Psychological Test and Assessment Modeling, Volume 62, 2020 (1), 11-27

Using a multilevel random item Rasch model to examine item difficulty variance between random groups

Johannes Hartig[1], *Carmen Köhler*[2] *& Alexander Naumann*[2]

Abstract

In educational assessments, item difficulties are typically assumed to be invariant across groups (e.g., schools or countries). We refer to variances of item difficulties on the group level violating this assumption as random group differential item functioning (RG-DIF). We examine the performance of three methods to estimate RG-DIF: (1) three-level Generalized Linear Mixed Models (GLMMs), (2) three-level GLMMs with anchor items, and (3) item-wise multilevel logistic regression (ML-LR) controlling for the estimated trait score. In a simulation study, the magnitude of RG-DIF and the covariance of the item difficulties on the group level were varied. When group level effects were independent, all three methods performed well. With correlated DIF, estimated variances on the group level were biased with the full three-level GLMM and ML-LR. This bias was more pronounced for ML-LR than for the full three-level GLMM. Using a three-level GLMM with anchor items allowed unbiased estimation of RG-DIF.

Keywords: Multilevel Rasch Model, Random Item Effects, Measurement Invariance

[1] *Correspondence concerning this article should be addressed to:* Johannes Hartig, PhD, DIPF | Leibniz Institute for Research and Information in Education, Rostocker Str. 6, 60323 Frankfurt, Germany; email: hartig@dipf.de

[2] DIPF | Leibniz Institute for Research and Information in Education, Germany

Data from educational assessment studies are oftentimes used to compare examinee test scores across different groups. In the U.S., educational assessments are frequently used to monitor student competence level and student progress in different schools and districts (DePascale, 2003). As such, these studies have a major impact on educational policies and reforms (see, e.g., Baird, et al., 2011; DePascale, 2003). In cross-national surveys such as PISA (Programme for International Student Assessment), TIMSS (Third International Mathematics and Science Study), or PIAAC (Programme for the International Assessment of Adult Competencies), competencies are assessed in numerous countries in order to investigate, for example, skill acquisition in relation to educational systems (Foshay, Thorndike, Hotyat, Pidgeon, & Walker, 1962). In order to draw valid inferences from evaluation results or comparative studies, the measurement instrument – that is, the test items that were constructed to measure specific skills – should be measurement invariant across groups (e.g. Jöreskog, 1971; Schweig, 2014). Measurement invariance implies that the item parameters of a measurement model are constant across groups, so that individuals from different groups but with equal ability levels have the same probability of answering the item correctly (Millsap & Everson, 1993). If the item response is not independent of group membership, the item shows differential item functioning (DIF). DIF can occur, for example, when one group is more familiar with specific item content than other groups. DIF can be either uniform, meaning that the difference in item parameter estimates across groups is constant across the entire ability spectrum, or nonuniform, meaning that the difference in item parameter estimates across groups depends on the ability level. In this study, the focus lies on uniform DIF modeled with a multilevel generalization of the Rasch model.

Before comparisons between two or more groups can be drawn, item parameters are often examined for DIF and items with DIF are excluded from analyses or revised in test development processes. Alternatively, DIF can be included in psychometric models and results of interest (e.g., group differences) can be obtained while modeling DIF simultaneously. In the literature, various DIF detection methods that examine whether item parameters are invariant across groups exist. In the traditional approaches, the groups are assumed to be fixed. The idea is to match the trait between the fixed groups and test for each item whether it is more or less difficult for the members of a specific group. In the context of item response theory (IRT), multi-facet IRT analysis can be used to examine, for example, item-by-gender interactions (Wu, Adams, Wilson, & Haldane, 2007). In these models, both the items and group-membership are included as fixed effects, with an additional interaction effect. By including group-membership as a main effect, the different ability levels between males and females are controlled for. An additional interaction effect between gender and an item indicates that an item is easier or harder for a particular group. Such item-by-gender interactions were investigated, for example, in PISA 2000 (Adams & Wu, 2002). The method also allows investigating item-by-country interactions. In PISA 2015 and 2018, comparability between countries is ensured by means of concurrent calibration with fixed-item parameter linking. Country-specific item parameters are fixed to be equal by default and set free for individual items and countries if DIF is detected (OECD, 2017). However, such analyses are quite extensive, as they involve inspecting every item in each country separately.

