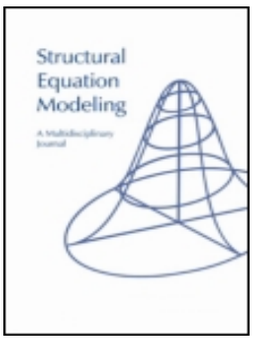
## Measurement bias in multilevel data

Jak, S.; Oort, F.J.; Dolan, C.V.

**DOI**
10.1080/10705511.2014.856694
**Publication date**
2014
**Document Version**
Final published version
**Published in**
Structural Equation Modeling



Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

# Measurement Bias in Multilevel Data

Suzanne Jak , Frans J. Oort & Conor V. Dolan

**To cite this article:** Suzanne Jak , Frans J. Oort & Conor V. Dolan (2014) Measurement Bias in Multilevel Data, Structural Equation Modeling: A Multidisciplinary Journal, 21:1, 31-39, DOI: 10.1080/10705511.2014.856694

**To link to this article:** https://doi.org/10.1080/10705511.2014.856694

*Structural Equation Modeling: A Multidisciplinary Journal*, 21: 31–39, 2014

ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2014.856694

Routledge
Taylor & Francis Group

# Measurement Bias in Multilevel Data

Suzanne Jak, Frans J. Oort, and Conor V. Dolan
*University of Amsterdam, The Netherlands*

Measurement bias can be detected using structural equation modeling (SEM), by testing measurement invariance with multigroup factor analysis (Jöreskog, 1971; Meredith, 1993; Sörbom, 1974), MIMIC modeling (Muthén, 1989), or restricted factor analysis (Oort, 1992, 1998). In educational research, data often have a nested, multilevel structure, for example when data are collected from children in classrooms. Multilevel structures might complicate measurement bias research. In 2-level data, the potentially "biasing trait" or "violator" can be a Level 1 variable (e.g., pupil sex), or a Level 2 variable (e.g., teacher sex). One can also test measurement invariance with respect to the clustering variable (e.g., classroom). This article provides a stepwise approach for the detection of measurement bias with respect to these 3 types of violators. This approach works from Level 1 upward, so the final model accounts for all bias and substantive findings at both levels. The 5 proposed steps are illustrated with data of teacher–child relationships.

**Keywords**: cluster bias, measurement invariance, multilevel structural equation modeling

In the presence of measurement bias, systematic differences between observed test scores are not completely attributable to true differences in the trait(s) that the test is supposed to measure. Suppose given male and female respondents have the same score on a latent trait. In the absence of bias, the expected observed test scores of these respondents (conditional on their common latent trait score) are equal. In the presence of sex bias, this does not hold and we consider the test to lack measurement invariance with respect to sex. Sex is a nominal variable, but measurement bias can be tested with respect to any variable. Measurement bias can be detected using structural equation modeling (SEM), by testing measurement invariance with multigroup factor analysis (MGFA; Jöreskog, 1971; Meredith, 1993; Sörbom, 1974), MIMIC modeling (Muthén, 1989), or with restricted factor analysis (RFA; Oort, 1992, 1998).

With multilevel data structures, the investigation of measurement bias is not straightforward. For instance, consider the case of pupils nested in classes. First, the standard SEM approaches need to be adjusted to account for the multilevel structure. Second, the variable with respect to which measurement bias is to be investigated might be defined at different levels. For example, a Level 1 variable could be sex of the pupils; a Level 2 variable could be sex of the teachers. The biasing variable might also be class itself (i.e., the clustering variable, which we view as a special kind of Level 2 variable).

Here, we propose a five-step procedure to investigate measurement bias (or to establish measurement invariance) in the two-level case. First, we give a short description of multilevel SEM and the investigation of measurement invariance. Then, we describe the situations in which measurements are biased with respect to a Level 1 variable, a Level 2 variable, or with respect to the clustering variable itself. We present our five-step procedure to detect bias in these three situations, and illustrate the procedure with an analysis of data of teacher–pupil relationships.

## MULTILEVEL SEM

In educational and psychological research, cluster sampling methods are often used. Cluster sampling refers to randomly selecting higher level units, and consequently selecting lower level units within these higher level units. Common multilevel data structures are two-level structures, for example, children nested in classrooms or employees nested in teams.

Correspondence should be addressed to Suzanne Jak, Department of Education, University of Amsterdam, Nieuwe Prinsengracht 130, 1018 VZ, Amsterdam, the Netherlands. E-mail: S.Jak@uva.nl

Individuals who are members of the same group share group-level characteristics, and might therefore be more similar to members of their own group than to members of different groups. Multilevel models take into account the dependence of observations in nested data sets (see Bryk & Raudenbush, 1992; Goldstein, 1995; Longford, 1993; Snijders & Bosker, 1999).

