Article

Investigating Measurement Invariance by Means of Parameter Instability Tests for 2PL and 3PL Models

Educational and Psychological Measurement
2019, Vol. 79(2) 385–398

DOI: 10.1177/0013164418777784
journals.sagepub.com/home/epm
SAGE

Rudolf Debelak[1] and Carolin Strobl[1]

Abstract

M-fluctuation tests are a recently proposed method for detecting differential item functioning in Rasch models. This article discusses a generalization of this method to two additional item response theory models: the two-parametric logistic model and the three-parametric logistic model with a common guessing parameter. The Type I error rate and the power of this method were evaluated by a variety of simulation studies. The results suggest that the new method allows the detection of various forms of differential item functioning in these models, which also includes differential discrimination and differential guessing effects. It is also robust against moderate violations of several assumptions made in the item parameter estimation.

Keywords

item response theory, Monte Carlo simulation, differential item functioning, measurement invariance

Introduction

A general assumption of item response theory (IRT) models is that the item parameters are constant over the population of test takers. In practical assessments, violations of this assumption are often summarized as differential item functioning (DIF; Holland & Wainer, 1993; Millsap, 2012; Osterlind & Everson, 2009).

[1]University of Zurich, Zurich, Switzerland

Corresponding Author:
Rudolf Debelak, Department of Psychology, University of Zurich, Binzmühlestrasse 14, Zurich 8050, Switzerland.
Email: rudolf.debelak@psychologie.uzh.ch

Among the methods for the detection of DIF, methods relying on an IRT framework can be discerned from those not relying on an IRT framework (Magis, Béland, Tuerlinckx, & De Boeck, 2010), which are also named nonparametric DIF methods by some authors. Examples for nonparametric DIF methods include the Mantel-Haenszel method (Mantel & Haenszel, 1959; see also Holland & Thayer, 1988), the simultaneous test bias method and its variations (e.g., Li & Stout, 1996), and methods based on logistic regression (Swaminathan & Rogers, 1990).

A recently proposed method for detecting DIF in an IRT framework are M-fluctuation tests (Komboz, Strobl, & Zeileis, 2018; Strobl, Kopf, & Zeileis, 2015; T. Wang, Strobl, Zeileis, & Merkle, 2018; Zeileis & Hornik, 2007). Approaches based on M-fluctuation tests aim at detecting DIF effects by considering the relationship between the item parameters and person covariates, such as age, gender, or educational level, in a data-driven way. They test the null hypothesis that the item parameters are constant over all values of person covariates. M-fluctuation tests can be thus considered as an alternative to DIF tests, which are based on focal and reference groups (for an overview, see Magis et al., 2010) and mixture distribution IRT models (Rost, 1990, 1991; Rost & von Davier, 1995; von Davier & Yamamoto, 2004, 2007).

M-fluctuation tests are further related to work of Glas on the Lagrange-multiplier test (Glas, 1999, 2010; Glas & Suárez-Falcón, 2003; Glas & van der Linden, 2010). Glas (1998, 2001) already described the Lagrange-multiplier test as a tool for the detection of DIF in the two-parametric logistic (2PL) and three-parametric logistic (3PL) models (Birnbaum, 1968). However, his method requires the a priori definition of reference and focal groups. As a consequence, continuous person covariates (e.g., age) have to be discretized to allow the definition of reference and focal groups. The approach discussed here can be applied to detect DIF in covariates of any type, including metric covariates such as age, without previous discretization.

The purpose of this article is the derivation and evaluation of M-fluctuation tests, which are based on marginal maximum likelihood (MML) estimation, in order to broaden the scope of IRT models for which this class of tests can be applied. We have analytically derived the score contributions necessary for formulating the test statistics and will evaluate these tests as a new method for testing DIF in the widely used 2PL model and a constrained 3PL model. Besides the new derivation of M-fluctuation tests in the MML framework, the main contribution of this article is the investigation of their power and Type I error rate for these models under various conditions. We also discuss additional results on the robustness of M-fluctuation tests to violations of the distributional assumptions, which is an important aspect of their practical application.