A crucial assumption inherent in the definition of DIF is that differences in item difficulties are independent from ability differences between groups. As a consequence, methods to model and detect DIF require identification restrictions with respect to DIF and/or ability differences between groups. One typical restriction is to define anchor items that are assumed to be free from DIF (e.g., Kopf, Zeileis, & Strobl, 2015). If the DIF for at least one item is fixed to zero, ability differences between groups and DIF for the remaining items can be estimated. Another approach is to restrict average DIF or average item difficulties within groups to zero (e.g., Wang, 2004), assuming that DIF is balanced across all items. Another traditional approach to test measurement invariance originates in the structural equation modeling (SEM) framework, where multiple group confirmatory factor analysis (MG-CFA) is used to test for DIF. The basic idea of MG-CFA models is to constrain factor intercepts and/or loadings in all groups, and compare this constrained model to a model where some or all item parameters are freely estimated across groups (Jöreskog, 1971). If the unconstrained model fits better, DIF is implied. Modification indices can inform about which parameters exactly differ across the groups and should thus be freely estimated. A disadvantage of this approach is that several different models need to be estimated and compared. Also, the assumption of exact item parameter invariance across groups is rather strict. A more recent enhancement of the MG-CFA approach is a Bayesian MG-CFA approximate measurement invariance approach (Muthén & Asparouhov, 2013). It is more lenient in its assumptions: It allows for some differences in item parameters, assuming that only the mean of the differences across groups should be zero. This approach seems very promising, but raises the question for adequate priors (Cieciuch, Davidov, Schmidt, Algesheimer, & Schwartz, 2014).

DIF between random groups

In the aforementioned traditional approaches, the prevalent idea is to compare fixed item parameters between fixed groups. Instead of treating group membership as a fixed factor, the data can also be analyzed in a hierarchical framework with individual students nested in groups, treating schools or countries as random factors (French & Finch, 2010). Throughout this paper, we will refer to DIF related to random groups as random group DIF (RG-DIF). This approach is appealing when the number of groups is large, such as in evaluations with many schools or in cross-national surveys with many countries. In these scenarios, the individual groups and specific item-by-group interactions (i.e. identifying items functioning differentially in specific groups) might be less interesting than obtaining comprehensive statistics quantifying invariance violations on the item level (i.e. identifying items violating measurement invariance across all groups).

Recently, Bayesian random item effects models (RIEM) have been proposed, in which the country specific item parameters are considered random deviations from the international item parameters (De Jong, Steenkamp, & Fox, 2007; Fox & Verhagen, 2010; Verhagen, 2012). The variations in country-specific item parameters can be interpreted as indicators for measurement invariance violations. The models provide a very comprehensive and flexible framework to deal with DIF in multilevel data. The models allow for variances of the item parameters (both difficulty and discrimination) on the group level. Within this

approach, some models contain independent, item-specific variance components (e.g., Verhagen & Fox, 2013); others include covariances of those random effects on the group level (Verhagen, Levy, Millsap, & Fox, 2016). Including a covariance structure allows RG-DIF to be correlated between items on the group level. In cross-national surveys, for instance, this would mean that the direction of RG-DIF can be systematically related between specific items. These correlations can be of substantive interest, since they inform about whether pairs or groups of items show similar DIF. This may be the case, for instance, when two items with related item content favor certain groups of examinees similarly. To our knowledge, the implications and practical relevance of the group level covariance structure of random item effects has not yet been examined in detail.

Aim of the study

The aim of this paper is to examine different methods for estimating uniform RG-DIF in applied settings, taking effects of the group level covariance structure into account. This is done using a Generalized Linear Mixed Model (GLMM) framework, which can be implemented in a wide variety of software packages. Group level variances and covariances of item parameters can be estimated using three-level GLMMs with responses nested in individuals and individuals nested in groups (e.g., countries). As this method is computationally intensive, we propose a second approach, namely multilevel logistic regression (ML-LR) as a screening method for uniform DIF in multilevel data. Logistic regression (LR) is a traditional DIF detection method (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Zumbo, 1999), which has been extended to the multilevel context (French & Finch, 2010; Moineddin, Matheson, & Glazier, 2007; Swanson, Clauser, Case, Nungester, & Featherman, 2002). So far, the two methods have not been compared with regard to their performance to detect uniform DIF in multilevel data. Particularly new to this study is the consideration of possible covariances between items that show DIF across countries. Note that, unlike the three-level GLMM approach, multilevel logistic regression (ML-LR) cannot take dependencies between DIF items into account. The assumption that two items show similar DIF is rather plausible, however, and this covariation might need to be taken into account in order to correctly identify DIF.

In the following, both methods are presented in detail, starting with the GLMM approach. Subsequent sections outline the simulation study we conducted to compare the approaches. Note that the current article focuses solely on effects of group membership on item difficulties, that is, uniform DIF.