Multilevel SEM allows for different models for variances and covariances of within-group differences and between-group differences (Muthén, 1994). We limit our presentation to two-level structures of individuals (Level 1 or the within level) in groups (Level 2 or the between level). Consider the multivariate response vector $\mathbf{y}_{ij}$, with scores from subject $i$ in group $j$, which is decomposed into a group mean ($\boldsymbol{\mu}_j$), and an individual deviation from the group mean ($\boldsymbol{\eta}_{ij}$):

$$\mathbf{y}_{ij} = \boldsymbol{\mu}_j + \boldsymbol{\eta}_{ij}, \tag{1}$$

where $\boldsymbol{\mu}_j$ and $\boldsymbol{\eta}_{ij}$ are independent. The overall covariances of $\mathbf{y}_{ij}$ ($\boldsymbol{\Sigma}_{\text{TOTAL}}$) can be written as the sum of the covariances of $\boldsymbol{\mu}_j$ ($\boldsymbol{\Sigma}_{\text{BETWEEN}}$) and the covariances of $\boldsymbol{\eta}_{ij}$ ($\boldsymbol{\Sigma}_{\text{WITHIN}}$):

$$\boldsymbol{\Sigma}_{\text{TOTAL}} = \boldsymbol{\Sigma}_{\text{BETWEEN}} + \boldsymbol{\Sigma}_{\text{WITHIN}} \tag{2}$$

One can postulate separate models for the within (Level 1) and between (Level 2) matrices. The within model describes the covariance structure within groups and the between model describes the covariance and mean structure between groups. For example, these might be common factor models:

$$\boldsymbol{\Sigma}_{\text{BETWEEN}} = \boldsymbol{\Lambda}_{\text{B}}\, \boldsymbol{\Phi}_{\text{B}}\, \boldsymbol{\Lambda}_{\text{B}}{}' + \boldsymbol{\Theta}_{\text{B}}, \tag{3}$$

$$\boldsymbol{\mu}_{\text{BETWEEN}} = \boldsymbol{\tau}_{\text{B}} + \boldsymbol{\Lambda}_{\text{B}}\, \boldsymbol{\kappa}_{\text{B}}, \tag{4}$$

$$\boldsymbol{\Sigma}_{\text{WITHIN}} = \boldsymbol{\Lambda}_{\text{W}}\, \boldsymbol{\Phi}_{\text{W}}\, \boldsymbol{\Lambda}_{\text{W}}{}' + \boldsymbol{\Theta}_{\text{W}}, \tag{5}$$

Here, $\boldsymbol{\Phi}_{\text{B}}$ and $\boldsymbol{\Phi}_{\text{W}}$ are covariance matrices of the common factors at the between and within level, respectively; $\boldsymbol{\Theta}_{\text{B}}$ and $\boldsymbol{\Theta}_{\text{W}}$ are (diagonal) matrices with variance of the residual factors at the between and within level, respectively; $\boldsymbol{\kappa}_{\text{B}}$ is a vector with common factor means at the between level; $\boldsymbol{\Lambda}_{\text{B}}$ and $\boldsymbol{\Lambda}_{\text{W}}$ are matrices with factor loadings at the between and within level, respectively; and $\boldsymbol{\tau}_{\text{B}}$ is a vector with intercepts at the between level. As within-level scores are deviations from the group mean, there is no mean structure at the within level. The dimensions of the matrices and the parameter estimates can differ across levels. For example, one could combine a three-factor model at the within level with a single-factor model at the between level.

## MEASUREMENT BIAS IN SINGLE-LEVEL SEM

We define *measurement bias* as a violation of measurement invariance (Mellenbergh, 1989). Consider some unobserved trait ($T$), which is assumed to be measured with observed indicators ($X$). Measurements are invariant with respect to some variable ($V$), if $V$ is associated with the observed indicators ($X$) only indirectly via the trait ($T$) that $X$ is supposed to measure. Measurement invariance holds if the conditional distribution of $X$ given values of $T$ and $V$ is equal to the conditional distribution of $X$ given values of $T$ but for different levels of $V$:

$$f_1\,(X \mid T = t, V = v) = f_2(X \mid T = t)\,. \tag{6}$$

Note that given this formal definition, we can distinguish two kinds of bias (Mellenbergh, 1989). If the violator $V$ has a direct relationship with any indicator $X$, then this is called uniform bias: a main effect of $V$ on $X$. The second kind of bias involves a direct effect of an interaction of the violator $V$ and the trait $T$ on the indicator $X$. This is called nonuniform bias. Throughout this article we adopt the terminology of Oort (1991), and call $V$ a (potential) violator, because it is a variable that possibly violates measurement invariance.

In the definition of measurement bias, $X$, $T$, and $V$ could be nominal, ordinal, interval, or ratio variables, they could be latent or manifest, and their relationships could be linear or nonlinear. The choice of a statistical technique to detect measurement bias partly depends on the distribution of the observed scores. With discrete observed scores, multigroup item response theory (IRT) models are obvious choices to test the equality of discrimination and difficulty parameters across groups (Meade & Lautenschlager, 2004; Raju, Laffitte, & Byrne, 2002). Across-group differences in difficulty parameters indicate uniform bias, and additional across-group differences in discrimination parameters indicate nonuniform bias (Mellenbergh, 1989).

