Cluster bias: Testing measurement invariance in multilevel data

Jak, S.

Publication date
2013
Document Version
Final published version

CLUSTER BIAS *Testing measurement invariance in multilevel data*

Suzanne Jak

CLUSTER BIAS

Testing measurement invariance in multilevel data

Suzanne Jak

Cluster bias

Testing measurement invariance in multilevel data

Suzanne Jak

Cover design: Esther Ris, proefschriftomslag.nl

Printed by: Ipskamp Drukkers B.V.

Cluster bias

Testing measurement invariance in multilevel data

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. D.C. van den Boom

ten overstaan van een door het college voor promoties ingestelde

commissie, in het openbaar te verdedigen in de Agnietenkapel

op vrijdag 27 september 2013, te 14:00 uur

door

Suzanne Jak

geboren te Hengelo

Promotiecommissie

Promotores:	Prof. dr. F.J. Oort
	Prof. dr. C.V. Dolan
Overige leden:	Dr. M.E. Timmerman
	Dr. J.M. Wicherts
	Prof. dr. J.J. Hox
	Prof. dr. D. Borsboom
	Dr. A.J. Verhagen

Faculteit der Maatschappij- en Gedragswetenschappen

CONTENTS

INTRODUCTION

This thesis is about using structural equation modelling to detect and account for measurement bias in multilevel data. The basic concepts and their importance will be illustrated below, by using an example from educational research.

Suppose, a researcher is interested in the influence of students' motivation on their mathematical ability. After weeks of calling schools, she finds 200 teachers and 700 students willing to participate in her study. The students complete a motivation questionnaire with 10 items such as "I think learning math is good for me" and "I like math", scored on a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree). The children also make a mathematical ability test consisting of 60 items that can be answered correctly or incorrectly.

Before the researcher can test any hypothesis on the relation between motivation and mathematical ability, the researcher needs to know: are these measurements valid? Are differences in motivation and mathematical ability reflected by differences in the associated item scores (Borsboom, Mellenbergh & van Heerden, 2004)? A related issue is the question of measurement invariance: do the items measure the same attributes for different (groups of) respondents (Mellenbergh, 1989; Meredith, 1993; Oort, 1992, 1993)? If the mathematical ability test items indeed measure the same attribute in boys and girls, then boys and girls with equal mathematical ability should, on average, have identical observed scores. If this is not the case, we speak of measurement bias. For example, an item with a worded math problem may be easier to solve for girls, because girls are generally better in reading than boys (Wei et al., 2012). For that reason, with equal levels of mathematical ability, girls may have more correct answers on this item than boys will.

This thesis presents models and methods to investigate and account for measurement bias in multilevel data, such as data from children in school classes. One difficulty is that we do not have a direct measure of the (latent) variables of interest, such as mathematical ability or motivation, we have to work with observed item scores. The relationship between the observed item scores and motivation or mathematical ability can be represented by a measurement model, such as a linear factor model (Mellenbergh, 1994; Spearman, 1904, 1928). In this thesis we use factor models as measurement models, in which the variables that were intended to be measured are represented by continuous latent common factors, that capture all common variance in the observed scores. Each item is also affected by a unique factor that has a structural part (causing item specific variance), and a random part (measurement error) (Bollen, 1989).

The investigation of measurement bias should always be preceded by the establishment of a sensible measurement model. Chapter 1 serves as an introduction to the concept of measurement bias. Using two examples with data from a cognitive ability test, we show

that measurement bias and multidimensionality are closely related. An item that shows measurement bias is multidimensional, as it taps into a dimension that was not intended to be measured. If this dimension is related to the variable with respect to which measurement bias is tested (often variables like gender, ethnicity, age), then the item is said to be biased with respect to this variable.

Another question that the researcher in the example above might ask is: Are motivation and mathematical ability measured identically in different classrooms? As she collected data from school children, who are clustered in classes, the data have a multilevel structure. Children's scores are affected by class level variables, such as teacher quality and classroom composition. Differences in these variables may lead to differences in the average scores of children in different classrooms, that are not accounted for by the common factor (mathematical ability or motivation in the example). In Chapter 2 we propose a test for cluster bias, which can be used to examine whether measurements are biased with respect to school class. This test can be used in more situations than with children in classes only. It can be used to investigate bias with respect to any clustering variable in multilevel data (e.g. data from people in countries, from patients in hospitals, or from children in families), hence the general term "test for cluster bias".

The motivation items from the example were scored on 7-point scales, which can be treated as continuous scores in a linear factor model (Dolan, 1994). The answers to the math questions were dichotomous (right/wrong), which needs to be taken into account in the measurement model. Chapter 3 extends the test for cluster bias to situations with dichotomous and ordinal item responses.

The cause of cluster bias in the mathematical ability test or in the motivation items is a class level variable, such as the mathematical ability of the teacher. If the researcher also measured the mathematical ability of teachers, she may test whether the differences between the classroom level math scores can be explained by mathematical ability of the teacher. This involves testing for measurement bias with respect to a class level variable. In the population, if there is no cluster bias, there is no bias with respect to any other class level variable. In Chapter 4 it is investigated whether the test for cluster bias indeed detects all bias that is caused by specific class level variables.

The researcher from the example has three types of so-called violators with respect to which the tests may be biased: student level variables (e.g. student's gender, student's ethnicity), the clustering variable (class) and class level variables (e.g. teacher quality, average student SES). In Chapter 5 we propose a 5-step approach to investigate bias with respect to these three types of violators.

This thesis will help researchers who analyse multilevel data to evaluate measurement bias in their research instruments in a systematic and valid way.

CHAPTER 1

Measurement bias and multidimensionality; an illustration of bias detection in multidimensional measurement models

Abstract Restricted factor analysis can be used to investigate measurement bias. A prerequisite for the detection of measurement bias through factor analysis is the correct specification of the measurement model. We applied restricted factor analysis to two subtests of a Dutch cognitive ability test. These two examples serve to illustrate the relationship between multidimensionality and measurement bias. We conclude that measurement bias implies multidimensionality, whereas multidimensionality shows up as measurement bias only if multidimensionality is not properly accounted for in the measurement model.

Based on: Jak, S., Oort, F.J. & Dolan, C.V. (2010). Measurement bias and multidimensionality; an illustration of bias detection in multidimensional measurement models. *Advances in Statistical Analysis, 94*, 129-137.

INTRODUCTION

In the presence of measurement invariance, systematic differences between observed test scores are attributable to true differences in the trait(s) that the test measures. A test is measurement invariant with respect to V, if the following conditional independence holds:

$$f_1(X \mid T = t, V = v) = f_2(X \mid T = t), \quad (1)$$

where X is a set of observed variables, T is a set of attributes measured by X, and V is a set of variables other than T, possibly violating measurement invariance. Function f_1 is the conditional distribution function of X given values of t and v, and f_2 is the conditional distribution function of X given t. If the conditional independence does not hold (i.e., if $f_1 \neq f_2$), the measurement of T by X is said to be biased with respect to V. In the presence of measurement bias, differences between observed test scores may not represent true differences between respondents.

The principle of conditional independence (PCI) was introduced by Mellenbergh (1989) to define item bias (or differential item functioning), with X representing a test item, T a latent trait, and V some group membership. Yet Mellenbergh emphasized the generality of the definition: X, T, and V may be measured on the nominal, ordinal, interval or ratio level, they may be latent or manifest, and their relationships may be linear or nonlinear. In their review of statistical methods for the detection of measurement bias, Millsap and Everson (1993) distinguished between latent variable methods (with latent T) and observed variable methods (with observed T), but they only considered group membership as possible V. Oort (1991) showed that a whole range of measurement issues can be subsumed under the PCI. Relevant measurement issues only differ in what is substituted for X (e.g., item responses, test scores), T (e.g., one or more latent traits), and V (e.g., other items, other latent traits, group membership, time of measurement occasion, socio-demographic variables). Oort called variables V potential violators of unbiased measurement (hence the symbol V). Meredith (1993) used the PCI to define weak measurement invariance and factorial invariance across populations defined by V, and called V a selection variable.

Structural equation modeling (SEM) with latent variables provides flexible means to test measurement invariance, i.e., measurement issues related to the PCI-based definition of unbiased measurement can be investigated using SEM. Most typically, the X variables are observed variables (item scores or test scores) and the T variables are continuous latent variables. The V variables can be group membership in multigroup data, time index in longitudinal data (see King-Kallimanis, Oort & Garst (2010) for an example), or any other

variable, observed or latent. Different SEM methods to detect measurement bias with respect to each of these types of V have been proposed.

If measurement bias is investigated with respect to a nominal V representing groups (e.g., treatment versus control group, men versus women), then we can use multigroup factor analysis (MGFA) with structured means (Sörbom, 1974). In the multigroup method, specific manifestations of bias can be investigated by testing across group constraints on intercepts (uniform bias) and factor loadings (nonuniform bias); see Vandenberg and Lance (2000) for a review. Similarly, measurement bias in longitudinal data (e.g., response shift) can be investigated using longitudinal factor analysis (Oort, 2005).

Another way to detect bias, with respect to any variable (e.g., age, gender, personality trait, attitude, mood), is by conducting restricted factor analysis (RFA) as proposed by Oort (1992, 1998). In the RFA method, uniform bias can be investigated by testing the significance of direct effects of exogenous variables (V) on the observed variables (X). In effect, the RFA method is equivalent with using multiple indicator multiple cause (MIMIC) models to detect measurement bias (Muthén, 1989), the only difference being that in MIMIC models the V variables have causal effects on the T variables, whereas in the RFA method V and T variables are merely associated. Advantages of RFA (and MIMIC analysis) over multigroup factor analysis (MGFA) are that it is not necessary to categorize continuous V variables into groups, and that bias can be investigated with respect to several violators simultaneously.

A prerequisite for the detection of measurement bias through any of these SEM methods is the correct specification of the measurement model. The definition of unbiasedness based on PCI features distributions of X conditional on T. This requires the relationship between X and T, including the dimensionality of T, to be correctly specified. Misspecification of the dimensionality of T in the measurement model may lead spurious bias results (Ackerman, 1992).

In this paper, we present two examples of measurement bias detection through RFA. We focus on the specification of the measurement model, and discuss explicitly the relationship between multidimensionality and measurement bias.

METHOD

The RFA method is used to study measurement invariance of the "Q1000 Capaciteiten Hoog" with respect to age and gender. This is a commercial test, designed to measure cognitive abilities of highly educated people (Meurs HRM, Woerden, The Netherlands). The test consists of seven subtests, with a total of 137 dichotomous items (scored 0 for incorrect, 1 for correct). The test was administered to 1617 respondents (961 men and 656

women, 17 to 63 years of age, m = 37.9, sd = 9.0) as part of a selection procedure for a traineeship in Dutch government. All respondents were highly educated (BA level at least). Here we present the results for two subtests, Mathematical ability and Spatial visualization ability. Prior to investigating measurement bias, we first established the measurement model. Subsequently, we applied the RFA method to investigate bias with respect to gender and age.

ESTABLISHING THE MEASUREMENT MODEL

We first fitted a one-factor model in both subtests. Standardized residuals and modification indices (MIs, this is equivalent to using Lagrange Multiplier tests; Muthén & Muthén, 2006) were used to guide specification search. To guard against capitalizing on chance, the MIs were tested at a Bonferroni adjusted level of significance (nominal alpha of 5% was divided by $p(p-1)/2$, where p is the number of items in the subtest). We only permitted modifications that were amendable to substantive interpretation.

DETECTING MEASUREMENT BIAS

Once we established the measurement models, we added gender and age to the model as exogenous variables. Gender and age were allowed to correlate with each other and with the ability factor(s), but all direct effects of gender and age on the test items were fixed to zero. Measurement bias was evaluated by testing these zero direct effects, using MIs. If the largest of the MIs was significant at a Bonferroni adjusted alpha level (nominal alpha of 5% was divided by pq, where p and q are numbers of items and exogenous variables), the direct effect was set free to be estimated. The associated item was then considered biased. This procedure was repeated until none of the remaining fixed direct effects was significant (at a re-adjusted level of significance, i.e., dividing nominal alpha by $pq - r$, where r is the number of direct effects set free).

STATISTICAL ANALYSIS

As the items of the ability tests are dichotomous, we fitted our models to a matrix of tetrachoric correlations, using weighted least squares with adjusted mean and variance (WLSMV) as implemented in Mplus 4.2 (Muthén & Muthén, 2006). WLSMV provides asymptocically correct standard errors and an adjusted χ^2 statistic (Muthén, du Toit and Spisic 1997). All MIs and χ^2 difference tests were re-scaled to improve the approximation of the χ^2 distribution (Satorra & Bentler, 2001).

In addition to the adjusted χ^2 statistic, the root mean squared error of approximation (RMSEA) and the expected cross validation index (ECVI) were used as measures of

overall goodness-of-fit (Browne & Cudeck, 1993). RMSEA values smaller than 0.05 indicate close fit, and values smaller than 0.08 are still considered satisfactory. Confidence intervals around the RMSEA values and ECVI values were calculated with the freely available computer program NIESEM (Dudgeon, 2003).

RESULTS

MATHEMATICAL ABILITY

Mathematical ability is measured with 12 worded, four-choice math problems. Although the overall goodness-of-fit of the one-factor model was reasonable ($\chi^2 = 329.55$, $df = 48$, $p < .01$, RMSEA= .060 [90% CI: .053, .067], ECVI = .241 [90% CI: .208, .279]), significant MIs identified correlated residuals. All items with correlated residuals were at the end of the test. Apparently, time constraints caused respondents to hurry through the last part of the test, so that the results were affected by speed as well as mathematical ability. We added a second factor, labeled "Speed", to account for the extra shared variance in the last six items. The fit of this two-factor model is close ($\chi^2 = 63.50$, $df = 43$, $p = .02$, RMSEA= .017 [90% CI: .007, .026], ECVI = .083 [90% CI: .072, .099]).

Using this measurement model, we added gender and age as exogenous variables (Figure 2). We found a positive correlation between gender and mathematical ability ($r = 0.34$), indicating higher mathematical ability for men, and a negative correlation between age and speed ($r = -0.20$), indicating that older people are slower, which may have affected their test performance. Two items showed bias. Age had a significant direct effect on Item 1 ($\beta = .12$), indicating that the item is easier for older people: In a subgroup of equally able respondents, older respondents perform better on Item 1. Item 2 was found to be biased with respect to both age ($\beta = -.12$) and gender ($\beta = -.13$): For respondents with equal ability, this item was easier for women, and easier for younger people.

