The third application, which aims to assess DIF at the single-item level, can, obviously, only be done by imposing additional restrictions on μ .

Revealing All DIF-Relevant Information

A simple inspection of the DIF structure reveals a lot about the functioning of the test items in the two groups, indeed, from a formal point of view all essential aspects of DIF can be inferred from it. First, because μ was set to zero during estimation, $-\hat{\Delta}\beta_i$ directly gives the estimated difference in the expected means of person parameters in the two groups, that is, the parameter μ , which would be obtained if the specific item was used (or all items with the same value of $\hat{\Delta}\beta_i$) as an anchor. Anchor items with $\hat{\Delta}\beta_i = 0$ predict no difference between the means of the reference and the focal group, anchor items with negative $\hat{\Delta}\beta_i$ imply a positive μ , and those with positive $\hat{\Delta}\beta_i$ predict a negative μ . Second, the higher the absolute value of a $\hat{\Delta}\beta_i$, the more the participants from the two groups differ, on average, in their responses to the corresponding item. This might be due to DIF or due to a nonzero impact μ , or to both. Third, $\hat{\Delta}\beta_i$ was examined to see if there are groups of items that have equal or at least very similar values of $\hat{\Delta}\beta_i$, indicating that these groups might be suitable anchor item sets or constitute subtests free of DIF. Such groups can be identified using the confidence intervals for the $\hat{\Delta}\beta_i$. For each possible assumed true value μ_r of μ , the group of items that can be assumed to be free of DIF consists of all items whose confidence interval of $\hat{\Delta}\beta_i$ includes $-\mu_r$. Specifically, a nonempty intersection of the confidence intervals of all $\hat{\Delta}\beta_i$ is an indicator that the test as a whole is DIF-free. However, because this is a multiple testing problem, it might be reasonable to adjust for the

Figure 1. Illustration of the DIF structure for an hypothetical example of four items and for three different identification constraints ($\mu = 0$, $\mu = 1$, $\mu = -1$).

Note. Although the structure of the estimated $\Delta\beta_i$ and their confidence intervals remains the same, this is not true for the assignment of DIF. DIF items are as follows: For $\mu = 0$, Items 2 and 3; for $\mu = 1$, Items 1, 3, and 4; and for $\mu = -1$ Items 2 and 4. DIF = differential item functioning.

size of the confidence intervals, for example, by a Bonferroni correction. Furthermore, confidence intervals corresponding to any other desirable hypothesis test can also be used.

As an illustrative example, the dichotomized version of the verbal aggression data was used (Smits, De Boeck, & Vansteelandt, 2004; Vansteelandt, 2000), which has been extensively reanalyzed (e.g., De Boeck & Wilson, 2004; Magis et al., 2010). It contains item-level data of self-reported verbal aggression and reports the outcomes of 316 respondents (243 women and 73 men), each answering 24 items. The reference group consists of the female, the focal group of the male respondents. On the whole, men endorse 1.17 items more, but the difference is not significant on the level of the total scores (two-sided Welch $t = -1.61$ on ≈ 125.8 df , $p = .11$).

For the DIF structure analysis, the values of the $\hat{\Delta}\beta_i$ were calculated as described above, thereby identifying the model by setting $\mu = 0$. Using the inverses of the squared standard errors as weights, a weighted mean of the $\hat{\Delta}\beta_i$ of -0.301 (= intercept of weighted least squares; $t = -2.54$, $p = .018$) was determined, affirming the difference observed in the total scores of women and men.

Figure 2. Graphical representations of the DIF structure of the verbal aggression data.
Note. A DIF structure plot (top) and a difficulty-DIF plot (bottom) are shown, and in both plots the bigger the square, the lower the standard error of $\hat{\Delta\beta}_i$. Items with a ★ in the top panel are flagged for DIF by the logistic regression method, Lord's χ^2 , and Raju's area method (cf. Magis, Béland, Tuerlinckx, & De Boeck, 2010). DIF = differential item functioning.

The graphical representation of the DIF structure is inspired by meta-analysis, where this type of plot is known as a *forest plot*. Figure 2 shows the DIF structure for the verbal aggression data on the left and a difficulty-DIF plot on the right. *R* code using the packages TAM (Kiefer,

Robitzsch, & Wu, 2014) and metafor (Viechtbauer, 2010) for estimating the DIF structure of this data set and reproduction of Figure 2 is in the online appendix.

