AStA Adv Stat Anal (2010) 94: 117–127
DOI 10.1007/s10182-010-0126-1

ORIGINAL ARTICLE

Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias; a simulation study

M.T. Barendse · F.J. Oort · G.J.A. Garst

Received: 2 November 2009 / Accepted: 30 January 2010 / Published online: 26 May 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract Factor analysis is an established technique for the detection of measurement bias. Multigroup factor analysis (MGFA) can detect both uniform and nonuniform bias. Restricted factor analysis (RFA) can also be used to detect measurement bias, albeit only uniform measurement bias. Latent moderated structural equations (LMS) enable the estimation of nonlinear interaction effects in structural equation modelling. By extending the RFA method with LMS, the RFA method should be suited to detect nonuniform bias as well as uniform bias. In a simulation study, the RFA/LMS method and the MGFA method are compared in detecting uniform and nonuniform measurement bias under various conditions, varying the size of uniform bias, the size of nonuniform bias, the sample size, and the ability distribution. For each condition, 100 sets of data were generated and analysed through both detection methods. The RFA/LMS and MGFA methods turned out to perform equally well. Percentages of correctly identified items as biased (true positives) generally varied between 92% and 100%, except in small sample size conditions in which the bias was nonuniform and small. For both methods, the percentages of false positives were generally higher than the nominal levels of significance.

Keywords Measurement bias · Differential item functioning · Nonlinear structural equation modelling · Factor analysis · Latent moderated structures

M.T. Barendse · F.J. Oort (✉) · G.J.A. Garst
Department of Education, University of Amsterdam, Nieuwe Prinsengracht 130,
1018 VZ Amsterdam, The Netherlands
e-mail: F.J.Oort@uva.nl

G.J.A. Garst
e-mail: G.J.A.Garst@uva.nl

1 Introduction

Measurement bias may jeopardise all research, especially behavioural and social science research in which subjective measures are used. Respondents with the same ability (or trait, attitude, mood, etc.) should get equal test scores, but structural bias may prevent this. In the presence of measurement bias, observed differences in item and test scores do not reflect true differences between respondents. Therefore, it is important to investigate measurement bias in all tests, with respect to all relevant variables, to improve test validity and to establish fairness in tests for all respondents.

Measurement bias (or item bias, or differential item functioning (DIF)) can formally be defined as a violation of measurement invariance (after Mellenbergh 1989):

$$f_1(X|T=t, V=v) = f_2(X|T=t), \tag{1}$$

where X is a set of observed variables (e.g. test items or questionnaire scales), T is the concept of interest measured by X, and V is a set of variables other than T, possibly violating conditional independence. Function f_1 is the conditional distribution function of X given values t and v, and f_2 is the conditional distribution function of X given t. If the conditional independence does not hold, that is, if $f_1 \neq f_2$, then the measurement of T by X is said to be biased with respect to V.

The formal definition of (1) can be explained with the graphical display in Fig. 1. In Fig. 1(a), there is no bias, but in Fig. 1(b) variable V explains variance in measurement X in addition to what is already explained by the concept of interest T. In other words, measurement X is biased with respect to variable V because X does not just measure T but V as well. In Fig. 1(b), this bias is uniform, but in Fig. 1(c) there also is an interaction effect of T and V on X, indicating nonuniform bias, where the extent of bias varies with levels of T. For example, if T is mathematical ability, X is a worded mathematical problem, and V is verbal ability, then in Fig. 1(a), mathematical ability and verbal ability are correlated, but verbal ability does not directly affect X, whereas in Fig. 1(b) it does. So, in Fig. 1(b), the worded mathematical problem does not just measure mathematical ability, but verbal ability as well. If the effect of verbal ability on X varies with different levels of mathematical ability (e.g. only above a certain threshold), or if verbal ability affects X only if the verbal ability is insufficient, then the bias in X is nonuniform, as in Fig. 1(c).