We did not find an immediate explanation for Item 1, which was about chicken farmers and their relative numbers of chickens. Item 2 was a worded problem about employees' preferences of what to do at an upcoming office party. To solve the item, one must assume that half of the male employees prefer dancing over bowling. Perhaps the older male respondents have been distracted more than other respondents by the unusual gender role behaviour.

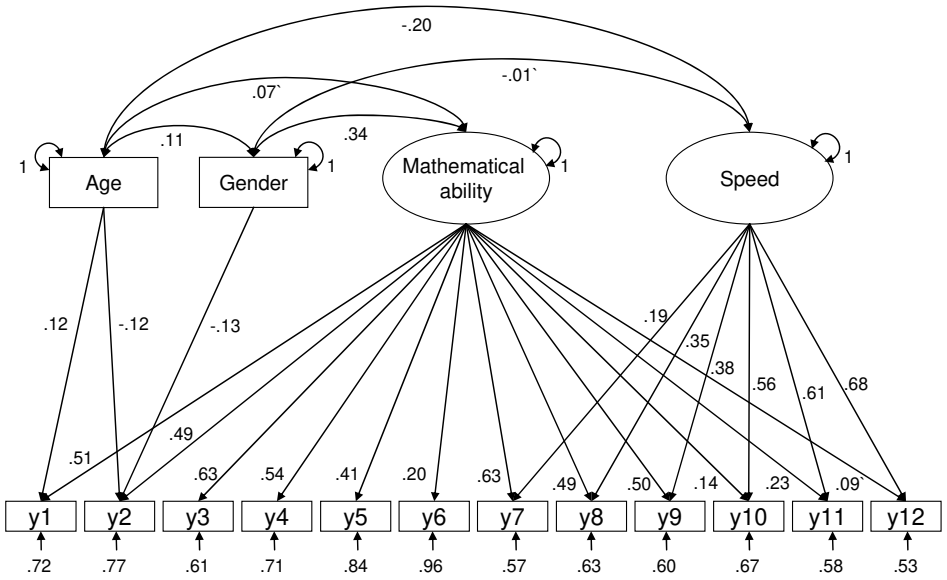


Figure 1 Mathematical ability measured by worded problems.

Notes: All figures denote standardized parameter estimates; apostrophes indicate non-significance; N = 1617; model fit: $\chi^2 = 103.79$, df = 58, $p < .01$, RMSEA= .022 [90% CI: .015, .029], ECVI = .122 [90% CI: .108, .143].

SPATIAL VISUALIZATION ABILITY

The Spatial visualization ability test consists of 17 items. Each item pictures a three-dimensional cube with different patterns on each of its planes. Through mental rotation, respondents have to choose from four options which other cube is a rotation of the first cube.

The overall goodness-of-fit of the one-factor model is reasonable: $\chi^2 = 750.64$, $df = 95$, $p < .01$, RMSEA= .065 [90% CI: .061, .070], ECVI = .537 [90% CI: .485, .594]). However, MIs identified 15 covariances among the item residuals of three subsets of items. Inspection of item content showed that the three groups of items differed in the number of mental rotations needed to solve the items. We modeled this property by adding three factors to the general ability factor, hypothesizing that different mental capacities are required to solve problems that require different numbers of rotations. The fit of this four-factor model was good: $\chi^2 = 133.02$, $df = 87$, $p < .01$, RMSEA= .018 [90% CI: .012, .024], ECVI = .165 [90% CI: .148, .187]).

We added gender and age as exogenous variables to the revised measurement model (Figure 2). Significant positive correlations between gender and general visual-spatial ability (r = .15), specific single rotation ability (r = .12), and double rotation ability (r = .13) indicated that men do slightly better than women. Negative correlations between age and general visual-spatial ability r = -.24), single rotation ability (r = -.18) and triple rotation ability *(r* = -.10) seemed to indicate that the associated skills deteriorate with increasing age. None of the items was found to be biased with respect to age or gender.

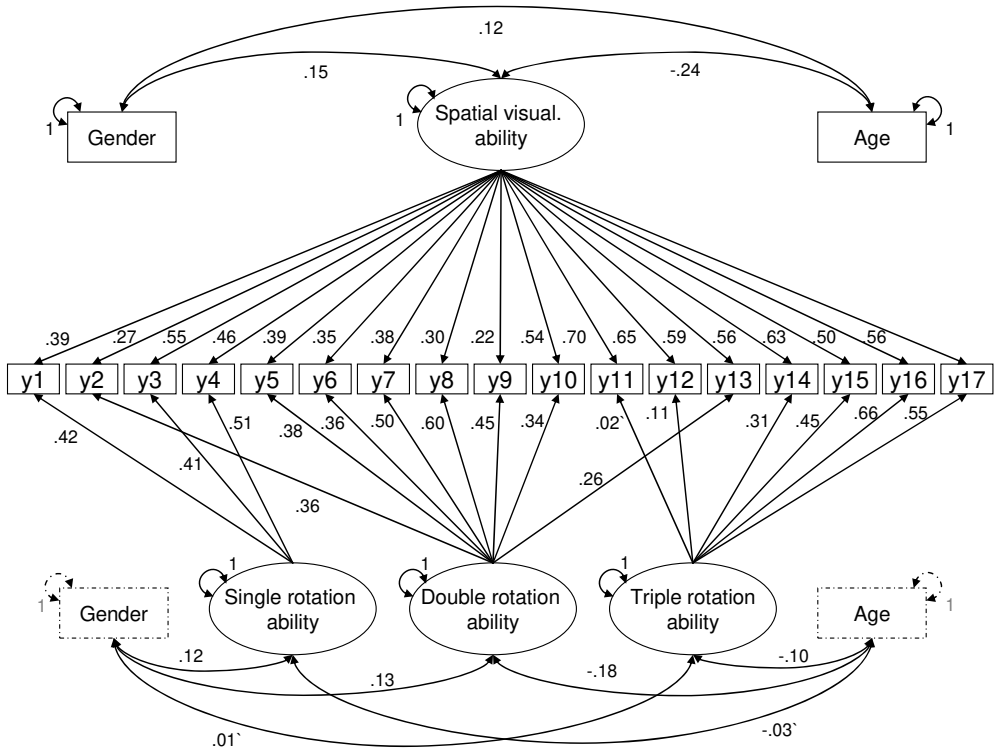


Figure 2 Spatial visualization ability measured by cube rotation problems.

Notes: All figures denote standardized parameter estimates; apostrophes indicate non-significance; for visual clarity, residual variances are not shown, and variables gender and age are pictured twice; N = 1617; model fit: χ^2 = 165.54 with df = 107, p < .05, RMSEA= .018 [90% CI: .013, .023], ECVI = .206 [90% CI: .187, .231].

DISCUSSION

We applied RFA to detect measurement bias with respect to age and gender to two subtests of a Dutch cognitive ability test. We also applied the MGFA method to the cognitive ability data, categorizing age into two age groups and conducting separate

analyses to investigate bias with respect to gender and age. Here, the MGFA and RFA methods yielded very similar results, but the MGFA method does have some disadvantages. In our example, gender and age were correlated (men were older). When we use MGFA to separately investigate bias with respect to gender and age then it might be difficult to distinguish gender bias from age bias. Investigation of gender and age group bias simultaneously in MGFA would involve the comparison of at least four smaller groups (younger women, older women, younger men, older men). Besides complicating the procedure and the interpretation of the results, this also means less precise parameter estimates and loss of statistical power.

Limitations of the RFA method generally come from the measurement bias definition being far more general. For example, in the RFA method T is operationalized as a continuous latent variable, whereas in the definition T can be a discrete latent variable, as in latent class analysis (also incorporated in SEM; Muthén & Muthén, 2006), or T can be an observed variable, as in some of the older bias detection methods such as the Mantel-Haenszel procedure (Holland & Thayer, 1988) and the logistic regression procedure (Swaminathan & Rogers, 1990). Furthermore, in the RFA method only linear conditional independence can be tested, and the method is not readily suited to detect nonuniform bias (although the RFA method can be extended with latent moderated structures; see Barendse, Oort & Garst, 2010). In the MGFA method nonuniform bias can be investigated by testing across group constraints on factor loadings. Still, when we applied the MGFA method to our cognitive ability data we did not find any nonuniform bias.

In the present research we relied on modification indices for model modification, and we tested these at a Bonferroni adjusted level of significance to prevent chance results. Saris, Satorra, & Van der Veld, (2009) suggested to use modification indices in combination with the expected parameter change, and to take the statistical power of the modification index into account as well. This is generally worthwhile, but does not lead to other results in our examples, as the model modifications were already justified substantively and we checked whether the modifications changed the parameter estimates substantially.

In practice it may be difficult to find the true cause of apparent bias, because there may be many possible violators of the measurement model operating simultaneously. Even if all possible violators are known, it will not be possible to operationalize and measure all possible causes of measurement bias. For example, in the worded math problem about office parties we conjecture that the apparent sex and age bias is really caused by the unusual gender role behaviour in the text of the worded problem. As we have no measure of "familiarity with unusual gender role behaviour" available, we can only detect bias with respect to sex and age. Researchers of measurement bias should be aware of this problem, and always try to investigate bias with respect to as many possible violator variables as

available. One of the advantages of the RFA method is that bias can be detected with respect to multiple possible violators simultaneously.

MEASUREMENT BIAS AND MULTIDIMENSIONALITY

The present examples serve to illustrate the relationship between measurement bias and multidimensionality. In both examples we rejected the one-dimensional factor model in favour of a multidimensional factor model. In the first example, if we ignored the speed factor, we found age bias in the last items of the test, which would have been difficult to interpret. In the multidimensional model it is clear that the last items (also) measure speed and that age is correlated with speed. In the second example, the specific rotation factors that vary in their correlations with gender and age could have been mistaken for bias in the associated items. In one of the other Q1000 cognitive ability tests, a 37-item vocabulary test, measurement bias detection yielded multiple items that favoured younger respondents (results not shown here). Inspection of item content showed that these biased items all inquired after the meaning of words with English origin. The biasing factor was therefore taken to be familiarity with English language, which is assumed to be inversely related with age.

In general, the interpretation of apparent measurement bias involves reflection on possible biasing factors. In the one-dimensional model, all items are really affected by two factors: the single common factor and an item-specific residual factor, as in Spearman's (1928) original "two-factor theory". If all residual variance was really only random error variance then measurement bias would be absent by definition. But if the residual variance also contains structural variance then this may stem from a biasing factor. If multiple items in a test are affected by the same biasing factors, these factors may surface as additional common factors, as was the case with speed in the mathematical ability test, the specific rotation factors in the spatial-visual test, and English language familiarity in the vocabulary test. However, if the residual factors do not share any structural variance, then the hypothesis of unidimensionality will not be rejected, although measurement bias may still be present. Oort (1991) used the definition of measurement bias to define unidimensionality as the absence of measurement bias with respect to any variable that might be relevant in whatever context the test is used. Following Lord and Novick's (1968) notion of "complete latent space", we can define k-dimensionality as the number of dimensions of T that is needed to achieve statistical independence of all items X. Modeling all k dimensions guarantees the absence of measurement bias.

With the RFA method, if we operationalize the biasing factor as one of the variables V, we can detect bias with respect to the nuisance factor itself. In the mathematical ability example, we might consider speed to be a biasing factor, and the effects of the speed factor on Items 7 through 12 as measurement bias. Instead of the speed factor as an

additional *T* in a multidimensional measurement model, the speed factor then features as a latent *V* in a model with a unidimensional *T*. This once more shows that multidimensionality and measurement bias really address the same problem. Measurement bias in a unidimensional model may disappear in a multidimensional model. The other way around, misspecification of the dimensionality of *T* in the measurement model may lead to spurious findings of bias.

In conclusion, measurement bias and multidimensionality are related, but not equivalent. Measurement bias implies multidimensionality, but multidimensionality shows up as measurement bias only if multidimensionality is not properly accounted for in the measurement model.

CHAPTER 2

A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data

Abstract We present a test for cluster bias, which can be used to detect violations of measurement invariance across clusters in 2-level data. We show how measurement invariance assumptions across clusters imply measurement invariance across levels in a 2-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. The test is illustrated with an example from school research. In a simulation study, we show that the cluster bias test has sufficient power, and the proportions of false positives are close to the chosen levels of significance.

Based on: 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.

INTRODUCTION

Measurement invariance (or the absence of measurement bias) of a given instrument across groups is a necessary condition for the comparisons of groups with respect to the latent variables that the instrument purports to measure. The importance of measurement invariance is widely recognized (Mellenbergh, 1989; Millsap & Everson, 1991; Meredith, 1993; Vandenberg & Lance, 2000). A method that is often used to investigate measurement invariance is Multigroup Factor Analysis (MGFA), which involves testing the equality of measurement parameters (specifically, factor loadings, intercepts, and residual variances) over groups (e.g., Wicherts, et al., 2004, Smits, et al, 2011). This approach is applicable to both the linear factor model (Reise, Widaman & Pugh, 1993) and the ordinal factor model (Millsap & Yun-Tein, 2004). MGFA becomes unwieldy or even unfeasible if the number of groups is large (Selig, Card and Little, 2008). Measurement invariance issues in a large number of groups are common in cross-cultural research (Byrne & van de Vijver, 2010) and in teacher evaluation studies (Marsh & Hocevar, 1984).

The aim of the present paper is to present a multilevel approach to investigate measurement invariance in a large number of groups. We circumvent the limitations of standard MGFA in this context by treating group membership as random. In view of this, we refer to groups as clusters, and we refer to violations of measurement invariance across clusters as cluster bias. We present a test for cluster bias in the two-level factor model. The test is illustrated with an empirical example from educational research. In addition, using simulated data, we investigate the performance of the test for cluster bias in terms of detection rate (power), false positives (Type 2 error), and estimation bias for different types and sizes of bias.

MEASUREMENT BIAS

Consider a measurement instrument X that was designed to measure a trait T (e.g., intelligence). Measurement bias with respect to a variable V (e.g., gender) implies that systematic differences in test scores on X over the levels of V (boys and girls) are not only attributable to differences in trait T but also to differences in variable V or variables associated with V (e.g., motivation). Measurement invariance (i.e. absence of bias) therefore is a necessary condition for the substantive interpretation of systematic differences between observed test scores. Mellenbergh (1989) defines measurement bias as a violation of measurement invariance. A test X is measurement invariant with respect to V if the following conditional independence holds:

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

with X representing measurements, T representing the trait of interest, and V representing any other variable. Function f_1 is the conditional distribution of X given values t and v, and f_2 is the conditional distribution of X given t. If the conditional independence does not hold (i.e., if $f_1 \neq f_2$), the measurement of T by X is biased with respect to V. Mellenbergh distinguishes between uniform bias and non-uniform bias, depending on whether the distribution of X is uniformly affected by V or not. Figure 1 gives a graphical representation of unbiased measurement, uniform bias and non-uniform bias.

