On the power of the test for cluster bias

Jak, S.; Oort, F.J.

DOI
10.1111/bmsp.12053
Publication date
2015
Document Version
Final published version
Published in
British Journal of Mathematical & Statistical Psychology

Link to publication

Citation for published version (APA):
Jak, S., & Oort, F. J. (2015). On the power of the test for cluster bias. *British Journal of Mathematical & Statistical Psychology*, *68*(3), 434-455. https://doi.org/10.1111/bmsp.12053

British Journal of Mathematical and Statistical Psychology (2015), 68, 434–455
© 2015 The British Psychological Society

The British Psychological Society

www.wileyonlinelibrary.com

On the power of the test for cluster bias

Suzanne Jak[1,2]* and Frans J. Oort[3]

[1]Utrecht University, The Netherlands
[2]National University of Singapore, Singapore
[3]University of Amsterdam, The Netherlands

Cluster bias refers to measurement bias with respect to the clustering variable in multilevel data. The absence of cluster bias implies absence of bias with respect to any cluster-level (level 2) variable. The variables that possibly cause the bias do not have to be measured to test for cluster bias. Therefore, the test for cluster bias serves as a global test of measurement bias with respect to any level 2 variable. However, the validity of the global test depends on the Type I and Type II error rates of the test. We compare the performance of the test for cluster bias with the restricted factor analysis (RFA) test, which can be used if the variable that leads to measurement bias is measured. It appeared that the RFA test has considerably more power than the test for cluster bias. However, the false positive rates of the test for cluster bias were generally around the expected values, while the RFA test showed unacceptably high false positive rates in some conditions. We conclude that if no significant cluster bias is found, still significant bias with respect to a level 2 violator can be detected with an RFA model. Although the test for cluster bias is less powerful, an advantage of the test is that the cause of the bias does not need to be measured, or even known.

1. Introduction

The importance of establishing measurement invariance of research instruments is widely recognized; a measurement instrument should function identically in different groups of respondents (see Cheung & Rensvold, 1998; Meredith, 1993; Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000). If measurement invariance does not hold with respect to some variable (e.g., gender), then two respondents with identical values on the latent trait that the test is supposed to measure may have different expected scores, depending on their value on the other variable (e.g., depending on being a man or a woman). When a test is not invariant with respect to gender, we call the test biased with respect to gender, and gender is then called a *violator* (violator of measurement invariance; Oort, 1992). Measurement bias is also referred to as differential item functioning (DIF) in the item response theory (IRT) literature. Within structural equation modelling (SEM), the two prevalent models to investigate measurement bias are multigroup models (Horn & McArdle, 1992; Little, 1997; Sörbom, 1974; Widaman & Reise, 1997) and multiple-indicator multiple-cause (MIMIC) models (Muthén, 1989) or, equivalently, restricted factor analysis (RFA; Oort, 1992, 1998). We will use RFA models to detect measurement invariance with

*Correspondence should be addressed to Suzanne Jak, Padualaan 14, 3584 CH Utrecht, The Netherlands (email: s.jak@uu.nl).

DOI:10.1111/bmsp.12053

respect to specific violators in this study. The RFA method will be explained in Section 1.3.

In the present study we consider the investigation of measurement invariance in two-level data. Two-level data are data with a clustered structure, such as children in school classes or patients in hospitals. Measurement invariance with respect to clusters can be tested with a multigroup model with a large number of groups. This is, however, a cumbersome strategy due to the large number of parameters involved. Alternative approaches can be found by using a random effects approach as described by Verhagen and Fox (2012) using multilevel random effects models in the IRT framework (De Jong, Steenkamp, & Fox, 2007; Fox & Verhagen, 2010) or in the SEM framework (Jak, Oort, & Dolan, 2013; Muthén, 1990; Rabe-Hesketh, Skrondal, & Pickles, 2004; see Muthén & Asparouhov, 2013; for an overview of these and other methods). In this study we use the method described by Jak *et al.* (2013) who introduced the term 'cluster bias' to refer to measurement bias across clusters. Cluster bias can be interpreted as measurement bias with respect to any cluster-level variable.

With two-level data, the lower level (e.g., student or patient level) is called *level 1* or the *within* level. The higher level (e.g., class or hospital level) is called *level 2* or the *between* level. With two-level data, measurement bias can be present at the within level or at the between level. Most applications of testing measurement invariance in multilevel data are about testing invariance with respect to specific variables at the between level. For example, Davidov, Dülmer, Schlüter, Schmidt, and Meuleman (2012) used multilevel SEM to test for measurement invariance across countries, and Muthén, Khoo, and Gustafsson (1997) tested for invariance across two school types. Spilt, Koomen, and Jak (2012) tested for measurement bias with respect to both a specific level 1 variable (student gender) and a specific level 2 variable (teacher gender), and also investigated cluster bias. In this paper we focus on measurement bias at level 2. For methods to test for measurement bias with respect to level 1 variables specifically, see Kim, Yoon, Wen, Luo, and Kwok (2015) and Ryu (2014).

The purpose of this study is to compare the performance of two methods to investigate measurement bias at the between level. One method is the test for cluster bias, which can be considered a global test of measurement bias at the between level, in which the violating variable(s) is (are) not necessarily measured or even known. The other method is the RFA method, which requires the operationalization of possible violators of measurement invariance, in order to include them as exogenous variables in multilevel factor analysis. The principle of two-level SEM and these two approaches are explained below.

1.1. Measurement bias at level 2

With two-level SEM, the covariance matrix is modelled as the sum of the within level (level 1) and the between level (level 2) covariance matrices (Muthén, 1990; Rabe-Hesketh *et al.*, 2004):

$$\boldsymbol{\Sigma}_{\text{total}} = \boldsymbol{\Sigma}_{\text{between}} + \boldsymbol{\Sigma}_{\text{within}}. \tag{1}$$

For example, consider data concerning the closeness of teacher–child relations, obtained using a 5-item questionnaire, completed by teachers for several of their pupils. The (pooled, within-class) differences between children are modelled by $\boldsymbol{\Sigma}_{\text{within}}$. Teachers also differ in the general closeness of their relations with children. The differences between teachers are modelled by $\mathbf{\Sigma}_{\text{between}}$. At the within and between levels, distinct measurement models can be used to describe the covariances between the item scores. In the present study we employ the linear factor model as the measurement model (Mellenbergh, 1994).