Mellenbergh (1989) introduced the principle of conditional independence to define item bias (or DIF). In this definition, the concept of interest T can be operationalised with either a latent variable, as in item response models, or with an observed variable that serves as a proxy for the latent trait, as in contingency table models. Millsap and Everson (1993) reviewed statistical methods for the detection of measurement

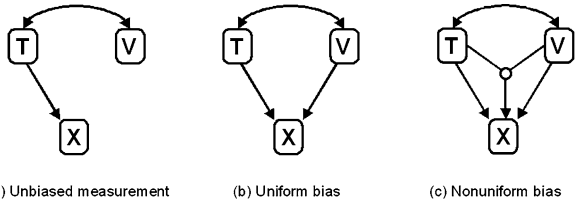

Fig. 1 Graphical representation of (**a**) unbiased measurement of T with respect to V, (**b**) uniform bias in X with respect to V, and (**c**) nonuniform bias in X with respect to V

bias with both types of models. The early latent variable methods relied on item response theory (Lord 1980), but Meredith (1993) applied Mellenbergh's definition to multigroup factor analysis (MGFA) to define weak measurement invariance, strong factorial invariance, and strict factorial invariance. These hypotheses of invariance are generally tested through MGFA (reviewed by Vandenberg and Lance 2000), but Oort (1992, 1998) suggested the use of restricted factor analysis (RFA) as a means to investigate bias with respect to group membership (and other variables; Oort 1991).

In both MGFA and RFA, the concept of interest T is operationalised as a (latent) common factor with multiple measures X as (observed) indicators. In MGFA, uniform and nonuniform bias can be detected by testing across group constraints on intercepts and factor loadings. If intercepts vary, the *difficulty* of the associated measurement X varies across groups (uniform bias), and if factor loadings vary then the *discrimination* between different levels of T varies across groups (nonuniform bias). In RFA, the data of different groups are taken together and group membership is added to the model as an exogenous variable V that covaries with T. Measurement bias is indicated by direct effects of this V variable on the X variables. The RFA method to detect measurement bias is equivalent to the multiple indicator multiple cause (MIMIC) analysis, but in MIMIC models the V variables have causal effects on the T variables (Muthén 1989).

Possible advantages of RFA (and MIMIC analysis) over MGFA when investigating measurement invariance are that in RFA variables V can be continuous or discrete, observed or latent, and measurement bias can be investigated with respect to multiple variables V simultaneously. Moreover, as it is not necessary to divide the sample into sub-samples by V, RFA is also believed to yield more precise parameter estimates and to have more statistical power to detect measurement bias.

A disadvantage of RFA is that it is not readily suited to detect nonuniform bias. In the RFA model, nonuniform bias would appear as a nonlinear interaction effect, violating the assumption of multivariate normality. There are two main classes of approaches for analysis of interaction effects (Moosbrugger et al. 2009; Schermelleh-Engel et al. 2010): the product indicator approaches and the distribution-analytic approaches The product indicator approaches, as first described by Kenny and Judd (1984), require a measurement model for the nonlinear products of the observed variables. The distribution-analytic approaches are based on the analysis of the multivariate density function of the indicator variables that takes the non-normal distribution into account. The two distribution-analytic approaches that have been proposed are known as the latent moderated structures (LMS) approach (Klein and Moosbrugger 2000) and the quasi-maximum likelihood approach (Klein and Muthén 2007). LMS has been implemented in the computer program M-plus (Muthén and Muthén 2001). In a simulation study, the LMS estimates proved to be consistent, unbiased and efficient (Klein and Moosbrugger 2000).

The purpose of the present paper is to investigate whether RFA with LMS enables the detection of nonuniform measurement bias. In a simulation study of uniform and nonuniform measurement bias detection, the performance of the RFA/LMS method will be compared with the MGFA method.

2 Methods

Measurement bias in simulated data will be detected with both the RFA/LMS method and the MGFA method. Keeping in line with other simulation studies of bias detection, and to not disadvantage the MGFA method, we will consider a dichotomous violator representing two groups. Consequently, the MGFA model will be used for the generation of data.