The remainder of this article is organized as follows: The second section briefly reviews the statistical framework underlying the M-fluctuation tests and provides technical details on M-fluctuation tests for the 2PL model and the 3PL model with a common pseudo-guessing parameter. The third section outlines three Monte Carlo simulation studies that aimed at evaluating the new method. The fourth section provides a summarizing discussion.

M-Fluctuation Tests for Detecting DIF in the 2PL Model and a Constrained 3PL Model

We begin our discussion with the 2PL model of Birnbaum (1968). In the context of this article, we use a parametrization of this model that is based on the following item response function:

$$P(u_{ji} = 1 | \theta_j, a_i, d_i) = \frac{\exp(a_i \cdot \theta_j + d_i)}{1 + \exp(a_i \cdot \theta_j + d_i)} \quad (1)$$

In this equation, θ_j denotes the person parameter of respondent j, and a_i and d_i are the slope (or discrimination) and intercept parameters for item i. Similar forms of the 2PL model were used by McDonald (1999) and Fox (2010). In contrast to other parametrizations, higher d_i values increase the probability of a correct response.

The inclusion of an additional parameter c_i for the lower asymptote, which is called the guessing or pseudo-guessing parameter in the literature, leads to the 3PL model. However, for small to moderate sample sizes, the item parameters of the 3PL model are usually difficult to estimate (De Ayala, 2009). We therefore consider a constrained 3PL model with a common lower asymptote c for all items:

$$P(u_{ji} = 1 | \theta_j, a_i, d_i, c) = (1 - c) + c \cdot \frac{\exp(a_i \cdot \theta_j + d_i)}{1 + \exp(a_i \cdot \theta_j + d_i)} \quad (2)$$

The application of M-fluctuation tests for testing the stability of the Rasch model generally consists of two steps: First, the item parameters are estimated based on the whole sample. Second, the stability of the item parameter estimates with regard to a person covariate of interest is investigated.

We start with a brief discussion of the first step. The M-fluctuation tests for Rasch-type models that were presented so far (Komboz et al., 2018; Strobl et al., 2015) used conditional maximum likelihood estimates for the first step, which are not available for the 2PL and constrained 3PL models. A possible alternative approach is joint maximum likelihood estimation. However, a fundamental problem of this approach is that it leads to inconsistent item parameter estimates (Baker & Kim, 2004); a formal proof for the special case of the Rasch model was provided by Ghosh (1995). In this study, we therefore used MML estimation of the item parameters, which assumes a normal distribution of the person parameters.

In the second step, the individual contributions to the score function (i.e., the derivative of the log-likelihood for each individual person) are ordered with respect to a covariate of interest (e.g., age). If this person covariate does not affect the item parameters, the individual contributions should show no clear pattern but lead to random fluctuations of the cumulative score process that summarizes the individual contributions over the range of the covariate. However, any relationship between the person covariate and the item parameters would lead to systematic fluctuations in the cumulative score process (as was illustrated, e.g., in Strobl et al., 2015). Let $0 \leq t \leq 1$

denote a relative proportion of the overall sample size N, then the cumulative score process is defined as (Zeileis, Hothorn, & Hornik, 2008)

$$W_l(t) := \hat{V}^{-1/2} N^{-1/2} \sum_{j=1}^{\lfloor N \cdot t \rfloor} \psi(u_{(j|l)}, \hat{b})$$

In this equation, $\hat{b}$ denotes the vector of the estimated item parameters, $u_{(j|l)}$ denotes the jth ordered observation with respect to the lth covariate, and $\psi(u_{(j|l)}, \hat{b})$ denotes the individual score function, which is the derivative of the individual contributions to the log-likelihood $\Psi(\boldsymbol{u}_j)$ for a person j. $\hat{V}$ is an estimate of the covariance matrix for the score function.

