UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Cluster bias: Testing measurement invariance in multilevel data

Jak, S.

Publication date
2013

Link to publication

Citation for published version (APA):
Jak, S. (2013). *Cluster bias: Testing measurement invariance in multilevel data*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:22 Jun 2025

SUMMARY AND GENERAL DISCUSSION

In this thesis we presented methods and procedures to test and account for measurement bias in multilevel data. Multilevel data are data with a clustered structure, for instance data of children grouped in classrooms, or data of employees in teams. For example, with data of children in classes, we can distinguish two levels in the data: we denote the child level Level 1 or the within level, and the class level Level 2 or the between level. Children in the same class share class level characteristics, such as the teacher, classroom composition, and class size. Such class level characteristics may affect child level variables, leading to structural differences between the responses of children from different classes. With multilevel structural equation modeling (multilevel SEM), we can accommodate such differences by specifying models at the different levels of multilevel data. Such models can be constrained to test substantive and psychometric hypotheses. In this thesis, we considered specifically the psychometric hypothesis of measurement invariance.

Measurement bias is defined as a violation of measurement invariance (Mellenbergh, 1989). Suppose that item X is designed to measure latent attribute T. Measurement invariance with respect to a variable V holds if the conditional distribution of X, given T and V, is equal to the conditional distribution of X, given T. In other words, measurement invariance holds if all influence of V on X runs via T. Within (single level) structural equation modeling, the two prevalent models to investigate measurement bias are multigroup models (Sörbom, 1974; Horn & McArdle, 1992; Little, 1997; Widaman & Reise, 1997) and Restricted Factor Analysis (RFA; Oort, 1992, 1998) or, equivalently, MIMIC (Muthén, 1989) models.

This thesis focusses on the combination of measurement bias and multilevel data. In Chapter 1 we introduced the concept of measurement bias, and in Chapter 2 we presented a test for cluster bias, which serves as an overall test of measurement bias with respect to any Level 2 variable. We extended the test for cluster bias to discrete or ordinal data in Chapter 3. In Chapter 4, we compared the performance of the test for cluster bias with the RFA test. To conclude, in Chapter 5, we presented a five step procedure facilitating the investigation of measurement bias with respect to Level 1 and Level 2 violators of measurement invariance. In the next section, we summarize the main findings of these five chapters, and we discuss the outcomes, contributions, and limitations of this thesis.

MEASUREMENT BIAS AND MULTIDIMENSIONALITY

Chapter 1 shows two examples of measurement bias detection using RFA. We investigated measurement bias with respect to age and gender in a mathematical ability test and in a spatial visualization test. Preceding the detection of measurement bias, examination of the

dimensionality of the measurement models led to two multidimensional measurement models. We stressed the importance of establishing the correct measurement model, as omitting important dimensions from the measurement model may lead to spurious findings of measurement bias. We ended the chapter with the conclusion that measurement bias and multidimensionality are closely related, but not equivalent. Measurement bias implies multidimensionality, but multidimensionality appears as measurement bias only if multidimensionality is not properly accounted for in the measurement model.

A TEST FOR CLUSTER BIAS

In Chapter 2 we presented a test to investigate measurement bias with respect to the clustering variable in multilevel data. We showed how measurement invariance assumptions across clusters imply measurement invariance across levels in a two-level factor model. Cluster bias is investigated by testing whether the within level factor loadings are equal to the between level factor loadings, and whether the between level residual variances are zero. We illustrated the test with an example from educational research. In a simulation study, we showed that with continuous data from five items, the chi-square difference test has sufficient power to detect cluster bias, given a large enough number of clusters. With 50 clusters with 25 observations per cluster, the power to detect cluster bias was sufficient if the bias accounted for 3% or more of the total variance of the indicator. With only 20 clusters with 25 observations each, power to detect cluster bias was still sufficient, if bias accounted for at least 5% of the total variance. The proportions of false positives were higher than the nominal level of significance in conditions with 100 clusters, but lower in conditions with 20 clusters.

TESTING FOR CLUSTER BIAS USING TWO-LEVEL ORDINAL FACTOR ANALYSIS

In Chapter 3 we extended the test for cluster bias to ordinal item responses, using the ordinal two-level factor model (Grilli & Rampichini, 2007). Based on a simulation study, we concluded that cluster bias can be tested in ordinal data with the likelihood ratio test and Wald test. Both tests demonstrated sufficient power to detect large bias, and show acceptable false positive rates. The scaled likelihood ratio test, as implemented in the program Mplus, is not recommended for cluster bias testing, as substantive numbers of inadmissible results were obtained in all conditions. The chapter included an illustration of the test with data concerning research on teacher – student relations.