Here we use SEM to detect measurement bias. Within SEM, $X$ is typically observed continuous, but can also be ordinal (Flora & Curran, 2004; Jöreskog & Moustaki, 2001; Millsap & Tein, 2004); $T$ is a continuous unobserved common factor; and $V$ can be continuous, ordinal or nominal, and observed or unobserved. One possible way of testing measurement invariance in the case of a nominal variable $V$ (e.g., sex) is through MGFA. In MGFA, measurement invariance is tested by determining whether factor loadings and intercept are equal across the groups. Violations of the equality (over groups) of intercepts are interpreted as uniform bias, and violations of the equality (over groups) of the factor loadings and intercepts are interpreted as nonuniform bias. Equality of residual variances over groups can be tested as well, but is not required for correct comparisons of common factor means across groups. As explained in conceptual terms in Dolan, Roorda, and Wicherts (2004), these constraints can be shown to follow from Equation 6. For an overview of the use of MGFA for measurement invariance testing, see Vandenberg and Lance (2000), Millsap and Everson (1991), Millsap and Tein (2004), and Little (1997).

Another, more flexible approach is the use of the RFA model (Oort, 1992, 1998) or the MIMIC model (Muthén,

1989). These models differ only in the treatment of the violator *V.* In the MIMIC model, *T* is regressed on *V,* whereas in the RFA model, the violator *V* is correlated with *T.* Measurement bias is detected by testing the significance of direct effects of the violator *V* on the measurements *X.*

Advantages of the RFA method over MGFA are that with RFA, continuous violators can be incorporated without the need to create groups, whereas multigroup analysis needs a split of the continuous variable into subgroups. Bias investigation with respect to several violators simultaneously is also more straightforward with RFA. With MGFA, testing more violators involves creating more subgroups with smaller sample sizes, whereas in RFA, it only involves the addition of covariates. A disadvantage of the RFA method is that the detection of nonuniform bias is less straightforward. However, recent developments using latent interaction terms or moderated factor analysis provide a viable method to investigate nonuniform bias in the RFA framework (Barendse, Oort, & Garst, 2010; Barendse, Oort, Werner, Ligtvoet, & Schermelleh-Engel, 2011; see also Molenaar, Dolan, Wicherts, & van der Maas, 2010).

In this article, we apply the RFA method, and restrict ourselves to testing uniform measurement bias only. Testing uniform bias is the first step in testing measurement bias with the RFA or MIMIC method and the power to detect nonuniform bias is generally lower than for uniform bias (Barendse et al., 2010; Woods, 2009).

## MEASUREMENT BIAS IN TWO-LEVEL SEM

In our two-level SEM procedure for bias detection, we consider a potential violator at Level 1 or Level 2. In the latter case, one possibility is that the Level 2 violator is the cluster identifier itself (i.e., a nominal variable with as many values as there are groups or classes). We treat the cluster identifier as a special type of violator. The different levels of the violator variable require different models for bias detection.

### Violator Is a Level 1 Variable

The violator is a Level 1 variable if it has variance within clusters. If data come from children within classrooms, possible Level 1 violators are all variables that vary over children within classes. Examples are children's sex, children's ethnicity, or education level of the parents.

### Violator Is the Clustering Variable

We call measurement bias with respect to the clustering variable *cluster bias* (Jak, Oort, & Dolan, 2013). If data come from children within classrooms, cluster bias means that the test does not measure the same construct in all the classes. In this case, two pupils from different classrooms with identical values of the latent trait might differ with respect to their expected observed test score. The presence of cluster bias can be tested by imposing specific constraints in the models for $\boldsymbol{\Sigma}_{\text{WITHIN}}$ and $\boldsymbol{\Sigma}_{\text{BETWEEN}}$. These constraints ensure that differences between the cluster means are exclusively attributable to differences in the common factor means.

Cluster bias can only be caused by Level 2 variables. Therefore, if cluster bias is not present, we can assume that there is no measurement bias with respect to any Level 2 variable. Testing for cluster bias thus serves as a first step before the investigation of bias with respect to specific Level 2 variables.

### Violator Is a Level 2 Variable

Violators at Level 2 have variance between clusters. Level 2 violators can be aggregates of Level 1 violators, such as the proportion of boys in the class, the proportion of children from a minority group, or average socioeconomic status. Level 2 violators can also be specific to Level 2, such as teacher sex, teacher age, or number of pupils in a class. These violators can only violate measurement invariance at the between level, as they do not vary within clusters. For example, children in classes with a male teacher might show different response behavior to a certain test than children in classes with a female teacher. Teacher sex has no direct influence on the within level, because children within the same class have the same teacher.

## THE FIVE-STEP PROCEDURE

To facilitate the practice of bias investigation with respect to the three types of violators, we propose a five-step procedure for the investigation of measurement bias in two-level data. This procedure includes the detection of measurement bias with respect to Level 1 violators and Level 2 violators, among which is the cluster identifier. The five steps we propose are the following:

1. Test whether there is Level 2 variance and covariance.
2. Establish a measurement model at Level 1.
3. Investigate bias with respect to Level 1 violators.
4. Investigate cluster bias.
5. Investigate bias with respect to Level 2 violators.

In this procedure, Step 3 also comprises the findings from Step 2, and Step 5 also comprises the findings from Step 4. Of course, there are other conceivable procedures. For example, one could test for cluster bias first, and subsequently investigate bias with respect to the Level 1 violators. Alternatively, one could investigate bias with respect to the Level 2 violators with a saturated Level 1 model. However, a convenient property of this five-step approach is that the final model from Step 5 includes all relevant results from the previous steps. Starting the analysis at Level 1 and

then working upward to Level 2 is in line with Bryk and Raudenbush's (1992) two-phase approach in ordinary multilevel regression, and with the stepwise modeling approach of multilevel mediation effects of Preacher, Zyphur, and Zhang (2010).