Here we focus on the linear common factor model as the measurement model (Mellenbergh, 1994), and on structural equation modeling (SEM) as the method of measurement bias detection.

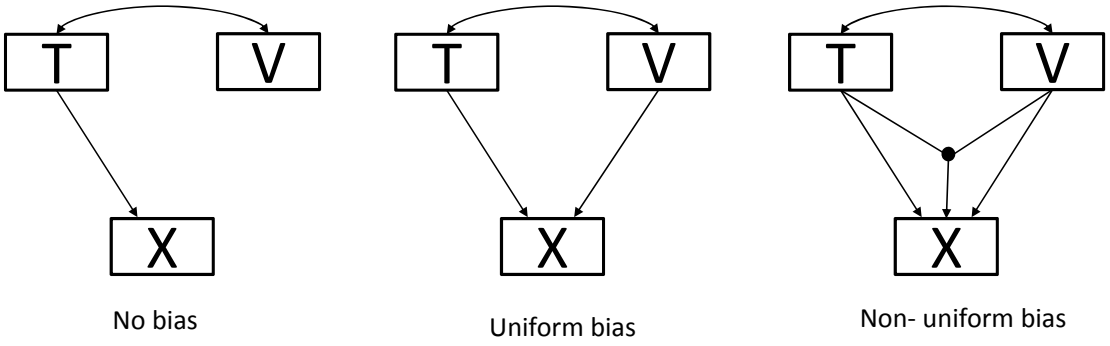


Figure 1. Graphical representation of unbiased measurement, uniform bias, and non-uniform bias.

USING SEM TO DETECT MEASUREMENT BIAS

SEM provides various methods to test for measurement invariance, such as *multiple indicator multiple cause* analysis (MIMIC; Muthén, 1989) and *restricted factor analysis* (RFA; Oort, 1992, 1998). In these methods the trait of interest is operationalized as a latent variable, measured by multiple observed variables. Uniform bias can be investigated by testing the significance of direct effects of exogenous variables (V) on the observed variables (X), and non-uniform bias by testing the significance of the product of the trait of interest and the exogenous variable ($T*V$) on the observed variables (Barendse, Oort & Garst, 2010; Barendse, Oort, Werner, Ligtvoet & Schermelleh-Engel, 2012). The only difference between MIMIC and RFA is that in MIMIC, the exogenous variable V is assumed to have a causal effect on latent variable T, whereas in RFA the two variables are merely correlated.

If V is a nominal variable, indicating group membership, multigroup factor analysis (MGFA; Sörbom, 1974) can also be used to investigate bias, as mentioned above. Meredith (1993) introduced the term *weak measurement invariance* in his operationalization of

measurement invariance within linear MGFA. With MGFA, a series of increasingly restrictive models can be fitted to test different levels of measurement invariance. In the terminology of Meredith (1993; Meredith & Teresi, 2006), we distinguish *configural invariance,* with equal patterns of factor loadings across groups, *weak factorial invariance,* with equal values of factor loadings, *strong factorial invariance,* with equal intercepts in addition to equal values of factor loadings, and *strict factorial invariance*, with equal residual variances in addition to equal factor loadings and intercepts. In MGFA, uniform bias and non-uniform bias are associated with violations of strong and weak factorial invariance, respectively. We do not consider strict factorial invariance here, as strong factorial invariance suffices for meaningful across group comparison of common factor means.

MEASUREMENT BIAS IN MULTILEVEL DATA

In educational and psychological research, data often have a hierarchical multilevel structure, such as data from children in classrooms, employees in teams, or patients from physicians. With multilevel data, the grouping variable typically has many levels, which are called clusters. In contrast to standard MGFA, in which the grouping variable is viewed as a fixed variable, group membership is considered a random variable in multilevel factor analysis. Violations with respect to this random variable, that is, violations of measurement invariance across clusters, will be referred to as *cluster bias*. Multilevel SEM is suited to test for cluster bias, as first suggested by Rabe - Hesketh, Skrondal, and Pickles (2004), and Muthén (1990). We will show the implications of various across cluster constraints, yielding a test of cluster bias. The cluster bias test will be illustrated with an empirical example, and further evaluated through a small scale simulation study.

In the next section, we will show how three increasingly restrictive assumptions across *clusters* (i.e. configural, weak, and strong factorial invariance across clusters) lead to testable invariance hypothesis across *levels* in a two-level factor model.

METHOD

TWO-LEVEL SEM WITH INVARIANCE RESTRICTIONS

Consider the multivariate response vector $\mathbf{y}_{ij}$, containing p test scores of subject i in cluster j. The scores can be decomposed into p cluster means $\boldsymbol{\mu}_j$ and p individual deviations $\boldsymbol{\eta}_{ij}$ from the cluster means:

$$\mathbf{y}_{ij} = \boldsymbol{\mu}_j + \boldsymbol{\eta}_{ij} . \qquad (2)$$

As the individual deviations are independent of the cluster means, the overall variances and covariances of $\mathbf{y}_{ij}$, denoted $\mathbf{\Sigma}$, can also be decomposed into independent parts,

$$\begin{aligned}\mathbf{\Sigma} &= \mathrm{COV}(\mathbf{y}_{ij}, \mathbf{y}_{ij}) \\ &= \mathrm{COV}(\boldsymbol{\mu}_j, \boldsymbol{\mu}_j) + \mathrm{COV}(\boldsymbol{\eta}_{ij}, \boldsymbol{\eta}_{ij}) \\ &= \mathbf{\Sigma}_B + \mathbf{\Sigma}_W ,\end{aligned} \tag{3}$$

where $\mathbf{\Sigma}$, $\mathbf{\Sigma}_B$, and $\mathbf{\Sigma}_W$ are $p \times p$ variance-covariance matrices. $\mathbf{\Sigma}_B$ contains the variances and covariances of the cluster means, and $\mathbf{\Sigma}_W$ contains the pooled within-cluster variances and covariances of the individual deviations from the cluster means. COV( ) denotes covariance, and the B and W subscripts denote *between clusters* and *within clusters*.

CONFIGURAL INVARIANCE ACROSS CLUSTERS

We assume a common factor model for $\mathbf{y}_{ij}$, where q common factors are measured by the p tests,

$$\mathbf{y}_{ij} = \boldsymbol{\tau}_j + \mathbf{\Lambda}_j \boldsymbol{\xi}_{ij} + \boldsymbol{\varepsilon}_{ij}, \tag{4}$$

where vector $\boldsymbol{\xi}_{ij}$ contains the scores on the q common factors of individual i in cluster j, vector $\boldsymbol{\varepsilon}_{ij}$ contains the scores on the p residual factors of individual i in cluster j, vector $\boldsymbol{\tau}_j$ contains p intercepts, and matrix $\mathbf{\Lambda}_j$ is a $p \times q$ matrix containing the factor loadings. The residual factors have zero means, and are mutually independent and independent of the common factors. Intercepts and factor loadings can be considered measurement parameters, characteristic for the tests or measurement instruments. Intercepts $\boldsymbol{\tau}_j$ indicate the attractiveness or (reverse) difficulty of the tests in cluster j, and factor loadings $\mathbf{\Lambda}_j$ indicate how well the tests discriminate between subjects with different common factor values in cluster j.

As in each cluster j, the p residual factors have zero means, the cluster means of the p observed variables are given by

$$\begin{aligned}\boldsymbol{\mu}_j &= \mathrm{E}(\boldsymbol{\tau}_j + \mathbf{\Lambda}_j \boldsymbol{\xi}_{ij} + \boldsymbol{\varepsilon}_{ij}) \\ &= \boldsymbol{\tau}_j + \mathbf{\Lambda}_j \mathrm{E}(\boldsymbol{\xi}_{ij})\end{aligned}$$

$$= \boldsymbol{\tau}_j + \boldsymbol{\Lambda}_j \boldsymbol{\kappa}_j \,, \tag{5}$$

where E( ) denotes the expected value, and vector $\boldsymbol{\kappa}_j$ contains the q common factor means in cluster j. Substitution of Equations 4 and 5 into Equation 2 yields

$$\begin{aligned} \boldsymbol{\eta}_{ij} &= \mathbf{y}_{ij} - \boldsymbol{\mu}_j \\ &= \boldsymbol{\Lambda}_j \boldsymbol{\xi}_{ij} - \boldsymbol{\Lambda}_j \boldsymbol{\kappa}_j + \boldsymbol{\varepsilon}_{ij} \,. \end{aligned} \tag{6}$$

As a result, the between cluster variances and covariances of $\boldsymbol{\mu}_j$ can be expressed as

$$\begin{aligned} \boldsymbol{\Sigma}_B &= \mathrm{COV}(\boldsymbol{\mu}_j, \boldsymbol{\mu}_j) \\ &= \mathrm{COV}(\boldsymbol{\tau}_j + \boldsymbol{\Lambda}_j \boldsymbol{\kappa}_j, \boldsymbol{\tau}_j + \boldsymbol{\Lambda}_j \boldsymbol{\kappa}_j) \\ &= \mathrm{COV}(\boldsymbol{\tau}_j, \boldsymbol{\tau}_j) + \mathrm{COV}(\boldsymbol{\tau}_j, \boldsymbol{\Lambda}_j \boldsymbol{\kappa}_j) + \mathrm{COV}(\boldsymbol{\Lambda}_j \boldsymbol{\kappa}_j, \boldsymbol{\tau}_j) + \mathrm{COV}(\boldsymbol{\Lambda}_j \boldsymbol{\kappa}_j, \boldsymbol{\Lambda}_j \boldsymbol{\kappa}_j) \,, \end{aligned} \tag{7}$$

and the pooled within cluster variances and covariances of $\boldsymbol{\eta}_{ij}$ as

$$\begin{aligned} \boldsymbol{\Sigma}_W &= \mathrm{COV}(\boldsymbol{\eta}_{ij}, \boldsymbol{\eta}_{ij}) \\ &= \mathrm{COV}(\boldsymbol{\Lambda}_j \boldsymbol{\xi}_{ij} - \boldsymbol{\Lambda}_j \boldsymbol{\kappa}_j + \boldsymbol{\varepsilon}_{ij}, \boldsymbol{\Lambda}_j \boldsymbol{\xi}_{ij} - \boldsymbol{\Lambda}_j \boldsymbol{\kappa}_j + \boldsymbol{\varepsilon}_{ij}) \\ &= \mathrm{COV}(\boldsymbol{\Lambda}_j \boldsymbol{\xi}_{ij}, \boldsymbol{\Lambda}_j \boldsymbol{\xi}_{ij}) + \mathrm{COV}(\boldsymbol{\varepsilon}_{ij}, \boldsymbol{\varepsilon}_{ij}) \,. \end{aligned} \tag{8}$$

WEAK FACTORIAL INVARIANCE ACROSS CLUSTERS

Equality of factor loadings $\boldsymbol{\Lambda}_j$ (i.e. discrimination parameters) over clusters, or weak invariance, is one constraint that follows from the definition of measurement invariance. If we introduce this constraint, $\boldsymbol{\Lambda}_j = \boldsymbol{\Lambda}$ for all j, and if we assume that the intercepts $\boldsymbol{\tau}_j$ (i.e. difficulty parameters) are uncorrelated with the common factor means $\boldsymbol{\kappa}_j$ (i.e. means of subject ability parameters), then Equations 7 and 8 simplify to

$$\begin{aligned} \boldsymbol{\Sigma}_B &= \boldsymbol{\Lambda}\, \mathrm{COV}(\boldsymbol{\kappa}_j, \boldsymbol{\kappa}_j)\, \boldsymbol{\Lambda}' + \mathrm{COV}(\boldsymbol{\tau}_j, \boldsymbol{\tau}_j) \\ &= \boldsymbol{\Lambda}\, \boldsymbol{\Phi}_B\, \boldsymbol{\Lambda}' + \boldsymbol{\Theta}_B \,, \end{aligned} \tag{9}$$

and

$$\begin{aligned}\boldsymbol{\Sigma}_W &= \boldsymbol{\Lambda}\,\mathrm{COV}(\boldsymbol{\xi}_{ij}, \boldsymbol{\xi}_{ij})\,\boldsymbol{\Lambda}' + \mathrm{COV}(\boldsymbol{\varepsilon}_{ij}, \boldsymbol{\varepsilon}_{ij}) \\ &= \boldsymbol{\Lambda}\,\boldsymbol{\Phi}_W\,\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_W\,, \end{aligned} \tag{10}$$

Where the $q \times q$ matrix $\boldsymbol{\Phi}_B$ contains the variances and covariances of the common factor cluster means $\boldsymbol{\kappa}_j$, the $p \times p$ matrix $\boldsymbol{\Theta}_B$ contains the variances and covariances of the intercepts $\boldsymbol{\tau}_j$, the $q \times q$ matrix $\boldsymbol{\Phi}_W$ contains the pooled within variances and covariances of the common factors $\boldsymbol{\xi}_{ij}$, and the $p \times p$ matrix $\boldsymbol{\Theta}_W$ contains the pooled within variances of the residual factors $\boldsymbol{\varepsilon}_{ij}$. As the residual factors are assumed to be independent, $\boldsymbol{\Theta}_W$ is a diagonal matrix. Matrix $\boldsymbol{\Theta}_B$, however, is not necessarily diagonal. As the random intercepts in $\boldsymbol{\tau}_j$ may share common variance, matrix $\boldsymbol{\Theta}_B$ may contain some off-diagonal elements.

STRONG FACTORIAL INVARIANCE ACROSS CLUSTERS

Equality of intercepts $\boldsymbol{\tau}_j$ (i.e. difficulty parameters) over clusters, or strong invariance, is another constraint that follows from the definition of measurement invariance. If we add this constraint, $\boldsymbol{\tau}_j = \boldsymbol{\tau}$ for all j, then $\boldsymbol{\Theta}_B = 0$ and Equation 9 further simplifies to

$$\boldsymbol{\Sigma}_B = \boldsymbol{\Lambda}\,\boldsymbol{\Phi}_B\,\boldsymbol{\Lambda}'\,. \tag{11}$$

So, if we assume invariance of $\boldsymbol{\tau}_j$ across clusters, $\boldsymbol{\Theta}_B$ disappears altogether.