In a first step, we therefore determined the individual score functions for the 2PL and constrained 3PL models. A derivation of these score functions will be made available as an online document (https://doi.org/10.5167/uzh-151192) accompanying this study. Given these score functions, an important decision concerns the choice of the test statistic, which aims to summarize the resulting cumulative score process. An overview on possible choices is given by T. Wang, Merkle, and Zeileis (2014). As in the study of Strobl et al. (2015), the double maximum test statistic was used in this study. Based on the choice of this summary statistic, p values can be calculated by comparing its value with its distribution under the null hypothesis.

A possible problem concerns the presence of actual mean differences in the distribution of the person parameters between several groups of respondents, which is usually called an impact effect (e.g., Chen, Chen, & Shih, 2014; DeMars & Jurich, 2015; Kopf, Zeileis, & Strobl, 2015a, 2015b; W.-C. Wang, Shih, & Sun, 2012). Ability differences between groups of respondents are ubiquitous in psychological research (e.g., between female and male respondents; Halpern, 2012), and should not be confused with DIF.

A suitable general estimation framework in the context of IRT was developed by Bock and Zimowski (1997). The purpose of this framework is to estimate item parameters from data obtained from multiple known groups. In all groups, the person parameters are usually assumed to follow a normal distribution; for calibration purposes, a standard normal distribution is typically assumed for the first group. The multiple group framework generally allows the definition of different constraints on the item and person parameters. The means and variances of the person parameter distributions can be estimated freely for each group of respondents based on the data, or be constrained to be equal for all groups (i.e., $\mu = 0$ and $\sigma^2 = 1$). Likewise, it is possible to estimate each item parameter freely for each group of respondents, or to constrain it to be equal across all respondent groups. This approach was also independently suggested by T. Wang et al. (2018) in a study on M-fluctuation tests for the two-parametric normal-ogive (2PNO) model.

In summary, M-fluctuation tests for the 2PL model and the constrained 3PL model can be calculated by the following steps: First, estimate the item parameters using an

MML approach in combination with a multiple group framework to account for the possibility that the data were obtained from groups with different ability distributions. Second, calculate the scores based on the parameter estimates and the response data. Third, calculate an appropriate test statistic to detect systematic deviations of the cumulative score process and calculate the corresponding p value.

Simulation Studies

For practical applications, it is important to investigate the power of these tests with regard to the instability of various item parameters. For the 2PL model and the constrained 3PL model, this point was evaluated by three simulation studies. We will summarize the first of those in the following subsection, and will then report the results of the remaining two, which investigated the robustness of the method, in a separate subsection.

Simulation Study I: Principal Results on the Power and Type I Error Rate

Aim of Simulation Study I. The principal goal of Simulation Study I was to investigate the Type I error rate and power of M-fluctuation tests against DIF effects in the various model parameters of the 2PL model and constrained 3PL model under a limited set of conditions, which will be described below. We used a multiple-group IRT framework to estimate the item parameters. This framework allows the modeling of ability differences between known groups of respondents, and therefore also allows to model impact effects. We expected that impact effects that are not modeled would lead to a bias in the item parameters, and thus to false-positive results in the M-fluctuation tests. The second goal of this study was to test this assumption. Finally, the third goal was to investigate the effect of constraints on the item parameters on the power of M-fluctuation test. We expected the M-fluctuation tests to be most sensitive against uniform and nonuniform DIF effects if all item parameters were constrained to be equal for all respondents in the parameter estimation.

This study is an important extension of the previous studies of Strobl et al. (2015) and Komboz et al. (2018) that investigated the Type I error and power for dichotomous and polytomous Rasch models under the conditional maximum likelihood framework. It further complements the study of T. Wang et al. (2018), which investigated M-fluctuation tests for the 2PNO model.