1.2. The test for cluster bias

Testing for cluster bias can be seen as a global test for measurement bias with respect to all possible level 2 violators. In the case of cluster bias, one or more indicators measure different constructs in different clusters. In the closeness example from the previous subsection, if cluster bias is present, this means that cluster-level variables other than closeness are causing differences between the teachers' scores. Jak *et al.* (2013) showed that in the absence of cluster bias, for p observed variables and k common factors, the following model holds:

$$\begin{aligned}\mathbf{\Sigma}_{\text{between}} &= \Lambda\mathbf{\Phi}_{\text{between}}\mathbf{\Lambda}',\\ \mathbf{\Sigma}_{\text{within}} &= \mathbf{\Lambda}\mathbf{\Phi}_{\text{within}}\mathbf{\Lambda}' + \mathbf{\Theta}_{\text{within}}.\end{aligned} \tag{2}$$

In this model, $\mathbf{\Phi}_{\text{between}}$ and $\mathbf{\Phi}_{\text{within}}$ are $k \times k$ covariance matrices, $\mathbf{\Theta}_{\text{within}}$ is a $p \times p$ (diagonal) matrix with residual variances, and $\mathbf{\Lambda}$ is a $p \times k$ matrix with factor loadings. The factor loadings are equal across the within and between level. Cluster bias appears as residual variance at the between level, and can be modelled by estimating a (diagonal) $p \times p$ matrix with residual variance at the between level ($\mathbf{\Theta}_{\text{between}}$), so that $\mathbf{\Sigma}_{\text{between}} = \mathbf{\Lambda}\,\mathbf{\Phi}_{\text{between}}\,\mathbf{\Lambda}' + \mathbf{\Theta}_{\text{between}}$. The test for cluster bias involves testing whether parameters in $\mathbf{\Theta}_{\text{between}}$ are zero, in a model with equal factor loadings across levels. Although Jak *et al.* (2013) showed which model holds if there is uniform cluster bias, it is not clear what the correct model is with non-uniform bias. In addition, the cluster-bias model can test whether measurement bias is absent, but cannot differentiate between uniform and non-uniform bias. If $\mathbf{\Theta}_{\text{between}} \neq \mathbf{0}$, this can be caused by a difference in factor loadings across clusters (non-uniform bias) or by a difference in intercepts across clusters (uniform bias). In this study we restrict ourselves to the evaluation of uniform measurement bias.

Testing for cluster bias is very useful in situations where it is important that a measurement instruments functions identically across a large number of groups, but the variables that potentially cause bias across groups are not measured. For example, Jak (2014) tested the measurement invariance of a dyscalculia screening instrument, and found that three scales showed measurement bias at the school class level. Inspection of the item contents gave rise to the hypothesis that the instruction a teacher gives prior to administering the test influences the test scores. However, the quality of the instruction was not measured in the study, so this hypothesis could not be tested with an RFA model.

1.3. The RFA method

If the interest is in testing for measurement bias with respect to specific between level variables that are measured in the study, these variables can be added to the model. Measurement bias with respect to such variables can be investigated using RFA (Oort, 1992). The RFA model is statistically equivalent to MIMIC models to detect measurement bias (Muthén, 1989), the only difference being that in MIMIC models, the violator variables have causal effects on the common factor, whereas in the RFA model these variables are merely associated. In both models, all influence of the potential violator on the factor indicators runs through the common factor. Uniform measurement bias is represented by a direct effect of the violator on the indicators. Non-uniform bias would indicate that the amount of bias is not equal across different values of the common factor, and is represented by an interaction effect of the common factor and the violator on the indicators. A disadvantage of the RFA (and MIMIC) method is that the detection of non-uniform bias is not straightforward. However, recent developments using latent interaction terms or moderated factor analysis provide a viable method to investigate non-uniform bias in the extended RFA framework (Barendse, Oort, & Garst, 2010; Barendse, Oort, Werner, Ligtvoet, & Schermelleh-Engel, 2012; see also Molenaar, Dolan, Wicherts, & van der Maas, 2010). Still, because we want to be able to compare the RFA method with the test for cluster bias, we do not consider non-uniform measurement bias in this study. Henceforth, if we write 'measurement bias', this refers to *uniform* measurement bias.

1.4. Combining the test for cluster bias and the RFA method

Jak, Oort, and Dolan (2014) describe a five-step procedure to investigate measurement bias in multilevel data. In this procedure, step 1 involves testing the necessity of applying multilevel analysis, step 2 consists of establishing a measurement model at level 1, step 3 involves testing for measurement bias at level 1, step 4 refers to testing for cluster bias, and step 5 concerns explaining the cluster bias with observed level 2 variables using RFA modelling. The present study focuses on steps 4 and 5 of this procedure.

Testing for measurement bias with respect to specific violators (step 5) is only appropriate if there is variance in the indicators that is not explained by the common factor(s). So, if there is no cluster bias, that is, if the residual level 2 variance is zero in step 4, investigating measurement bias with respect to possible between-level violators is superfluous. The test for cluster bias thus serves as a global test of measurement invariance at the between level. However, the cluster-bias test is subject to Type I errors (false positives) and Type II errors (false negatives). If the power of the overall cluster-bias test is larger than the power of the RFA test, then not detecting cluster bias will render the RFA test unnecessary. In that case, the RFA test will not detect bias that the test for cluster bias would not detect. However, if the power of the test for cluster bias is smaller than the power of the RFA test, it is possible that a researcher will not detect cluster bias with the cluster-bias test, but will detect measurement bias with respect to particular level 2 variables with the RFA test. In this study we use simulated data to compare the power and false positive rates of the test for cluster bias and the RFA test in several conditions, varying the size of the bias, the intraclass correlation (ICC) and the sample sizes at both levels. As including the violating variable in the model adds information about the bias, we expect the RFA method to be more powerful than the test for cluster bias.

2. Method

To compare the performance of the test for cluster bias and the RFA test, we generated 500 data sets for each of 54 conditions, according to a factorial design with the following factors: bias effect size (none, small, large); size of ICC (.10, .20, .30); between-level sample size (50, 100 clusters); and within-level sample size (2, 5, 25 observations per cluster). In all conditions, the population model was a two-level, one-factor model with five indicators, with one observed covariate (violator) at the between level. Population values are given in Figure 1. In the population, factor loadings are equal across levels, and there is no residual variance at the between level.

2.1. Size of intraclass correlation

For conditions with ICC = .10, .20 and .30 respectively, 10, 20 and 30% of the variance in an unbiased indicator is at the between level. Biased indicators have larger ICCs because the variance caused by the violator adds to the between-level variance of an indicator. The range of the ICCs was identical to the population values in the study of Maas and Hox (2005) and is based on an empirical investigation of ICCs in health research (Gulliford, Ukoumunne, & Chinn, 1999).