Each row of the DIF structure plot on the top half of Figure 2 refers to a single item, where the name of the item is on the left, followed by a 95% confidence interval for the corresponding $\hat{\Delta}\beta_i$. The size of the filled squares used to mark $\hat{\Delta}\beta_i$ is proportional to the inverse of the squared standard error of $\hat{\Delta}\beta_i$. The estimate $\hat{\Delta}\beta_i$ and its confidence interval are also given numerically. All test items are ordered by their value of $\hat{\Delta}\beta_i$. The collection of the $\hat{\Delta}\beta_i$ is clearly heterogeneous (there are confidence intervals that do not overlap), which is confirmed by a large Cochran Q of 69.8 ($df = 23$, $p < .0001$). Negative $\hat{\Delta}\beta_i$ indicate that men are more likely to endorse the item, while positive $\hat{\Delta}\beta_i$ suggest that women agree to the item more often. Note that if the R-DIF model was identified by assuming no differences between the expectations for women and men, that is, by setting $\mu = 0$, then more DIF items are obtained (whose confidence intervals do not include zero) than, for example, with the logistic regression method, Lord's χ^2 or Raju's area method (items flagged as DIF with these methods are marked with a ★ in Figure 2, the R-DIF model is here implicitly identified with a corresponding impact of $\mu \approx -0.25$).

Inspecting the DIF structure plot immediately delivers clues for content-related aspects of DIF in this data set: Items with large values of $\hat{\Delta}\beta_i$ are those which ask respondents if they would *want* to do a certain action (e.g., item S2WantShout: "I miss a train because a clerk gave me faulty information. I would want to shout."). Items with low values of $\hat{\Delta}\beta_i$ ask if the respondent would actually *do* a certain action (e.g., item S2DoCurse: "I miss a train because a clerk gave me faulty information. I would curse."). Furthermore, items with large $\hat{\Delta}\beta_i$ more often ask for actions involving *shouting*, while many items with low values of $\hat{\Delta}\beta_i$ ask for scolding. Thus, men more often report that they would actually behave aggressively, whereas women state that they would want to be aggressive but would probably not be. At the same time, men are more likely to scold, women more likely to shout.

The difficulty-DIF plot in the bottom half of Figure 2 allows to see whether there is any systematic relationship between DIF and the estimated difficulty parameters in the reference group. As the plot is roughly symmetric, this does not seem to be the case for the verbal aggression data. Random-effects metaregression of the $\hat{\Delta}\beta_i$ on the difficulties in the reference group can be used to investigate the asymmetry of a difficulty-DIF plot, similar to the random-effects variant of Egger's regression test for funnel plot asymmetry in meta-analysis (cf. Sterne & Egger, 2005). For the verbal aggression data, there is no evidence of asymmetry ($p = .255$).

Graphical representations of the DIF structure, like the one in Figure 2, are quite intuitive and can easily be interpreted by practitioners. Of course, they still cannot solve the problem of defining DIF at the single-item level, as the x axis (whose zero point determines which items show DIF and which do not) can be shifted to the left or to the right ad libitum, depending on the identification constraint used for μ (cf. Figure 1). Nevertheless, all other DIF-related aspects that are not affected by the identification problem can be inferred from the DIF structure: First, representations such as Figure 2 provide a starting point for content-related questions such as the following: What have items in common that have similar $\hat{\Delta}\beta_i$ values? Is there a reason for it? What tells them apart from items with different $\hat{\Delta}\beta_i$? These questions might uncover aspects of the data that are highly relevant for the theory of the concept that the items are supposed to measure. Second, plots of the DIF structure directly show the consequences of choosing a specific value for μ in a DIF analysis: A specific value of μ corresponds to a vertical line through $-\mu$ and if this line intersects a confidence interval, the corresponding item would be tested DIF-free under the assumption that this μ is the true impact. For all other items, the further away their $\hat{\Delta}\beta_i$ values are from $-\mu$, the higher their amount of DIF. Furthermore, the sign of DIF, that is, which group is favored, also becomes immediately obvious.

DIF Indices—Quantifying the Amount of DIF in a Test

Although DIF is not well-defined at the single-item level, it is nevertheless possible to define DIF on the test level. According to the DIF structure, a test can be defined as DIF-free if and only if all $\Delta\beta_i$ have the same value. In this case, μ can be set to $-\Delta\beta_i$. One could still argue that it might be equally plausible that the true μ would in fact be some other (arbitrary) value and that *all* items favor one of the two groups to exactly the same amount, that is, that *all* items show exactly the same amount of DIF. Yet, the above definition is the best can be done when a well-defined notion of test DIF is needed, that is, one that is not affected by the identification problem.