Design of Simulation Study I. In this study, we simulated the responses of a sample of two groups to an item set of 30 items based on the 2PL model or the constrained 3PL model. Each simulated dataset contained two groups of respondents, of which one could be affected by DIF. In each simulated dataset, each respondent was assigned to one of the two groups by the result of a Bernoulli experiment, with both outcomes being equally likely. The simulated datasets shared the following characteristics:

- *Number of items and item parameters*: All datasets contained the responses to 30 items. Our method for creating the item parameters resembles that chosen in earlier simulation studies on DIF (Cao, Tay, & Liu, 2017; Tay, Huang, & Vermunt, 2016). The intercept parameters were first drawn from a uniform distribution $\mathcal{U}(-1.45, 1.45)$ and then centralized to a mean of 0, with the same values being used under all conditions. The slope parameters were drawn from a normal distribution $\mathcal{N}(1, 0.25)$, with one negative value being replaced by 0.3. For the constrained 3PL model, the pseudo-guessing parameter was set to 0.1. The item parameters were generated independently, and there was no systematic relationship between the item parameters and the position of the item.
- *Number of respondents*: The simulated samples consisted of 500, 1,000, or 3,000 respondents, similar to earlier studies on tests for DIF effects (e.g., Kopf et al., 2015a, 2015b).
- *Presence of an impact effect*: If no impact effect was present, the ability parameters in both groups were drawn from a standard normal distribution. If an impact effect was simulated, the ability parameters in the first groups were drawn from a standard normal distribution, whereas the ability parameters in the second group were drawn from a normal distribution $\mathcal{N}(1, 1)$. Very similar conditions were investigated in a variety of simulation studies, for example, DeMars and Jurich (2015), Kopf et al. (2015a), and W.-C. Wang et al. (2012).
- *Presence of a DIF effect*: In addition to a possible impact effect, the item parameters could also differ for the same groups of respondents. Overall, these conditions varied with regard to the following factors:
  - *Percentage of items with DIF effect*: The last 12 items (40%) were DIF items. Similar conditions were used in the studies of Chen et al. (2014), Kopf et al. (2015a), and W.-C. Wang et al. (2012). As there was no systematic relationship between the position of the items and their item parameters, the DIF items did not differ systematically from the rest of the item set. For DIF items, at least one item parameter (intercept, slope, and, if the constrained 3PL model was the data generating model, the pseudo-guessing parameter) was changed for a part of the sample. All DIF items were manipulated in the same way, as will be described in the following.
  - *Type and size of DIF effect*: If the item slope parameter was affected by DIF, it was altered by 0.3. If the intercept parameter was affected by DIF, it was altered by 0.6; and if the pseudo-guessing parameter was affected by DIF, it was altered by 0.1. Similar sizes of DIF effects on the intercept parameters were used in the studies of Chen et al. (2014), Kopf et al. (2015a), and Kopf et al. (2015b). A similar effect for the slope parameter was used in the study of W.-C. Wang et al. (2012). Changes in the slope or pseudo-guessing parameters led to nonuniform DIF effects, whereas changes which affected only the intercept parameters led to uniform DIF effects.

- *Direction of DIF*: Under the first simulated condition, the change of the item parameters was the same for all DIF items, leading to an unbalanced DIF effect. In the second simulated condition, item parameters were increased for the even-numbered DIF items, and decreased for the odd-numbered DIF items, leading to a balanced DIF effect. If the DIF was unbalanced and if an impact effect was present, the intercept, slope, and the pseudo-guessing parameters were increased in the group with a higher mean ability parameter. As can be seen from Equations (1) and (2), an increase of the intercept and pseudo-guessing parameters increases the probability of a correct response. This corresponds to earlier simulation studies, where unbalanced DIF effects favored the more able group (e.g., Chen et al., 2014; Kopf et al., 2015a, 2015b; W.-C. Wang et al., 2012).

Under each condition, 1,000 datasets were simulated. Each dataset was analyzed with several variations of M-fluctuation tests, which differed with regard to the constraints of the item parameter estimation. As outlined in the introduction, we used a multiple group framework for estimating the item parameters. In this framework, both the item parameters and the parameters of the person parameter distribution are allowed to differ between the two groups of respondents. We focus here on model specifications which assumed the item parameters to be constant for both respondent groups. Under this constraint, the distribution of the person parameters could be either assumed to be standard normal for each group, or a normal distribution with group-specific μ and σ^2 parameters.