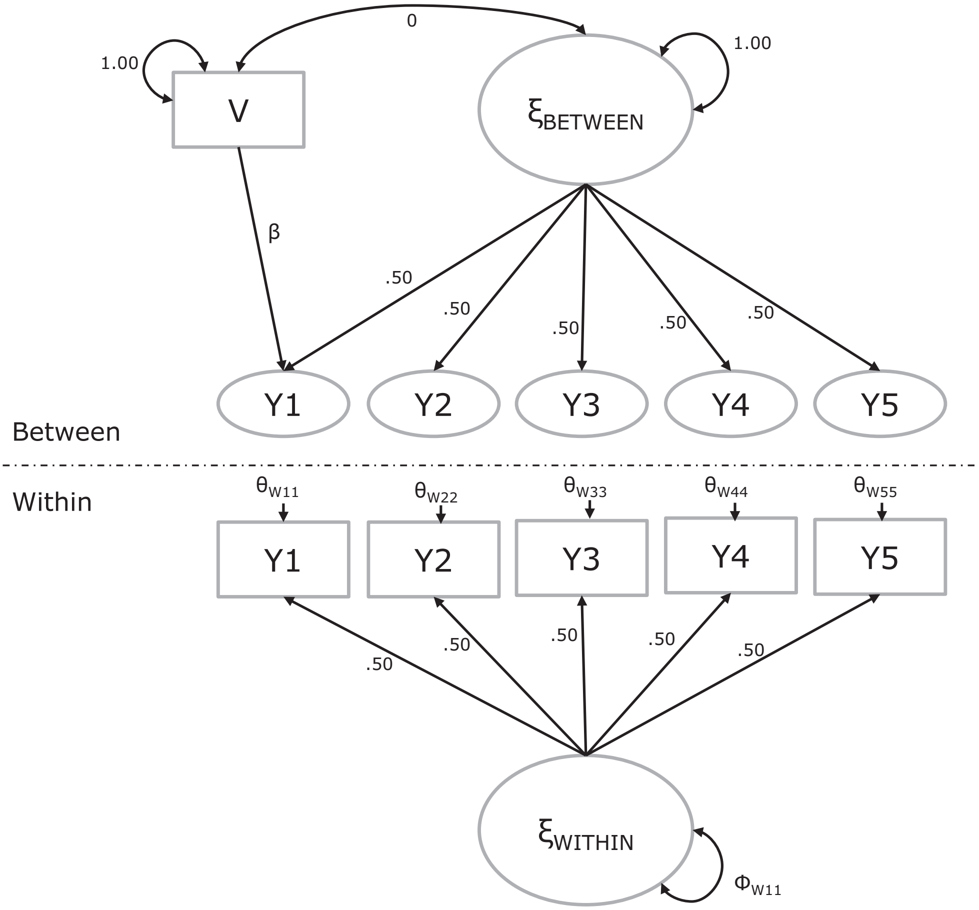

Figure 1. Two-level measurement model with population parameter values. *Note*. In conditions with ICC = .10, θ_{W11}–θ_{W55} = 1.25, Φ_{W11} = 4.00 and β = 0/.159/.363 with 0, 1 and 5% bias, respectively. In conditions with ICC = .20, θ_{W11}–θ_{W55} = 0.50, Φ_{W11} = 2 and β = 0/.112/.257 with 0, 1 and 5% bias, respectively. In conditions with ICC = .30, θ_{W11}–θ_{W55} = 0.33, Φ_{W11} = 1 and β = 0/.092/.210 with 0, 1 and 5% bias, respectively.

2.2. Bias effect size

Uniform bias was introduced in the first indicator, by including a direct effect of the violator on this indicator. Small-sized bias was defined as a direct effect of a size which corresponds to 1% of the total variance of the indicator being caused by the violator. Large-sized bias was defined as a direct effect which amounts to 5% of the total variance being caused by the violator. The absolute values of the direct effects depend on the size of the ICC (see Figure 1). We only considered conditions with bias in one of the five indicators, corresponding to 20% of the indicators being biased. Cheung and Rensvold (1998) stated that usually only one or two items per construct are biased. We chose an amount of biased items that is similar to the conditions with many (20%) biased items in the study of Magis and Facon (2012), and similar to the simulation study of Kim and Yoon (2011) who imposed bias in one out of six items.

2.3. Between-level sample size

We considered conditions with 100 and with 50 clusters. One hundred is the minimum number of clusters for which the chi-square statistic follows its expected asymptotic distribution to a reasonable approximation (Hox, Maas, & Brinkhuis, 2010). As in practice the numbers of clusters are often smaller than 100, we also considered conditions with 50 clusters.

2.4. Within-level sample size

A within-level sample size of 25 corresponds to the typical size of a school class (Elffers, 2012; Thoonen, Sleegers, Peetsma, & Oort, 2011). Group sizes of five are common in data from organizational research where it is a typical size of a working team (Jackson & Joshi, 2004; Koman & Wolff, 2008). Cluster sizes of two correspond to a typical cluster size in data from family research (Duncan, Alpert, & Duncan, 1998; Voorpostel & Blieszner, 2008).

2.5. Data generation

We generated continuous multivariate normally distributed data using the same procedure as Jak *et al.* (2013), using the mvtnorm package in R (Genz *et al.*, 2012). An example R script can be found in Data S1.

2.6. Likelihood ratio test and Wald test

The likelihood ratio test (LRT) and the Wald test were used to test the significance of parameters. The likelihood ratio equals the difference in minus twice the log-likelihoods of a model with and without the parameter(s) of interest. This difference follows a chi-square distribution with degrees of freedom equal to the difference in numbers of parameters between the two models, assuming the parameter of interest is zero. If the chi-square is significant, given the chosen alpha level, then the hypothesis of the parameter(s) of interest being zero is rejected. In the present study we used the default estimator for multilevel data in M*plus* (Muthén & Muthén, 2007), which is the robust maximum likelihood estimator. This estimator is called ‘robust’ because it provides a test statistic and standard errors that are robust against non-normality of the data. M*plus* provides a test statistic that is asymptotically equivalent to the Yuan–Bentler test statistic (T_2; Yuan &

Bentler, 2000). The differences in minus twice the log-likelihoods of models that are estimated using the robust maximum likelihood estimator theoretically need a correction to approximate the chi-square distribution (Satorra & Bentler, 2001). However, simulation studies have shown that conducting the LRT with this correction often leads to untrustworthy results and that the corrected LRT does not perform better than the uncorrected LRT (Cham, West, Ma, & Aiken, 2012 Jak *et al.*, 2013). Moreover, we generated multivariate normal data, and in that case it is found that the unscaled LRT performs better than the scaled LRT (Hox *et al.*, 2010). In this study we therefore apply the uncorrected LRT.

The Wald test is based on the parameter estimate divided by its standard error, and tests the hypothesis that the parameter is zero. M*plus* provides standard errors using Huber–White sandwich estimators (Huber, 1967; White, 1982). Hox *et al.* (2010) found that when normality holds, the normal standard errors are more accurate than the robust standard errors. The robust standard errors are, however, the only standard errors available in M*plus*, and will therefore be used in this simulation study.

As the Wald test is asymptotically equivalent to the LRT (Engle, 1984), we expect the two tests to do equally well in terms of power and false positive rates.

2.7. Testing for level 2 bias with the cluster-bias test and the RFA test

Table 1 gives an overview of the three models and the three outcomes that we consider in the simulation study. We gave each combination a label (A, …, I) to organize the presentation of the results. We looked at the power, the false positive rate, and the false positive rate with a misspecified model, for each of three tests: the test for cluster bias and two versions of the RFA test. Examples of Mplus scripts to fit the cluster bias and RFA models are provided in Data S1. We explain the individual cases below.

To investigate the power of the test for cluster bias and the RFA test, we considered the conditions in which bias was introduced in indicator 1. Cluster bias is tested in a model, as depicted in Figure 2 (case A). In a one-factor model with equal factor loadings across levels, we tested the significance of the between-level residual variance of indicator 1, with the between-level residual variance of the other indicators fixed at zero.