There are several possibilities for testing the null hypothesis that all items have the same value of $\Delta\beta_i$. Bechger and Maris (2015) have proposed the squared Mahalanobis distance $\chi^2_{\Delta R}$, which is not directly calculated from the $\hat{\Delta}\beta_i$, but from the estimated pairwise differences between the difficulties of the test items and the difficulty of an arbitrary chosen reference item, calculated separately in both groups, and a (generally nondiagonal) covariance matrix. Alternatively, (Bonferroni-corrected) confidence intervals according to a suitable significance test can be calculated. Under the null hypothesis that there is no DIF, these would have a nonempty intersection.

However, in many applications a given test will show significant DIF and in these cases it might be informative to quantify the *amount of DIF*. Four statistics are proposed that should be useful for this purpose.

The MN(T) statistic. To motivate the first index, consider the following two definitions:

Definition 1: Let T be a significance test that is suitable for testing single items for DIF when the value of μ in the R-DIF model is known. An item is called (μ^*, T) -DIF-free if and only if the item is considered DIF-free when μ^* is assumed to be the true μ and T is used for significance testing.

Definition 2: Given an appropriate significance test T , a set of items A is called T -DIF-homogeneous if and only if there exists a value μ^* such that all items in A are (μ^*, T) -DIF-free. μ^* is called a witness of the T -DIF-homogeneity of A .

The most important example of a T that could be used for testing is the two-sided Wald test for $H_0 : \Delta\beta_i = -\mu^*$. Having a graphical representation of the DIF structure in mind, where the confidence intervals of the $\Delta\beta_i$ are dual to the significance test T , a set of items is T -DIF-homogeneous if and only if the intersection of the confidence intervals of the corresponding $\hat{\Delta}\beta_i$ is not empty (a witness μ^* can then be defined by choosing an arbitrary value x from this intersection and setting $\mu^* = -x$). Because the above definition does not exclude the case that a T -DIF-homogeneous set comprises just a single item, the set of all items of a test can always be partitioned into a set of T -DIF-homogeneous subsets.

If the items of a test as a whole form a T -DIF-homogeneous set this indicates the absence of any DIF. On the contrary, if the only T -DIF-homogeneous subsets of test items are singletons, all items function differently to a very high degree so that there are not even two items, which could be tested DIF-free. In real applications, neither of these two extreme cases is likely to occur. Instead, one can expect that the test items can be partitioned into T -DIF-homogeneous sets, where one expects this partition to consist of *fewer sets when we have low DIF* in the data and of *more sets as the amount of DIF in the data increases*.

A partition of all items of a test with T -DIF-homogeneous sets is generally not unique, however, the *minimal number* of T -DIF-homogeneous sets that are needed for this purpose is an invariant:

Theorem 1: *For a given test conforming to the R-DIF model and a significance test T with dual two-sided confidence intervals, the minimal number k of T -DIF-homogeneous sets that are necessary to cover all items can be calculated with the following algorithm: Calculate the DIF structure, that is, all $\hat{\Delta}\beta_i$ and their confidence intervals corresponding to T . Set $n=1$ and let A_1 be the set of all test items. Repeat the following until the algorithm terminates: If the intersection of the confidence intervals of the $\hat{\Delta}\beta_i$ of items in A_n is not empty, set $k=n$ and stop the algorithm. Otherwise, let L_n be the set of all left end points of all confidence intervals of the $\hat{\Delta}\beta_i$ corresponding to items in A_n .*

- i. Choose $\mu_n := -\max(L_n)$.
- ii. Let $B_n \subset A_n$ maximal such that B_n is T -DIF-homogeneous as witnessed by μ_n .
- iii. Let $A_{n+1} := A_n/B_n$.
- iv. Increase n by one.