Random group DIF in a GLMM framework

Within this paper, GLMMs with a logit link function for dichotomous variables were applied:

$$\text{logit}\left(Y_{pci}=1\right)=\eta_{pgi}\ , \tag{1}$$

where Y_{pci} is the response of person p in group g to item i, and η_{pci} is the linear predictor part of the GLMM. A multilevel Rasch model (e.g., Kamata & Cheong, 2007) is given by

$$\eta_{pgi} = \theta_g^B + \theta_p^W - b_i \ , \tag{2}$$

where θ_g^B represents the ability level of group g, θ_p^W represents the within-class between-person ability part of person p, and b_i is the difficulty of item i. θ_g^B and θ_p^W are random effects with $\theta_g^B \sim \mathcal{N}\left(0, \sigma^2_{\theta^B}\right)$ and $\theta_p^W \sim \mathcal{N}\left(0, \sigma^2_{\theta^W}\right)$. To allow for DIF, the items are treated as group specific:

$$\eta_{pgi} = \theta_g^B + \theta_p^W - b_{gi} \ . \tag{3}$$

The item difficulties can be treated as fixed or as random effects (De Boeck, 2008). Treating the group specific item difficulties b_{gi} as fixed corresponds to the traditional multi-group approaches with item-by-group interactions. The more parsimonious alternative is a multilevel random item Rasch model, with independent, item-specific distributions $b_{gi} \sim \mathcal{N}\left(\mu_{b_i}, \sigma^2_{b_i}\right)$. The item specific mean μ_{b_i} corresponds to the overall (average) difficulty of an item. The item specific variances $\sigma^2_{b_i}$ represent the amount of RG-DIF. A variance close to zero means an item is invariant; a large variance indicates that an item is functioning differentially across groups.

To allow for dependencies between RG-DIF, item difficulties can be modeled as correlated random effects with a joint multivariate normal distribution

$$b_{gi} \sim \mathcal{N}\left(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b\right) \ , \tag{4}$$

where $\boldsymbol{\mu}_b$ is the vector of average item difficulties and $\boldsymbol{\Sigma}_b$ is a covariance matrix with the RG-DIF variances $\sigma^2_{b_i}$ on the diagonal. The covariances $\sigma_{b_{i,j}}$ on the off-diagonal contain information regarding the correlations of RG-DIF between items on the group level.

Separation of RG-DIF and ability differences between groups

The identification problem of separating ability differences between groups from differences in item difficulties also exists in the case of random groups. Specifically, when estimating RG-DIF, the item specific variance components $\sigma^2_{b_i}$ need to be separated from the between-group variance $\sigma^2_{\theta^B}$ of the measured ability. With uncorrelated RG-DIF this separation is straightforward, since group specific item difficulties b_i are centered to a grand mean of zero and their variances $\sigma^2_{b_i}$ correspond to residual variation in

item difficulty that cannot be accounted for by group differences in overall ability. With correlated RG-DIF, however, the challenge to simultaneously estimate the between-group differences in the general ability dimension and the full (residual) covariance $\boldsymbol{\Sigma}_b$ in a model with random effects for all items arises. This can be addressed by introducing identification restrictions on b_{gi} (e.g., fixing the sums of residuals across groups to zero). Within Bayesian estimation, restrictions could be placed on the covariance matrix $\boldsymbol{\Sigma}_b$ itself (e.g., having the non-diagonal elements sum up to zero). Identification would also be given if one item is used as a reference category and only $I-1$ random effects are estimated. However, this would mean using the reference item as a DIF-free anchor and all variation in the other items' difficulties would be relative to the reference. The option of using anchor items is introduced below as a separate method. Within a GLMM framework, another option to estimate random effects for all items and thus obtain a full covariance structure is to center predictors. Usually, to implement IRT models in a GLMM framework, dummy variables $d_{pgi} \in \{0,1\}$ are generated that code which item a response was given to. To ensure that item effects cannot account for between group variances, those dummy variables can be group centered:

$$d^*_{pgi} = d_{pgi} - \bar{d}_{.gi} \ , \tag{5}$$

where $\bar{d}_{.gi}$ is the mean of the original dummy variable across all persons and responses within one group, thus reflecting the proportion of responses that were given to the specific item i. Instead of values of zero and one, d^*_{pgi} takes on positive values of $1-\bar{d}_{.gi}$ for responses given to item i and negative values of $-\bar{d}_{.gi}$ for responses to other items. Using the centered dummy variables and separating the fixed item effects from the random between-group-variation, the final GLMM for the estimation of correlated RG-DIF is given by

$$\eta_{pgi} = \theta^B_g + \theta^W_p - b_i - \sum_{i=1}^{I} d^*_{pgi} b_{gi} \, , \tag{6}$$

with b_i denoting the fixed average item difficulties and b_{gi} denoting between-group variation with

$$b_{gi} \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}_b\right) . \tag{7}$$

In this model, the probability of all responses depends on all group specific item difficulty variations b_{gi}.

Anchor items

When implementing the model to estimate RG-DIF in a GLMM framework, different items can be treated in different ways – more specifically, not all items have to be mod-

eled as random. If only a subset J of all items I is modelled as random, the remaining items $I \setminus J$ are used as anchor items, meaning that their difficulty is treated as invariant across all groups:

$$\eta_{pgi} = \theta_c^B + \theta_p^W - b_i - \sum_{j=1}^{J \subseteq I} d_{pgj}^* b_{gj} \,. \tag{8}$$

The model with anchor items includes fixed effects for all items; a covariance structure between groups is only estimable for items with possible DIF, that is, $J \subseteq I$. For these items, the variance-covariance matrix is estimated: $b_{gj} \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}_b\right)$. In Equation (6) all items are treated as DIF items and no anchor items are defined. Equation (6) is thus a special case of Equation (8) with $J = I$.