If the interest is in Level 1 violators only, one can stop the analysis after Step 3. If the interest is in Level 2 variables only, one can limit the modeling to the $\mathbf{\Sigma}_{\text{BETWEEN}}$ covariance matrix, and specify a saturated model for $\mathbf{\Sigma}_{\text{WITHIN}}$.

After explaining the five steps in the next subsections, we illustrate the approach with data of teacher–child relationship research.

### Step 1: Test Whether There Is Level 2 Variance and Covariance

Multilevel modeling is only required if there is variance at Level 2. Fitting structural equation models to Level 2 is only relevant if there is covariance on Level 2. The intraclass correlation (ICC) of a given variable gives the proportion of the variance that can be attributed to Level 2. A common rule of thumb is that ICCs over .05 indicate the necessity of multilevel analysis. One might also want to statistically test whether the Level 2 variance deviates significantly from zero. The significance of the Level 2 variance and covariance can be tested by fitting a null model ($\mathbf{\Sigma}_{\text{BETWEEN}} = 0$) and independence model ($\mathbf{\Sigma}_{\text{BETWEEN}}$ is diagonal) to the between covariance matrix, while specifying a saturated model for $\mathbf{\Sigma}_{\text{WITHIN}}$ (Hox, 2002; Muthén, 1994). If the $\chi^2$ test statistic of the null model is significant, we conclude that there is significant Level 2 variance. If the $\chi^2$ test statistic of the independence model is significant, we conclude that there is significant Level 2 covariance. Testing significance of variances and covariances in this manner is common, but not strictly correct (Stoel, Garre, Dolan, & van den Wittenboer, 2006). Correct testing requires the derivation of an asymptotic distribution of the likelihood ratio test statistic, which can be a complex mixture of multiple different $\chi^2$ distributions. As this is beyond the scope of this work, we accept that the testing procedure is not correct, and keep in mind that it leads to an overly conservative test. That is, the conclusion will too often be that the Level 2 variance or covariance is not significant.

If there is no Level 2 variance, single-level techniques could be used. If there is Level 2 variance, but no Level 2 covariance, Step 2 can still be performed using the pooled within covariance matrix, with the sample size set equal to $M - N$, where $M$ is the total number of subjects and $N$ is the number of clusters (Muthén, 1994). Steps 3, 4, and 5 are redundant in this case.

### Step 2: Establish a Measurement Model at Level 1

In the second step, we establish a measurement model for $\mathbf{\Sigma}_{\text{WITHIN}}$, leaving $\mathbf{\Sigma}_{\text{BETWEEN}}$ unconstrained. So, both levels are analyzed simultaneously while specifying a saturated model at the between level.

### Step 3: Investigate Bias With Respect to Level 1 Violators

In Step 3, we take the measurement model that we established in Step 2, and using this model, we investigate bias with respect to Level 1 violators. In this step, we still do not model the Level 2 covariance matrix; that is, the Level 2 model remains saturated.

MGFA is not suitable for bias investigation with respect to Level 1 violators. This is because by creating groups based on a Level 1 violator, part of the clustering structure in the model is lost. For example, if we split children in classes into a group with boys and a group with girls, we disregard that some boys and girls have the same teacher. Considering this, the RFA method is better suited to investigate bias at the within level. Therefore, the Level 1 violators of interest are added as covariates, and the direct effects of the violators on the indicators are tested. All direct effects that are considered significant and relevant should be added to the model. The significance of direct effects could be tested one by one by likelihood ratio tests of models with and without the estimated direct effect. Alternatively, modification indexes of the (fixed) direct effects in the most constrained model could be used (Sörbom, 1989). Modification indexes reflect the expected decrease in the model's chi-square, if the associated parameter (direct effect) would be freely estimated.

### Step 4: Investigate Cluster Bias

The fourth step involves establishing measurement invariance with respect to the cluster variable by the imposition of appropriate constraints in the two-level model. We refer to measurement bias with respect to the cluster variable as cluster bias. Cluster bias is caused by one or more Level 2 variables. These variables could be measured or unmeasured, perhaps even unknown, but cluster bias can still be investigated. Investigation of cluster bias can thus be seen as an overall test for measurement bias with respect to all possible Level 2 violators. As explained in Jak et al. (2013), in the absence of cluster bias, the following two-level model holds:

$$\mathbf{\Sigma}_{\text{BETWEEN}} = \mathbf{\Lambda}\,\mathbf{\Phi}_{\text{B}}\,\mathbf{\Lambda},'$$

and

$$\mathbf{\Sigma}_{\text{WITHIN}} = \mathbf{\Lambda}\,\mathbf{\Phi}_{\text{W}}\,\mathbf{\Lambda}' + \mathbf{\Theta}_{\text{W}}; \tag{6}$$

that is, a model with the same factor loadings at Level 1 and Level 2, and no residual variance at Level 2. If the factor loadings are not equal over levels, the common factors do

not have the same interpretation over levels (Muthén, 1990; Rabe-Hesketh, Skrondal, & Pickles, 2004), so the Level 2 common factor scores cannot be interpreted as the simple cluster means of the Level 1 common factor scores. If the residual variance of a given indicator variable is found to be greater than zero, then the indicator is affected by cluster bias. If cluster bias in several indicators is caused by the same (possibly unobserved) violator, then the residual factors at Level 2 will covary.

