Cluster bias: Testing measurement invariance in multilevel data

Jak, S.

Publication date
2013

CHAPTER 5

Measurement bias in multilevel data

Abstract Measurement bias can be detected using structural equation modeling (SEM), by testing measurement invariance with multi group factor analysis (MGFA; Jöreskog, 1971; Sörbom, 1974; Meredith, 1993), MIMIC modeling (Muthén, 1989) or with restricted factor analysis (RFA; 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 may complicate measurement bias research. In two-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). In this paper, we provide a stepwise approach for the detection of measurement bias with respect to these three types of violators. We propose working from Level 1 upwards, so the final model accounts for all bias and substantive findings at both levels. The 5 proposed steps are illustrated with data concerning teacher-child relationships.

Based on: Jak, S., Oort, F.J. & Dolan, C.V. (in press). Measurement bias in multilevel data. *Structural Equation Modeling.*

INTRODUCTION

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 of these respondents (conditional on their common latent trait score) is equal. In the presence of sex bias, this does not hold and we consider the test biased with respect to sex. Sex is a nominal variable, but measurement bias may be tested with respect to any variable. Measurement bias can be detected using structural equation modeling (SEM), by testing measurement invariance with multi-group factor analysis (MGFA; Jöreskog, 1971; Meredith, 1993; Sörbom, 1974), MIMIC modeling (Muthen, 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 in order to account for the multilevel structure. Second, the variable with respect to which measurement bias is to be investigated may be defined at different levels. For example, a Level 1 variable may be sex of the pupils; a Level 2 variable may be sex of the teachers. The biasing variable may also be class itself, i.e., the clustering variable, which we view as a special kind of Level 2 variable.

Here, we propose a 5-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 5-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, e.g., children nested in classrooms or employees nested in teams. Individuals who are members of the same group, share group level characteristics, and may 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

datasets (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) in groups (Level 2). 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 covariances of $\mathbf{y}$ ($\mathrm{COV}(\mathbf{y}, \mathbf{y}) = \boldsymbol{\Sigma}_{\mathrm{TOTAL}}$) can be written as the sum of the covariances of $\boldsymbol{\mu}$ ($\mathrm{COV}(\boldsymbol{\mu}, \boldsymbol{\mu}) = \boldsymbol{\Sigma}_{\mathrm{BETWEEN}}$) and the covariances of $\boldsymbol{\eta}$ ($\mathrm{COV}(\boldsymbol{\eta}, \boldsymbol{\eta}) = \boldsymbol{\Sigma}_{\mathrm{WITHIN}}$):

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

As $\boldsymbol{\eta}_{ij}$ represents the individual deviations from the group mean, the expected value of $\boldsymbol{\eta}_{ij}$ ($\boldsymbol{\mu}_{\mathrm{WITHIN}}$) is zero, and the overall mean ($\boldsymbol{\mu}_{\mathrm{TOTAL}}$) equals the expected value of $\boldsymbol{\mu}_j$ ($\boldsymbol{\mu}_{\mathrm{BETWEEN}}$):

$$\boldsymbol{\mu}_{\mathrm{TOTAL}} = \boldsymbol{\mu}_{\mathrm{BETWEEN}}. \tag{3}$$

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 may be common factor models:

$$\boldsymbol{\Sigma}_{\mathrm{BETWEEN}} = \boldsymbol{\Lambda}_{\mathrm{B}} \boldsymbol{\Phi}_{\mathrm{B}} \boldsymbol{\Lambda}_{\mathrm{B}}' + \boldsymbol{\Theta}_{\mathrm{B},} \tag{4}$$

$$\boldsymbol{\mu}_{\mathrm{BETWEEN}} = \boldsymbol{\tau}_{\mathrm{B}} + \boldsymbol{\Lambda}_{\mathrm{B}} \boldsymbol{\kappa}_{\mathrm{B}}, \tag{5}$$

$$\boldsymbol{\Sigma}_{\mathrm{WITHIN}} = \boldsymbol{\Lambda}_{\mathrm{W}} \boldsymbol{\Phi}_{\mathrm{W}} \boldsymbol{\Lambda}_{\mathrm{W}}' + \boldsymbol{\Theta}_{\mathrm{W}}, \tag{6}$$