With the RFA test, we included the violating variable at the between level as an exogenous variable that is correlated with the common factor. Subsequently, we tested the significance of the direct effect of the violator on indicator 1. We used the RFA test in two ways; see Figure 2 (cases B and C) for a graphical representation of these models. In the first model we fixed all the residual variance at the between level at zero, hypothesizing that the violator explains all cluster bias (case B). In the second model,

Table 1. An overview of the combinations of tests and outcomes

Outcome test	True positive rate (power)	False positive rate	False positive rate with misspecified model
Test for cluster bias	Case A	Case D	Case G
RFA test	Case B	Case E	Case H
RFA test accounting for cluster bias	Case C	Case F	Case I

Note. In addition, in case J, we investigated false positives by testing the residual variance in indicator 1, while the bias was already accounted for in the RFA model.

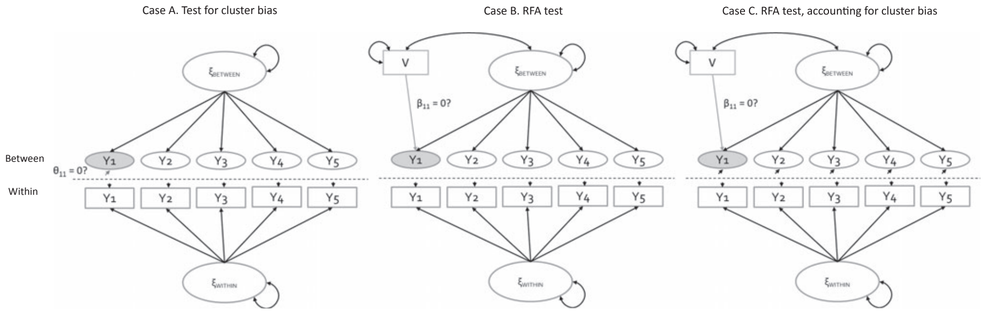
Case A. Test for cluster bias
Case B. RFA test
Case C. RFA test, accounting for cluster bias
ξ BETWEEN
V
β11 = 0?
Between
θ11 = 0?
Within
Y1
Y2
Y3
Y4
Y5
ξ WITHIN

Figure 2. The three models used to investigate the power of three tests (corresponding to cases A, B and C).

residual variance was freely estimated for all indicators, allowing for possible cluster bias in the indicators that is not explained by the violator (case C).

We investigated the false positive rates of all tests in three ways. Firstly, we tested for bias in the conditions where no bias was introduced (cases D, E and F). Secondly, we tested for bias in indicator 2 (an unbiased indicator), in conditions where the bias was in indicator 1 (cases G, H and I). So in these cases we investigate the false positive rates with a misspecified model.

In the conditions with bias in indicator 1, we investigated a third type of false positives (case J). In these cases we accounted for the bias by letting the violator have a direct effect on indicator 1. We then tested the residual variance in indicator 1. As the violator is the only cause of the cluster bias, significance of the residual variance represents a false positive result.

We test against levels of significance of alpha of 5 and 10%. The test for cluster bias involves testing a variance parameter, which cannot be negative by definition. Therefore, in line with Stoel, Garre, Dolan, and van den Wittenboer (2006), we employ one-tailed levels of significance of .05 and .10 with the test for cluster bias. Direct effects can be either negative or positive, so with the RFA tests we use two-tailed tests. This implies that in order to obtain an alpha level of .05 we used a critical chi-square value of $\chi^2_{\text{crit}} = 2.71$ with the test for cluster bias and $\chi^2_{\text{crit}} = 3.84$ with the RFA tests. With an alpha level of .10, these critical values are $\chi^2_{\text{crit}} = 1.64$ for the test for cluster bias and $\chi^2_{\text{crit}} = 2.71$ for the RFA tests. Critical values for the Wald tests are obtained in the same manner (with $\alpha = .05$, $z_{\text{crit}} = 1.28$ for the cluster-bias test and 1.64 for the RFA test, and with $\alpha = .10$, $z_{\text{crit}} = 0.84$ for the cluster-bias test and $z_{\text{crit}} = 1.28$ the RFA test).

2.7.1. Expectations about the cluster-bias model versus the RFA model

Because inclusion of the violating variable adds information into the analysis, we expect the RFA tests to have more power than the test for cluster bias. This is also in line with the finding that in multilevel regression, the cross-level interaction between a level 1 and a level 2 variable can be statistically significant, even if the associated random slope variance is not significant (Snijders & Bosker, 1999).[1] With respect to the false positive rates we

[1] We would like to thank one of the reviewers for this suggestion.

expect the RFA models to do worse than the cluster-bias model when the model is misspecified, and the RFA model with residual variance for all indicators to do better than the RFA model without residual variance.

3. Results

In order to save space, we only report tables with results from the conditions with ICC = .10 and ICC = .30, with alpha levels of 5%. The results in conditions with ICC = .20 are shown in the figures. The complete set of tables, including results with alpha levels of 10%, can be found in Supporting information. We consider a power level of .80 or greater as acceptable.

3.1. Power

Results of the true positive rates of the three tests are shown in Table 2. The true positive rates for all models are higher in the conditions with a higher ICC, but the patterns of true positives of the different models across sample-size conditions are similar. Figure 3 shows the results from the condition with ICC = .20 graphically. We first discuss the results of the conditions where ICC = .20. With large bias, all bias is detected by all three models, provided the total sample size is large (100 or 50 clusters with 25 observations per cluster). With smaller samples, the two RFA tests still have adequate power, but the power of the test for cluster bias drops considerably with 50 clusters of five observations per cluster, and to even lower levels with two observations per cluster. The power of the Wald test is generally greater than the power of the LRT in all conditions.

With small-sized bias, the test for cluster bias detects 86% (with the LRT) or 90% (with the Wald test) of the bias in the condition with the largest sample size, and detects <10% of the bias in conditions with small sample sizes. The two RFA tests perform better, with acceptable power in conditions with 25 observations per cluster, and in the condition with 100 clusters with five observations. With just 50 clusters of five observations, the RFA tests detect around 58% (LRT) and around 70% (Wald test) of the bias, which drops to around 20 and 40% with 50 clusters of two observations per cluster. Overall, the power of the RFA tests is considerably greater than the power of the test for cluster bias. The power of the two RFA test with residual variance is similar to the power of the RFA test without residual variance.

A larger ICC was associated with higher power for all tests. With ICC = .10, the test for cluster bias only had acceptable power levels when the bias was large, and the total sample size was large (100 or 50 clusters of 25 observations). With small bias, the power was only acceptable with 100 clusters of 25 observations. With ICC = .30, 100 clusters with five observations also led to acceptable power for the test for cluster bias. Although a larger ICC leads to more power, the sample-size conditions in which the two RFA tests had acceptable power rates were identical across ICC conditions.

3.2. False positive rates

In conditions without bias, the expected false positive rate is the chosen alpha level of significance. Observed false positive rates for all tests in conditions with ICC = .10 and ICC = .30 are given in Tables 3 and 4. With the Wald test, almost all false positive rates are >.05, while with the LRT most false positive rates are around or under the expected .05.