Three issues about the model specification in the test of cluster bias require attention. The first concerns the scaling of the common factors. With freely estimated factor loadings at both levels, the common factors on Level 1 and Level 2 can be given a metric by fixing their variances at unity. With equality constrained factor loadings, and the factor variances at Level 1 fixed at unity, the factor variances at Level 2 are identified by the equality constraints on the factor loadings and can be freely estimated.

The second issue concerns correlated residuals. The test for cluster bias is based on the factor structure established in Step 2. If this factor model includes correlated residuals, the model should be reparameterized. The reason is that in the test of cluster bias, the residual variance on Level 2 has to be zero, and the same structure is imposed on the within and between level (Equation 6). Instead of correlated residuals, an additional common factor can be introduced. With the two factor loadings fixed at 1, the estimate of the common factor's variance is equal to the (possibly negative) estimate of the covariance between the residuals. Note that this common factor should be uncorrelated to the other factors in the model, and its variance should be estimated at both levels.

The third issue concerns testing the significance of the Level 2 residual variance. Because variances are on the boundary of the parameter space under the hypothesis that they are zero, the omnibus likelihood ratio test could be a complex mixture of $\chi^2$ distributions (Stoel et al., 2006). This pertains to the same problem as in Step 1. However, in the test of cluster bias we can simplify the distribution of the likelihood ratio statistic by testing a single variance parameter at a time. The distribution of this likelihood ratio is a simple 50/50 mixture of a $\chi^2$ distribution with 0 *df* (so half of the area under the curve equals zero) and a $\chi^2$ distribution with 1 *df*. When testing whether a single residual variance equals zero, the likelihood ratio test requires only a simple adjustment of the chosen alpha level. In this case alpha is multiplied by two, which is similar to the procedure in one-sided instead of two-sided testing. For example, with 1 *df,* the critical $\chi^2$ value associated with an alpha level of .05 is 3.84 for a two-sided test and 2.71 for a one-sided test.

### Step 5: Investigate Bias With Respect to Level 2 Violators

The model we propose to use in Step 5 is the final model of Step 4 with all Level 1 and Level 2 violators as covariates. At Level 1, this corresponds to the final RFA model from Step 3. If the factor loadings are still equal across Level 1 and Level 2, the common factor(s) have the same interpretation at both levels.

With respect to Level 2 violators, the pros and cons of MGFA and RFA (or the MIMIC model) coincide with those of single-level analysis. We apply the RFA method, because it facilitates the investigation of uniform bias with respect to all aggregated Level 1 violators and the specific Level 2 violators simultaneously. See Muthén, Khoo, and Gustafsson (1997) and Spilt, Koomen, and Jak (2012) for examples of MGFA with Level 2 violators.

If bias with respect to Level 2 violators has been found, it can be tested whether all cluster bias is explained by the Level 2 violators. This implies testing cluster bias again, but now controlling for the detected bias at Level 2.

## ILLUSTRATION

### Data

The Closeness scale of a Dutch translation of the Student–Teacher Relationship Scale (STRS; Koomen, Verschueren & Pianta, 2007; Koomen, Verschueren, van Schooten, Jak, & Pianta, 2011; Pianta, 2001) includes 11 items. Closeness refers to the degree of warmth and open communication. The closeness items are given in the Appendix. Data of 1,493 students (Level 1) were gathered from 659 primary school teachers (Level 2; 182 men, 477 women) from 92 regular elementary schools. One hundred eighty-two male teachers reported on 242 boys and 227 girls; 477 female teachers reported on 463 boys and 561 girls. The children were in Grades 1 through 6. Responses were given on a 5-point scale ranging from 1 (*definitely does not apply*) to 5 (*definitely does apply*).

### Statistical Analysis

Measurement bias was investigated with respect to pupil sex (Level 1) and teacher sex (Level 2). For simplicity, we treat the item responses as continuous, although in fact they are ordinal. For examples of fitting multilevel models to ordinal item responses we refer to (among others) Grilli and Rampichini (2007), Ansari and Jedidi (2000), and Goldstein and Browne (2005). We used robust maximum likelihood estimation (MLR) in M*plus* (Muthén & Muthén, 2007) to obtain parameter estimates. This estimation method provides a test statistic that is asymptotically equivalent to the Yuan–Bentler T2 test statistic (Yuan & Bentler, 2000), and standard errors that are robust for nonnormality. A correction factor for the chi-squares is used to calculate chi-square differences between nested models (Satorra & Bentler, 2001).

In addition to the adjusted $\chi^2$ statistic, the root mean square error of approximation (RMSEA; Steiger & Lind, 1980) and the comparative fit index (CFI; Bentler, 1990) were used as measures of overall goodness of fit. RMSEA

values smaller than .05 indicate close fit, and values smaller than .08 are still considered satisfactory (Browne & Cudeck, 1992). CFI values over .95 indicate reasonably good fit (Hu & Bentler, 1999).