TESTING FOR CLUSTER BIAS

If there is no cluster bias, then a two-level factor model given by Equations 10 and 11 should fit the data (to reasonable approximation). Absence of cluster bias in two-level factor analysis is similar to strong invariance in multigroup factor analysis (Meredith, 1993). We refer to the model given by Equations 10 and 11 as the *cluster invariance model*.

If there is uniform cluster bias, then the two-level factor model given by Equations 9 and 10 should fit the data. We refer to this model as the *cluster bias model*. Presence of uniform cluster bias violates strong invariance, while weak invariance may still hold.

If the model given by Equations 9 and 10 does not fit the data, then there is non-uniform cluster bias, assuming that at least the pattern of factor loadings is correct and invariant across clusters. This is similar to configural invariance in multigroup factor models.

When testing for cluster bias, one could start with fitting the cluster invariance model, i.e., a model in which the factor loadings are constrained to be equal across levels, and the

covariance matrix $\boldsymbol{\Theta}_B$ is constrained to be zero, as follows from strong invariance across clusters. To investigate uniform cluster bias, one can use the likelihood ratio test (or chi-square difference test) to compare the fit of the cluster invariance model with the fit of the cluster bias model (with a diagonal $\boldsymbol{\Theta}_B$), and use the likelihood ratio test to test whether $\boldsymbol{\Theta}_B$ equals zero. It is not possible to conduct this test with a symmetric $\boldsymbol{\Theta}_B$ with all elements free to be estimated, as such a model is not identified. Therefore, alternatively to this omnibus test, one can consider the modification indices (Sörbom, 1989) to explore possible non-zero diagonal and off-diagonal elements in $\boldsymbol{\Theta}_B$. Between level residual covariances (i.e. covariances between random intercepts) may originate from a common cause of cluster bias in multiple indicators.

After testing uniform cluster bias, one can investigate non-uniform bias by testing the omnibus hypothesis that the factor loadings are equal across levels, using the likelihood ratio test or by considering modification indices. If the factor loadings are not equal across levels, the common factors do not have the same interpretation across clusters (Muthén, 1990; Rabe-Hesketh, Skrondal & Pickles, 2004).

A cautionary note concerns the scaling of the common factors. With freely estimated factor loadings at both levels, the common factors at both levels can be scaled by fixing their variances a non-zero value (e.g., $\text{diag}(\boldsymbol{\Phi}_W) = \text{diag}(\mathbf{I})$, $\text{diag}(\boldsymbol{\Phi}_B) = \text{diag}(\mathbf{I})$). With the factor loadings constrained to be equal over the levels, and the factor variances of ξ_{ij} at the within level fixed (e.g., $\text{diag}(\boldsymbol{\Phi}_W) = \text{diag}(\mathbf{I})$), the factor variances of κ_j at the between level are identified by the equality constraints on the factor loadings and can be freely estimated. In such a model there is no reason to assume equality of ξ_{ij} and κ_j variances.

Similar to usual measurement invariance testing in single level data, there may be measurement bias with respect to some but not all indicators. This is referred to as partial invariance (Byrne, Shavelson & Muthén, 1989). With respect to cluster bias we will use the term *partial cluster invariance*. The tests for cluster bias will be illustrated in the next section.

ILLUSTRATIVE EXAMPLE

DATA

We illustrate the test for cluster bias with data from Thoonen, Sleegers, Peetsma, and Oort (2010). Participants in this study are 2814 students from 121 school classes, from fourth to sixth grade (ten through twelve years old). Students' attitude to mathematics was measured with five items, such as "I have no trouble focusing my attention during mathematics", with four response options.

PROCEDURE

First, we verify the necessity of multilevel analysis of the nested data structure. The intra class correlation (ICC) reflects the proportion of a single variable's variance that can be accounted for by the between level. The statistical significance of the collective between level variance of all observed variables can be tested by fitting a null-model ($\boldsymbol{\Sigma}_B = 0$) to the between level covariance matrix, while specifying a saturated model for $\boldsymbol{\Sigma}_W$. If the null model does not fit we conclude that there is significant between level variance. The statistical significance of the covariances can be tested by fitting the independence model (with diagonal $\boldsymbol{\Sigma}_B$) to the between level covariance matrix, while specifying a saturated model for $\boldsymbol{\Sigma}_W$. If the independence model does not fit, we conclude that there is significant between level covariance.

Second, we establish a measurement model for $\boldsymbol{\Sigma}_W$**,** while specifying a saturated model for $\boldsymbol{\Sigma}_B$, and third, we test for cluster bias by imposing the cluster invariance model of Equations 10 and 11 for $\boldsymbol{\Sigma}_W$ and $\boldsymbol{\Sigma}_B$, and compare its fit with alternative models allowing for cluster bias.

STATISTICAL ANALYSIS

We refer to the student level as the 'within level', and to the classroom level as the 'between level'. As the item responses are scored on a four-point scale, we should evaluate the maximum likelihood of the ordinal item responses, which however is often not feasible because of the large computational demands (Grilli & Rampichini, 2007). In our example, Mplus (Muthén & Muthén, 2007) indeed did not converge to a solution. As our example only serves to illustrate the cluster bias test, we treat the responses to the four-point scale as approximately continuous. We used robust maximum likelihood estimation (MLR) in Mplus to obtain parameter estimates. This estimation method provides a test statistic that is asymptotically equivalent to the Yuan-Bentler test statistic (T_2, Yuan & Bentler, 2000), and standard error estimates that are robust for non-normality.

In addition to this test of exact fit, we calculate the root mean square of approximation (RMSEA) as an index of approximate fit. RMSEA values smaller than .05 indicate close fit, and values smaller than .08 are still considered satisfactory (Browne & Cudeck, 1992). We also calculate level specific RMSEA's as described by Ryu and West (2009). For example, the model fit at the within level, $RMSEA_W$, is given by

$$RMSEA_W = \sqrt{\frac{\chi^2_W - df_W}{df_W(M)}}, \tag{12}$$

where χ^2_W and df_W are the chi-square test statistic and degrees of freedom obtained from fitting a model with saturated between models, and M is the total sample size. The $RMSEA_B$ for model fit at the between level is calculated in a similar way,

$$RMSEA_B = \sqrt{\frac{\chi^2_B - df_B}{df_B(N)}}\ , \tag{13}$$

with the chi-square test statistic and degrees of freedom from a model with saturated within part, and using the number of clusters, N, instead of the total sample size.

When establishing a measurement model for $\boldsymbol{\Sigma}_W$, we want to avoid correlated residual factors, because when we subsequently test for cluster bias, we want to model the same dimensionality for $\boldsymbol{\Sigma}_W$ and $\boldsymbol{\Sigma}_B$. With $\boldsymbol{\Theta}_B$ equal to zero, additional dimensions cannot be modeled trough $\boldsymbol{\Theta}_W$ and $\boldsymbol{\Theta}_B$. The solution is to reparameterize the model by adding (uncorrelated) factors, one for each residual covariance. For such additional factors, fixing both factor loadings at unity (or any other non-zero value) and estimating the (possibly negative) factor variance, leads to a model that is statistically equivalent to a model with residual covariances.

When testing for cluster bias we use modification indices (MI's) to guide model specification. To guard against chance results, we test MI's at a Bonferroni corrected level of significance. That is, we use as a critical value the chi-square that is associated with a two-sided level of significance (α) of 0.05 divided by the number of possible modifications under consideration. When testing the significance of residual variances we choose a one-sided level of significance, to account for the problem of bounded parameter space explained by Stoel, Garre, Dolan and Van den Wittenboer (2006).

RESULTS

The intra class correlations of the five items are respectively .043, .037, .033, .028 and .020. A null model (no variances and covariances) at the between level does not fit the data (χ^2 = 153.65, df = 15, p < .05, RMSEA = .057, $RMSEA_B$ = .276), indicating that there is significant between level variance. The independence model (no covariances at the between level, so $\boldsymbol{\Sigma}_B$ is diagonal) does not fit the data either (χ^2 = 43.51, df = 10, p < .05, RMSEA = .035, $RMSEA_B$ = .166), indicating that there are significant covariances at the between level. Therefore, these data require multilevel modeling.

A MEASUREMENT MODEL AT THE STUDENT LEVEL

A one factor model for $\mathbf{\Sigma}_W$ does not fit the data (χ^2 = 85.14, *df* = 5, $p < .05$, RMSEA = .075, $RMSEA_W$ = .075). MI's indicate a possible covariance between the residuals of Items 4 and 5, probably because both items concern the ability to concentrate on math. Adding the residual covariance results in a model with satisfactory fit ($\chi 2$ = 19.04, *df* = 4, $p < .05$, RMSEA = .037, $RMSEA_W$ = .037). Another significant MI indicates a large residual correlation between Item 2 and Item 3, which may be due to the fact that both items focus on "working hard during math". Adding this residual covariance results in excellent model fit ($\chi 2$ = 0.35, *df* = 3, p = .95, RMSEA = 0). We reparameterize the model by adding two uncorrelated factors, one for Item 4 and 5, and one for Item 2 and 3. For each of these factors, we fix both factor loadings at 1, so the model is equivalent with the model containing the correlated residuals.

TESTING FOR CLUSTER BIAS

The model with the between level and within level factor loadings constrained to be equal, and no residual variances at the between level, fits the data reasonably well ($\chi 2$ = 45.56, *df* = 15, $p < .05$, RMSEA = .027). MI's show that the zero residual variance of Item 3 causes misfit, indicating cluster bias in Item 3. Freeing the residual variance for this item results in better model fit ($\chi 2$ = 24.56, *df* = 14, $p < .05$, RMSEA = .017). Another significant MI reveals a non-zero residual variance for Item 1. Freeing the residual variance results in excellent model fit ($\chi 2$ = 12.89, *df* = 13, p = .46, RMSEA = 0). Allowing a covariance between the residual factors at the between level does not improve model fit. All factor loadings can be considered equal across levels. The final model with free residual variance at the between level for Item 1 and 3 is depicted in Figure 2. For Item 1, 1.6% of the total variance is explained by cluster bias. For Item 3, this percentage is 5.5 %. These percentages are calculated by dividing the residual variance at the between level by the total variance of the item.

CONCLUSION

The presence of cluster bias in Items 1 and 3 implies that, given the same attitude to mathematics, school classes still have different mean scores on Item 1 and Item 3. Item 1 is about immediately starting to work on mathematics after the assignment is given. It is possible that this 'immediately starting' aspect does not only depend on pupil's attitude to mathematics, but also on other between level characteristics such as the teaching style of the teacher. For Item 3, an item about trying one's best at math, an explanation could be that the item scores do not just depend on pupil's attitude but also on between level variables such as classroom climate or the teacher's enjoyment in teaching mathematics.

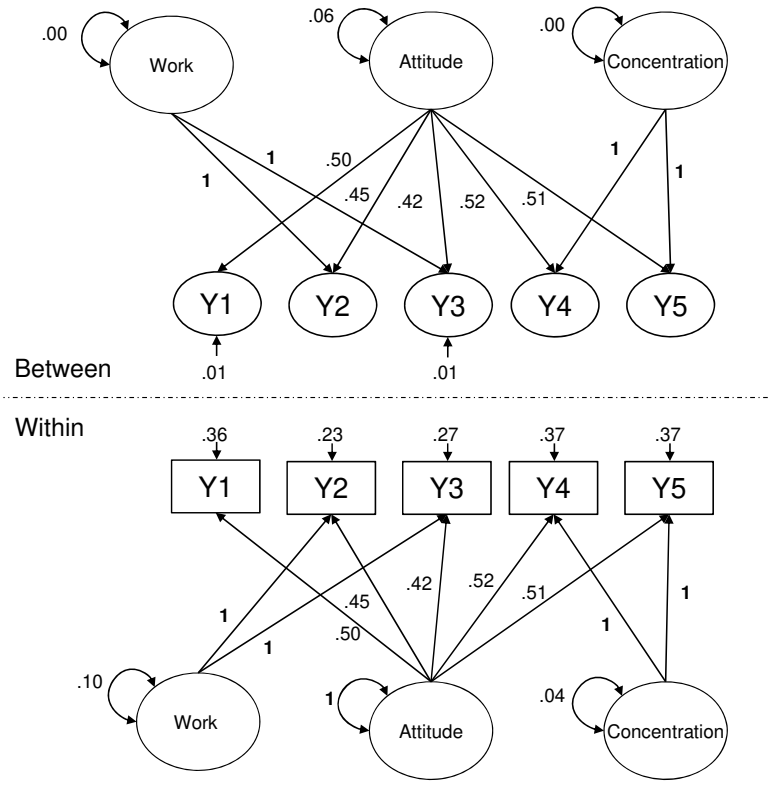


Figure 2. Partial cluster invariance model with parameter estimates. **Note:** All parameter estimates are significantly larger than zero (with $\alpha = 0.05$), with exception of the variances of the 'Concentration' and 'Work' factors at the between level.

SIMULATION STUDY

A simulation study can give an idea of the power of the cluster bias test. The data are generated according to the model as depicted in Figure 3. This model comprises a one-factor model with five indicators and one between-level exogenous variable (V). The exogenous variable V is used to introduce bias in the first indicator variable. We introduce uniform bias by regressing the first indicator variable on V, and we introduce non-uniform bias, by regressing the first indicator variable on the product of the common factor and V. For unbiased indicators, 50% of the total variance is residual variance, and 10% of the total variance is between level variance (ICC = 0.10). The population parameter values are given in Figure 3.

We generate multivariate normal data in two steps, using the computer program R (R Development Core Team, 2010). First, cluster means are generated according to the following equation:

$$\mu_{hj} = \tau_h + \lambda_h \varkappa_j + b_h v_j + c_h (\varkappa_j v_j) , \quad (14)$$

where μ_{hj} is the cluster mean of observed indicator h in cluster j, $\varkappa_j$ is the cluster mean of the common factor, v_j is the cluster mean of V, τ_i is the intercept for indicator h, λ_h is the factor loading of indicator h, and b and c are regression coefficients. The cluster scores $\varkappa_j$ and v_j are drawn from the bivariate standard normal distribution, with zero means, unity variances, and zero covariance. As shown in Figure 3, indicator variables 2 through 5 are unbiased.

In the next step, we draw data from the multivariate normal distribution with means corresponding to the associated cluster means from the previous step, and covariance matrix $\mathbf{\Sigma}_W$ which is calculated as $\mathbf{\Sigma}_W = \mathbf{\Lambda}\ \mathbf{\Phi}_W\ \mathbf{\Lambda'} + \mathbf{\Theta}_W$, with the parameter values given by Figure 3.