Logistic multilevel regression as screening method

Logistic regression is a well-established method to analyze DIF with respect to fixed groups (Zumbo, 1999). The probability of answering an item correctly is predicted by a fixed predictor (e.g., gender) while controlling for an ability estimate $\hat{\theta}_p$. If the predictor has an impact on the item response while controlling for ability, the item shows DIF. This idea can be transferred to ML-LR analysis in order to examine the effect of a random group membership while controlling for ability. Since this analysis is conducted for each individual item, the data structure is reduced to two levels with persons nested within groups. The linear predictor part of the logistic regression model can be expressed as

$$\eta_{pg} = \gamma_0 + \gamma_1 \hat{\theta}_p + u_{0g} \,, \tag{9}$$

where γ_0 is the intercept corresponding to the average difficulty of the respective item, γ_1 is the regression weight for the ability estimate $\hat{\theta}_p$, and u_{0g} is the group level residual with

$$u_{0g} \sim \mathcal{N}\left(0, \sigma_u^2\right). \tag{10}$$

In ML-LR, RG-DIF is estimated separately for each item. The variance σ_u^2 can be used as an indicator of RG-DIF, as it represents the between-group variance in item responses when ability differences are taken into account. For $\hat{\theta}_p$, any type of IRT-based or classical test score can be used. Since $\hat{\theta}_p$ is based on all items, it is implicitly assumed that group differences based on the entirety of the items are unbiased. Due to the reduction to a two-level model with a single random effect, this method is computationally undemanding. It can be implemented easily in a wide variety of established software packages (e.g., lme4 in R, M*Plus*, HLM, SPSS, and SAS). It thus offers an attractive way to

screen for RG-DIF even in large data sets. One specific aim of this study is to examine the performance of ML-LR to detect RG-DIF in comparison with the computationally more intensive three-level model. The item-wise analysis by means of ML-LR implies that RG-DIF is treated as independent between items, since interdependencies between RG-DIF as represented in the off-diagonal of $\boldsymbol{\Sigma}_b$ in Equation (7) are not accounted for. Disregarding potential covariances between RG-DIF might result in biased variance parameter estimates.

Method

Simulation design

The simulation study was conducted to investigate how the three methods – the full three-level GLMM, the three-level GLMM with anchor items, and the ML-LR approach – perform in detecting multi-group DIF when the DIF covaries between items. Given the problem of separating between-group differences in item difficulties and between-group differences in ability, increasing bias in RG-DIF estimations has to be expected for the full three-level GLMM when more positive correlations between RG-DIF effects are present. Bias is also expected for ML-LR, as it does not take dependencies between DIF items into account at all. Using anchor items within the three-level GLMM, which are known in the simulation, should allow unbiased estimation of RG-DIF for the remaining items. The focus of the simulation study thus lies on the covariance structure of item difficulties across countries, which was the main factor we manipulated. Another manipulated but not extensively investigated factor was the variation in ability between groups. Fixed factors were item size, number of students per group, number of groups, and the variances of item difficulties across countries, that is, the multi-group DIF.

Fixed design factors

The number of items in each condition was fixed to $I = 18$. The amount of items is typically larger in large-scale assessment settings. However, there is no reason to assume that a larger number of items would affect the fundamental pattern of results with respect to the factors we are interested in. The number of groups and group size were fixed at 50 and 100, respectively[3]. In practice, these sizes could represent evaluation studies with students nested in schools. It is not representative for large-scale assessments like PISA, where analyses of complex models based on the whole sample (often exceeding 500,000 cases) are limited by computational power.

Considering first the variances of the variance-covariance matrix of the item difficulties, $\boldsymbol{\Sigma}_b$, the variances for the 18 items were chosen to represent different sizes of DIF. Items

[3]We ran simulations with larger sample sizes (up to 200 groups with 500 individuals within each group) for selected conditions, and the pattern of results and size of effects didn't change substantially.

one to six had a variance of zero, thus showing no RG-DIF; they can be considered anchor items. Items seven to twelve had a variance of $\sigma_i^2 = 0.3$, representing medium RG-DIF. The last six items had large DIF, with item difficulty parameters varying at $\sigma_i^2 = 0.6$. Note that these variances are the main focus of the simulation study, since the performance of the two approaches was evaluated based on how accurately they retrieved these parameters.