We used RFA (Oort, 1992, 1998) to investigate measurement bias with respect to pupil's sex and teacher's sex. Sex was entered as an exogenous variable that is correlated with the common factor, and that has no direct effects on the item scores. Direct effects were added if the modification index was significant at a Bonferroni corrected level of significance (two-sided $\alpha = .05$/number of possible effects). When testing cluster bias, we started with a fully constrained model, and freed parameters if needed. We tested the residual variances one by one at a Bonferroni corrected one-sided level of significance of .05 (i.e., .10 two-sided) divided by the number of constrained variances at the between level. The equality of factor loadings across levels was tested at $\alpha = .05$ divided by the number of constrained factor loadings.

## Results

*Step 1: Test whether there is Level 2 variance and covariance.* The ICCs for the closeness items varied between .13 (for Item 8) and .28 (for Items 3 and 5). The Level 2 variance and covariance was significant, indicated by a significant $\chi^2$ for the null model, $\chi^2(66) = 702.16$, $p < .05$, RMSEA = .080, CFI = .87; and for the independence model, $\chi^2(55) = 178.35$, $p < .05$, RMSEA = .039, CFI = .98. Although the RMSEA and the CFI of the independence model indicate satisfactory fit, the $\chi^2$ indicates that there is significant covariance.

*Step 2: Establish a measurement model at the within level.* A one-factor model fitted closely to the Level 1 covariance matrix, $\chi^2(44) = 111.15$, $p < .05$, RMSEA = .032, CFI = .99.

*Step 3: Investigate measurement bias with respect to pupil's sex.* The RFA model with pupil's sex as an exogenous variable fitted well, $\chi^2(54) = 174.91$, $p < .05$, RMSEA = .039, CFI = .98. However, modification indexes suggested direct effects of pupil's sex on Item 2 and Item 3. Adding these direct effects significantly improved model fit, $\Delta\chi^2(2) = 34.96$, $p < .05$. The correlation between the common factor closeness and pupil's sex was positive and significant ($r = .25$, $p < .05$). As boys were scored 0 and girls 1, this means that teachers experience more closeness with girls than with boys. The standardized direct effects on Item 2 and Item 3 were both positive ($\beta = .10$), indicating that for equal levels of closeness, girls obtained higher scores than boys on these items.

*Step 4: Test for cluster bias (are we measuring the same over teachers?).* The model with equal factor loadings at the within and between level and no residual variance at the between level did not fit the data satisfactorily, $\chi^2(109) = 831.67$, $p < .05$, RMSEA = .067, CFI = .85. One by one freeing of the Level 2 residual variance of the indicators with the highest modification indexes resulted in a model with all Level 2 residual variances estimated. This model fitted satisfactorily, $\chi^2(98) = 322.77$, $p < .05$, RMSEA = .039, CFI = .95. However, for three indicators, the factor loadings could not be considered equal across Level 1 and Level 2. Therefore, the factor loadings of Item 5, Item 8, and Item 10 were freely estimated. This resulted in a very good fitting model, $\chi^2(95) = 275.23$, $p < .05$, RMSEA = .036, CFI = .96. Items 5 and 10 were more indicative (i.e., had higher factor loadings) of closeness at Level 2, and Item 8 was more indicative of closeness at Level 1. Therefore, the Level 2 common factor cannot simply be interpreted as the aggregated version of the Level 1 factor.

The presence of cluster bias in all closeness items shows that there are other factors than teacher's closeness with pupils that cause differences on the closeness items. Teacher sex could be one explanation for these differences.

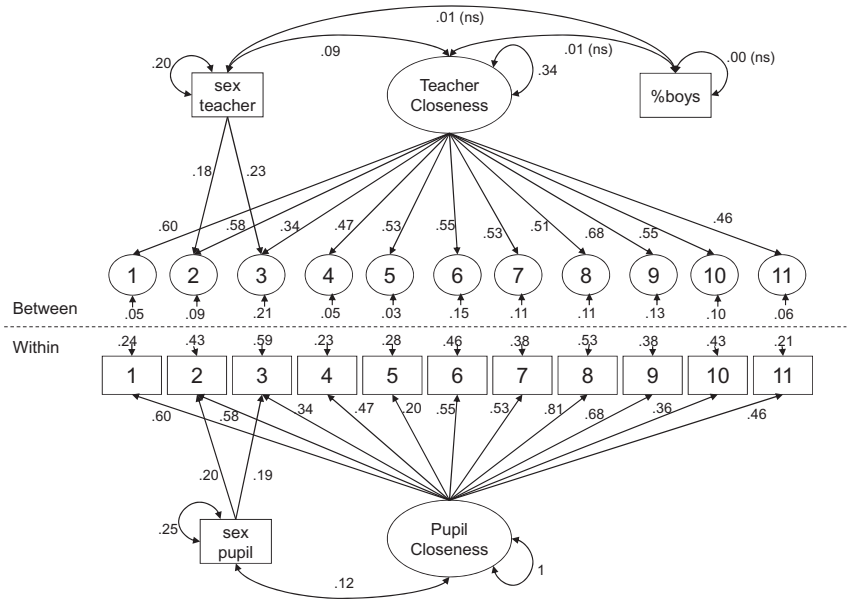
*Step 5: Investigate measurement bias with respect to teacher's sex.* An RFA model with teacher's sex and the proportion of boys in the classroom as exogenous variables at the between level and the final RFA model from Step 3 at the within level fitted the data well, $\chi^2(123) = 351.36$, $p < .05$, RMSEA = .035, CFI = .96. In this model, all factor loadings, except for Items 5, 8, and 10 were constrained to be equal across Level 1 and Level 2, and all residual variance at Level 2 was estimated. Step-by-step inspection of modification indexes pointed to teacher sex bias in Items 2 and 3. Addition of two direct effects of teacher sex on these items resulted in good model fit, $\chi^2(121) = 330.47$, $p < .05$, RMSEA = .034, CFI = .96. A graphical representation with parameter estimates of this model is given by Figure 1. The correlation between closeness and teacher sex is .34, indicating that female teachers experience more closeness than male teachers. The standardized direct effects were both positive, $\beta = .17$ for Item 2 and $\beta = .19$ for Item 3.These items are thus considered more applicable by female teachers; that is, with equal levels of closeness, female teachers give higher scores on these items than male teachers.