Results of Simulation Study I. Overall, the power and Type I error rate were heavily influenced by the choice of the model specification. We first focus on the results obtained with a model specification in which the impact groups were allowed to differ with regard to their person parameter distribution, but all item parameters were assumed to be equal for all groups. When the data were generated from a 2PL model without DIF, the rate of significant results was between 0.019 and 0.040 under all conditions and therefore slightly below the nominal alpha level of 0.05, which indicates a conservative behavior of the M-fluctuation tests. For the constrained 3PL model, the Type I error rate was between 0.017 and 0.047 under all conditions. Similar results had been found by Strobl et al. (2015) and Komboz et al. (2018) for M-fluctuation tests for dichotomous and polytomous Rasch models, and by T. Wang et al. (2018) for the 2PNO model.

Figure 1 presents results on the power of M-fluctuation tests when either the slope, the intercept, or both parameters were affected by DIF.

The corresponding results for the constrained 3PL model are presented in Figure 2. When all three parameters (slope, intercept, and pseudo-guessing) were affected by DIF, the power rates were near 1 under all conditions for this model.

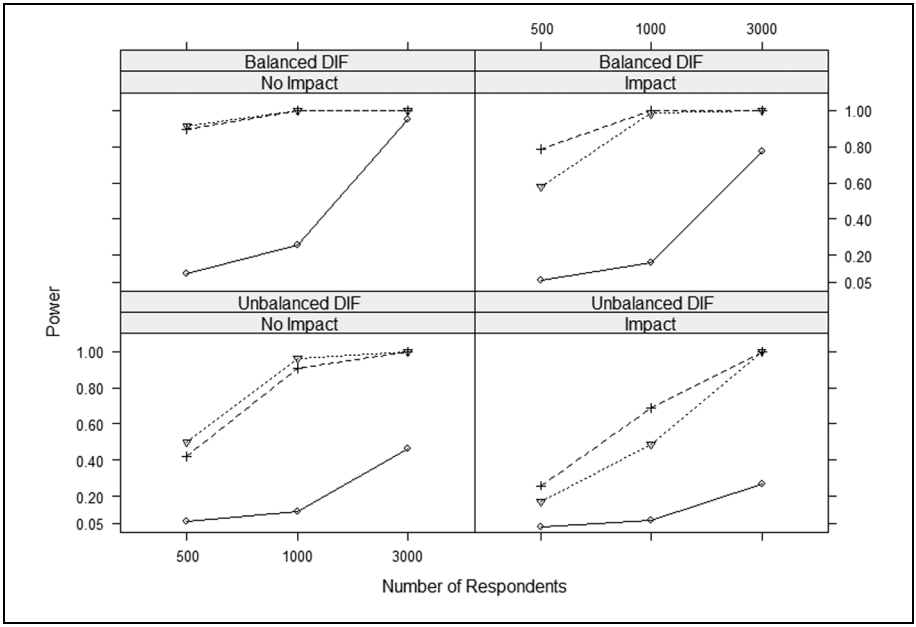

Figure 1. Results of M-fluctuation tests for the 2PL model if the slope (solid lines), intercept (dotted line), or both parameters (broken line) are affected by DIF for various conditions of sample size, balanced (top row) and unbalanced (bottom row) DIF effects when impact was present (right column) or not (left column). All results were obtained using model constraints that modeled impact effects.

When alternative model specifications were used in the item parameter estimation, worse results were obtained. Model specifications that did not model an impact effect in the data usually led to a highly inflated Type I error rate, which reached values of up to 1 for both models under some conditions. Model specifications that allowed item parameters to differ between the respondent groups led to lower power rates and did not seem promising for practical applications. We omit details for brevity.

Conclusions From Simulation Study I. Simulation Study I aimed at three goals. The first goal was the investigation of the power of M-fluctuation tests when the item parameters are estimated under a wide range of different model constraints. Generally, we found that M-fluctuation tests had power against uniform and nonuniform DIF effects. The power of the M-fluctuation tests increased with the size of the sample of simulated respondents and was generally larger for balanced than for unbalanced DIF effects.