Proof: Assume the algorithm determines the value k and assume, toward a contradiction, that there are $k-1$ T -DIF-homogeneous sets that already cover all items of the test. Let $\mu_1 < \dots < \mu_k$ be the witnesses of the T -DIF-homogeneity of the sets determined by the above algorithm and let $\tilde{\mu}_1 < \dots < \tilde{\mu}_{k-1}$ be the witnesses of the assumed smaller T -DIF-homogeneous cover. Because the item in A_1 with the largest left end point of the corresponding confidence interval of its $\hat{\Delta}\beta_i$ needs to be covered in at least one of the T -DIF-homogeneous sets witnessed by the $\tilde{\mu}_1 < \dots < \tilde{\mu}_{k-1}$, and because $-\mu_1$ is defined as exactly this left end point of the items confidence interval, it is concluded that $-\tilde{\mu}_1 \geq -\mu_1$ and, hence, $\tilde{\mu}_1 \leq \mu_1$. Analogously, also the item in A_2 with the smallest left end point of the corresponding confidence interval of its $\hat{\Delta}\beta_i$ needs to be covered as witnessed by one of the $\tilde{\mu}_1 < \dots < \tilde{\mu}_{k-1}$. Because $\tilde{\mu}_1 \leq \mu_1$, the T -DIF-homogeneity of this item is not witnessed by $\tilde{\mu}_1$. Because $-\mu_2$ is exactly the left end point of the item's confidence interval, it is concluded that $\tilde{\mu}_2 \leq \mu_2$. By induction, $\tilde{\mu}_i \leq \mu_i$ for all $i=1, \dots, k-1$. This leads to a contradiction as it implies that there is no witness among the $\tilde{\mu}_1 < \dots < \tilde{\mu}_{k-1}$ for any item in A_k .

With these preparations, a unique DIF statistic is defined as follows:

Definition 3: *Given a significance test T , $MN(T)$, the minimal number index with respect to T , is defined as the minimal number of T -DIF-homogeneous sets that are needed to cover all items.*

The $MN(T)$ statistic not only provides a quantification of the amount of DIF in a test, it might also be useful when the DIF structure is used to analyze content-related aspects of the concept the test is supposed to measure, for example, when the concept comprises different subcomponents. In addition to techniques such as (exploratory) multidimensional IRT (Reckase, 2009), the $MN(T)$ index can offer further clues to determine the number of these components. The $MN(T)$ index is likely to provide relevant information that is not yet captured by other methods.

The $ML(T)$ statistic. Although the $MN(T)$ statistic of a test is intuitively interpretable and interesting for exploring content-related structure in the data, it can happen that two tests have the same $MN(T)$ but that the witnesses of the corresponding T -DIF-homogeneous sets are close to one another for the first test but far away from each other for the second (at least on average, witnesses are generally not unique). In this case, a statistic that assigns a larger amount of DIF to the second test is of interest.

Definition 4: Given a significance test T , $ML(T)$, the minimal length index with respect to T , is defined as the length of the minimal interval I such that for each item there exists a $\mu^* \in I$ so that the item is $(-\mu^*, T)$ -DIF-free.

If T is dual to a two-sided confidence interval, the left end point of the minimal interval is given by the minimum of all right end points of all confidence intervals corresponding to all the $\hat{\Delta}\beta_i$ of all the items in the test, whereas the right end point is defined by the maximum of all left end points of all those confidence intervals. Therefore, the $ML(T)$ statistic is unique and straightforward to calculate once the DIF structure has been estimated.

Inspired by the definition of $ML(T)$, the idea of an interval of μ values might also be appealing for further considerations. It is mentioned that there are two necessary steps in each procedure that aims to detect DIF at the single-item level, determining μ and, subsequently, conducting appropriate significance tests. However, as elaborated in more detail in the last section, the existence of a true μ is more than doubtful. Therefore, it might be a reasonable alternative to replace the first step, determining a specific value for μ , by defining an appropriate interval in which μ is allowed to vary.

$ML(T)$ could also be used to obtain a less restrictive definition of test DIF, that is, one that allows to consider more tests as DIF-free: One can evaluate $ML(T)$ and set it in relation to the standard deviation of the person parameters. Then a test could be regarded as DIF-free if the corresponding ratio does not exceed a predefined small value, depending on how liberal one is willing to be with DIF.

The DIF- I^2 statistic. The I^2 statistic (Higgins & Thompson, 2002) is commonly reported in meta-analysis to quantify the amount of heterogeneity in a collection of studies. The very same statistic applied to a DIF structure quantifies the amount of inconsistency in a collection of $\hat{\Delta}\beta_i$. Let $s_{\Delta\beta}^2$ be the mean squared standard error of the $\hat{\Delta}\beta_i$ and let τ^2 be the variance of the $\hat{\Delta}\beta_i$. Both quantities are independent of the choice of the impact μ .

Definition 5: The I_{DIF}^2 statistic is defined by

$$I_{DIF}^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{s}_{\Delta\beta}^2}, \quad (10)$$

where $\hat{\tau}^2$ and $\hat{s}_{\Delta\beta}^2$ are estimators of the respective parameters.