Fixing the Level 2 residual variance at zero for the two biased items significantly deteriorated model fit, $\Delta\chi^2(2) = 185.58$, $p < .05$. Not all cluster bias in these items, therefore, is explained by teacher sex.

## Conclusion

The bias with respect to pupil's sex in Item 2 and Item 3 shows that the difference between boys and girls on these items is larger than would be expected based on their common factor scores. In other words, even if the levels of closeness were equal, girls get somewhat higher scores on

**Unstandardized parameter estimates**




**Standardized parameter estimates**

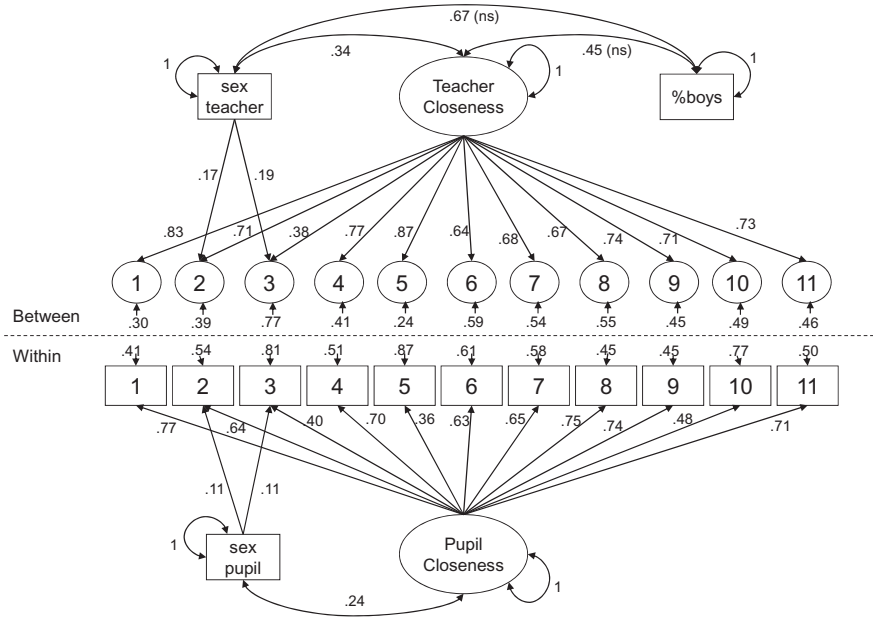



FIGURE 1 Restricted factor analysis model from Step 5. The top shows the unstandardized parameter estimates, the bottom shows the standardized parameter estimates (standardized within Level 1 and within Level 2). Nonsignificant parameter estimates are indicated by "(ns)."

these items. Item 2 is about the child seeking comfort when he or she is upset. Apparently, in the perception of teachers, girls seek more comfort than boys do, given equal levels of closeness. Item 3 is about the children's reaction to physical affection or touch from the teacher. So, with equal levels of closeness, girls seem to be more comfortable with physical affection than boys (in the perception of teachers).

Items 2 and 3 were also biased with respect to teacher sex in the same direction. An explanation for this bias in Item 2 is that female teachers in general experience more comfort seeking from children. For Item 3, it is hypothesized

that male teachers show their closeness less with physical affection or touch than female teachers do. A possible explanation could be that male teachers fear being accused of touching children in inappropriate ways (Jones, 2004).

If one would not control for the bias in the two items, the correlation between closeness and sex would be slightly overestimated, (.26 instead of .24 for pupil sex, and .36 instead of .34 for teacher sex). In all items, cluster bias was still present, even after controlling for teacher sex bias. Apparently, other Level 2 violators are causing differences in the closeness items, so that not all differences between teachers can be attributed to differences in the average closeness of the teachers with their pupils.

## DISCUSSION

This article proposes a five-step approach to the detection of measurement bias with respect to Level 1 violators, Level 2 violators, and the clustering variable. We illustrated the approach using data from teacher–child interactions. The five steps of the approach were suggested based on the idea of working upward from Level 1, so that the final model accounts for all bias and substantive findings at both levels. The five-step approach seems the most obvious approach to us. However, we are not claiming this is the only way. The order of Step 3 (investigating bias with respect to Level 1 violators) and Step 4 (testing cluster bias) can be reversed without consequences for the final model in Step 5. Another possibility could be not to work upward from Level 1, but analyze the two levels separately, by investigating Level 2 bias with an unrestricted model at Level 1. When we analyzed our data in this way, we found no Level 2 bias. This is probably the result of diminished statistical power. In general, the results in a multistep analysis might depend on the details of the procedure. In most situations, a universally optimal procedure might not exist. We expect that different procedures will generally identify the same items as being biased, but the power to detect bias might vary. If one is unsure whether the bias finding should be taken seriously, being able to explain the bias substantively might be the ultimate check.