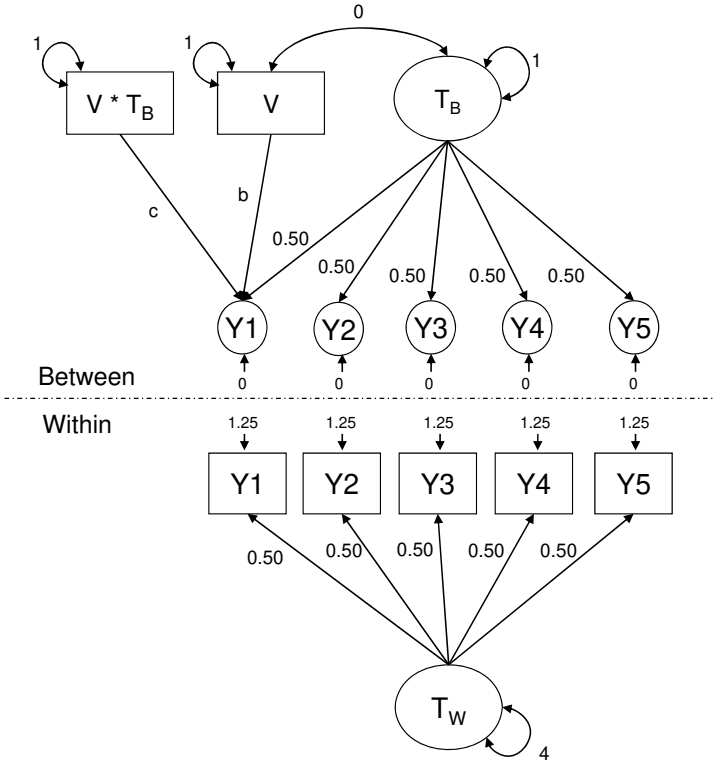


Figure 3. Two-level measurement model with population parameter values. **Note:** In conditions with 0, 1, 3 and 5 % bias, the corresponding values for *b* and *c* were 0, 0.159, 0.278 and 0.363 respectively.

CONDITIONS

We vary the size of uniform bias, the size of non-uniform bias, and the number of clusters. We choose the numbers of clusters to be 20, 50 or 100. The size of 100 is sometimes mentioned as the minimum number of clusters that ensures that the chi-square statistic

(obtained using MLR estimation) follows its expected asymptotic distribution to reasonable approximation (Hox, Maas & Brinkhuis, 2010). We also consider 50 and 20 clusters, as in practice the numbers of clusters are often (much) smaller than 100.

We vary the size of bias by choosing different values for *b* and *c* (see Figure 3). In the uniform bias conditions, we introduce bias by choosing a non-zero value of *b*. In the non-uniform bias conditions, we introduce bias by choosing a non-zero value of *c*. Values for *b* and *c* are chosen 0, 0.159, 0.278 and 0.363, which correspond to 0, 1, 3 and 5% of the total variance in the first indicator variable if only one type of bias is present. For example, using the parameter values from Figure 3, a *b* value of 0.159 yields a proportion of explained variance of $0.159^2 / (0.159^2 + 0.50^2 + 0.50^2 \times 4 + 1.25) = 0.01$ (1 %). In the conditions with both uniform and non-uniform bias we retain the values for *b* and *c* mentioned above. Non-zero values for *b* and *c* cause additional variance, so that biased indicators have larger variance and larger ICC's than unbiased indicators. Hereinafter we refer to bias percentages of 1%, 3%, and 5% as small, medium, and large bias, respectively.

We do not vary the intra class correlation. For the unbiased indicator variables, we set the ICC value at 0.10, which value we consider typical for school data. For biased indicators, the ICC's then vary from 0.10 to 0.19 (depending on the size of the biases). Snijders and Bosker (1999) qualify 0.05 to 0.20 as common in educational data. Hox (2002) notes that ICC's of 0.10 to 0.15 are often found. Preacher, Zyphur and Zhang (2010) consider ICC's of 0.10 to represent a medium sized effect.

For simplicity, we only consider balanced designs, in which the group sizes of all clusters are equal. We choose group sizes of 25 because this is the typical size of a school class in the Netherlands. Total sample sizes are 2500 observations in conditions with 100 clusters, 1250 observations in conditions with 50 clusters, and 500 observations in conditions with 20 clusters.

Varying the sizes of uniform and non-uniform bias and the numbers of clusters results in $4 \times 4 \times 3 = 48$ conditions. For each condition, we generate 500 datasets. To each data set, we fit the measurement invariance model, a partial invariance model that allows estimation of the between level residual variance of the first (biased) indicator variable, and a partial invariance model that allows estimation of the between level residual variance of the second (unbiased) variable.

RESULTS

Overall fit of the cluster invariance model

We evaluate the rejection rate of the cluster invariance model by the chi-square test of overall goodness-of-fit, at the 0.05 level of significance. Table 1 gives the results for the cluster invariance model in the 48 conditions.

The expected chi-square in the no bias condition is equal to the degrees of freedom (*df* = 19). The expected standard deviation of the central chi-square distribution with 19 degrees of freedom is $\sqrt{(19 \times 2)} = 6.16$. In our simulations, the means of the chi-squares in the no bias conditions are 24.78, 21.02, and 20.07 for N = 20, N = 50, and N = 100, respectively. The standard deviations are 11.59, 9.11, and 8.51for N = 20, N = 50, and N = 100, respectively. The rejection rates were 0.27, 0.17 and 0.13, which is substantially higher than the expected 0.05.

The mean chi-square and rejection rate increases with size of the bias and with the number of clusters. In the conditions with 20 clusters, rejection rates vary from .30 in the small uniform bias condition to 0.94 in the conditions with combinations of medium and large bias. In the conditions with 50 clusters, rejection rates vary from 0.32 in the small non-uniform bias condition to 1.00 in the conditions with combinations of medium and large bias. In the conditions with 100 clusters, the rejection rate is around .45 or .41 in conditions with small uniform or small non-uniform bias. With medium uniform or medium non-uniform bias the rejection rates are 0.99 and 0.97, respectively. One deviation from the general pattern of results is that the mean chi-square and rejection rate for the small non-uniform bias condition is higher with N = 20 than with N = 50.

Overall fit of the partial cluster invariance model

To all data sets, we fit the model with between level residual variance for the biased indicator free to be estimated. The fit results and rejection rates are given in Table 2. The expected chi-square is equal to the degrees of freedom (*df* = 18). The expected standard deviation of the chi-square distribution with 18 degrees of freedom is $\sqrt{(18 \times 2)} = 6.00$. The chi-square values, standard deviations and rejection rates decrease with numbers of clusters, but do not vary with size of the bias. In the N = 20 conditions, the mean varies between 23.39 and 28.78. In the N = 50 conditions, the mean chi-square varies between 19.65 and 21.67. In the N = 100 conditions, the mean chi-square varies between 18.63 and 19.97.

Table 1. Cluster invariance model (df = 19), 500 replications: Mean χ^2, standard deviation (SD) and rejection rate (*p*) at $\alpha = 0.05$ (two-sided).

Size of bias in first indicator		20 clusters		50 clusters		100 clusters	
Uniform	*Non-uniform*	*Mean χ^2 (SD)*	*p*	*Mean χ^2 (SD)*	*p*	*Mean χ^2 (SD)*	*p*
zero	*zero*	24.78 (11.59)	.27	21.02 (9.11)	.17	20.07 (8.51)	.13
zero	*small*	28.53 (14.89)	.36	27.59 (12.83)	.32	29.87 (12.88)	.41
zero	*medium*	53.25 (81.77)	.68	57.24 (37.70)	.83	81.90 (37.68)	.97
zero	*large*	93.83 (191.97)	.81	113.25 (106.80)	.97	173.90 (94.33)	1.00
small	*zero*	27.22 (12.33)	.30	26.99 (11.39)	.33	30.26 (12.24)	.45
small	*small*	34.41 (21.35)	.48	37.55 (18.05)	.59	51.43 (22.24)	.86
small	*medium*	64.03 (118.17)	.73	80.22 (52.61)	.93	120.51 (56.92)	1.00
small	*large*	117.75 (189.85)	.86	160.70 (372.97)	.98	236.97 (137.78)	1.00
medium	*zero*	44.66 (27.71)	.68	55.84 (22.53)	.89	83.63 (29.69)	.99
medium	*small*	60.36 (57.36)	.76	78.42 (40.73)	.94	122.00 (50.87)	1.00
medium	*medium*	98.41 (141.95)	.86	137.23 (126.98)	.99	206.40 (95.27)	1.00
medium	*large*	198.65 (753.96)	.93	268.49 (658.89)	1.00	373.04 (246.74)	1.00
large	*zero*	72.93 (53.42)	.89	108.41 (49.49)	1.00	176.19 (55.55)	1.00
large	*small*	95.82 (102.37)	.90	139.41 (90.41)	.99	224.05 (93.13)	1.00
large	*medium*	177.64 (447.527)	.94	238.23 (297.33)	1.00	353.43 (191.64)	1.00
large	*large*	335.45 (1964.43)	.94	433.97 (1016.59)	1.00	627.20 (727.86)	1.00

Table 2. Partial cluster invariance model (df = 18), 500 replications: Mean χ^2, standard deviation (SD) and rejection rates (*p*) at $\alpha = 0.05$ (two-sided).

Size of bias in first indicator		*20 clusters*		*50 clusters*		*100 clusters*	
Uniform	*Non-uniform*	*Mean χ^2 (SD)*	*p*	*Mean χ^2 (SD)*	*p*	*Mean χ^2 (SD)*	*p*
zero	*zero*	28.78 (22.74)	.38	21.67 (11.55)	.21	19.61 (7.94)	.14
zero	*small*	25.88 (18.33)	.31	21.32 (9.75)	.18	18.67 (7.69)	.10
zero	*medium*	25.73 (14.00)	.31	20.49 (8.44)	.15	19.63 (8.09)	.11
zero	*large*	24.96 (12.10)	.30	21.13 (9.08)	.19	19.97 (8.22)	.13
small	*zero*	24.33 (11.21)	.27	21.07 (8.56)	.17	19.03 (7.84)	.10
small	*small*	23.69 (11.06)	.24	20.79 (8.48)	.15	19.72 (8.47)	.14
small	*medium*	25.02 (11.42)	.31	21.24 (8.34)	.16	19.76 (7.88)	.12
small	*large*	25.33 (12.42)	.28	20.89 (8.81)	.16	19.95 (8.43)	.14
medium	*zero*	23.39 (10.05)	.25	20.48 (8.25)	.16	18.63 (6.78)	.07
medium	*small*	24.02 (10.95)	.27	20.45 (8.38)	.14	18.79 (8.43)	.09
medium	*medium*	23.80 (10.70)	.26	21.06 (8.06)	.17	19.06 (7.32)	.10
medium	*large*	25.79 (11.73)	.32	20.77 (7.97)	.14	18.99 (6.90)	.08
large	*zero*	24.29 (11.00)	.27	19.65 (7.70)	.12	19.55 (8.32)	.11
large	*small*	24.71 (10.96)	.31	20.64 (7.78)	.16	19.11 (7.58)	.11
large	*medium*	24.48 (10.77)	.28	20.47 (7.87)	.13	19.59 (7.16)	.10
large	*large*	24.50 (11.03)	.29	20.64 (8.11)	.14	19.61 (7.94)	.14

Power of the chi-square difference test

For each of the 24000 datasets, the difference in -2 × log likelihood values of the invariance model and the partial invariance model is calculated. This results in a chi- square difference test with one degree of freedom. With MLR estimation, the differences in -2 log likelihood values do not follow the chi-square distribution. However, the Mplus program provides scaling correction factors, which can be used for correct difference testing. The scaled chi-square differences sometimes produce a negative value, which is an invalid result (this is a well-known problem; see Satorra & Bentler (2010) for a possible solution). In our analysis, the negative differences are consistently associated with small estimates of the residual variance at the between level. In calculating proportions of significant chi-square tests, we therefore consider the negative values for scaled chi-square differences as indications of non-significant differences. We test the chi-square differences against critical values of 2.71, 3.84, and 6.63, associated with two-sided alphas of 0.10, 0.05, and 0.01. Proportions of significant chi-square differences are reported in Table 3.

The power of the chi-square difference test increases with the size of the bias and with the number of clusters. Here we discuss the results for alpha = 0.10, for the other results we refer to Table 3. In the N = 20 condition, power is low, 0.22 and 0.21 for small uniform bias and small non-uniform bias. For medium bias, power increases to 0.74 and 0.64. Acceptable power of over 0.80 is achieved when uniform bias is large. In the N = 50 conditions the power to detect small uniform or non-uniform bias is 0.45. Power increases to 0.98 and 0.95 for medium uniform or non-uniform bias, and to 0.80 for the combination of small uniform and small non-uniform bias. For larger amounts of bias almost all chi-square differences are significant. In the N = 100 conditions, the power to detect small uniform or non-uniform bias is 0.75 and 0.67 respectively. Power is already 1.00 with medium uniform or non-uniform bias.

False positives of the chi-square difference test

Proportions of false positives (Type 2 error) are calculated by testing the significance of the between level residual variance of the second indicator, which is unbiased. These results are reported in Table 4. Overall, the proportion of false positives increases with sample size and with size of the bias. In the conditions without bias, all proportions of false positives are lower than the chosen alpha levels. In the N = 20 conditions, the proportion of false positives never exceeds the alpha level. In the N = 50 conditions they are generally equal to the alpha level. In most of the N = 100 conditions, the proportions are higher than the nominal alpha levels.

Table 3. Proportions of bias detection in (biased) Indicator 1, using the chi-square difference test at α = 0.01, 0.05, and 0.10.