Manipulated factors

For the covariances of the variance-covariance matrix of the item difficulties, four conditions with varying correlation patterns were simulated: (1) Uncorrelated random effects, (2) pairwise correlations, (3) correlations between half the RG-DIF items, and (4) correlations between two distinct clusters of RG-DIF items. Figure 1 illustrates the four patterns graphically. Note that items one to six showed no RG-DIF and thus group level

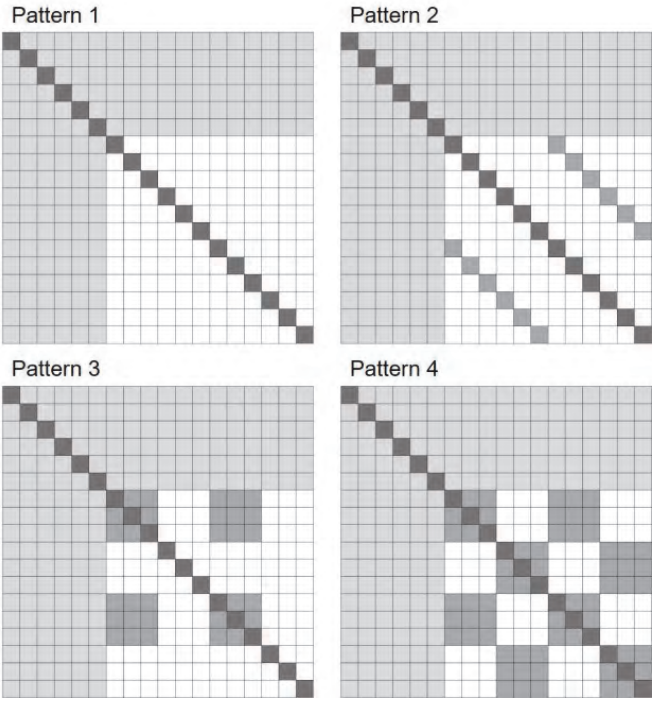

Figure 1:
The four correlation patterns used to generate the covariances of item difficulties on the group level. White: $r = 0.0$; light grey = correlations including items with variances of zero, medium grey: $r = 0.7$; dark grey: main diagonal ($r = 1.0$).

covariances with all other items had to be zero in any of the four conditions. All correlations in all conditions were set to .7, thus reflecting substantial dependencies between the RG-DIF of the item difficulty parameters. Positive residual correlations are more likely to occur for RG-DIF, as they can be caused by any factor affecting group-specific item difficulties in the same direction, such as common content or format. Negative correlations, that is, effects in opposite directions are far less likely and were thus not included in the simulation. At the same time, having only positive covariances exacerbates the identification problem of separating item specific covariance from between-group ability differences.

We also varied the ability level of the group by imposing a low intraclass correlation coefficient (ICC) of .1 ($\sigma^2_{\theta^B} = 0.1$ and $\sigma^2_{\theta^W} = 0.9$) in one condition and a high ICC of .5 ($\sigma^2_{\theta^B} = 0.5$ and $\sigma^2_{\theta^W} = 0.5$) in another. A 50% variance in ability explained by the group can be considered a high, yet realistic amount. In PISA 2015, for example, the proportion of variance in science performance between schools was 30% on average across OECD countries, but more than 50% e.g. in Bulgaria, Hungary, and the Netherlands (OECD, 2016).

Data generation and data analysis

The data was generated under the three-level GLMM model in the open source software R 3.4.0 (R Core Team, 2018). The number of replications in each condition was 100.

For the GLMM approach, the data sets were analyzed using the lme4 package version 1.1-13 (Bates et al., 2017). The package allows estimating the variance-covariance matrix $\mathbf{\Sigma}_b$ for all items simultaneously. The dummy variables d_{pgi} were group centered to ensure that between group variances remain at the person level and cannot be accounted for by the items (see Equation 11). To investigate whether the anchoring of items has an effect on the DIF detection, we analyzed the data sets (1) treating all items as random and (2) using the first six items as anchor items and all others as random (see Equation 12). To speed up estimation, the tolerance for declaring convergence in the penalized iteratively weighted residual sum-of-squares was raised from the default of 1E-7 to 1E-3, the computation of the gradient and Hessian of nonlinear optimization solution was suppressed, and penalized iteratively reweighted least squares were used for optimization instead of the default Laplace approximation (using the options *nAGQ=0, control=glmerControl(calc.derivs=F, tolPwrss=1e-3)* in the glmer function).

For the ML-LR approach, the data sets were first scaled under a unidimensional Rasch model using the R package TAM (Robitzsch, Kiefer, & Wu, 2017). In a second step, the Weighted Likelihood Estimates (WLE; Warm, 1989) from the Rasch scaling were used as ability estimate $\hat{\theta}$ in the ML-LR analyses (see Equation 9). Note that one ML-LR was estimated for each item.

Results

Each analysis with the full three-level GLMM took about 3.5 hours of computing time, analyses with six anchor items about 30 minutes each. The ML-LR analyses, including the estimation of the Rasch Model to obtain the WLEs, took about 30 minutes for the entire 100 replications. All model estimations converged within the default of 1000 iterations.