Table 2. Power (cases A, B and C). True positive rates of the LRT and Wald test, using the cluster-bias model and the RFA model, based on 500 replications per condition, α = .05 (one-sided for the test for cluster bias and two-sided for the RFA tests)

			ICC = .10						ICC = .30					
			Cluster bias		RFA		RFA free Θ		Cluster bias		RFA		RFA free Θ	
Size bias	*N* between	*N* within	LRT	Wald test	LRT	Wald test	LRT	Wald test	LRT	Wald test	LRT	Wald test	LRT	Wald test
Large	100	25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		5	.686	.762	1.000	1.000	1.000	1.000	.838	.878	1.000	1.000	1.000	1.000
		2	.150	.242	.966	.982	.962	.982	.226	.334	.990	.996	.988	.996
	50	25	1.000	1.000	1.000	1.000	1.000	.988	.998	.998	1.000	1.000	1.000	.998
		5	.438	.532	.994	.996	.988	.992	.516	.582	.994	.998	.992	.996
		2	.102	.198	.792	.890	.780	.878	.164	.234	.882	.942	.870	.936
Small	100	25	.756	.826	1.000	1.000	1.000	1.000	.806	.866	1.000	1.000	1.000	1.000
		5	.108	.174	.808	.888	.792	.884	.122	.186	.896	.950	.882	.946
		2	.074	.134	.418	.558	.404	.550	.078	.142	.538	.654	.512	.652
	50	25	.454	.552	.982	.992	.978	.982	.580	.646	.996	.998	.992	.994
		5	.082	.100	.508	.662	.462	.638	.088	.140	.582	.720	.550	.702
		2	.044	.076	.208	.314	.194	.344	.038	.116	.298	.424	.270	.426

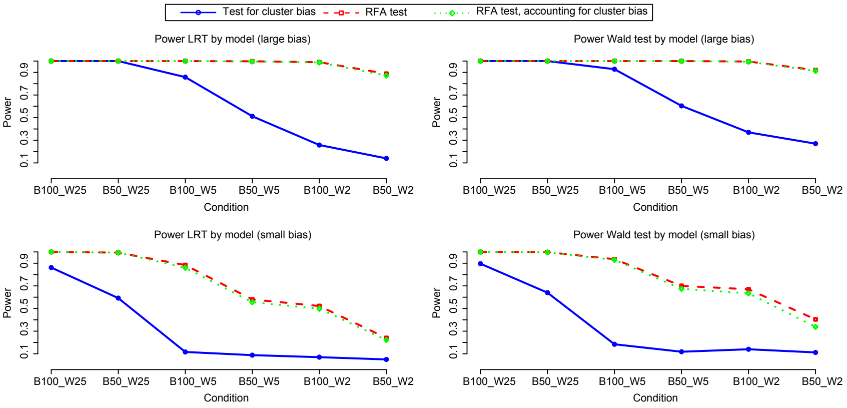

Figure 3. Power (cases A, B and C). True positive rates of the test for cluster bias and the two RFA tests with various sample sizes in conditions with large bias (top) and small bias (bottom), with the LRT (left) and the Wald test (right) in the condition with ICC = .20. *Note*. On the *x*-axis, B100_W25 refers to the condition with 100 clusters with 25 observations per cluster, B50_W25 to the condition with 50 clusters with 25 observations per cluster, and so on.

We do not see any structural differences between the ICC conditions or between the cluster-bias and RFA tests.

3.3. *False positive rates with misspecified models*

We obtained interesting results in conditions where we introduced the bias in indicator 1, but we tested bias in indicator 2. In this case the model is effectively misspecified. The results in the three ICC conditions are very similar. Here we will discuss the results from the ICC = .20 conditions, which are depicted in Figure 4.

The test for cluster bias shows the smallest false positive rates, followed by the RFA test with residual variance. The RFA test without residual variance has the largest false positive rates. The Wald test shows larger false positive rates than the LRT in all conditions.

When the bias was small, the false positive rates of the test for cluster bias were generally acceptable for the LRT, but between 5 and 11% for the Wald test. An exception is the condition with 50 clusters of five observations, in which the LRT has a false positive rate of .17. The RFA test without residual variance identified indicator 2 as biased in 42% (LRT) and 54% (Wald test) of the samples in the condition with the largest sample size. With smaller sample sizes these percentages drop considerably, and with 50 clusters with two observations the false positive rates are around 6% for all LRTs, and around 18% with the Wald test. The RFA test with residual variance also showed unacceptably large false positive rates, but smaller than the RFA test without residual variance.

With large bias in indicator 1, the RFA test without estimated residual variance identified indicator 2 as biased in almost all cases (94% with the LRT and 96% with the Wald test) with the largest sample size, while the RFA test with residual variance identified 25 and 40% of the cases as biased, and the test for cluster bias falsely detected bias in only 23 and 31% of the cases. With smaller sample sizes the false positive rates for all tests decreased, leading to acceptable false positive rates for the test for cluster bias (with the LRT), but not for the RFA tests. In the RFA models, the significant direct effects on indicator 2 were all negative.

Table 3. False positives (cases D, E and F). False positive rates of the LRT and Wald test, using the cluster-bias model and the RFA model, based on 500 replications per condition, α = .05 (one-sided for the test for cluster bias and two-sided for the RFA tests)

			ICC = .10						ICC = .30					
			Cluster bias		RFA		RFA free **Θ**		Cluster bias		RFA		RFA free **Θ**	
Size bias	*N* between	*N* within	LRT	Wald test	LRT	Wald test	LRT	Wald test	LRT	Wald test	LRT	Wald test	LRT	Wald test
None	100	25	.038	.068	.044	.086	.042	.086	.034	.060	.054	.124	.058	.124
		5	.034	.080	.056	.134	.048	.130	.026	.058	.048	.112	.044	.106
		2	.024	.072	.032	.088	.024	.092	.060	.108	.060	.114	.048	.110
	50	25	.028	.042	.056	.120	.052	.120	.022	.030	.052	.134	.048	.132
		5	.034	.062	.068	.164	.068	.156	.026	.050	.042	.138	.038	.134
		2	.046	.090	.046	.094	.032	.118	.030	.084	.054	.136	.048	.132

Table 4. False positives with misspecified model (cases G, H and I). False positive rates of the LRT and Wald test in conditions with bias, using the cluster-bias model and the RFA model, based on 500 replications per condition, $\alpha = .05$ (one-sided for the test for cluster bias and two-sided for the RFA tests)

			ICC = .10						ICC = .30					
			Cluster bias		RFA		RFA free **Θ**		Cluster bias		RFA		RFA free **Θ**	
Size bias	*N* between	*N* within	LRT	Wald test	LRT	Wald test	LRT	Wald test	LRT	Wald test	LRT	Wald test	LRT	Wald test
Large	100	25	.164	.242	.910	.950	.236	.380	.176	.242	.948	.980	.194	.304
		5	.054	.088	.330	.464	.238	.370	.058	.098	.346	.460	.204	.314
		2	.052	.104	.120	.208	.100	.206	.038	.074	.202	.312	.174	.292
	50	25	.082	.136	.636	.770	.152	.262	.096	.142	.750	.822	.158	.258
		5	.038	.070	.210	.332	.144	.272	.038	.066	.488	.358	.132	.268
		2	.050	.102	.116	.178	.098	.180	.054	.088	.124	.224	.112	.200
Small	100	25	.054	.112	.330	.464	.206	.334	.054	.082	.368	.520	.230	.380
		5	.048	.084	.122	.214	.112	.208	.036	.072	.140	.208	.128	.200
		2	.034	.092	.088	.176	.070	.262	.054	.104	.088	.172	.072	.170
	50	25	.040	.072	.196	.310	.132	.248	.034	.056	.210	.386	.098	.250
		5	.034	.052	.106	.202	.088	.190	.030	.046	.082	.202	.066	.184
		2	.044	.086	.054	.114	.052	.124	.034	.088	.074	.158	.068	.154