Size of bias in first indicator		*20 clusters*			*50 clusters*			*100 clusters*		
Uniform bias	*Non-uniform bias*	α			α			α		
		.01	.05	.10	.01	.05	.10	.01	.05	.10
zero	*zero**	.00	.00	.02	.00	.01	.02	.00	.02	.03
zero	*small*	.07	.15	.21	.21	.37	.45	.39	.59	.67
zero	*medium*	.41	.57	.64	.84	.94	.95	.99	.99	1.00
zero	*large*	.67	.80	.83	.97	.99	.99	1.00	1.00	1.00
small	*zero*	.06	.14	.22	.17	.34	.45	.44	.65	.75
small	*small*	.25	.38	.47	.58	.74	.80	.90	.96	.97
small	*medium*	.52	.68	.75	.92	.96	.98	1.00	1.00	1.00
small	*large*	.75	.82	.87	.99	1.00	1.00	1.00	1.00	1.00
medium	*zero*	.48	.65	.74	.91	.97	.98	1.00	1.00	1.00
medium	*small*	.63	.76	.82	.96	.98	.99	1.00	1.00	1.00
medium	*medium*	.78	.87	.91	1.00	1.00	1.00	1.00	1.00	1.00
medium	*large*	.85	.93	.95	1.00	1.00	1.00	1.00	1.00	1.00
large	*zero*	.80	.92	.94	1.00	1.00	1.00	1.00	1.00	1.00
large	*small*	.82	.90	.94	1.00	1.00	1.00	1.00	1.00	1.00
large	*medium*	.90	.95	.96	1.00	1.00	1.00	1.00	1.00	1.00
large	*large*	.92	.95	.96	1.00	1.00	1.00	1.00	1.00	1.00

*In the no bias condition, results are the proportion of false positives.

Table 4. False positives: Proportions of bias detection in (unbiased) Indicator 2, using the chi-square difference test at α = 0.01, 0.05, and 0.10.

Size of bias in first indicator		*20 clusters*			*50 clusters*			*100 clusters*		
Uniform bias	*Non-uniform bias*	α .01	.05	.10	α .01	.05	.10	α .01	.05	.10
zero	*zero*	.00	.01	.04	.01	.02	.04	.00	.02	.04
zero	*small*	.00	.02	.04	.01	.01	.05	.01	.02	.06
zero	*medium*	.01	.02	.04	.01	.03	.05	.01	.05	.08
zero	*large*	.01	.02	.05	.02	.06	.11	.03	.09	.15
small	*zero*	.00	.01	.03	.00	.03	.05	.00	.03	.05
small	*small*	.01	.03	.04	.00	.03	.05	.01	.05	.08
small	*medium*	.00	.03	.06	.01	.04	.07	.01	.07	.12
small	*large*	.01	.04	.07	.03	.06	.11	.04	.11	.17
medium	*zero*	.00	.02	.05	.01	.04	.06	.01	.05	.09
medium	*small*	.00	.01	.05	.02	.05	.08	.02	.06	.12
medium	*medium*	.01	.05	.07	.01	.06	.09	.03	.11	.17
medium	*large*	.01	.05	.09	.02	.08	.15	.06	.17	.24
large	*zero*	.01	.02	.04	.02	.06	.09	.02	.07	.15
large	*small*	.00	.05	.08	.02	.08	.12	.04	.10	.17
large	*medium*	.01	.03	.07	.02	.10	.15	.06	.15	.21
large	*large*	.01	.05	.10	.03	.10	.17	.07	.18	.27

Estimation bias

We also examine the accuracy of parameter estimation in cluster invariance model. The percentage of estimation bias is calculated as 100 × (mean estimated value – population value) / population value. See Figure 3 for the population values. According to Muthén, Kaplan and Hollis (1987), estimation bias less than 10% can be considered negligible.

With the cluster invariance model, all percentages of estimation bias are smaller than 10%. The largest estimation bias is found in the N = 100 condition, for the common factor variance on the between level, which is – 4.5 %.

With the partial cluster invariance model, the residual between level variance for the biased item is consistently underestimated, but never more than 10 %. The population value for this residual variance in the uniform bias conditions is calculated as the square of *b*. The highest bias percentages are – 8.3 % in the N = 20 conditions, – 7.3 % in the N = 50 conditions, and – 2.2 % in the N = 100 conditions. For all other parameters in the partial cluster bias model, sizes of estimation bias are identical with the results from the cluster invariance model.

DISCUSSION

We have presented a test of measurement invariance across cluster in the two-level common factor model. The simulations show that the chi-square difference test and the overall model fit test both have sufficient power to detect cluster bias, given a large enough number of clusters. With 50 clusters, the power to detect cluster bias is sufficient if the bias accounts for 3% or more of the total variance of the indicator. With only 20 clusters, power to detect cluster bias is still sufficient if bias accounts for at least 5% of the total variance. The proportions of false positives are higher than the nominal level of significance in conditions with 100 clusters, but lower in conditions with 20 clusters.

The present approach to detect cluster bias is a viable way to investigate measurement invariance across clusters. In our illustrative example, we have investigated measurement invariance across school classes. This is relevant, for example, when teachers are evaluated based on the performance of their classes. The investigation of cluster bias may be of importance in other fields as well, for example in cross-cultural research or organizational research, were multilevel structures are very common.

We have considered data structures with two levels. However, the model can be extended to three-level structures, such as with data from children within school classes within schools. The same invariance constraints will apply to higher levels.

A requirement for the models presented in this paper is that the highest level units should be independent. These models can therefore not be used with classified nesting structures, where units at the same level of the hierarchy are classified by more than one factor (Rasbash and Goldstein, 1994).

If cluster bias is detected, the next step would be to understand the cause of the bias. In our illustration we hypothesize that the detected bias might be due to teaching style of the teacher. Unfortunately, we do not have an actual measure of teaching style to test this hypothesis. If teaching style is indeed the cause of the cluster bias, regression of the intercept on teaching style should explain the bias, and should render the residual variance at the between level equal to zero.

As illustrated, cluster bias is caused by a between level variable. The bias implies that something else than the construct we intend to measure is causing cluster differences in the observed scores. So cluster bias, as modeled here, is actually bias with respect to a between level variable. Even if we do not have actual measures of substantive between level variables, the present model still allows us to investigate bias with respect to such variables. A similar point concerning bias with respect to unmeasured variables in the context of single level MGFA is made by Lubke, Dolan, Kelderman & Mellenbergh (2003).

Absence of cluster bias suggests that there are no between level variables causing cluster bias. Fitting the cluster invariance model thus serves as a quick and easy first step before the investigation of bias with respect to specific between level variables. Suppose we are interested in the effect of teacher sex in the measurement of children's interest in mathematics. If there is no cluster bias, we know that all group differences are explained by differences in average interest in mathematics, and there is no variance to be explained by teacher sex. Of course, the power to detect bias may be greater when testing cluster bias with respect to specific measured between level variables, compared to the power of the present approach.

CONCLUSION

We presented a framework for testing cluster bias. Violations of measurement invariance across clusters can readily be tested. The cluster bias test has sufficient power to detect cluster bias, and its specification with any multilevel structural equation modeling software is straightforward.

CHAPTER 3

Using two-level ordinal factor analysis to test for cluster bias in ordinal data

Abstract The test for cluster bias is a test of measurement invariance across clusters in two-level data. The present paper examines the true positive rates (empirical power) and false positive rates of the test for cluster bias using the Likelihood Ratio Test (LRT) and the Wald test with ordinal data. A simulation study indicates that the scaled version of the LRT, that accounts for non-normality of the data, gives untrustworthy results, while the unscaled LRT and the Wald test perform well in terms of empirical power rate if the amount of cluster bias is large, and have acceptable false positive rates. The test for cluster bias is illustrated with data from research on teacher – student relations.

Based on: Jak, S., Oort, F.J. & Dolan, C.V. (under review). Using two-level ordinal factor analysis to test for cluster bias in ordinal data.

INTRODUCTION

If psychometric data have a two-level structure, as is the case with data from students in school classes, it is important to ensure that an instrument measures the same construct(s) across students in different clusters. In the case of cluster bias, differences in test scores between students from different clusters cannot be attributed exclusively to differences in the construct(s) measured at the student level. For example, in the case of students' test scores on a motivation questionnaire, differences between students from different classes can be fully explained by differences in motivation if cluster bias is absent. In the presence of cluster bias however, variables other than motivation appear to contribute to differences in students' scores.

Cluster bias is a special case of measurement bias, which can be defined as a violation of measurement invariance. Measurement invariance holds if all measurement parameters are equal across different groups (Mellenbergh, 1989). In the present study, the factor model is the measurement model of interest (Mellenbergh, 1994). In this case, the notion of measurement invariance is denoted *factorial invariance* (Meredith, 1993). The measurement parameters in the factor model are factor loadings (regression coefficients relating the indicator to the common factor), intercepts (the means of the residual factors) and residual variances (variance in the indicators that is not explained by the common factor(s)). Measurement invariance with respect to some grouping variable can be tested using multigroup factor models with a mean structure (Sörbom, 1974). In the terminology of Meredith, we distinguish the following forms of invariance: *Configural invariance,* comprising equal patterns of factor loadings across groups, *weak factorial invariance,* comprising equal values of factor loadings, *strong factorial invariance,* comprising equal intercepts in addition to equal values of factor loadings, and *strict factorial invariance*, comprising equal residual variances in addition to equal factor loadings and intercepts (Meredith & Teresi, 2006).

To test measurement invariance across clusters in multilevel data, the test for cluster bias can be used (Jak, Oort & Dolan, 2013). If one considers the clustering variable a fixed variable, multigroup factor analysis is an obvious choice to investigate measurement bias. When the clusters are viewed as a sample from a population of clusters, random effects modeling is suitable. With large numbers of groups, the random effects approach of multilevel structural equation modeling offers clear advantages. One advantage is that the model fitting procedure is simpler than it is in the case of a multigroup model with a large number of groups. A second advantage is that with multilevel structural equation modeling, the possible causes of clusterbias can be investigated by regressing the parameters representing the bias on potential causes, if these have been measured. Statistical methods to investigate measurement bias across clusters in continuous data have been developed (Muthén, 1990; Rabe-Hesketh, Skrondal & Pickles, 2004) and have been found to perform well with continuous item responses (Jak, Oort, & Dolan, 2013). As in educational and

psychological testing, item responses are often ordinal, e.g., 5-point Likert scales in attitude measures or binary, correct/incorrect, responses in mathematical tests, it is important to establish that this method works well with such data as well. The purpose of the present paper is therefore to extend the test for cluster bias to ordinal data, using the multilevel factor model for ordinal data (Grilli & Rampichini, 2007).

TESTING FOR CLUSTER BIAS IN THE ORDINAL TWO-LEVEL FACTOR MODEL

Ordinal two-level factor models can be used to investigate cluster bias in ordinal data (Grilli & Rampichini, 2007). With p observed variables or items, the p-dimensional vector of observed discrete item responses $\mathbf{y}_{ij}$ of individual i in cluster j can be viewed as originating from a p -dimensional vector of underlying (unobserved) continuous response variables $\mathbf{y}_{ij}^*$. It is assumed that for each variable y_{pij} with a number of C_p categories, a set of C_p -1 threshold parameters exists, such that y_{pij} takes on values {1, 2, …, C_p } if a certain threshold on the underlying variable y_{pij}^* is passed (see Lord & Novick, 1968; Muthén, 1984; Olsson, 1979; Christofferssen, 1975). For example, given a variable with five response options, there are four threshold parameters τ, such that:

$$y_{pij} = \begin{cases} 1 \text{ if } y_{pij}^* < \tau_1 \\ 2 \text{ if } \tau_1 < y_{pij}^* < \tau_2 \\ 3 \text{ if } \tau_2 < y_{pij}^* < \tau_3 \\ 4 \text{ if } \tau_3 < y_{pij}^* < \tau_4 \\ 5 \text{ if } y_{pij}^* > \tau_4 \end{cases} \quad (1)$$

.

This model is extended to a two-level model by decomposing the vector of underlying continuous response variables $\mathbf{y}_{ij}^*$, into a vector of cluster means ($\boldsymbol{\mu}_j$), and a vector of individual deviations from the cluster means ($\boldsymbol{\eta}_{ij}$):

$$\mathbf{y}_{ij}^* = \boldsymbol{\mu}_j + \boldsymbol{\eta}_{ij}. \quad (2)$$

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

$$\boldsymbol{\Sigma}_{\text{TOTAL}} = \boldsymbol{\Sigma}_{\text{BETWEEN}} + \boldsymbol{\Sigma}_{\text{WITHIN}}. \quad (3)$$

Any structural equation model can be fitted to the within and between level covariance matrices. A two-level factor model for p observed variables and k common factors is given by:

$$\boldsymbol{\Sigma}_{\text{BETWEEN}} = \boldsymbol{\Lambda}_{\text{BETWEEN}}\, \boldsymbol{\Phi}_{\text{BETWEEN}}\, \boldsymbol{\Lambda}_{\text{BETWEEN}}' + \boldsymbol{\Theta}_{\text{BETWEEN}},$$
$$\boldsymbol{\Sigma}_{\text{WITHIN}} = \boldsymbol{\Lambda}_{\text{WITHIN}}\, \boldsymbol{\Phi}_{\text{WITHIN}}\, \boldsymbol{\Lambda}_{\text{WITHIN}}' + \boldsymbol{\Theta}_{\text{WITHIN}}, \quad (4)$$

where $\boldsymbol{\Phi}_{\text{BETWEEN}}$ and $\boldsymbol{\Phi}_{\text{WITHIN}}$ are k by k covariance matrices, $\boldsymbol{\Theta}_{\text{BETWEEN}}$ and $\boldsymbol{\Theta}_{\text{WITHIN}}$ are p by p (diagonal) matrices with residual variances, and $\boldsymbol{\Lambda}_{\text{BETWEEN}}$ and $\boldsymbol{\Lambda}_{\text{WITHIN}}$ are p by k matrices with factor loadings at the between- and within-level, respectively.

Grilli and Rampichini (2007) outlined the specification and fitting procedures for multilevel factor models with ordinal data using maximum likelihood estimation via an EM (Expectation - Minimization) algorithm using adaptive numerical quadrature (denoted by MLR estimation in Mplus, Muthén & Muthén, 2007). Although theoretically the estimation of ordinal multilevel factor models poses no problems, estimation of the model parameters is computationally demanding. The maximum likelihood method is therefore restricted to the estimation of simple models with a small number of random effects. Fortunately, the model that is used to investigate cluster bias is quite restrictive, so that its parameters can usually be estimated using MLR estimation. As explained by Jak, Oort and Dolan (2013, in press), in the absence of cluster bias, the following model holds:

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

and

$$\boldsymbol{\Sigma}_{\text{WITHIN}} = \boldsymbol{\Lambda}\, \boldsymbol{\Phi}_{\text{WITHIN}}\, \boldsymbol{\Lambda}' + \boldsymbol{\Theta}_{\text{WITHIN}}. \quad (5)$$

If there is no cluster bias, the factor loadings are equal across levels, and there is no residual variance at the between level. The test for cluster bias implies constraining factor loadings to be equal across levels and testing whether the residual variances at the between level are zero. If the factor loadings are not equal across levels, the common factors do not have the same interpretation across levels (Muthén, 1990; Rabe-Hesketh, Skrondal & Pickles, 2004). If the between level residual variance of a given indicator is found to be greater than zero, then the indicator is judged to be affected by cluster bias.