Recovery of RG-DIF

The main criterion of interest is the RG-DIF, that is, the estimated σ_i^2, by the three different methods (1) three-level GLMM, (2) three-level GLMM with anchor items, and (3) ML-LR. Specifically, the bias of the estimated variances σ_i^2 (i.e., the estimated RG-DIF minus the data generating value) is used to evaluate the methods. To display the results comprehensively, they are aggregated across five bundles of items. The items within each bundle have identical variances and correlations in each condition. Items 1-6 have variances of zero in all conditions; the other item bundles were items 7-9, 10-12, 13-15, and 16-18. Thus, the between-replication factors examined are correlation pattern and intraclass correlation, while the item bundle is a within-replication factor that contains the information about the level of RG-DIF and the correlation of the RG-DIF with other items. Table 1 summarizes the properties of the items within the five bundles used to aggregate results.

A first result was that the ICC of θ had no effect on the estimated RG-DIF and the respective bias. Results for the other factors are therefore pooled across the two conditions low and high ICC.

Figure 1 summarizes the mean bias of σ_i^2 in each condition. The varying conditions were the correlation pattern, the item bundle (i.e., the item specific variances and group level correlations), and the estimation method.

Table 1:
RG-DIF ($\boldsymbol{\sigma_i^2}$) and correlations of items within the five item bundles in the four correlation patterns.

		Correlation of RG-DIF with other items in correlation patterns			
Items	σ_i^2	Pattern 1	Pattern 2	Pattern 3	Pattern 4
1-6	0.0	no	no	no	no
7-9	0.3	no	pairwise	within cluster 1	within cluster 1
10-12	0.3	no	pairwise	no	within cluster 2
13-15	0.6	no	pairwise	within cluster 1	within cluster 1
16-18	0.6	no	pairwise	no	within cluster 2

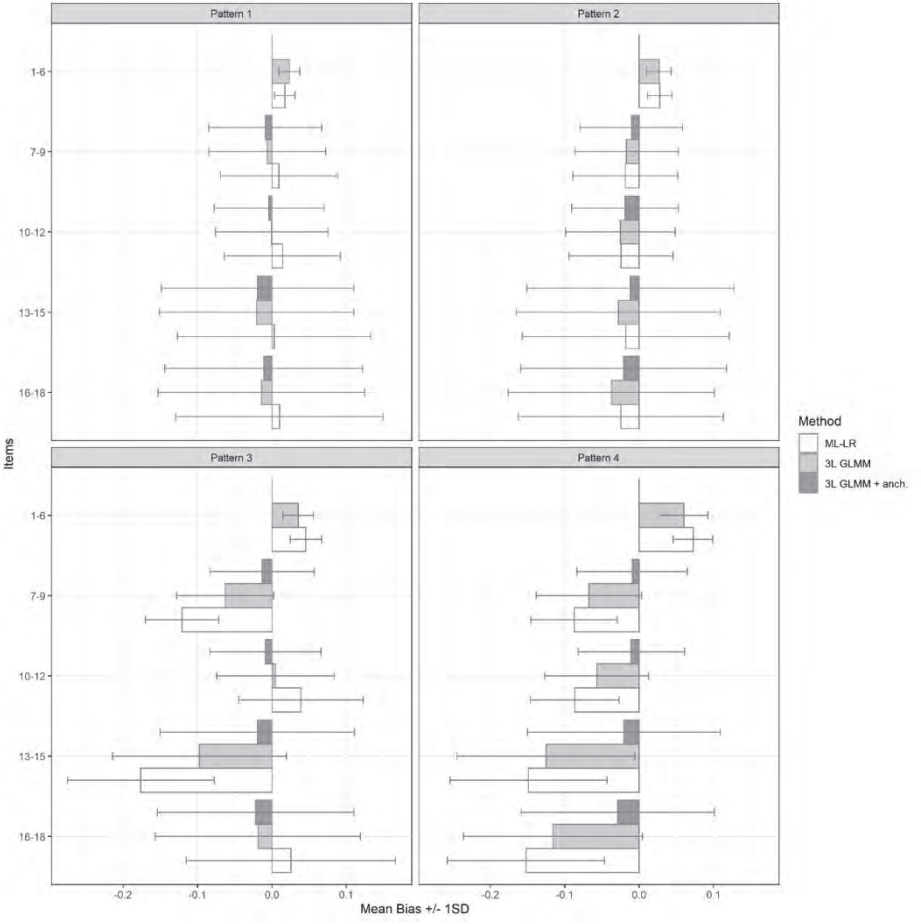

Figure 2:
Bias for estimated RG-DIF (σ_i^2) depending on the correlation pattern, the item bundle, and the estimation method. ML-LR = multilevel logistic regression; 3L GLMM three-level GLMM with freely estimated full covariance structure $\boldsymbol{\Sigma_b}$; 3L GLMM + anch. = three-level GLMM with items 1-6 as anchor items.