Here, $\boldsymbol{\Phi}_B$ and $\boldsymbol{\Phi}_W$ are covariance matrices of the common factors at the between and within level respectively, $\boldsymbol{\Theta}_B$ and $\boldsymbol{\Theta}_W$ are (diagonal) matrices with variance of the residual factors at the between and within level respectively, $\boldsymbol{\kappa}_B$ is a vector with common factor means at the between level, $\boldsymbol{\Lambda}_B$ and $\boldsymbol{\Lambda}_W$ are matrices with factor loadings at the between and within level, respectively, and $\boldsymbol{\tau}_B$ is a vector with intercepts at the between level. The dimensions of these matrices and the parameter estimates can be different over the two levels. For example, one may 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 influences 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:

$$f_1(X \mid T = t, V = v) = f_2(X \mid T = t). \tag{7}$$

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 non-uniform bias. Throughout this paper 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 may be nominal, ordinal, interval or ratio variables, they may be latent or manifest, and their relationships may be linear or nonlinear. Within SEM, X is typically observed continuous or 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, observed or unobserved. One possible way of testing measurement invariance in the case of a nominal variable V, e.g., sex, is through multi group factor analysis (MGFA). In this model, 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, violations of the equality (over groups) of the factor loadings and intercepts are interpreted as non-uniform 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 eq. 7. For an overview of the use of MGFA for measurement invariance testing, see Vandenberg & Lance (2001), Millsap & Everson (1993), 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 (Muthen, 1989). These models differ only in the treatment of the violator V. In the MIMIC model, T is regressed on V, while 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 of creating groups, while 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, while in RFA, it only involves the addition of covariates. A disadvantage of the RFA method is that the detection of non-uniform bias is less straightforward. However, recent developments using latent interaction terms or moderated factor analysis provide a viable method to investigate non-uniform 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 paper, 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 non-uniform bias is generally lower than for uniform bias (Barendse et al., 2010; Woods, 2009). Besides this, non-uniform bias is often hard to interpret, as it involves an effect of the interaction of V and T on X.

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 over the classes. In this case, two pupils in different classes with identical values of the latent trait, may differ with respect to their expected observed test score. As explained in Jak et al. (2013), the presence of cluster bias can be tested by imposing specific constraints on 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, it is suggested 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. Of course, one should bear in mind that the power to detect bias with respect to specific measured Level 2 variables may be greater than the power of the overall test for cluster bias.

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 socio economic 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 may 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 5-STEP PROCEDURE

To facilitate the practice of bias investigation with respect to the three types of violators, we propose a 5-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, cluster bias, and measurement bias with respect to Level 2 violators. The five steps we propose are:

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 comprises the findings from Step 2, and Step 5 comprises the findings from Step 4. As there are several issues that should be considered, there are other procedures that could be followed. 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 5-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 upwards 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 from teacher-child relationship research in Section 3.

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 intra class correlation of a given variable (ICC) reflects the proportion of the variance that can be attributed to Level 2. Besides qualifying the magnitude of the between variance, one may

wish to test whether the Level 2 variance deviates significantly from zero. The significance of the between variance and covariance can be tested by fitting a null-model ($\mathbf{\Sigma}_{BETWEEN} = 0$) and independence model ($\mathbf{\Sigma}_{BETWEEN}$ is diagonal) to the between covariance matrix, while specifying a saturated model for $\mathbf{\Sigma}_{WITHIN}$ (Hox, 2002; Muthén, 1994). If the χ^2 test statistic of the null model is significant, we conclude that there is significant Level 2 variance. If the χ^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 may be a complex mixture of many different χ^2 distributions. In this stage, we accept that the testing procedure is not correct, and keep in mind that it leads to an over-conservative test, so 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 may 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). Step 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 $\mathbf{\Sigma}_{WITHIN}$, while leaving $\mathbf{\Sigma}_{BETWEEN}$ 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, i.e., 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 in 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 on the within level. So, 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 between a model with and without the estimated direct effect.