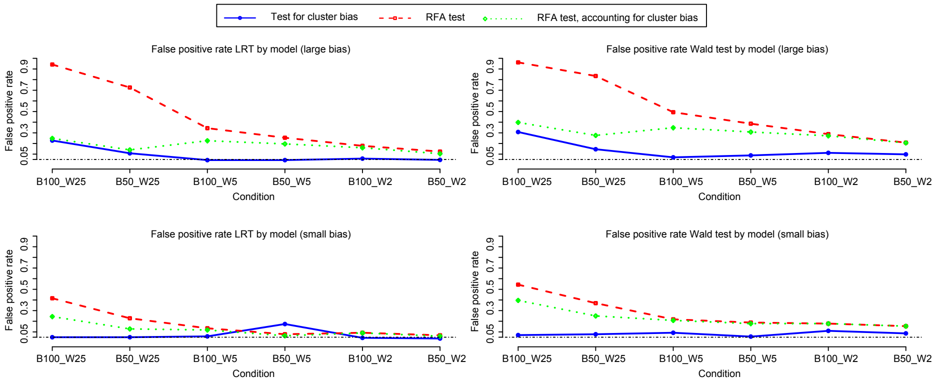

Figure 4. False positives with a misspecified model (cases G, H and I). False positive rates of the test for cluster bias and the two RFA tests with various sample sizes in conditions with large bias (top) and small bias (bottom) for ICC = .20, with the LRT (left) and the Wald test (right). The bias was in indicator 1, while we tested bias in indicator 2. *Note*. On the *x*-axis, B100_W25 refers to the condition with 100 clusters with 25 observations per cluster, B50_W25 to the condition with 50 clusters with 25 observations per cluster, and so on.

3.4. False positive rates when accounting for bias

The false positive rates when testing for cluster bias, when the bias is already accounted for by the violator, are given in Table 5. The false positive rates seem to increase when the within-level sample size decreases. The false positive rates of the LRT are generally under the nominal level of significance and tend to become closer to 5% as the within sample size decreases. The false positive rates of the Wald test are generally higher than the expected

Table 5. False positive rates (case J) of the LRT and Wald test in conditions with bias, using the cluster-bias test after accounting for the bias in the RFA model, based on 500 replications per condition, α = 0.05

			Cluster bias			
			ICC = .10		ICC = .30	
Size bias	*N* between	*N* within	LRT	Wald test	LRT	Wald test
Large	100	25	.014	.040	.040	.064
		5	.032	.058	.042	.070
		2	.030	.078	.048	.086
	50	25	.030	.058	.026	.036
		5	.022	.054	.018	.048
		2	.032	.066	.052	.096
Small	100	25	.026	.046	.038	.076
		5	.038	.056	.044	.066
		2	.038	.094	.040	.090
	50	25	.026	.042	.030	.048
		5	.036	.050	.010	.048
		2	.034	.058	.028	.076

value, and become more inaccurate as the within sample size decreases. The false positives are generally higher in the ICC = .30 condition than in the ICC = .10 condition.

3.5. The effect of cluster size

The results show that the true and false positive rates of all tests vary with the total sample size. To check whether the within-level or between-level sample size has a larger effect on the performance of the cluster-bias test, we additionally investigated the effect of the within sample size relative to the between sample size. We expected that the false positive rates would increase with smaller cluster sizes, because more random error would be aggregated to the between level. For each ICC condition, we investigated the false positive rates and the power to detect small bias of the three tests in four conditions with a total sample size of 1,250 (50 clusters with 25 observations, 125 clusters with 10 observations, 250 clusters with 5 observations, and 625 clusters with 2 observations). The results from the ICC = .10 and ICC = .30 conditions are shown in Table 6. Figure 5 shows the false positive rates and true positive rates as obtained with the LRT and the Wald test in the four conditions with ICC = .20 (the results were very similar in the other two ICC conditions). With the Wald test, but not with the LRT, there seems to be an upward trend in the false positive rates of the test for cluster bias, but not for the RFA tests. However, the false positive rates of the test for cluster bias are still below the nominal level of significance in all conditions, except for the condition with 625 clusters with two observations (where the false positive rate is 7%). The RFA tests detect almost all bias in all conditions, but the power of the test for cluster bias shows a gradual decrease when cluster size becomes smaller.

4. Discussion

The results of the simulation study show that the inclusion of the violating variable in the analysis adds considerably to the likelihood of detecting the bias. So, in fitting a series of models in order to investigate measurement bias in multilevel data (Jak *et al.*, 2014), the finding that cluster bias is absent does not exclude the possibility that this is a false negative and that significant bias with respect to a level 2 violator may be found using an RFA model. Of course, the RFA model requires the availability of a violating variable. So, although the test for cluster bias is less powerful, an advantage of the test is that the cause of the bias does not need to be operationalized, or even known.

Another advantage of the test for cluster bias is that the false positive rates were generally acceptable, while the RFA tests had high false positive rates in conditions where the bias was in an indicator other than the one actually subject to the test. The high false positive rates with the RFA test show that when the model does not account for measurement bias, the common factor is contaminated by the bias. For example, suppose that the trait of interest is closeness of the relation between teachers and students, and indicator 1 is biased by gender of the teacher, meaning that for equal levels of closeness, women on average attain higher scores on this indicator than men do. Indicator 1 is then not only an indicator of closeness, but also an indicator of gender (and gender-related characteristics). Not accounting for this bias results in the contamination of the closeness factor with gender. The interpretation of the factor is then closeness and (probably to a smaller extent) being a woman. Indicator 2 is actually not an indicator of gender, so an

Table 6. Power and false positive rates of the LRT and Wald test, using the cluster-bias model and the RFA model with a total sample size of 1,250, based on 500 replications per condition, α = .05 (one-sided for the test for cluster bias and two-sided for the RFA tests)

			ICC = .10						ICC = .30					
			Cluster bias		RFA		RFA free **Θ**		Cluster bias		RFA		RFA free **Θ**	
Size bias	*N* between	*N* within	LRT	Wald test	LRT	Wald test	LRT	Wald test	LRT	Wald test	LRT	Wald test	LRT	Wald test
Power														
Small	50	25	.454	.552	.982	.992	.978	.982	.580	.646	.996	.998	.992	.994
	125	10	.276	.346	.996	1.000	.992	1.000	.374	.444	.998	.998	.998	.998
	250	5	.188	.260	.994	.996	.994	.996	.184	.288	1.000	1.000	1.000	1.000
	625	2	.106	.190	1.000	1.000	1.000	1.000	.128	.230	.996	1.000	.996	1.000
False positive rate														
None	50	25	.028	.042	.056	.120	.052	.120	.022	.030	.052	.134	.048	.132
	125	10	.044	.070	.038	.124	.036	.122	.040	.066	.042	.084	.036	.084
	250	5	.040	.088	.060	.126	.052	.126	.026	.060	.056	.126	.052	.122
	625	2	.068	.142	.058	.122	.056	.120	.060	.106	.052	.110	.052	.110