Results show that for uncorrelated RG-DIF (pattern 1), all methods work equally well, showing a bias very close to zero. With pairwise correlated RG-DIF (pattern 2), bias is still small. Yet, it is visible that for ML-LR and the full three-level GLMM, there is a slight positive bias for the variance of the item bundle with a true RG-DIF of zero and a slightly higher negative bias for the RG-DIF of the other items. When RG-DIF is correlated within one bundle of items (pattern 3), there is a pronounced negative bias in RG-

DIF estimated for the items with group level correlations (items 7-9 and 13-15) with ML-LR and the full three-level GLMM. At the same time, RG-DIF for items with zero DIF is overestimated. When all RG-DIF is correlated within item clusters (pattern 4), both ML-LR and full three-level GLMM are underestimating RG-DIF for the items with correlated RG-DIF and overestimating RG-DIF for the items with no actual group level variances. For patterns 3 and 4, bias in RG-DIF is larger with ML-LR than with the full three-level GLMM. The three-level GLMM with items 1-6 as anchor items performed equally well in all conditions, meaning that estimates of RG-DIF were not affected by group-level correlations.

Recovery of correlation patterns

The full three-level GLMM and the three-level GLMM with anchor items provide estimates for the covariance matrix $\boldsymbol{\Sigma}_b$, allowing to examine the recovery of the correlations between DIF on the group level. Since those are not the focus of our study, these results are only briefly summarized. Generally, the recovery of the general RG-DIF correlation patterns was good for the items with RG-DIF (items 7 to 18), while the absolute size of the correlations was biased. With no or few correlations (patterns 1 and 2), the difference between the true and estimated correlations did not exceed absolute values of .1. With stronger true correlations, the estimated correlations had a negative bias of up to -.25 in pattern 4, with the largest bias occurring for true correlations of zero. Correlations for items 1-6, which had true variances and covariances of zero, were highly positively biased among those items (up to .77 in pattern 4), moderately positively biased (up to .24) between items 1-6 and items with correlated RG-DIF, and moderately negatively biased (up to -.24) between items 1-6 and items with uncorrelated RG-DIF. In the three-level GLMM with anchor items, only the correlations between items with RG-DIF were estimated. Those were practically unbiased: Absolute differences between the true and estimated correlations never exceeded .03.

Discussion

The aim of this paper was to examine detection of uniform RG-DIF with different methods and under different conditions regarding the correlational structure of RG-DIF. We consider this structure an important yet neglected research topic, since the relation of DIF between items contains valuable information beyond that of individual items. In summary, RG-DIF can be recovered well by all investigated methods when there are no or only few correlations on the group level. If RG-DIF is correlated on the group level, however, correlated RG-DIF is underestimated while that of other items tends to be overestimated. This illustrates that DIF generally is a relative concept. Per definition, DIF means an item is functioning differently than the totality of items in a test or than a set of anchor items. If all items within a test had the same “DIF” across all groups, this variation in examinee responses would not be detected or interpreted as DIF but would be reflected in between-group differences in ability. The fact that (RG-) DIF is relative to

the behavior of the totality of the items demonstrates the importance of being aware that DIF is not only the property of a single item but also of relations between DIF across items.

With respect to the examined methods, the full three-level GLMM performed better in recovering RG-DIF than ML-LR, but was still substantially affected by DIF correlated on the group level. It obviously remains challenging to separate between-group differences in the general ability dimension and the (residual) covariance $\mathbf{\Sigma}_b$. Using anchor items clearly offers a solution to this problem. However, while pure anchor items with zero RG-DIF existed and were known in the simulation, both is not necessarily the case in empirical applications.

Limitations

A general limitation of the methods under study is that the GLMM framework restricts the models to generalizations of the Rasch model. This obviously limits comparability with studies using other IRT models. We nevertheless consider the approach promising to study the structure of DIF across multiple groups due to flexibilities given within the GLMM framework. The models can, for instance, easily be extended to include predictor variables on the group, person, or item level.

Furthermore, as in all simulation studies, the findings are limited by the specific conditions realized in the study. We limited the design by keeping the number of groups, the group size, the number of items and the number of anchor items fixed. The bias observed in our study may decrease in absolute size with larger sample sizes, but we assume that the pattern of results, particularly the effect of group level correlations, remains the same. Another specific limitation is the restriction to positive residual correlations of RG-DIF, which we consider are more realistic to find in practice. For more heterogeneous correlations, particularly patterns with positive and negative dependencies cancelling each other out, the bias of estimated RG-DIF might be smaller. This is, however, merely a hypothesis, which we can neither deny nor confirm based on our study.

A more severe limitation is the applied estimation technique. We decided to implement the model in a general multilevel software package and to adjust the estimation method to allow for a faster computation. It remains open to further investigation whether other estimation techniques (e.g., the default Laplace approximation implemented in lme4 or Bayesian estimation) that also allow for other ways of solving the identification problem when estimating a full covariance matrix perform better in recovering RG-DIF. Again, we do not assume the general pattern of results to be different, but the absolute size of the bias may differ between methods.