I_{DIF}^2 is similar to an intraclass correlation coefficient. Note that I_{DIF}^2 is independent of a specific scaling: A linear transformation of the difficulty parameters will influence any reasonable estimates for τ^2 and $s_{\Delta\beta}^2$ by the same factor, and thus I_{DIF}^2 will stay the same. Values of I_{DIF}^2 are in $[0, 1)$, where higher values represent a large inconsistency in the DIF structure.

Given an estimated DIF structure, a wide variety of well-studied estimators of $\hat{\tau}^2$ and $\hat{s}_{\Delta\beta}^2$ are available from the meta-analysis literature (Sánchez-Meca & Marín-Martínez, 2008; Viechtbauer, 2007). Viechtbauer (2005) recommendation is followed to use the restricted maximum likelihood (REML) estimator for τ^2 and the estimator proposed by Higgins and Thompson (2002) for $s_{\Delta\beta}^2$:

$$\hat{s}_{\Delta\beta}^2 = \frac{(I - 1) \sum_i w_i}{(\sum_i w_i)^2 - \sum_i w_i^2}, \quad (11)$$

where I is the number of items and $w_i = 1/s_{\Delta\beta_i}^2$ are inverse variance weights for the $\hat{\Delta}\beta_i$.

Note that the REML estimator of τ^2 is not available in closed form, and thus the DerSimonian and Laird (1986) moment estimator for τ^2 has seen more widespread use. Plugging in the DerSimonian and Laird (1986) estimator into Equation 10 leads to the variant

$$I_{DIF}^{2*} = \frac{\max(0, Q - (I - 1))}{Q}, \quad (12)$$

where Q is the Cochran Q -statistic given by

$$Q = \sum_i \frac{(\hat{\Delta}\beta_i - \Delta)^2}{s_{\Delta\beta_i}^2} \quad (13)$$

and $\Delta = (\sum_i w_i \hat{\Delta}\beta_i) / (\sum_i w_i)$ is the inverse variance weighted mean of the $\hat{\Delta}\beta_i$. Note that if the true τ^2 is 0, Q has expectation $I - 1$ if the estimates for the individual $\Delta\beta_i$ are independent and normally distributed, which is approximately the case for MML estimation. Thus, I_{DIF}^{2*} can be interpreted as a ratio of excess heterogeneity and total heterogeneity in the DIF structure.

The standardized DIF variance statistic. The last index proposed compares the variance of the $\Delta\beta_i$, which is a measure of the amount of DIF in the test, to the mean of the variances of the person parameters in the reference and the focal group:

Definition 6: *The standardized DIF variance (S-DIF-V) of a test is defined as*

$$\text{S-DIF-V} = \frac{(J_R + J_F - 2)\hat{\tau}^2}{(J_F - 1)\hat{\sigma}_F^2 + (J_R - 1)\hat{\sigma}_R^2}, \quad (14)$$

where $\hat{\tau}^2$ is the REML estimator of the variance of the $\Delta\beta_i$, J_R and J_F are the sample sizes in the reference and focal groups, respectively, and $\hat{\sigma}_F^2$ and $\hat{\sigma}_R^2$ are estimates of the variance of the person parameters.

Standardizing by the variance of only one of the groups yields a variant of this effect size that might be useful if the variance in the focal group differs a lot from the variance in the reference group:

$$\text{RS-DIF-V} = \frac{\hat{\tau}^2}{\hat{\sigma}_R^2} \quad \text{and} \quad \text{FS-DIF-V} = \frac{\hat{\tau}^2}{\hat{\sigma}_F^2}. \quad (15)$$

In contrast to MN and I^2 the standardized DIF variance and ML are effect sizes in the sense that they are (mostly) independent of the sample size (though random sampling error of course yields different values of S-DIF-V).

Computation of the four DIF indices for the verbal aggression. Table 1 shows the values of the proposed DIF indices for the verbal aggression data. Wald statistic was used to compute MN(T) and ML(T). The MN(Wald) statistic is 2, indicating that there might be at least two clusters of differentially functioning items. The value for ML(Wald) of 0.680 is the difference between the upper bound of the Wald confidence interval for $\hat{\Delta}\beta_{\text{S2DoScold}}$ and the lower bound of the interval for $\hat{\Delta}\beta_{\text{S2WantShout}}$. On the Rasch scale, this means that at least some of the items are quite far away from each other in terms of DIF. A considerable amount of DIF in the data is also confirmed by $I_{DIF}^2 = .669$. Because $\hat{\tau}_{REML}^2 = 0.223$ (SE = 0.099), the I_{DIF}^2 is clearly nonzero. The standardized DIF variance of 0.114 (with a slightly lower value when standardized by the reference

Table 1. DIF Statistics for the Verbal Aggression Data.