TESTING FOR CLUSTER BIAS AS A GLOBAL TEST OF MEASUREMENT BIAS

The cause of cluster bias is by definition a cluster level variable. For example, in the case of data of children in classes, cluster bias may be caused by bias with respect to the teacher's teaching ability. In the test of cluster bias, the actual violator of measurement invariance (if any) does not have to be measured. Therefore, the test for cluster bias can serve as a global test of measurement bias with respect to all class level variables. In Chapter 4, we compared the power and false positive rate of the test for cluster bias and the RFA test. As was expected, the RFA test has more power than the test for cluster bias. The test for cluster bias showed a smaller false positive rate overall. We conclude that non-detection of cluster bias does not rule out the possibility that significant bias with respect to a Level 2 violator may be found using the RFA test.

A FIVE STEP APPROACH TO DETECT MEASUREMENT BIAS IN MULTILEVEL DATA

In the final chapter of this dissertation we proposed a step-wise approach for the detection of measurement bias with respect to Level 1 violators, Level 2 violators, and the clustering variable. In this procedure, Step 1 involves testing the necessity of applying multilevel modeling, Step 2 consists of establishing a measurement model at Level 1, Step 3 involves testing for measurement bias at Level 1, Step 4 concerns testing for cluster bias, and Step 5 refers to explaining the cluster bias with observed Level 2 variables. The five steps of the approach were based on the idea of working bottom-up from Level 1, so that the final model considers all bias and substantive findings at both levels. The five steps are illustrated with data about the closeness between teachers and students.

DISCUSSION

In this dissertation we presented a test for cluster bias, i.e. a test for measurement bias with respect to clusters in multilevel data. The cluster bias test is integrated in a framework to test for measurement bias with respect to specific Level 1 and Level 2 variables. The major contribution of this thesis is that it provides researchers guidance to investigate measurement bias in their multilevel data in a viable and systematic way. In the following section, we elaborate on the differences and similarities of our approach in comparison with existing approaches, we discuss multidimensionality in the light of cluster bias. Finally, we identify some limitations of the current work.

ALTERNATIVE APPROACHES TO THE INVESTIGATION OF MEASUREMENT BIAS IN LARGE NUMBERS OF GROUPS

The test for cluster bias is a useful addition to the existing set of structural equation modeling tools to investigate measurement bias. However, it is not the only test that can be used to investigate measurement invariance across clusters in multilevel data. One of the alternatives to the test for cluster bias is to test for measurement bias in a fixed effects model, i.e. in a multigroup model in which each cluster is a group. The equal factor loadings and intercepts across groups (clusters) in a multigroup model represent absence of cluster bias. Although this approach is possible in principle, it is hardly practical when the number of clusters is large or when the within cluster sample size is relatively small. The latter results in instability, the former results in tests with potentially prohibitively large number of degrees of freedom.

Muthén and Asparouhov (2013) describe an alternative way to circumvent the cumbersome strategy of multigroup modeling with large numbers of groups, using a 2-step procedure with Bayesian estimation. They introduce the concept of "approximate measurement invariance", referring to the analysis of measurement invariance across several groups using Bayesian SEM (BSEM). In Step 1 of the procedure (the analysis of approximate measurement invariance), in each group the difference between the group specific measurement parameter (factor loading or intercept) and the average of the particular parameter across all groups is estimated. The researcher can then identify the group with the largest difference between its measurement parameter and the average parameter as the most deviant group. In the next step, using BSEM, one estimates a model in which all factor loadings and intercepts are equal across groups, except for the groups that were identified as deviant in the previous step. This is similar to the use of modification indices with maximum likelihood estimation in a multigroup model, where the most deviant group will show the largest modification index in an analysis with equal factor loadings and intercepts. An advantage of the BSEM method is that it works well for the analysis of categorical variables, while maximum-likelihood estimation with categorical variables often leads to computational problems due to the numerical integration involved (a phenomenon that we encountered in the examples in Chapters 2 and Chapter 3). A disadvantage of the approximate measurement invariance approach is that it relies on prior distributions for the model parameters, and different priors may yield different outcomes. Muthén and Asparouhov recommend zero-mean, small-variance priors for the difference parameters. However, the optimal size of the small-variance of the priors is a subject of debate.