Alternatively, modification indices of the direct effects in the most constrained model could be used (Sorbom, 1989). Modification indices reflect the expected decrease in the models 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 (measured or unmeasured) Level 2 variables. Investigation of cluster bias can therefore 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 model holds:

$$\boldsymbol{\Sigma}_{\text{BETWEEN}} = \boldsymbol{\Lambda}\,\boldsymbol{\Phi}_{\text{B}}\,\boldsymbol{\Lambda}', \quad \text{and}$$

$$\boldsymbol{\Sigma}_{\text{WITHIN}} = \boldsymbol{\Lambda}\,\boldsymbol{\Phi}_{\text{W}}\,\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_{\text{W}}. \tag{8}$$

I.e. a model with equal factor loadings across Level 1 and Level 2, and no residual variance at Level 2. The test for cluster bias implies constraining factor loadings to be equal across levels and testing whether the residual variances at Level 2 are zero. 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(s) cannot be interpreted as the aggregate of the Level 1 common factor(s). If the residual variance of a given indicator is found to be greater than zero, then the indicator is affected by cluster bias.

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. This is because in the test of cluster bias, the residual variance on Level 2 has to be zero, while the same structure is imposed on the within and between level (Eq. 8). 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 may be a complex mixture of χ^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 relative simple 50/50 mixture of a χ^2 distribution with 0 degrees of freedom (so half of the area under the curve equals zero) and a χ^2 distribution with 1 degree of freedom. 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 one degree of freedom, the critical χ^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, but with residual variance at Level 2, and 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 constrained to be equal across Level 1 and Level 2, the common factor(s) have the same interpretation at both levels. We propose to estimate residual variance at Level 2 for all indicators here, even for indicators were cluster bias was not found in Step 4.

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 & Jak (2011) 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; Pianta, 2001) comprises 11 items. Closeness refers to the degree of warmth and open communication. The closeness items are given in Appendix A. Data of 1493 students were gathered from 659 primary school teachers (182 men, 477 women) from 92 regular elementary schools. 182 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, while 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 Mplus (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 non-normality. 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 χ^2 statistic, the root mean squared 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. CFI values over .95 indicate reasonably good fit (Hu & Bentler, 1999).

We used restricted factor analysis (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). However, we included direct effects only, if the standardized direct effect was larger than .10. 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 one-sided level of significance of .05 (i.e., .10 two-sided) divided by the number of constrained variances at the between

level. The one-sided level of significance is used here because we are testing a variance (Stoel et al. 2004). The equality of factor loadings over levels was tested at $\alpha = .05$ / number of constrained factor loadings.

RESULTS

Step 1: Test whether there is Level 2 variance and covariance

The intraclass correlations (ICC's) for the closeness items varied between .13 (for Item 8) and .28 (for Item 3 and Item 5). The Level 2 variance and covariance was significant, indicated by a significant χ^2 for the null model (χ^2 (66) = 702.16, $p < .05$, RMSEA = .080 and CFI=.87) and for the independence model (χ^2 (55) = 178.35, $p < .05$, RMSEA = .039 and CFI = .98). Although the RMSEA and the CFI of the independence model indicate satisfactory fit, the χ^2 shows that there is significant covariance.

Step 2. Establish a measurement model at the within level

A one-factor model fitted well to the Level 1 covariance matrix (χ^2 (44) = 111.15, $p < .05$, RMSEA = .032, and CFI = .99). The fit of this model could be further improved by adding a correlation between the residuals of Item 1 and Item 4. However, in previous research the closeness scale is always regarded to be unidimensional (Koomen , Verschueren, van Schooten, Jak & Pianta, 2011; Webb & Neuharth-Pritchett, 2010), and the RMSEA indicates close fit already. Therefore, we accept the one-factor model as the measurement model.