Statistic	Value
MN(Wald)	2
ML(Wald)	0.680
I^2_{DIF}	0.669
S-DIF-V	0.114
RS-DIF-V	0.109
FS-DIF-V	0.137

Note. DIF = differential item functioning; FS-DIF-V = focal group standardized DIF variance; ML = minimal length; MN = minimal number; RS-DIF-V = reference group standardized DIF variance; S-DIF-V = standardized DIF variance.

group and a slightly higher value for focal group standardization) indicates that DIF should be accounted for, as otherwise a substantial noise in person parameter estimates would result.

It is tempting to suggest specific values for the DIF statistics that characterize small, medium, and large amounts of DIF. However, giving sensible guidelines requires the careful evaluation of these indices in many real data sets, which would go beyond the scope of this article. Furthermore, all indices besides the standardized DIF variance also depend on the sample size, which should be taken into account in the interpretation. Therefore, more refined analyses and suggestions for possible reference values are left to further research.

Assessing DIF at the Single-Item Level

In the following, how the DIF structure can be used for traditional DIF analyses aiming to detect DIF at the single-item level is discussed. As a consequence of the identification problem this is only possible if additional constraints are assumed. In practical applications, this fact might raise serious doubts about the usefulness of the approach. However, this section was added due to two reasons: First, single-item DIF analyses have a long tradition, therefore, it is most likely that they will still be performed in the future. Second, there will still be situations where reasonable identification constraints can be made. For example, considering an university entrance test, where the main purpose of a DIF analysis is not to find out if men and women show different abilities in certain areas of expertise, but rather to make sure that they have equal chances to pass the test. In this case, the impact can be set to zero, which immediately solves the identification problem and which ensures that the test is fair.

As discussed above, there are two basic steps in any analysis focusing on DIF at the single-item level: Defining the impact μ , that is, solving the identification problem and, subsequently, performing appropriate significance tests. Because it is the first of these steps that is challenging, some exemplary suggestions are made to identify appropriate values for μ :

- Choosing $-\mu$ such that it equals the mean of all $\hat{\Delta}\beta_i$.
- Choosing $-\mu$ as a weighted sum of all the $\hat{\Delta}\beta_i$ (for example, with weights corresponding to their magnitudes of estimation error).
- Choosing μ such that a maximal number of items in the test are DIF-free. (This depends on the specific confidence intervals to calculate or other test statistics used for significance testing.)
- Discarding items with extreme values of $\hat{\Delta}\beta_i$ and then setting $-\mu$ to the trimmed mean, that is, the mean or weighted mean of the remaining items.

- Choosing μ such that corresponding estimates of the $\hat{\Delta}\beta_i$ minimize or maximize a certain target function.
- Choosing the value of μ to any arbitrary value. For example, μ was chosen to be zero to ensure test fairness between groups.

Single-item DIF analyses that are based on the DIF structure approach have two main advantages compared with other existing methods: First, it is made explicit on which identification constraint, that is, on which assumption about μ , the analysis is based on. Second, calculations become very simple as no iterative anchor item selection strategies or item purification procedures are needed. All there is to do is to fit the R-DIF model using the chosen identification constraint for μ and to calculate confidence intervals or intervals corresponding to some specified significance test.

Discussion

The identification problem of the R-DIF model makes it clear that the data alone can never tell us the whole story about DIF. For every data set, there are infinite options for μ , all implying the same likelihood of the data, but generally implying different DIF items. For any DIF detection procedure that aims to detect DIF at the single-item level, additional assumptions necessarily need to be made.

The approach proposed here explicitly takes the identification problem into account. Instead of searching for something that cannot be derived from the data, it uncovers the structure of the differential functioning of the test items. How to estimate this DIF structure and how it can be used in three different ways to perform DIF analyses are described.

First, the DIF structure reveals all essential DIF-relevant information that can be obtained from the data. It is sufficient in the sense that if it is assumed that the Rasch model holds in both groups and looking for uniform DIF, then there is no way to get more information about DIF out of the data than the DIF structure already reveals. It can easily be displayed graphically, thereby very intuitively revealing important aspects, for example, whether the test as a whole displays DIF, whether there is differential item pair functioning (Bechger & Maris, 2015), whether a group of items forms a cluster of items that could be considered DIF-free (when an appropriate identification constraint is used), whether there might be different such item clusters, whether an item favors the focal or the reference group, how items differ in their amount of DIF, and so on.