A framework for the detection of measurement bias across large numbers of groups within Bayesian Item Response Theory (IRT) is given by Verhagen and Fox (2012), using multilevel random item effects models (De Jong, Steenkamp & Fox, 2007; Fox &

Verhagen, 2010). Verhagen en Fox estimate a random effects parameter for all measurement parameters in the model (i.e. discrimination parameters and difficulty parameters in an IRT model), and test which of the measurement parameters have significant variance across clusters using Bayes factors or using the Deviance Information Criterion (DIC). Consequently, the cluster level variance in item parameters may be explained by adding covariates to the model. The approach of Verhagen en Fox is similar to the approach in this thesis in some respects. Both approaches treat groups as randomly drawn from a population of groups. Both approaches test the hypothesis of zero variance of parameters at the cluster level, and both allow for the explanation of non-zero variance by cluster level variables. The main differences between the two approaches relate to the modeling framework (multilevel IRT versus multilevel SEM), and the estimation method (Bayesian estimation versus frequentist (maximum likelihood) estimation). It is an interesting topic of future research to compare the outcomes of the two methods, for example by reanalyzing the data from Chapter 4 (about testing for cluster bias with ordinal data), using multilevel random item effects modeling.

MULTIDIMENSIONALITY AND CLUSTER BIAS

In Chapter 1 we showed that measurement bias and multidimensionality are closely related. We discussed that in a one-dimensional model, all items are really affected by two factors: the single common factor and an item-specific residual factor (Spearman, 1928). If all residual variance is only random error variance then measurement bias is absent by definition. If part of the residual variance represents structural variance, then this may stem from a biasing factor. In Chapters 2 to 5 we used a test of zero residual variance as an overall test for measurement bias at the between level in a two-level factor model. At the between level of a two-level factor model, non-zero residual variance always represents measurement bias. This is not the case in single level data (or at the within level), as we cannot distinguish variance caused by item specific factors from random measurement error variance. In the next paragraph we will explain the difference between residual variance in a single (or within) level model and residual variance in a between level model.

In a factor model, residual variance stems from a residual factor (**δ**) that consists of two components, a structural component, **s**, and a random component, **e** (Bollen, 1989). With VAR( ) denoting variance:

$$\mathrm{VAR}(\boldsymbol{\delta}) = \mathrm{VAR}(\mathbf{s}) + \mathrm{VAR}(\mathbf{e}) \quad , \tag{1}$$

in which **s** represents a specific component, that is unique to the indicator, causing systematic variance in the item score. The remaining part of the residual variance is caused

by a random component, **e**, representing measurement error. The expected value, denoted E( ) of the structural component **s** may be non-zero, and could be interpreted as the intercept in a factor model:

$$E(\mathbf{s}) = \boldsymbol{\tau}. \quad (2)$$

The random component is unsystematic and has an expected value of zero:

$$E(\mathbf{e}) = \mathbf{0}. \quad (3)$$

The residual variance of each indicator is thus equal to the sum of the variance of the two components, and the mean of the residual factor is equal to the mean of the structural component.

Zero structural residual variance represents invariance of the indicator with respect to all variables. As mentioned, in a single level model we cannot distinguish structural residual variance from measurement error variance, rendering it impossible to identify nonzero residual variance as measurement bias. At the second (and higher) level of a multilevel model, it *is* possible to test whether structural variance is present. Given that the cluster mean of the random component is expected to be zero (Equation 3), all residual variance at aggregated levels represents structural variance. Of course, if the number of observations per cluster is very small, some random error variance may be aggregated to the higher level. However, in Chapter 4 it appeared that the test for cluster bias did not falsely identify random residual variance as cluster bias even with cluster sizes as small as 2.

LIMITATIONS

The approaches to investigate measurement bias in multilevel data, that we presented in this thesis, were conceptualized within the framework of structural equation modeling. As such they are subject to the assumptions of multivariate normality of continuous data, or multivariate normality of the unobserved continuous responses underlying observed categorical data. Deviations from normality can lead to bias in the model parameters and in goodness of fit measures. Molenaar, Dolan & Verhelst (2010) and Molenaar, Dolan & De Boeck (2012) present models that take different sources of non-normality of the data into account. In future research, it would be interesting to find out how the models presented in this thesis may be combined with models to account for non-normality in the data.

All tests for measurement invariance in this thesis require that the majority of the indicators in a factor model are measurement invariant. For example, when testing the

invariance of the closeness scale with respect to child gender, the results are only valid if the majority of indicators is not biased with respect to gender. If all indicators were biased against boys, i.e. if for equal levels of closeness teachers report higher scores for girls, this bias will not be detected. Overall gender differences are captured by the common factor, so bias against boys in all indicators will lead to biased factor mean differences in a multigroup model, or biased correlations between gender and closeness in the RFA model. Such bias could be detected if we could identify one indicator that is truly invariant across gender. This indicator could then be used as an anchor indicator, by scaling the common factor's variance and mean with respect to this indicator. However, it is impossible to know which, if any, indicator is invariant.