Step 3. Investigate measurement bias with respect to pupil's sex

The RFA model with pupil's sex as an exogenous variable fitted well (χ^2 (54) = 174.91, $p < .05$, RMSEA = .039, and CFI = .98). However, modification indices 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 ($\beta = .10$ and $\beta = .10$), indicating that for equal levels of closeness, girls received higher scores than boys on these items.

Step 4. Test for cluster bias (are we measuring the same across 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 satisfactory (χ^2 (109) = 831.67, $p < .05$, RMSEA = .067, and CFI = .85). One by one freeing of the Level 2 residual variance of the indicators with the highest modification indices resulted in a model with all Level 2 residual variance estimated. This model fitted well (χ^2 (98) = 322.77, $p < .05$, RMSEA = .039, and 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 well fitting model, χ^2 (95) = 275.23, $p < .05$, RMSEA = .036, and CFI = .96. Items 5 and Item 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 directly be interpreted as the aggregated version of the Level 1 factor.

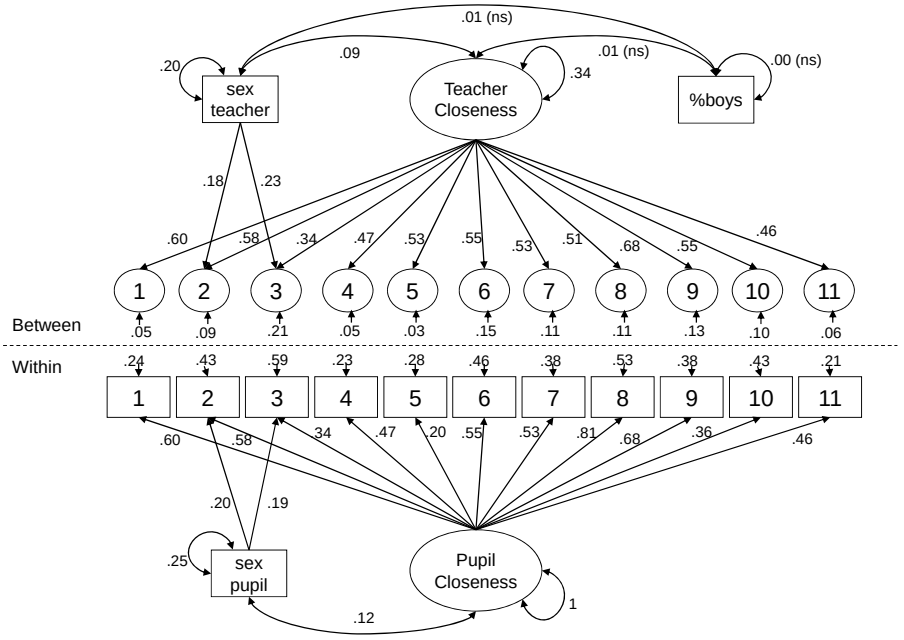
The presence of cluster bias in all closeness items shows that there are other factors than teacher's closeness with pupils that cause differences on the closeness items. Teacher sex could be one explanation for these differences.

Step 5. Investigate measurement bias with respect to teacher's sex

An RFA model with teachers sex and aggregated pupil's sex as exogenous variables at the between level and the final RFA model from Step 3 at the within level fitted the data well, χ^2 (123) = 351.36, p < .05, RMSEA = .035, and 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 indices and standardized parameter change, pointed to teacher sex bias in Items 2 and Item 3. Addition of two direct effects from teacher sex to these items resulted in good model fit, χ^2 (121) = 330.47, $p < .05$, RMSEA = .034, and CFI = .96. A graphical representation with parameter estimates of this model is shown in 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, i.e. 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, $p < .05$). So, not all cluster bias in these items is explained by teacher sex.