2.1 Data generation

Data were generated for two groups of subjects, using

$$x_j = \tau_g + \Lambda_g t_j + \delta_g \varepsilon_j, \tag{2}$$

as the model for the observed item scores of subject j in group g, where x_j is a vector of six item scores, t_j is the subject's score on the common factor (i.e. the trait of interest T), ε_j is the subject's score on the residual factor, τ_g is a vector of six intercepts, Λ_g is a vector of six common factor loadings, and δ_g is a vector of six residual factor loadings. Bias was introduced in the first of the six items, by introducing across group differences in intercepts and factor loadings. Factors that were varied included the size of uniform bias (0, 0.5, or 0.8 between group difference in the intercept of the biased item), size of nonuniform bias (0, 0.25, or 0.5 difference in the factor loading), sample size (2×100 or 2×500 subjects), and across group difference in ability distribution (0 or 0.5 standard deviation difference in the group mean). In a fully crossed design, these four factors would yield 36 different conditions, but we selected the 15 most interesting ones (see the first column of Table 1 for an overview). The number of replications was 100 in each of the 15 conditions.

Subject parameters t_j and ε_j were drawn from the normal distribution with mean zero and standard deviation 1: $t_j \sim N(0, 1)$ and $\varepsilon_j \sim N(0, 1)$. In the small sample size conditions, the number of subjects was 100 in each group, and in the large sample size conditions the number of subjects was 500 in each group. In conditions with a medium difference in the group means, t_j values for Group 2 subjects were drawn from a normal distribution with mean -0.5 and standard deviation 1, i.e. $t_j \sim N(-0.5, 1)$.

All intercepts τ were chosen equal to 0, except for the intercept for the first item in Group 2, which was chosen equal to 0 (no uniform bias), -0.5 (small uniform bias), or -0.8 (large uniform bias). All common factor loadings λ were chosen equal to 0.8, except for the factor loading for the first item in Group 2, which was chosen equal to 0.8 (no nonuniform bias), 0.55 (small nonuniform bias), or 0.3 (large nonuniform bias). All residual factor loadings δ were chosen equal to the square root of $(1 - \lambda_g^2)$.

2.2 Analyses

We used the computer program M-plus (version 4; Muthén and Muthén 2001) to generate data and to apply both the RFA/LMS method and the MGFA method to each of the 1500 data sets (100 replications in each of the 15 conditions).

For the purpose of the RFA/LMS method, the data of Group 1 and Group 2 were stacked, and a coding for group membership was added, yielding one hundred 200×7

matrices of observed item responses in each of the small sample conditions and one hundred 1000 × 7 matrices in each of the large sample conditions. In the RFA/LMS method, the observed scores on the six items are modelled as

$$x_j = \tau + \Lambda t_j + b v_j + c t_j v_j + \delta \varepsilon_j, \tag{3}$$

where t_j is the score on the common factor T, v_j is a dummy coding for group membership V of subject j, ε_j is the residual score of subject j, τ is a vector of six intercepts, Λ is a vector of six common factor loadings, δ is a vector of six residual factor loadings, and b and c are vectors containing six regression coefficients. A non-zero element in b indicates uniform bias and a non-zero element in c indicates nonuniform bias. In order to enable the estimation of the model parameters through RFA/LMS, group membership is modelled as a latent variable with a single observed indicator without residual variance and with the factor loading fixed at unity (however, to overcome identification problems the residual variance had to be fixed at a non-zero value; we chose 0.001). The parameters of the RFA/LMS model can be estimated with M-plus; see Appendix for an example script. Measurement bias is detected by comparing the fit of a null model in which both b and c are zero vectors ($b = \mathbf{0}$ and $c = \mathbf{0}$) with the fit of six alternative models in which for one of the items the corresponding b and c elements are set free to be estimated. The RFA/LMS method as implemented in M-plus utilises robust maximum likelihood estimation with a scaling correction to account for the violation of distributional assumptions (Muthén and Muthén 2001). For each item, the difference between the log-likelihood values associated with the null model and the alternative model has a chi-square distribution with two degrees of freedom, subject to the scaling correction factors of the two models (Satorra and Bentler 2001).