Second, the DIF structure can be used to define indices quantifying the amount of DIF in a test as a whole. In contrast to the definition of DIF at the item level, which lacks identifiability because of the identification problem, these indices do not depend on the assumption of a true μ . They are all derived from the invariant DIF structure and are, therefore, well-defined. The four indices proposed are meaningful measures of what the author intuitively has in mind when thinking about DIF.

Third, how the DIF structure can be utilized to perform traditional DIF analyses aiming to assess DIF at the item level is discussed. Several possible identification constraints are proposed, which might be useful in different situations. Calculations for the new approach are quite easy, as no complicated iterative algorithms are needed.

Comparing common DIF detection methods to the approach suggested here, the first thing to notice is that the traditional methods are generally limited to the third aspect, that is, they only focus on detecting DIF at the item level. Although it remains questionable if this is reasonable at all, as DIF cannot be defined at the item level, it is still worthwhile to compare these common methods to the approach suggested here for item-level analysis based on the DIF

structure, that is, the third application of the DIF structure. The new approach has the advantage that it is explicitly stated which identification constraints are used and that these constraints can be consciously chosen by the practitioner. This is in clear contrast to most other procedures, where the identification is not made explicit (by stating on which concrete assumptions about μ the DIF analysis is based on) but only implicit through the mathematical algorithms. More specifically, most procedures aim at determining anchor items, which in turn identify the R-DIF model. Many anchor item selection procedures have been proposed (Kopf et al., 2015a, 2015b), which typically rely on a ranking of items according to how likely they are DIF items. This ranking is usually based on an auxiliary DIF analysis, which is calculated using the whole test as an anchor.² There are different possibilities for defining the exact ranking of the items as well as of finally choosing the anchor (based on this ranking). Therefore, as each anchor identifies the R-DIF model, several DIF detection strategies result.

A related approach for assessing DIF at the item level, which also identifies the R-DIF model by fixing a set of anchor items, is an iterative approach called *item purification* (Magis et al., 2010): In the first step of this procedure, a certain selection strategy is used (with the whole test as an anchor) to determine a preliminary division of the test in a set of DIF-free items and a set of DIF items. In the second step, if the set of DIF items is not empty, the first step is repeated, but this time only the set of DIF-free items is used as an anchor. This procedure is repeated iteratively until it converges in the sense that the division of the test in DIF items and DIF-free items remains the same, or, if this does not happen, until a certain number of iterations is reached.

Summing up, regardless of the exact specifications of the anchor item selection procedure, all these customary methods have in common that they determine the anchor (and therefore, the identification of the model, i.e., the concrete value for μ) more or less by using the whole set of items as an preliminary anchor and then proceed with different selection strategies or iterative procedures. However, it generally remains unclear what the corresponding algorithms are intended to exactly search for. Indeed, knowing about the identification problem, it is impossible that they are aiming to try to find the true μ , as this cannot be inferred from the data alone. Instead, it seems that the ideas behind most of these algorithms are very close to the identification constraints proposed for μ in the last section. Therefore, why not state them directly and, additionally, profit from a better transparency concerning the identification of the model, as well as from much easier calculations?

Being aware of how the traditional methods work, it also becomes clear that any simulation study comparing the different methods cannot be meaningful, as results will depend on the identification constraint for μ under which DIF is simulated: The performance of a certain DIF detection algorithm will basically depend on how close its (explicit or implicit) identification constraint matches the one that is used in the data generating process. Furthermore, every simulation study comparing different DIF detection methods suffers from the conceptual problem that the data that is generated under a certain identification constraint is equally likely obtained under a different model, where the impact is shifted to the same amount as all the $\hat{\Delta}\beta_i$. Therefore, although the data generating process is the same for both models, the items which show DIF are not.

If the focus is not on DIF at the item- but rather on the test level, then the differential item pair functioning approach of Bechger and Maris (2015) is closest to the DIF structure approach proposed here. Bechger and Maris (2015) provided hypothesis tests for global DIF and differential item pair functioning. Although these hypothesis tests are useful from a decision-making perspective, the amount of DIF is not quantified. Also, small amounts of DIF might lead to a significant result if the sample size is large, so it is desirable to qualify a significant test result. The ML and S-DIF-V indices proposed here are effect sizes, which

are both largely independent of sample size and hence complement the global hypothesis test of Bechger and Maris (2015). The use of effect sizes has also been proposed for the logistic regression approach to DIF (Jodoin & Gierl, 2001; Zumbo & Thomas, 1997) as well as for the Mantel–Haenszel-type statistics (Hidalgo & López-Pina, 2004; Holland & Thayer, 1988), where the effect size is known as the ETS Delta Scale. However, both types of effect sizes quantify item-level DIF and are hence complementary to procedures that aim to identify DIF items.