In two-level data, the statistical power to detect measurement bias is of importance at both levels. The sample size at Level 1 will often be large enough for sufficient power. The sample size at Level 2 might often be too small. In our example there were 659 clusters, but data sets with just 100 clusters are very common. According to Maas and Hox (2005), parameter estimates and statistical tests are accurate from 100 clusters. Concerning the test of cluster bias, 50 clusters appeared to be sufficient to detect cluster bias of medium size (Jak et al., 2013). According to Maas and Hox, the number of Level 1 units within Level 2 units is not very influential in the results.

In our application, we do not test the absence of nonuniform measurement bias with respect to the Level 1 and Level 2 violators. As pointed out in the introduction, there are ways within RFA to test for nonuniform measurement bias (Barendse et al., 2010; Barendse et al., 2011; Molenaar et al., 2010). However, these methods have yet to be evaluated in multilevel models. Until these methods are available in multilevel situations, MGFA can be used to investigate nonuniform bias with respect to Level 2 violators. When applying MGFA to our data, we did not find nonuniform bias with respect to teacher sex, whereas the same uniform bias (in Item 2 and Item 3) was found.

Seeing that various choices can be made, when investigating measurement bias in multilevel data, we aimed at providing some guidance by presenting a five-step procedure that facilitates the investigation of measurement bias with respect to Level 1 and Level 2 violators. Using this approach, the final model takes all bias and substantive findings into account.

## ACKNOWLEDGMENTS

The authors would like to thank Helma Koomen for making her data available for secondary analysis.

## REFERENCES

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika, 65,* 475–496.

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

Barendse, M. T., Oort, F. J., Werner, C. S., Ligtvoet, R., & Schermelleh-Engel, K. (2011). *Measurement bias detection through factor analysis*. Manuscript submitted for publication.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*, 230–258.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Dolan, C. V., Roorda, W., & Wicherts, J. M. (2004). Two failures of Spearman's hypothesis: The GATB in Holland and the JAT in South Africa. *Intelligence*, *32*, 155–173.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491.

Goldstein, H. (1995). *Multilevel statistical models*. New York, NY: Halstead Press.

Goldstein, H., & Browne, W. (2005). Multilevel factor analysis models for continuous and discrete data. *Contemporary psychometrics: A festschrift for Roderick P. McDonald,* 453–475.

Grilli, L., & Rampichini, C. (2007). Multilevel factor models for ordinal variables. *Structural Equation Modeling*, *14*, 1–25.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

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.

Jones, A. (2004). Social anxiety, sex, surveillance and the "safe" teacher. *British Journal of Sociology of Education*, *25*, 53–66.

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

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*, 347–387.

Koomen, H. M. Y., Verschueren, K., & Pianta, R. C. (2007). *Leerling-Leerkracht Relatie Vragenlijst (LLRV): Handleiding* [Student–Teacher Relationship Scale: Manual]. Houten, The Netherlands: Bohn Stafleu van Loghum.

Koomen, H. M. Y., Verschueren, K., van Schooten, E., Jak, S., & Pianta, R. C. (2011). Validating the Student–Teacher Relationship Scale: Testing factor structure and measurement invariance across child gender and age in a Dutch sample. *Journal of School Psychology*. doi:10.1016/j.jsp.2011.09.001

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

Longford, N. T. (1993). *Random coefficient models*. Oxford, UK: Clarendon.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1,* 86–92.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*, 361–388.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Statistics*, *13*, 127–143.

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

Millsap, R. E., & Everson, H. (1991). Confirmatory measurement model comparison using latent means. *Multivariate Behavioral Research*, *26*, 479–497.

Millsap, R. E., & Tein, J. Y. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*, 479–515.

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.

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

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

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*, 376–398.

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

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

Oort, F. J. (1991). Theory of violators: Assessing unidimensionality of psychological measures. In R. Steyer, K. F. Wender, & K. F. Widaman (Eds.), *Psychometric methodology* (pp. 377–381). Stuttgart, Germany: Fischer.

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.

Pianta, R. C. (2001). *Student–Teacher Relationship Scale: Professional manual*. Lutz, FL: Psychological Assessment Resources.

Preacher, K., Zyphur, M., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*, 209–233.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*, 517–529.

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

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239.

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*, 371–384.

Spilt J. L., Koomen, H. M. Y., & 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.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA (Vol. 758).

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.

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*, *2*, 4–69.

Woods, C. M. (2009). Evaluation of MIMIC-model methods to DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*, 1–27.

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.

## APPENDIX
## CLOSENESS ITEMS

1. I share an affectionate, warm relationship with this child.
2. If upset, this child will seek comfort from me.
3. This child is uncomfortable with physical affection or touch from me (reverse scored).
4. This child values his/her relationship with me.
5. When I praise this child, he/she beams with pride.
6. This child tries to please me.
7. It is easy to be in tune with what this child is feeling.
8. This child openly shares his/her feelings and experiences with me.
9. My interactions with this child make me feel effective and confident.
10. This child allows himself/herself to be encouraged by me.
11. This child seems to feel secure with me.