In the MGFA method, a one-factor model is fitted to the separate 6 × 6 variance-covariance matrices and 6 × 1 mean vectors of the two groups, with across group equality constraints on intercepts and factor loadings. The common factor mean and variance are fixed for the first group and free to be estimated in the second group. The maximum likelihood estimation method is used to estimate all model parameters. Similar to the procedure in the RFA/LMS method, measurement bias is detected by comparing the fit of a null model with the fit of six alternative models. In the null model, all intercepts and factor loadings are constrained to be equal across groups, whereas in the alternative models the factor loadings and intercepts of one item are free to be estimated in both groups. An across group difference in intercepts indicates uniform bias ($\tau_1 \neq \tau_2$) and an across group difference in factor loadings indicates nonuniform bias ($\lambda_1 \neq \lambda_2$). Here, for each item, we use the difference in the chi-square values associated with the null and alternative model as a global two degrees of freedom test to detect uniform and/or nonuniform bias.

After applying both the RFA/LMS method and the MGFA method to each of the 1500 data sets, we determined how often the methods indicated bias in one of the items. We tested at 5%, 1%, and 0.1% levels of significance. For each of the 15 conditions and for each level of significance, we counted "true positives" and "false positives". A true positive is a biased item that was correctly detected as biased, and a false positive is an unbiased item that was incorrectly detected as biased.

3 Results

The results of bias detection using the RFA/LMS and MGFA methods are given in Table 1. The first column describes the condition parameters (sample size, ability distribution, and size of uniform and nonuniform bias). For each condition and for each method, the mean and the standard deviation of the chi-square difference tests of measurement bias are given, together with proportions of items detected as biased at varying levels of significance. These means, standard deviations, and proportions are calculated separately over 100 observations (i.e. 100 replications) for the first item (with varying levels of bias) and over 500 observations (100 replications $\times$ 5 items) for the other five items (Items 2 through 5 without bias).

From Table 1 it appears that in conditions without bias (Conditions 1, 8, 12), the means of the chi-square values for the first item are equal to those for the other five items, whereas in conditions with bias, the means of the chi-square values for the first item are clearly higher than those for the other five items. This is true for both the RFA/LMS method and the MGFA method, which two methods seem to perform equally well.

When testing at the 5% level of significance, the proportions of true positives in conditions with uniform bias were very high regardless the size of uniform bias and the sample size (92% to 100% with both methods). The methods performed worse when detecting nonuniform bias (RFA/LMS 41% to 100% true positives and MGFA 52% to 100% true positives). The low proportions of true positives were found in conditions where small sample size was combined with small nonuniform bias. In conditions with large nonuniform bias, the proportions of true positives were very high (92% to 100% with both methods). In conditions with a large sample size, the proportions of true positives were very high (RFA/LMS 96% to 100% and MGFA 99% to 100%), but in conditions with smaller sample sizes the proportions of true positives were lower (RFA 41% to 100% correct and MG 52% to 100% correct). As mentioned earlier, the lower percentages of true positives were found in conditions with small sample size in combination with small nonuniform bias. With both methods, group differences in ability did not systematically affect the proportions of true positives.

The proportions of false positives that we found when we tested at the 5% level of significance are generally larger than 0.05, especially in conditions with large sample sizes and large sizes of bias. Testing at lower levels of significance alleviates this problem, although the actual proportions of false positives are still higher than the nominal level of significance. This is true for both methods. Moreover, in conditions with small sample sizes and conditions with nonuniform bias, lowering the level of significance negatively affects the proportions of true positives, which is also true for both methods.