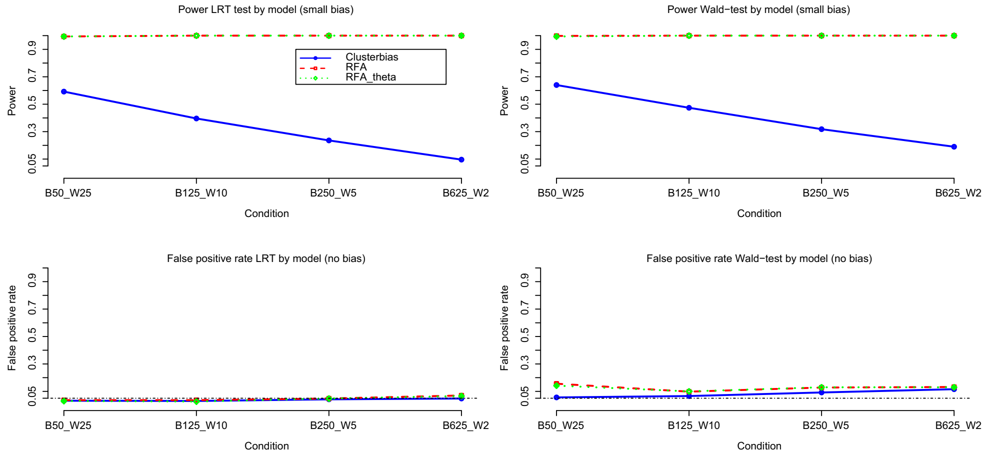

Figure 5. The effect of smaller cluster size. Power (top) and false positive rates (bottom) of the test for cluster bias and the two RFA tests using the LRT (left) and the Wald test (right), with various cluster sizes leading to a total sample size of 1,250 and ICC = .20. *Note*. On the *x*-axis, B50_W25 refers to the condition with 50 clusters with 25 observations per cluster, B125_W10 to the condition with 125 clusters with 10 observations per cluster, and so on. The dotted line indicates the nominal alpha level.

effect of gender on indicator 2 will be negative in order to compensate for the contamination of the common factor.

The false positive rates of the RFA test without residual variance were higher than the rates of the test with residual variance. This makes sense, as by estimating residual variance in the indicators, we account for part of the bias. Although false positive rates of the RFA test with residual variance are still higher than the chosen level of significance, based on the false positive rate, the RFA test with residual variance is preferred.

The Wald test generally showed more power to detect bias than the LRT. However, the Wald test also showed larger false positive rates in all conditions. In practice, one should never base conclusions about measurement bias on a significant test result only. The size of the bias as well as the possibility of interpreting the bias should be taken into account. In the RFA model, the size of the bias can be judged by looking at the size of the standardized direct effect of the violator on the indicator. In the cluster-bias model, the size of the bias in an indicator can be represented by the percentage of variance due to cluster bias (level 2 residual variance) in the total variance or in the total level 2 variance. More important than size and significance is theory. Only if detected bias can be interpreted substantively can the finding be taken as a real instance of measurement bias. Theory should always be used to counter false positive results, especially if the Wald test is used. The difference between the results from the LRT and the Wald test may also stem from the fact that the simulation study considered multivariate normal data. If inspection of the data supports multivariate normality, using normal theory rather than robust standard errors may be preferable (Hox *et al.*, 2010).

In practice, researchers may employ an iterative bias detection procedure. In an iterative procedure, a researcher starts by including a direct effect in the indicator that improves model fit most. The choice of which direct effect to include first can be based on testing direct effects on all indicators one by one, or by inspecting modification indices from the model without any direct effects. In our example, if bias were tested in the unbiased indicator, while the bias in indicator 1 was already accounted for, the factor would not be contaminated, and indicator 2 would not be marked as biased. In a simulation study, Barendse *et al.* (2012) showed that an iterative RFA procedure led to less false positive results than a single-run RFA procedure. Using modification indices or iterative testing is data-driven, however, and always bears the risk of capitalizing on chance characteristics of the data (MacCallum, Roznowski, & Necowitz, 1992). One may therefore adjust the nominal significance level to a more conservative value, and one should always take the substantive interpretation of possible bias into account.

This simulation study showed that with data with ICCs around .20, in order to have adequate power to detect large bias with the cluster-bias test, one needs a total sample size of at least 500, with sufficiently large cluster sizes. To detect small bias, the cluster-bias test needs a total sample size of at least 2,500, with cluster of sufficient size. If one can use the RFA test, smaller samples are sufficient to obtain adequate power. With large bias, a total sample size of 100 would be enough. With small bias, total sample sizes of 500 or more are required. With smaller ICCs larger sample sizes are needed to obtain adequate power. It should be noted that this holds assuming that the size of the bias is proportional to the ICC.

In conclusion, although the test for cluster bias has several advantages, this study showed that including the presumed cause of level 2 bias in the model to detect measurement bias is a more powerful approach than the test for cluster bias. We also showed that the RFA test with residual variance leads to less false positive results than the RFA test without residual variance. If a researcher's goal is to investigate measurement bias with respect to (level 1 and) level 2 violators, we advise following the five-step approach (Jak *et al.*, 2014) and testing for level 2 bias in step 5, while taking cluster bias into account. This study also showed that the power of the test for cluster bias is larger with a smaller number of clusters with a larger size than for more clusters with a smaller size.

Acknowledgement

We would like to thank Conor Dolan for valuable comments on this manuscript.

References

Barendse, M. T., Oort, F. J., & Garst, G. J. A. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and non-uniform measurement bias: A simulation study. *Advances in Statistical Analysis*, *94*, 117–127. doi:10.1007/s10182-010-0126-1

Barendse, M. T., Oort, F. J., Werner, C. S., Ligtvoet, R., & Schermelleh-Engel, K. (2012). Measurement bias detection through factor analysis. *Structural Equation Modeling*, *19*, 561–579. doi:10.1080/10705511.2012.713261

Cham, H., West, S. G., Ma, Y., & Aiken, L. S. (2012). Estimating latent variable interactions with non-normal observed data: A comparison of four approaches. *Multivariate Behavioral Research*, *47*, 840–876. doi:10.1080/00273171.2012.732901

Cheung, G. W., & Rensvold, R. B. (1998). Cross-cultural comparisons using non-invariant measurement items. *Applied Behavioral Science Review*, *6*, 93–110. doi:10.1016/S1068-8595 (99)80006-3

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, *43*, 558–575. doi:10.1177/0022022112438397

De Jong, M. G., Steenkamp, J. B. E., & 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

Duncan, T. E., Alpert, A., & Duncan, S. C. (1998). Multilevel covariance structure analysis of sibling antisocial behavior. *Structural Equation Modeling*, *5*, 211–228. doi:10.1080/10705519809540102

Elffers, L. (2012). One foot out the school door? Interpreting the risk for dropout upon the transition to post-secondary vocational education. *British Journal of Sociology of Education*, *33*, 41–61. doi:10.1080/01425692.2012.632866

Engle, R. F. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In M. D. Intriligator & Z. Griliches (Eds.), *Handbook of econometrics* (Vol. II, pp. 796–801). Amsterdam, the Netherlands: North-Holland.