The method under study is also clearly limited to moderate sample sizes because of its low computational power. It can be well implemented in evaluation studies with students in schools, and it seems particularly promising to model RG-DIF to gain insight about the correlational structure of DIF across items. It is, however, less easily applicable as a tool for DIF diagnosis in the context of large-scale assessments. If the correlational structure of DIF is of interest in contexts that involve large sample sizes, the method could be

applied to pilot studies with smaller samples or in (secondary) analyses with subsets of the complete data. A technically less elegant, yet practically easier-to-implement alternative could be to explore correlational structures of DIF in a two-stage approach. Group specific item difficulties can be estimated either in a multiple group model with fixed effects, or in separate analyses for each group. These estimated difficulties can be analyzed with respect to their correlational structure in a second step. Regardless of the specific methods, we consider the structure of DIF an important source of information that should be investigated in both evaluation studies and large-scale assessments. It can provide valuable information regarding the possible reasons for DIF that may not be detected by the mere inspection of individual items.

Outlook

In addition to addressing the limitations listed above, there are further possibilities extending the research on RG-DIF. First, we limited the modelling approach to uniform DIF within a GLMM framework. In a more general Bayesian random item effect framework, non-uniform DIF (i.e., group level covariances of random item discriminations) could be addressed as well. Another possibly interesting extension is the use of link functions for ordinal data. This would allow examining RG-DIF in questionnaire data, an area where measurement invariance is also often an important issue.

References

Adams, R. & Wu, M. (2002). *PISA 2000 technical report*. Paris: OECD.

Bates, D., Mächler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, D., Green, P. (2017). Package 'lme4': Linear Mixed-Effects Models using 'Eigen' and S4. R package version 1.1–10. Retrieved from https://cran.r-project.org/web/packages/lme4/

Baird, J., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T. & Daugherty, R. (2011). *Policy Effects of PISA.* Oxford University Centre for Educational Assessment (OUCEA) Report. Retrieved from http://oucea.education.ox.ac.uk/wordpress/wp- content/uploads/2011/10/Policy-Effects-of-PISA-OUCEA.pdf

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact versus an approximate (Bayesian) measurement invariance test: A cross-country illustration with a scale to measure 19 human values. *Frontiers in Psychology, 5*, 1-10.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533-559.

De Jong, M. G., Steenkamp, J. B. E. M., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34*, 260–278. doi:10.1086/518532

DePascale, C.A. (2003). The ideal role of large-scale testing in a comprehensive assessment system. *Journal of Applied Testing Technology, 5*, 1-11. Retrieved from http://www.testpublishers.org/assets/documents/volume%205%20issue%201%20The%20ideal%20role.pdf

Foshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievements of thirteen-year-olds in twelve countries: Results of an international research project, 1959–1961*. Hamburg: UNESCO Institute for Education.

Fox, J.-P., & Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 467–488). London, UK: Routledge Academic.

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement, 47*, 299–317. doi: 10.1111/j.1745-3984.2010.00115.x

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409-426.

Kamata, A. & Cheong, Y. F. (2007). Multilevel rasch models. In M. von Davier & C. H. Carstensen (Eds.). *Multivariate and Mixture Distribution Rasch Models. Extensions and Applications* (p. 217-232). New York, Berlin: Springer.

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement, 75,* 22-56.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing bias. *Applied Psychological Measurement, 17*, 297-334.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*, 257–274.

OECD (2009). *Pisa 2006 technical report*. Paris: OECD Publishing.

OECD (2012). *PISA 2009 technical report.* Paris: OECD Publishing.

OECD (2016). *PISA 2015 results (volume I): Excellence and equity in education*. Paris: OECD Publishing.

OECD (2017). *PISA 2015 Technical Report.* Paris: OECD Publishing. Retrieved from http://www.oecd.org/pisa/data/2015-technical-report

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Robitzsch, A., Kiefer, T., & Wu, M. (2017). TAM: Test analysis modules. R package version 2.4-9. Retrieved from http://cran.r-project.org/web/packages/TAM/index

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied PsychologicalMeasurement*, *17*, 105–116.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*, 53–75.

Verhagen, A. J. (2012). *Bayesian Item Response Theory models for measurement variance.* Phd Thesis University of Twente, Enschede. http://dx.doi.org/10.1990/1.9789036534697

Verhagen, A. J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*, 383–401. doi: 10.1111/j.2044-8317.2012.02059.x

Verhagen, J., Levy, R., Millsap, R. E., & Fox, J. P. (2016). Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models. *Journal of Mathematical Psychology, 72,* 171-182. doi: 10.1016/j.jmp.2015.06.005

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,* 221–261. doi: 10.3200/JEXE.72.3.221-261

Warm, T.A. (1989). *Weighted likelihood estimation of ability in item response theory. Psychometrika, 54*, 427-450. doi: 10.1007/BF02294627

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0. Generalised item response modeling software.* Victoria: ACER Press.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.