4 Discussion

When RFA was introduced as a method for measurement bias detection, it was stated that the method is only suited for the detection of uniform bias (Oort 1992). However, with the new possibilities of estimating interaction effects in structural equation

Table 1 Measurement bias detection results

Condition	RFA/LMS method					MGFA method				
	χ^2		Proportion of bias			χ^2		Proportion of bias		
	mean	st.d.	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$	mean	st.d.	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Small sample size										
No difference in ability										
1. 1 item; no bias	2.39	2.33	*0.070*	*0.020*	*0*	2.36	2.03	*0.070*	*0*	*0*
5 items; no bias	2.36	2.41	*0.062*	*0.022*	*0.008*	2.25	2.40	*0.070*	*0.018*	*0.006*
2. 1item; small unif bias	31.38	12.40	**1**	**0.980**	**0.920**	30.17	10.44	**0.990**	**0.980**	**0.940**
5 items; no bias	3.01	2.84	*0.122*	*0.028*	*0.008*	3.03	2.85	*0.134*	*0.036*	*0.004*
3. 1 item; large unif bias	68.96	19.45	**1**	**1**	**1**	66.59	14.78	**1**	**1**	**1**
5 items; no bias	3.78	3.33	*0.214*	*0.070*	*0.014*	3.82	3.35	*0.232*	*0.082*	*0.010*
4. 1 item; small nonunif bias	6.20	6.08	**0.410**	**0.210**	**0.060**	7.16	5.50	**0.520**	**0.300**	**0.070**
5 items; no bias	2.11	2.12	*0.056*	*0.012*	*0.004*	2.16	2.09	*0.068*	*0.012*	*0*
5. 1 item; large nonunif bias	17.04	11.26	**0.920**	**0.770**	**0.580**	17.82	8.55	**0.940**	**0.850**	**0.340**
5 items with no bias	2.14	2.12	*0.060*	*0.010*	*0.004*	2.28	2.13	*0.070*	*0.016*	*0.002*
6. 1 item; small unif/nonunif bias	28.09	12.14	**1**	**0.990**	**0.980**	27.75	9.18	**1**	**0.990**	**0.930**
5 items; no bias	2.48	2.43	*0.092*	*0.020*	*0.002*	2.56	2.45	*0.100*	*0.020*	*0.002*
7. 1 item; large unif/nonunif bias	60.46	20.93	**1**	**1**	**1**	57.27	13.58	**1**	**1**	**1**
5 items; no bias	2.47	2.38	*0.096*	*0.016*	*0.002*	2.67	2.50	*0.108*	*0.026*	*0.002*
Medium difference in ability										
8. 1 item; no bias	2.37	2.31	*0.090*	*0.020*	*0*	2.35	2.09	*0.009*	*0*	*0*
5 items; no bias	2.38	2.53	*0.066*	*0.016*	*0.010*	2.24	2.47	*0.064*	*0.016*	*0.008*

Table 1 (*Continued*)

Condition	RFA/LMS method					MGFA method				
	χ^2		Proportion of bias			χ^2		Proportion of bias		
	mean	st.d.	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$	mean	st.d.	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
9. 1 item; small unif bias	29.06	11.47	**0.990**	**0.960**	**0.900**	28.28	10.18	**0.990**	**0.960**	**0.920**
5 items; no bias	3.28	3.72	*0.140*	*0.032*	*0.010*	3.19	2.93	*0.148*	*0.040*	*0.008*
10. 1 item; small nonunif bias	6.40	6.28	**0.420**	**0.210**	**0.060**	7.57	6.05	**0.550**	**0.026**	**0.110**
5 items; no bias	2.10	2.10	*0.060*	*0.016*	*0*	2.18	2.12	*0.094*	*0.018*	*0*
11. 1 item; small unif/nonunif bias	21.78	10.43	**0.950**	**0.920**	**0.750**	20.74	8.96	**0.940**	**0.920**	**0.770**
5 items; no bias	2.43	2.53	*0.106*	*0.022*	*0.006*	2.42	2.33	*0.082*	*0.026*	*0.002*
Large sample size										
No difference in ability										
12. 1 item; no bias	2.17	2.03	*0.050*	*0.010*	*0*	2.20	2.14	*0.060*	*0.020*	*0*
5 items; no bias	2.12	1.93	*0.050*	*0.008*	*0*	2.17	2.06	*0.072*	*0.008*	*0*
13. 1 item; small unif bias	139.20	25.33	**1**	**1**	**1**	136.76	23.15	**1**	**1**	**1**
5 items; no bias	6.65	4.64	*0.484*	*0.262*	*0.078*	6.76	4.83	*0.482*	*0.272*	*0.094*
14. 1 item; small nonunif bias	18.82	8.25	**0.960**	**0.880**	**0.720**	25.02	9.37	**0.990**	**0.990**	**0.880**
5 items; no bias	2.32	2.23	*0.072*	*0.014*	*0.004*	2.70	2.58	*0.106*	*0.026*	*0.004*
15. 1 item; small unif/nonunif bias	119.33	23.30	**1**	**1**	**1**	123.83	22.16	**1**	**1**	**1**
5 items; no bias	4.09	3.39	*0.238*	*0.082*	*0.020*	4.61	3.78	*0.288*	*0.114*	*0.032*