There are different areas for further research. First, the proposed DIF indices should be further evaluated by detailed simulation studies, comparing their performances in different situations. Furthermore, more real data applications are needed and, based on the results, cut-off values quantifying small, medium, and high amounts of DIF should be deduced. Second, it seems worthwhile to investigate and extend the DIF structure approach for nonuniform DIF. So far the focus was on the Rasch model and, therefore, only considered differences in the difficulty parameters. Extending DIF structures to nonuniform DIF, that is, considering models with discrimination parameters, for example, the two parameter logistic (2PL) model, aggravates the identification problem as the discrimination parameters are relative to the variances of the latent variables in the two groups. Third, further research is also needed to address the case of more complex models, for example models for polytomous items or multidimensional data. Last, although all proposed approaches based on the DIF structure can easily be implemented in software, there is no precast software program specifically designed for DIF structure analyses yet. However, the DIF structure can be easily computed from output of standard software for Rasch modeling, as group-wise estimates of difficulty parameters and their standard errors are essentially all that is needed. For a conscious choice of the impact μ , MML estimation is preferable over conditional maximum likelihood (CML) estimation, as the latent means are natural parameters in MML and hence it is obvious how to shift the difficulties if a value of μ is set. Also, some implementations of CML fix one difficulty parameter to 0, so that no standard error results. In the appendix, in the online supplement, *R* code is supplied that demonstrates the estimation of the DIF structure, its graphical display, and calculation of the DIF indices.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. In practical applications a third step might be to quantify the size of differential item functioning (DIF), for example on the ETS Delta scale.
2. For example, it is possible to either determine the ranking of an item by using only one test statistic referring to a comparison with the whole test or by using separate tests statistics, each referring to a comparison with each other item of the test.

Supplemental Material

Supplemental material is available for this article online.

References

- Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, 80, 317-340.
- De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.
- DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7, 177-188.
- Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, 29, 278-295.
- Frick, H., Strobl, C., & Zeileis, A. (2015). Rasch mixture models for DIF detection: A comparison of old and new score specifications. *Educational and Psychological Measurement*, 75, 208-234.
- Hidalgo, M. D., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, 64, 903-915.
- Higgins, J., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21, 1539-1558.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.
- Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14, 329-349.
- Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70, 631-639.
- Kiefer, T., Robitzsch, A., & Wu, M. (2014). TAM: Test analysis modules (R package version 1.0-3). Retrieved from <http://CRAN.R-project.org/package=TAM>
- Kopf, J., Zeileis, A., & Strobl, C. (2015a). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement*, 39, 83-103.
- Kopf, J., Zeileis, A., & Strobl, C. (2015b). Anchor selection strategies for DIF analysis review, assessment, and new approaches. *Educational and Psychological Measurement*, 75(1), 22-56.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Magis, D., Béland, 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, 847-862.
- Müthen, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, 10, 121-132.
- Müthen, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, 54, 557-585.
- Osterlind, S., & Everson, H. (2009). *Differential item functioning*. Thousand Oakes, CA: Sage.
- Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14, 197-207.
- Reckase, M. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods*, 13, 31-48.
- Shih, C.-L., & Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators, multiple causes MIMIC method with a pure short anchor. *Applied Psychological Measurement*, 33, 184-199.
- Smits, D. J. M., De Boeck, P., & Vansteelandt, K. (2004). The inhibition of verbally aggressive behaviour. *European Journal of Personality*, 18, 537-555.
- Sterne, J. A., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99-110). Chichester, UK: Wiley.

- Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, 80, 289-316.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.
- 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. Braun (Eds.), *Test validity* (pp. 147-170). Hillsdale, NJ: Lawrence Erlbaum.
- Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, 80, 21-43.
- Vansteelandt, K. (2000). *Formal models for contextualized personality psychology* (Doctoral thesis). KU Leuven, Belgium.
- Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, 30, 261-293.
- Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, 26, 37-52.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48. Retrieved from <http://www.jstatsoft.org/v36/i03/>
- Wang, W.-C., Shih, C.-L., & Yang, C.-C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*, 69, 713-731.
- Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, 35, 339-361.
- Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.