The second goal aimed at testing our assumption that impact effects can lead to false-positive results if they are not modeled in the parameter estimation. The results of the simulation study indicate that this assumption is correct and therefore underline

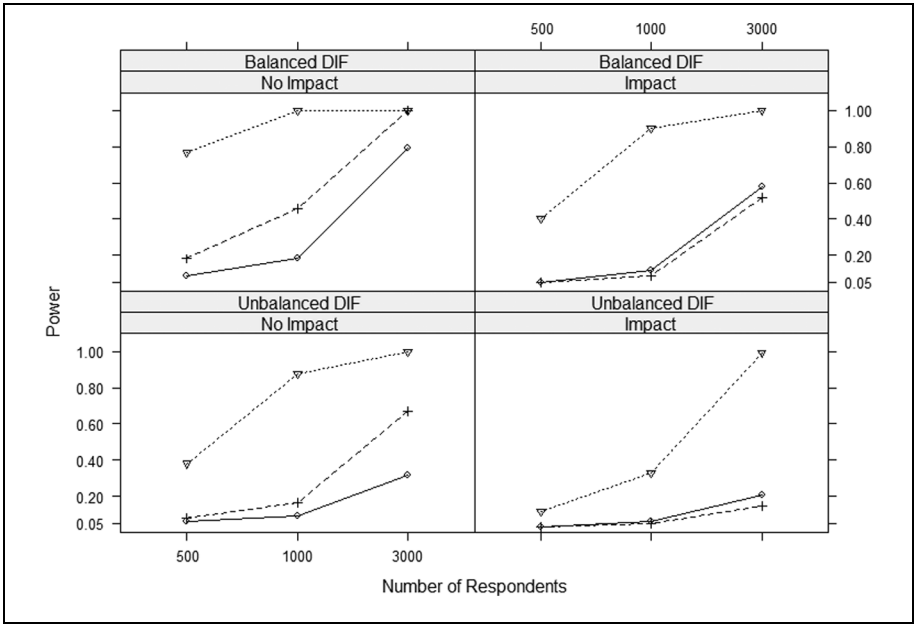

Figure 2. Results of M-fluctuation tests for the constrained 3PL model if the slope (solid lines), intercept (dotted line), or pseudo-guessing parameters (broken line) are affected by DIF for various conditions of sample size, balanced (top row) and unbalanced (bottom row) DIF effects when impact was present (right column) or not (left column). All results were obtained using model constraints that modeled impact effects.

our recommendation to model possible impact effects as part of the item parameter estimation. In our simulations, a model specification that allowed a free estimation of mean and variance for each impact group, but which also constrains the item parameters to be equal across the population, was found to be most useful.

The third goal was to test the assumption that the power of M-fluctuation tests is highest when all item parameters are constrained to be equal for all respondents. Again, this assumption was confirmed. In summary, the results of Simulation Study I agree with those of previous studies (Komboz et al., 2018; Strobl et al., 2015; T. Wang et al., 2018) and generalize these findings to the 2PL and 3PL models.

A Summary of Simulation Studies II and III: Investigating the Robustness of M-Fluctuation Tests

The conditions used in Simulation Study I can be regarded as rather ideal situations, which are usually not encountered in practical data analysis. In Simulation Studies II and III, we therefore aimed at investigating the power of M-fluctuation tests under a wide range of more realistic conditions. We focus on the principal results of these

studies here and only present a summary of the design and the main findings; additional details will be made available as a second online document (https://doi.org/10.5167/uzh-151192) accompanying this study.

Simulation Studies II and III included datasets which differed with regard to the following characteristics from the datasets investigated in Simulation Study I:

- The person parameters were drawn from a normal or from a skewed distribution (Simulation Studies II and III)
- DIF and impact effects were parallel or orthogonal (i.e., related to orthogonal person covariates) (Simulation Study II)
- Datasets for which the groups affected by impact were slightly misspecified in the item parameter estimation (Simulation Study III)

As in Simulation Study I, impact was modeled using a model constraint that allowed the impact groups to differ with regard to their person parameter distribution, but assumed all item parameters to be constant for all groups.