Notes: Proportion of bias is the proportion of items that is indicated as biased by the detection method at 5%, 1%, and 0.1% levels of significance; in bold typeset: proportions of true positives, calculated over 100 observations; in italics: proportions of false positives, calculated over $5 \times 100 = 500$ observations

models, the RFA method can be extended to also detect nonuniform bias. One of the advantages of the RFA method over the MGFA method is that it is not necessary to divide the sample into sub-samples. We therefore expected the RFA method to have more statistical power to detect measurement bias. However, in our study, the RFA/LMS and MGFA methods performed about equally well. A possible explanation is that in our MGFA procedure, we begin with across group constraints on all factor loadings and intercepts, thus limiting the difference between the two methods in the numbers of parameters to be estimated.

Another advantage of the RFA/LMS method over the MGFA method that has been mentioned is the possibility to investigate bias with respect to any violator variable, continuous or discrete, observed or latent. As a matter of fact, the LMS method is really suited for estimating interaction effects of latent variables only. We circumvented this problem by introducing group membership as a latent variable with a single indicator with a fixed factor loading and fixed residual variance. To overcome identification problems this residual variance had to be fixed at a non-zero value. Still, the RFA/LMS method performed very well, at least as well as the MGFA method. Yet another advantage of the RFA/LMS method is the possibility to investigate bias with respect to multiple violator variables simultaneously. In the MGFA method, this can only be done separately or by crossing factor levels and creating multiple smaller groups, which would complicate the analysis and yield less accurate parameter estimates.

The possibility of including multiple violator variables is especially important because in practise there may be many violators of the measurement model, some known and some unknown. Moreover, even if known, they may not be operationalised or available to the researchers. In such cases, we can still detect bias with respect to other variables (such as group membership) that are related to the actual biasing variables. For example, if worded math problems are biased with respect to verbal ability, but we did not measure verbal ability, then we can still detect bias with respect to group membership (e.g. with groups consisting of either native speakers or non-native speakers).

In the present research, we chose to combine the RFA method with LMS to estimate interaction effects, because it is implemented in M-plus and readily available. However, the newer quasi-maximum likelihood approach to the estimation of interaction effects (Klein and Muthén 2007) makes less stringent assumptions than LMS and may also be suitable for nonuniform bias detection. Future simulation studies of measurement bias detection should also include this newer method. In addition, future studies should conduct bias detection in an iterative manner. In the present study, we ran the detection procedures only once for every data set and then counted true positives and false positives. However, it has been demonstrated that it is better to conduct the RFA procedure iteratively (Oort 1998; Navas-Ara and Gómez-Benito 2002). That is, account for the item with the largest bias and rerun the bias detection procedure until no bias is found. Finally, future studies could investigate the behaviour of the RFA method with multiple violator variables, multiple biased items, and longer tests, to better represent the actual data sets that one generally encounters in substantive research (see Jak et al. 2010 for an example).