Unstandardized parameter estimates



Standardized parameter estimates

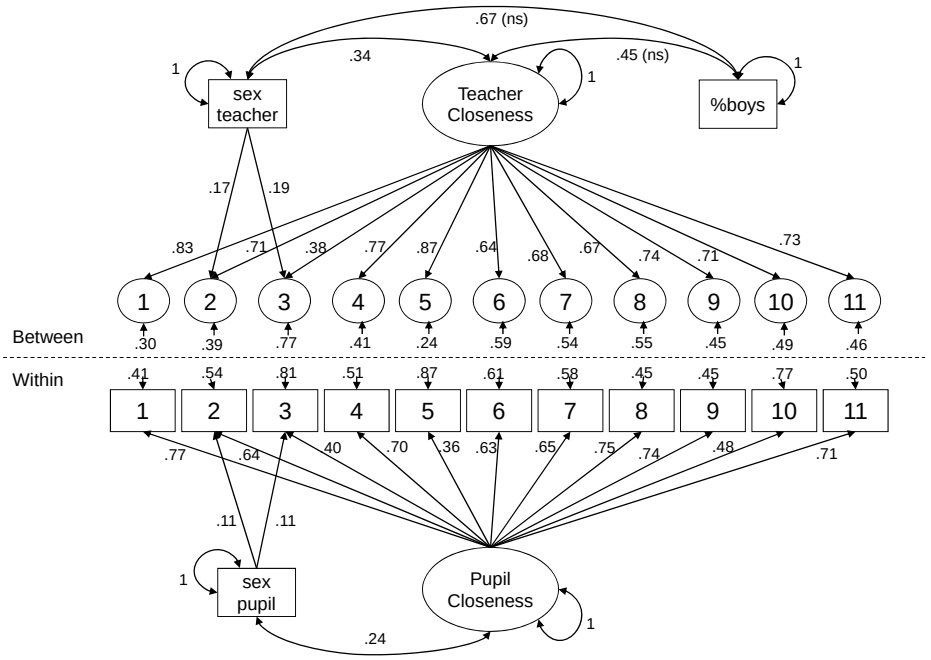


Figure 1. RFA model from Step 5. The upper figure shows the unstandardized parameter estimates, the lower figure shows the standardized parameter estimates (standardized within Level 1 and within Level 2). Non-significant parameter estimates are indicated by '(ns)'.

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 these items. Item 2 is about the child seeking comfort when he/she is upset. Apparently, in the perception of teachers, girls seek more comfort then boys do, given equal levels of closeness. Item 3 is about the children's reaction on 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 paper proposes a step-wise approach for 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 5 steps of the approach were suggested based on the idea of working upward from Level 1, so that the final model comprises all bias and substantive findings at both levels. The 5-step approach seems the most obvious approach to us. However, we are not claiming this is the only way. The order of Step 3 (investigate 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 decreased statistical power. In general, the results in a multi-step analysis may depend on the details of the procedure. In most situations, a universally optimal procedure is unlikely to exist.

We expect that different procedures will generally identify the same items as being biased, but the power to detect the bias may vary. If one is unsure whether the bias finding should be taken seriously, being able to explain the bias substantively may be the ultimate check.

In our application, we do not test the absence of non-uniform 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, Oort & Garst, 2010; Molenaar, Dolan, Wicherts & van der Maas, 2010). However, these methods have yet to be evaluated in the multilevel setup. Until these methods are available in multilevel situations, MGFA can be used to investigate non-uniform bias with respect to Level 2 violators. When applying MGFA to our data, we did not find non-uniform bias with respect to teacher sex, while the same uniform bias (in Item 2 and Item 3) was found.

Varying choices can be made, when investigating measurement bias in multilevel data. We aimed at providing some guidance by presenting a 5 step approach, which 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.

Appendix A. 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.