In summary, Simulation Studies II and III investigated conditions in which specific assumptions of the item parameter estimation, like the normal distribution of the person parameters or the groups for which impact is modeled, were violated. This is a novel aspect that has not been investigated in previous simulation studies by Strobl et al. (2015), Komboz et al. (2018), or T. Wang et al. (2018).

Central Findings From Simulation Studies II and III. Overall, M-fluctuation tests were found to have power against uniform and nonuniform DIF also under the more general conditions of Simulation Studies II and III. Again, the Type I error rate was close to the nominal alpha level if impact effects were modeled. In general, the model constraints have only little effect on the power of the M-fluctuation tests in both simulation studies. The only exception are constraints that do not account for impact effects that are present in the variable that is tested for DIF. This further supports our recommendation that it is crucial to model possible impact effects in variables that are tested for DIF in the estimation step of the procedure.

Orthogonal impact effects do not seem to increase the Type I error rate or to affect the power of the M-fluctuation tests, even if they are not modeled. Finally, it was found that M-fluctuation tests tend to be robust against a moderately skewed distribution of the person parameters and against misspecifications of the groups affected by impact effects.

General Discussion

This article described a method to calculate M-fluctuation tests for the 2PL model and a constrained 3PL model. As the results of the simulation studies showed, the M-fluctuation tests generally have power against DIF in the slope, intercept, and

pseudo-guessing parameters. The presented method therefore allows the detection of uniform and nonuniform DIF effects.

In contrast to the M-fluctuation tests for Rasch models (Komboz et al., 2018; Strobl et al., 2015), the presented method is based on a MML estimation of the item parameters, which assumes normally distributed person parameters. The results of Simulation Studies II and III indicated that M-fluctuation tests are robust against moderate violations of this assumption (i.e., a skewed person parameter distribution).

In empirical datasets, person covariates are often related to differences in ability. We proposed the application of the multiple-group IRT framework described by Bock and Zimowski (1997) to model possible impact effects. As the results of Simulation Study I suggest, an impact effect in a person covariate that is tested for DIF can increase the Type I error if it is not accounted for in the item parameter estimation step. However, we found that impact effects in orthogonal covariates do not lead to an increase in the Type I error rate. Our results also indicated that a slight misspecification of the groups affected by impact does generally not lead to a severe Type I error inflation.

Among the possible constraints on the item and person parameters in the multiple-group framework, we found a specification that constrains the item parameters to be equal across all groups, but which allows a free estimation of the person parameter distributions, to be most useful. This model constraint can therefore by recommended for practical data analyses.

As a conclusion of this article, we also want to present a brief outlook on future work. In empirical datasets, the distribution of latent traits may not only differ with regard to categorical person covariates (such as gender), but also with regard to continuous person covariates (such as age). In this case a discretization could be used to define groups of respondents in the multiple-group IRT framework with potentially different ability distributions. As the results presented in our simulations indicate, M-fluctuation tests are robust against a slight misspecification of the groups that are affected by impact effects. Nevertheless, it seems interesting to evaluate and compare possible strategies for addressing continuous person covariates.

The presented approach for developing M-fluctuation tests for the 2PL model and constrained 3PL model can be easily applied to develop M-fluctuation tests for additional IRT models, like multidimensional models (Reckase, 2009), and models for polytomous items, like the generalized partial credit model (GPCM; Muraki, 1992). We did not consider these models here, since they typically require larger samples. However, preliminary simulations, which we did not present in this article for brevity, indicate that M-fluctuation tests can also be used with the 3PL model, the two-dimensional 2PL model (Reckase, 2009), and the GPCM. An R package that allows the calculation of M-fluctuation tests for these models is currently in preparation. Finally, the presented M-fluctuation tests can be used as a foundation for developing methods analogous to Rasch trees (Strobl et al., 2015).