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Appendix: Example M-plus script for fitting the RFA/LMS model

```
title: detection of uniform and nonuniform bias in the first item
data: file is cond2r3.dat ;
variable: names are y1 y2 y3 y4 y5 y6 y7
analysis: type = random;
algorithm = integration;
model:
  ability by y1*0.5 y2*.5 y3*.5 y4*.5 y5*.5 y6*.5;
  violat by y7 @ 0.5;       !y7 is indicator for violator variable
  violat by y1;             !estimate regression on violator variable
  ability with violat;
  ability @1;
  violat @1;
  [ability @ 0];
  [violat @ 0];
  [y7 @ 1.5];       !alt option: set free to be estimated
  y7 @ .001;        !alt option: fix at .01 and set violat by y7 free
  abxvio | ability xwith violat; !introduce interaction term
  y1 on abxvio;     !estimate regression on interaction term
output: tech1 tech8 tech9;
savedata:
  results are cond2r3.res; !save the results
```

References

Jak, S., Oort, F.J., Dolan, C.V.: Measurement bias and multidimensionality; an illustration of bias detection in multidimensional measurement models. Adv. Stat. Anal. (2010). doi:10.1007/s10182-010-0128-z

Kenny, D., Judd, C.M.: Estimating the nonlinear and interactive effects of latent variables. Psychol. Bull. **96**, 201–210 (1984)

Klein, A.G., Moosbrugger, H.: Maximum likelihood estimation of latent interaction effects with the LMS method. Psychometrika **65**, 457–474 (2000)

Klein, A.G., Muthén, B.O.: Quasi maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. Multivar. Behav. Res. **42**, 647–673 (2007)

Lord, F.M.: Applications of Item Response Theory to Practical Testing Problems. Erlbaum, Hillsdale (1980)

Mellenbergh, G.J.: Item bias and item response theory. Int. J. Educ. Res. **13**, 127–143 (1989)

Meredith, W.: Measurement invariance, factor analysis, and factorial invariance. Psychometrika **58**, 525–543 (1993)

Millsap, R.E., Everson, H.T.: Methodology review: statistical approaches for assessing measurement bias. Appl. Psychol. Meas. **17**, 297–334 (1993)

Moosbrugger, H., Schermelleh-Engel, K., Kelava, A., Klein, A.G.: Testing multiple nonlinear effects in structural equation modelling: a comparison of alternative estimation approaches. In: Teo, T., Khine, M.S. (eds.) Structural Equation Modeling in Educational Research: Concepts and Applications, pp. 103–136. Sense, Rotterdam (2009)

Muthén, B.O.: Latent variable modeling in heterogeneous populations. Psychometrika **54**, 557–585 (1989)

Muthén, B.O., Muthén, L.K.: M-plus User's Guide: Statistical Analysis with Latent Variables. Muthén & Muthén, Los Angeles (2001)

Navas-Ara, M.J., Gómez-Benito, J.: Effects of ability scale purification on the identification of DIF. Eur. J. Psychol. Assess. **18**, 9–15 (2002)

Oort, F.J.: Theory of violators: assessing unidimensionality of psychological measures. In: Steyer, R., Wender, K.F., Widaman, K.F. (eds.) Psychometric Methodology, pp. 377–381. Fischer, Stuttgart (1991)
Oort, F.J.: Using restricted factor analysis to detect item bias. Methodika **6**, 150–166 (1992)
Oort, F.J.: Simulation study of item bias detection with restricted factor analysis. Struct. Equ. Model. **5**, 107–124 (1998)
Satorra, A., Bentler, P.M.: A scaled difference chi-square test statistic for moment structure analysis. Psychometrika **66**, 507–514 (2001)
Schermelleh-Engel, K., Werner, C.S., Klein, A.G., Moosbrugger, H.: Nonlinear structural equation modeling: is partial least squares an alternative? Adv. Stat. Anal. (2010). doi:10.1007/s10182-010-0132-3
Vandenberg, R.J., Lance, C.E.: A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organ. Res. Methods **2**, 4–69 (2000)