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.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2012) *mvtnorm: Multivariate normal and t distributions*. R package version 0.9-9992. Retrieved from http://cran.r-project.org/package=mvtnorm

Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies. *American Journal of Epidemiology*, *149*, 876–883. doi:10.1093/oxfordjournals.aje.a009904

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117–144. doi:10.1080/03610739208253916

Hox, J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*, 157–170. doi:10.1111/j.1467-9574.2009.00445.x

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In L. LeCam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 221–233). Berkeley, CA: University of California Press.

Jackson, S. E., & Joshi, A. (2004). Diversity in social context: A multi-attribute, multilevel analysis of team diversity and sales performance. *Journal of Organizational Behavior*, *25*, 675–702. doi:10.1002/job.265

Jak, S. (2014). Testing strong factorial invariance using three-level structural equation modeling. *Frontiers in Psychology*, *5*, 745. doi:10.3389/fpsyg.2014.00745

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling*, *20*, 265–282. doi:10.1080/10705511.2013.769392

Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling*, *21*, 31–39. doi:10.1080/10705511.2014.856694

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, *18*, 212–228. doi:10.1080/10705511.2011.557337

Kim, E. S., Yoon, M., Wen, Y., Luo, W., & Kwok, O. (2015). Within-level group factorial invariance in multilevel data: Multilevel factor mixture and multilevel MIMIC models. *Structural Equation Modeling*. doi:10.1080/10705511.2014.938217

Koman, E. S., & Wolff, S. B. (2008). Emotional intelligence competencies in the team and team leader: A multi-level examination of the impact of emotional intelligence on team performance. *Journal of Management Development*, *27*, 55–75. doi:10.1108/02621710810840767

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53–76. doi:10.1207/s15327906mbr3201_3

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. doi:10.1027/1614-1881.1.3.86

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504. doi:10.1037/0033-2909.111.3.490

Magis, D., & Facon, B. (2012). Angoff's delta method revisited: Improving DIF detection under small samples. *British Journal of Mathematical and Statistical Psychology*, *65*, 302–321. doi:10.1111/j.2044-8317.2011.02025.x

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*, 223–236. doi:10.1207/s15327906mbr2903_2

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

Molenaar, D., Dolan, C. V., Wicherts, J. M., & van der Maas, H. L. J. (2010). Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. *Intelligence*, *38*, 611–624. doi:10.1016/j.intell.2010.09.002

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585. doi:10.1007/BF02296397

Muthén, B. (1990). *Mean and covariance structure analysis of hierarchical data*, UCLA Statistics Series, no. 62. Los Angeles, CA: UCLA.

Muthén, B. & Asparouhov, T. (2013). *New methods for the study of measurement invariance with many groups*. Unpublished technical report. Retrieved from http://www.statmodel.com

Muthén, B., Khoo, S. T., & Gustafsson, J. E. (1997). *Multilevel latent variable modeling in multiple populations*. Unpublished technical report. Retrieved from http://www.statmodel.com

Muthén, L. K., & Muthén, B. O. (2007). *Mplus users guide* (5th ed.). Los Angeles, CA: The Author.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika*, *6*, 150–166.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling*, *5*, 107–124. doi:10.1080/10705519809540095

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika*, *69*, 167–190. doi:10.1007/BF02295939

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566.

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 172–194. doi:10.1111/bmsp.12014

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514. doi:10.1007/BF02296192

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, UK: Sage.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239. doi:10.1111/j.2044-8317.1974.tb00543.x

Spilt, J. L., Koomen, H. M., & Jak, S. (2012). Are boys better off with male and girls with female teachers? A multilevel investigation of measurement invariance and gender match in teacher–student relationship quality. *Journal of School Psychology*, *50*, 363–378. doi:10.1016/j.jsp.2011.12.002

Stoel, R. D., Garre, F. G., Dolan, C. V., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, *11*, 439–455. doi:10.1037/1082-989X.11.4.439

Thoonen, E. E. J., Sleegers, P. J. C., Peetsma, T. T.D., & Oort, F. J. (2011). Can teachers motivate students to learn? *Educational Studies*, *37*, 345–360. doi:10.1080/03055698.2010.507008

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70.

Verhagen, A. J., & Fox, J.-P. (2012). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*, 383–401. doi:10.1111/j.2044-8317.2012.02059.x

Voorpostel, M., & Blieszner, R. (2008). Intergenerational solidarity and support between adult siblings. *Journal of Marriage and Family*, *70*, 157–167. doi:10.1111/j.1741-3737.2007.00468.x

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25. doi:10.2307/1912526

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association. doi:10.1037/10222-009

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165–200. doi:10.1111/0081-1750.00078

Received 28 June 2013; revised version received 29 January 2015

Supporting Information

The following supporting information may be found in the online edition of the article:

Data S1. Example syntax for generating data in R and fitting models in M*plus* for the condition with N_{between} = 100, N_{within} = 25, large bias, ICC = .10.

Table S1. Case A. Case B and Case C: Power. ICC = .10.
Table S2. Case D. Case E and Case F: False positives. ICC = .10.
Table S3. Case G. Case H and Case I: False positives with misspecified model. ICC = .10.
Table S4. Case J. False positive rates of the likelihood ratio test (LRT) and the Wald test in conditions with bias. using the cluster bias test after accounting for the bias in the RFA model. ICC = .10.
Table S5. Power and false positive rates of the likelihood ratio test (LRT) and the Wald-test using the cluster bias model and the RFA model with a total sample size of 1250. ICC = .10.
Table S6. Case A. Case B and Case C: Power. ICC = .20.
Table S7. Case D. Case E and Case F: False positives. ICC = .20.
Table S8. Case G. Case H and Case I: False positives with misspecified model. ICC = .20.
Table S9. Case J. False positive rates of the likelihood ratio test (LRT) and the Wald test in conditions with bias using the cluster bias test after accounting for the bias in the RFA model. ICC = .20.
Table S10. Power and false positive rates of the likelihood ratio test (LRT) and the Wald-test using the cluster bias model and the RFA model with a total sample size of 1250. ICC = .20.
Table S11. Case A. Case B and Case C: Power. ICC = .30.
Table S12. Case D. Case E and Case F: False positives. ICC = .30.
Table S13. Case G. Case H and Case I: False positives with misspecified model. ICC = .30.
Table S14. Case J. False positive rates of the likelihood ratio test (LRT) and the Wald test in conditions with bias using the cluster bias test after accounting for the bias in the RFA model. ICC = .30.
Table S15. Power and false positive rates of the likelihood ratio test (LRT) and the Wald-test using the cluster bias model and the RFA model with a total sample size of 1250. ICC = .30.