Software Used in This Study

All analyses were carried out using the R environment for statistical computing (R Core Team, 2017), versions 3.3.2 and 3.3.3. The item parameters for the 2PL and 3PL models were estimated using the mirt package (Chalmers, 2012), version 1.20.1. The subsequent calculation of the M-fluctuation tests was carried out with the strucchange package (Zeileis, Leisch, Hornik, & Kleiber, 2002), version 1.5-1. The calculation of the scores was carried out with new R code, which was written specifically for this study.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by the Swiss National Science Foundation, grant number 100019_152548.

References

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Boca Raton, FL: CRC Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392-479). Reading, MA: Addison-Wesley.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433-448). New York, NY: Springer.

Cao, M., Tay, L., & Liu, Y. (2017). A Monte Carlo study of an iterative Wald test procedure for DIF analysis. *Educational and Psychological Measurement*, *77*, 104-118.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. Retrieved from http://www.jstatsoft.org/v48/i06/

Chen, J.-H., Chen, C.-T., & Shih, C.-L. (2014). Improving the control of Type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement*, *38*(1), 18-36.

De Ayala, R. (2009). *The theory and practice of item response theory*. New York, NJ: Guilford Press.

DeMars, C. E., & Jurich, D. P. (2015). The interaction of ability differences and guessing when modeling differential item functioning with the Rasch model. Conventional and tailored calibration. *Educational and Psychological Measurement*, *75*(4), 610-633.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

Ghosh, M. (1995). Inconsistent maximum likelihood estimators for the Rasch model. *Statistics & Probability Letters*, *23*(2), 165-170.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, *8*, 647-667.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, *64*, 273-294.

Glas, C. A. W. (2001). Differential item functioning depending on general covariates. In A. Boomsma, M. A. van Duijn & T. A. Sniders (Eds.), *Essays on item response theory* (pp. 131-148). New York, NY: Springer.

Glas, C. A. W. (2010). Item parameter estimation and item fit analysis. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 269-288). New York, NY: Springer.

Glas, C. A. W., & Suárez-Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*, 87-106.

Glas, C. A. W., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology*, *63*, 603-626.

Halpern, D. F. (2012). *Sex differences in cognitive abilities* (4th ed.). New York, NY: Psychology Press.

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

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. New York, NY: Taylor & Francis.

Komboz, B., Strobl, C., & Zeileis, A. (2018). Tree-based global model tests for polytomous Rasch models. *Educational and Psychological Measurement*, *78*, 128-166. doi: 10.1177/0013164416664394

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

Kopf, J., Zeileis, A., & Strobl, C. (2015b). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement*, *39*, 83-103.

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*, 647-677. doi:10.1007/BF02294041

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.

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

McDonald, R. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. New York, NY: Taylor & Francis.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Thousand Oaks, CA: Sage.

R Core Team. (2017). R: *A language and environment for statistical computing* [Computer software manual]. Retrieved from https://www.R-project.org/

Reckase, M. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271-282.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*, 75-92.

Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 257-268). New York, NY: Springer.

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.

Tay, L., Huang, Q., & Vermunt, J. K. (2016). Item response theory with covariates (IRT-C). *Educational and Psychological Measurement*, *76*, 22-42.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement*, *28*, 389-406.

von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 99-115). New York, NY: Springer.

Wang, T., Merkle, E. C., & Zeileis, A. (2014). Score-based tests of measurement invariance: use in practice. *Frontiers in Psychology*, *5*, 438. doi:10.3389/fpsyg.2014.00438

Wang, T., Strobl, C., Zeileis, A., & Merkle, E. C. (2018). Score-based tests of differential item functioning via pairwise maximum likelihood estimation. *Psychometrika*, *83*, 132-155. doi:10.1007/s11336-017-9591-8

Wang, W.-C., Shih, C.-L., & Sun, G.-W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, *72*, 687-708.

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*, 488-508.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*, 492-514.

Zeileis, A., Leisch, F., Hornik, K., & Kleiber, C. (2002). strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, *7*(1), 1-38. doi:10.18637/jss.v007.i02