Jak, Oort and Dolan (2013) 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 of 25 observations, 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.

In the next sections, we present a simulation study to investigate the performance of the test for cluster bias in ordinal data under various conditions. Finally, we illustrate the test with data from research on teacher-student relationships.

SIMULATION STUDY

We generated discrete scores on five items, representing responses of students in schools. The model we used to generate the data was a two-level factor model with one factor at each level, and a covariate at the between (second) level. Population parameter values are given in Figure 1. Factor loadings were equal across levels, and there was no residual variance at the between level. We introduced cluster bias in Item 1 by specifying a non-zero effect of the violator (covariate that possibly violates measurement invariance) on Item 1. Values that we chose were such that for unbiased items, 10% of the variance was at the between level (the intraclass correlation was .10). For unbiased items (Items 2, 3, 4 and 5), 50% of the total variance was common variance and 50% was residual variance. The size of the clusters was fixed at 25, which is a typical size of a school class.

CONDITIONS

Data were generated under various conditions. The size of the cluster bias was small, contributing 1% of the total variance, which corresponds to a small r^2 (Cohen, 1992), or large (contributing 5% of the total variance). We considered conditions with 100 clusters (total sample size is 100*25 = 2500) and conditions with 50 clusters (total sample size is 50*25 = 2500). We categorized the continuous normal data into 2 or 5 categories, and the observed score distributions were symmetrical or asymmetrical. Varying the factors size of the bias (none, small or large), number of clusters (50 or 100), number of categories (2 or 5), and frequency distribution (symmetrical or asymmetrical) yielded 3*2*2*2 = 24 conditions. We generated 500 samples per condition.

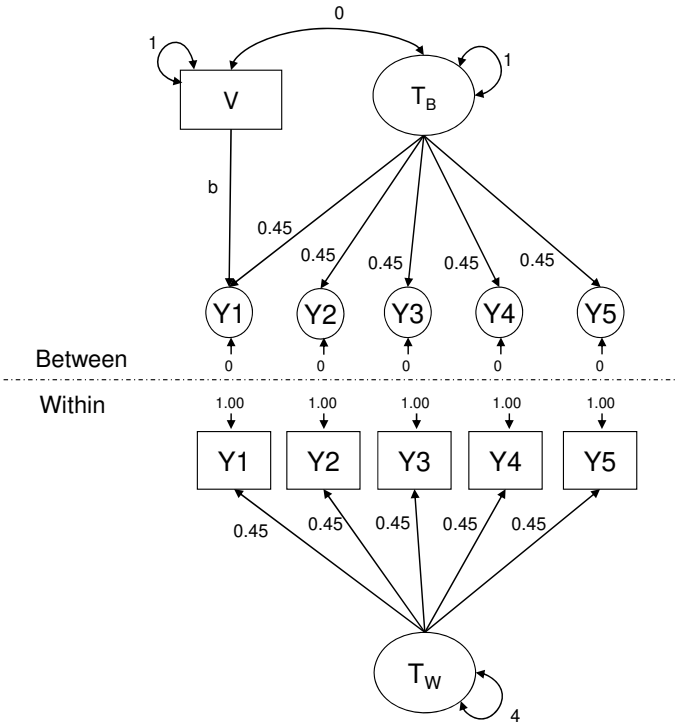


Figure 1. Two-level measurement model with population parameter values. In conditions with 0, 1, and 5 % bias, the corresponding values for *b* were 0, .142, and .324 respectively.

DATA GENERATION

We generated continuous multivariate normal data using the R program (R Development Core Team, 2011). First, cluster means were generated according to the following equation:

$$\mu_{ij} = \tau_i + \lambda_i t_j + b v_j, \quad (6)$$

where μ_{ij} is the mean of item *i* in cluster *j*, t_j is the cluster mean score on the common factor, v_j is the cluster score on the violator, τ_i is the intercept of item *i*, λ_i is the factor loading of item *i*, and *b* is a regression coefficient. The cluster scores t_j and v_j were drawn from the bivariate standard normal distribution, with means zero, unit variances and zero covariance.

In the next step, continuous data were drawn from the multivariate normal distribution with means corresponding to the associated cluster means from the previous step, and covariance matrix $\mathbf{\Sigma}_{WITHIN}$ that is calculated as $\mathbf{\Sigma}_{WITHIN} = \mathbf{\Lambda}\ \mathbf{\Phi}_{WITHIN}\ \mathbf{\Lambda'} + \mathbf{\Theta}_{WITHIN}$ (see Equation 5). We used the parameter values from Figure 1.

For unbiased items, the population values yield normally distributed continuous responses with a mean of 0 and a variance of 2.01. To obtain ordinal data, we categorized the continuous responses. Thresholds were chosen such that in conditions with symmetrically distributed scores, the population proportions for the two categories were .50 , .50 and the population proportions for five categories were .10, .20, .40, .20, .10 . Asymmetrical discrete distributions were created by assuming a mean of the underlying variable of -1, leading to population proportions of .76, .24 with two categories, and .28, .29, .32, .09, .02 with five categories. Biased items were given the same thresholds as unbiased items. The introduction of cluster bias increases the variance of the continuous variable with cluster bias. Greater variance leads to bigger tails in the continuous distribution, and more scores in the extreme categories of the categorical distribution.

ANALYSIS

We used robust maximum likelihood (MLR) estimation in Mplus (Muthén & Muthén, 2007) to fit the models to the generated datasets. MLR estimation of the parameters in the ordinal factor model is described by Grilli and Rampichini (2007). We investigated the effects of the various conditions on six outcomes: the proportions of true positives (empirical power) and the false positive rates of the likelihood ratio test, the likelihood ratio test with a correction factor (Satorra & Bentler, 2001), and of the univariate Wald test. The Wald test is the test that the parameter is zero, based on the parameter estimate divided by its standard error. We fitted three models to each sample:

Model 0: The cluster invariance model (Equation 5)

Model 1: A partial cluster invariance model with free Level 2 residual variance for Item 1 (a biased item)

Model 2: A partial cluster invariance model with free Level 2 residual variance for Item 2 (an unbiased item)

The true positives (power) of the likelihood ratio tests are associated with a significant difference in the likelihoods of Model 0 and Model 1, given the level of significance. The false positives of the likelihood ratio tests are indicated by a significant difference in fit between Model 0 and Model 2 in conditions without cluster bias. In conditions with an Item with cluster bias, a significant difference in fit between Model 0 and Model 2 indicates a false positive test with a misspecified model.

We investigated the true positives of the univariate Wald test by testing the significance of the Level 2 residual variance for Item 1 in Model 1. A false positive of the Wald test is found when in Model 2, the Level 2 residual variance for Item 2 is considered significant in conditions without cluster bias. False positive rates with misspecified models are also investigated in conditions with cluster bias. i.e. by testing the significance of the Level 2 residual variance of Item 2, while there is cluster bias in Item 1.

RESULTS

The results of the analyses with MLR estimation are shown in Table 1 and Table 2. The true and false positive rates in all conditions are shown for three tests: The uncorrected likelihood ratio test (LRT), the likelihood ratio test with a correction (scaled LRT), and the Wald test. Results are presented for alpha = .05 and alpha = .10, two-sided.

A graphical comparison of the results obtained with the three tests using alpha = .05 is shown in Figure 2. The Wald test is expected to give the same results as the LRT asymptotically (Engle, 1983). In our study, they indeed give similar results. Figure 2a shows the power of the three tests in the various conditions. It is striking that the scaled LRT shows decreasing power as the bias becomes larger. This points to a problem with this test. The last three columns of Table 1 show the proportions of cases where the three tests produced problematic results. Specifically, the scaled chi-square difference tests sometimes produce a negative value, which is invalid (this is a well-known problem; see Satorra & Bentler, 2010). The number of negative chi-square differences for the scaled LRT increased with the size of the bias. It seems that the estimation of the correction factor used in scaling the LRT is inaccurate with misspecified models. As the scaled LRT therefore is of limited use in testing for cluster bias, we limited our examination to the performance of the LRT and Wald test. The standard LRT also produced some negative values, indicating that the likelihood of the more restrictive model was higher than the likelihood of the least restrictive model. Our results show that the LRT produced these errors only if the size of the bias was small. Problems with the Wald test concerned untrustworthy standard errors due to non-positive definiteness of the first order derivative product matrix. These problems, while relatively rare overall, occurred more often in conditions with two response options than in conditions with five response options.

The power of the LRT and the Wald test exceeded .80 (marked in bold in Table 1) in all conditions were the bias was large (except for the asymmetrical condition with 50 clusters, with alpha = .05). In conditions with small bias, the power varied between .096 and .684. In general, power was higher in conditions with more response options and with a larger number of clusters. Figure 2b shows the false positive rates for the three tests with alpha = .05. While the LRT and the Wald test yield around 5% false positive rates in all

Table 1. Proportions of true positives and problems for all conditions, with α =.05 and α =.10.

Condition				Power α = .05			Power α = .10			Problems		
	N	***Size***	***Cat.***	***LRT***	***scaled LRT***	***Wald***	***LRT***	***scaled LRT***	***Wald***	***Negative LRT***	***Negative scaled LRT***	***Incorrect SE's***
Symmetrical	*50*	*small*	*2*	.136	.222	.116	.202	.274	.232	.254	.262	.078
			5	.262	.374	.236	.356	.454	.414	.150	.212	.018
		large	*2*	**.890**	.376	**.850**	**.936**	.388	**.948**	0	.566	.012
			5	**.998**	.064	**.996**	**1.00**	.064	**1.00**	0	.936	0
	100	*small*	*2*	.218	.324	.220	.300	.388	.364	.096	.096	.038
			5	.502	.608	.518	.618	.672	.684	.052	.080	.036
		large	*2*	**.996**	.396	**.990**	**.998**	.398	**.996**	0	.600	.002
			5	**1.00**	.014	**1.00**	**1.00**	.014	**1.00**	0	.986	.038
Asymmetrical	*50*	*small*	*2*	.096	.186	.180	.142	.236	.328	.330	.328	.152
			5	.224	.340	.228	.332	.398	.372	.142	.164	.034
		large	*2*	.794	.400	.734	**.844**	.434	**.850**	.010	.492	.020
			5	**.998**	.080	**.994**	**1.00**	.080	**.998**	0	.920	0
	100	*small*	*2*	.171	.274	.182	.252	.342	.312	.120	.136	.090
			5	.462	.588	.442	.568	.644	.632	.046	.056	.048
		large	*2*	**.984**	.524	**.974**	**.996**	.532	**.994**	0	.446	0
			5	**1.00**	.026	**1.00**	**1.00**	.026	**1.00**	0	.974	0

Note: ***N*** = number of clusters, ***Size*** = size of the clusterbias, ***Cat.=*** number of response categories, ***Negative LRT*** = the LRT results in a negative chi-square, ***Negative scaled LRT*** = the scaled LRT results in a negative chi-square, ***Incorrect SE's*** = Wald test is performed with untrustworthy standard errors.

Table 2. Proportions of false positives and problems for all conditions, with α = .05 and α =.10.

	Condition			False positives, α = .05			False positives, α = .10			Problems		
	N	***Size***	***Cat.***	***LRT***	***Scaled LRT***	***Wald***	***LRT***	***Scaled LRT***	***Wald***	***Negative LRT***	***Negative Scaled LRT***	***Incorrect SE's***
Symmetrical	*50*	*none*	*2*	.026	.078	.022	.044	.122	.078	.548	.490	.092
			5	.018	.120	.034	.050	.144	.122	.624	.574	.050
		small	*2*	.014	.066	.016	.028	.094	.086	.550	.532	.104
			5	.020	.092	.034	.032	.114	.100	.584	.550	.070
		large	*2*	.018	.074	.030	.046	.108	.090	.480	.456	.144
			5	.052	.170	.064	.098	.208	.164	.478	.446	.068
	100	*none*	*2*	.038	.076	.052	.046	.108	.098	.456	.446	.142
			5	.016	.078	.020	.036	.104	.060	.576	.538	.096
		small	*2*	.024	.072	.030	.044	.096	.078	.430	.442	.154
			5	.032	.070	.034	.044	.102	.078	.538	.504	.126
		large	*2*	.054	.136	.062	.114	.194	.148	.338	.348	.170
			5	.072	.142	.074	.110	.184	.156	.354	.356	.168
Asymmetrical	*50*	*none*	*2*	.026	.078	.022	.044	.122	.078	.548	.490	.102
			5	.028	.104	.046	.044	.142	.106	.572	.300	.001
		small	*2*	.012	.066	.024	.028	.094	.068	.564	.444	.136
			5	.016	.108	.020	.044	.156	.086	.568	.296	.114
		large	*2*	.020	.106	.024	.052	.124	.116	.558	.440	.130
			5	.054	.142	.050	.088	.170	.128	.460	.264	.076
	100	*none*	*2*	.024	.104	.044	.056	.150	.114	.340	.370	.086
			5	.018	.072	.028	.030	.100	.084	.562	.276	.200
		small	*2*	.022	.082	.024	.046	.108	.092	.312	.398	.060
			5	.021	.070	.036	.040	.116	.092	.492	.256	.116
		large	*2*	.028	.108	.040	.068	.156	.106	.284	.344	.074
			5	.044	.048	.142	.104	.190	.152	.354	.200	.124

Note: ***N*** = number of clusters, ***Size*** = size of the clusterbias, ***Cat.=*** number of response categories, ***Negative LRT*** = the LRT results in a negative chi-square, ***Negative scaled LRT*** = the scaled LRT results in a negative chi-square, ***Incorrect SE's*** = Wald test is performed with untrustworthy standard errors.

conditions, the scaled LRT is always yields around 10% false positive rate. The proportions of false positives were generally below or around the significance levels for the LRT and Wald test in conditions without bias. In conditions with cluster bias, the proportions of false positives of mainly the Wald test were higher if the size of cluster bias was large. The highest false positive rates were found in the symmetrical condition with large bias and 100 clusters. Asymmetry of the response distribution did not substantially affect power or false positive rate.

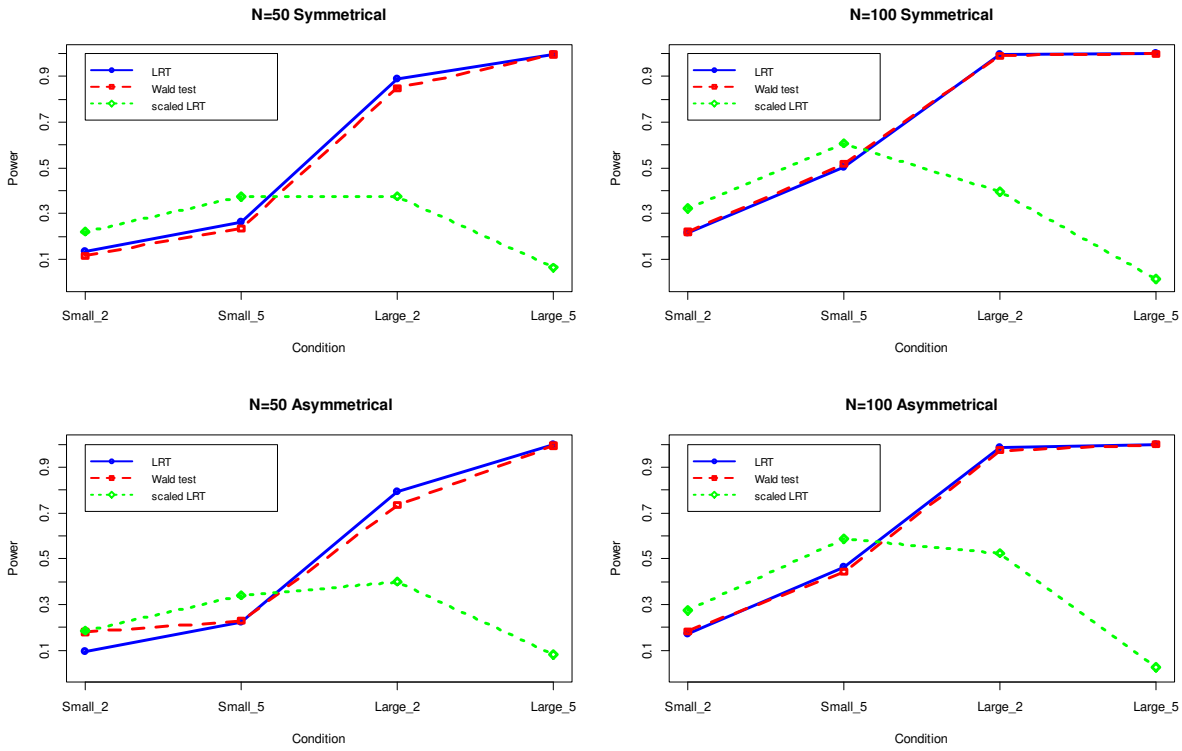


Figure 2a. A comparison of the power of the LRT, the Wald test, and the scaled LRT with alpha = .05 in different conditions.

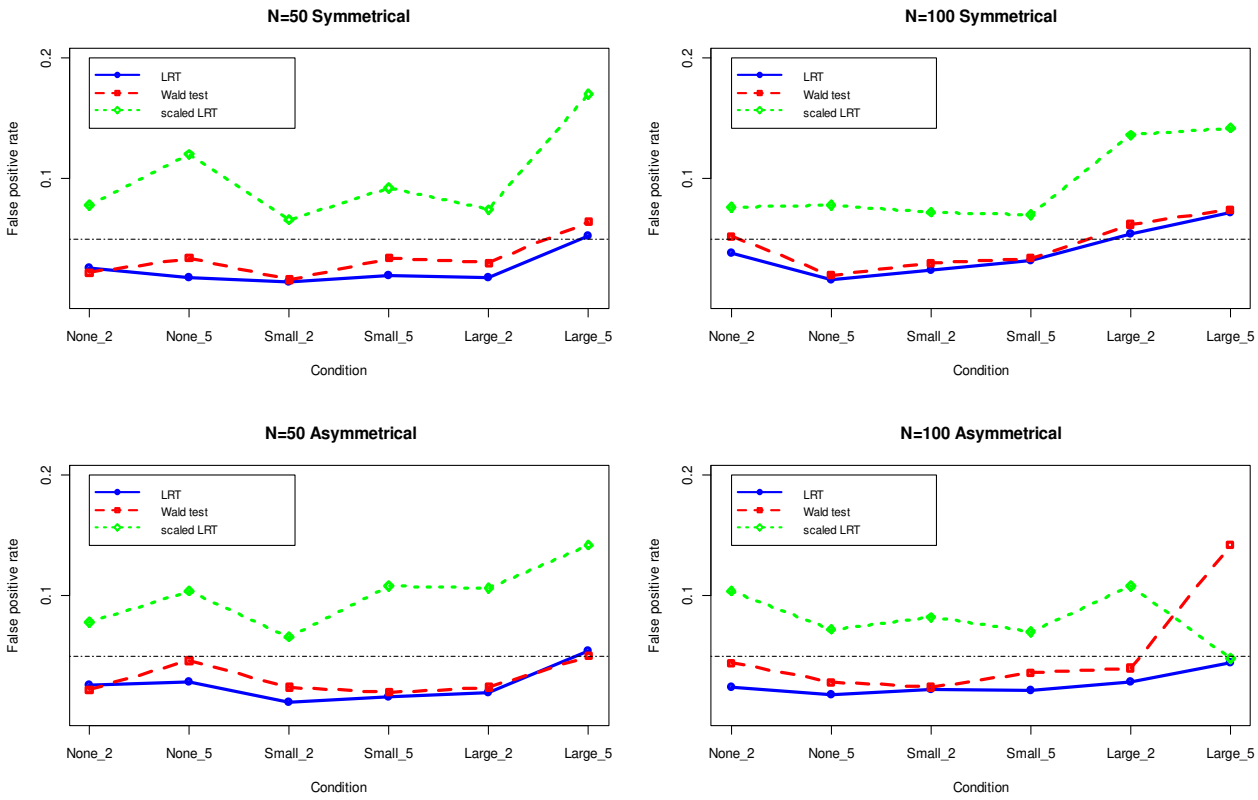


Figure 2b. A comparison of the false positive rate of testing cluster bias in Item 2 (unbiased item), while the cluster bias is in Item 1 (in the conditions with bias) of the LRT, the Wald test, and the scaled LRT with alpha = .05 in different conditions. The straight dotted lines denote the nominal alpha levels.

Note: None_2 : Condition without bias and 2 response options, None_5 : Condition without bias and 5 response options, Small_2 : Condition with small bias and 2 response options, Small_5 : Condition with small bias and 5 response options, Large_2 : Condition with large bias and 2 response options, Large_5 : Condition with large bias and 5 response options.

ILLUSTRATIVE EXAMPLE

DATA

We illustrate the test for cluster bias with data from the Dependency scale of a Dutch translation of the Student-Teacher Relationship Scale (STRS; Koomen, Verschueren & Pianta, 2007; Pianta, 2001). The scale comprises 6 items. Dependency refers to overly dependent and clingy child behavior. The dependency items are given in the note to Table 2. Data of 1493 students were gathered from 659 primary school teachers (182 men, 477 women) from 92 regular elementary schools. Each teacher reported on two or three students. 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

In earlier research, treating the responses as continuous outcomes, a one factor model was found to fit the item responses adequately (Koomen, Verschueren, van Schooten, Jak, Pianta, 2012; Spilt, Koomen & Jak, 2012). We use a one factor model with cluster invariance restrictions (see Equation 5) as the baseline model. An overall test for cluster bias was not feasible due to the number of parameters involved in this test. Therefore, we tested the residual variances one by one at a bonferroni corrected one-sided test with an alpha of .05. We used the one-sided test because we were testing the significance of a variance, that cannot have values below zero (Stoel, Garre, Dolan & van den Wittenboer, 2006).

RESULTS

Table 3 gives the -2 log likelihood of the cluster invariance model on the dependency data. The Level 2 residual variance of each indicator was freed one by one. For each model we calculated the chi-square value associated with the likelihood ratio test, the chi-square value associated with the scaled likelihood ratio test, and the Wald-statistic. For the Wald statistic, we test against a critical value of 2.39, (i.e. the z-value associated with an alpha level of .10, divided by the number of tests to be performed (six). For the LRT's, the critical value was 6.96, (i.e. the chi-square value associated with an alpha level of .10 / 6). Table 3 shows that all chi-square values were larger than this critical value, so, according to the likelihood ratio tests, there was cluster bias in all six indicators. The Wald statistic indicated there was significant cluster bias in all indicators except for Item 3. The proportions of cluster bias relative to the total variances are given in the last column. The most cluster bias is found in the first indicator, of which about one third of the variance is caused by other between factors that Dependency. For the other indicators, the percentages varied from .10 to .20 %.

CONCLUSION

Cluster bias implies that variables other than the common factor are causing differences in scores between clusters. The cluster bias was largest for the first indicator, i.e., the Item: "This child fixes his/her attention on me the whole day long". This item can be viewed as different from the others as it involves passive behavior of the child: focusing attention to the teacher, instead of actively attracting attention from the teacher. A possible explanation for the cluster bias could be found in teachers varying in the ability to perceive such behavior.

Table 3. Fit results of the cluster invariance model and six models with estimated Level 2 residual variance for one of the items.

Model	-2 Log likelihood	Scale factor	Scaled LRT Chi-square	LRT Chi-square	Wald test Estimate / SE	Proportion bias Level 2*	Proportion bias Total**
0. Invariance	25168.90	1.232					
1. $\theta_{BETWEEN,11}$	25035.44	1.208	287.66	133.46	5.933	.584	.333
2. $\theta_{BETWEEN,22}$	25148.30	1.218	26.29	20.61	2.966	.190	.091
3. $\theta_{BETWEEN,33}$	25157.46	1.229	10.08	11.45	2.232	.231	.098
4. $\theta_{BETWEEN,44}$	25142.84	1.212	44.03	26.07	3.798	.348	.166
5. $\theta_{BETWEEN,55}$	25075.82	1.201	387.84	93.08	5.656	.401	.209
6. $\theta_{BETWEEN,66}$	25101.42	1.207	156.24	67.50	4.530	.341	.166

* Calculated as: residual variance at Level 2 / total variance at Level 2

** Calculated as: residual variance at Level 2 / total variance at Level 1 + Level 2

Dependency items:

1. This child fixes his/her attention on me the whole day long.
2. This child reacts strongly to separation from me.
3. This child is overly dependent on me.
4. This child asks for my help when he/she really does not need help.
5. This child expresses hurt or jealousy when I spend time with other children.
6. This child needs to be continually confirmed by me.

DISCUSSION

From the simulation study we can conclude that cluster bias can be tested in ordinal data with the LRT and Wald-test. Both tests show good power to detect large bias, and show acceptable false positive rates. The scaled LRT, as implemented in Mplus, is not recommended for cluster bias testing as inadmissible results were obtained in all conditions, and their number increased with the amount of bias.

In the data from the illustration, the clusters were smaller than in the simulation study (average cluster size was around three in the illustration and 25 in the simulation). The test for cluster bias has yet to be evaluated for cluster sizes smaller than 25. We expect that with smaller cluster sizes, more within level random variance (as opposed to structural variance) in the indicators will be aggregated to the between level, leading to larger proportions of false positives in the test for cluster bias. However, even if this is the case, indicators with cluster bias will have more residual variance than other indicators at the between level, such as Item 1 from the illustration, which had twice as much between level residual variance as the other items.

In this paper we used MLR estimation. An alternative estimator, suitable for larger models, is the multilevel version of weighted least squares (denoted by WLSM in Mplus, Asparouhov & Muthén, 2007). This method replaces a complex model estimation with high dimensional numerical integration by multiple smaller models with low dimensional numerical integration. If the test for cluster bias cannot be performed by MLR estimation due to computational difficulties, WLSM may be a viable alternative, although simulation research is needed to verify this.

Measurement bias across clusters in discrete multilevel data could also be investigated using item response models for measurement bias. Verhagen and Fox (2012) show how to use Bayesian methods to test invariance hypothesis in the random item effects modeling framework.

In conclusion, this study showed that cluster bias can be tested in ordinal data, using ordinal factor analysis. We prefer and advice to use the unscaled LRT or the Wald test over the scaled version of the LRT, as the latter gave untrustworthy results. The unscaled LRT and the Wald test performed well in terms of empirical power rate if the amount of cluster bias is large, and showed acceptable false positive rates.

CHAPTER 4

On the power of the test for cluster bias

Abstract Cluster bias refers to measurement bias with respect to the clustering variable in multilevel data. Cluster bias can be investigated using two-level factor analysis, by constraining the factor loadings to be equal across levels, and testing the absence of residual variance at the cluster level (Level 2). The absence of cluster bias implies absence of bias with respect to any Level 2 variable (Jak, Oort, & Dolan, 2013). Therefore, the test for cluster bias serves as a global test of measurement invariance with respect to any Level 2 variable. However, the validity of the global test depends on its power. In this simulation study we evaluate whether not rejecting cluster invariance indeed implies absence of bias with respect to all Level 2 variables.

The performance of the test for cluster bias is compared with the performance of the RFA model (Oort, 1992) to detect the bias. It appeared that the RFA test has considerably more power than the test for cluster bias. However, the false positive rates of the test for cluster bias were generally around the expected values, while the RFA test showed unacceptably high false positive rates in some conditions. We conclude that if no significant cluster bias is found, there can still be significant bias with respect to a Level 2 violator in an RFA model. Although the test for cluster bias is less powerful, an advantage of the test is that the cause of the bias does not need to be measured, or even known.

Based on: Jak, S. & Oort, F. J. (under review). On the power of the test for cluster bias.

INTRODUCTION

The importance of establishing measurement invariance of research instruments is widely recognized; a measurement instrument should function identically in different groups of respondents (see Cheung & Rensvold, 1999; Meredith, 1993; Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000). If measurement invariance does not hold with respect to some variable (e.g. gender), then two respondents with identical values on the latent trait that the test is supposed to measure, may have different expected scores, depending on their value on the other variable (e.g. depending on being a man or a woman). When a test is biased with respect to gender, then gender is called a *violator* (of measurement invariance; Oort, 1992). Within 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.

In the present paper we consider the investigation of measurement invariance in two-level data. Two-level data are data with a clustered structure, such as children in school classes or patients in hospitals. In these cases there are two levels of analysis, the student or patient level is called *Level 1* or the *within* level. The class or hospital level is called *Level 2* or the *between* level. With two-level data, measurement bias can be present at the within level or at the between level. The purpose of this paper is to compare the performance of two methods to investigate measurement bias at the between level. One method is the test for cluster bias, which can be considered a global test of measurement bias at the between level, in which the violating variable(s) is (are) not necessarily measured or even known. The other method is the RFA test, which requires the operationalization of possible violators of measurement invariance, in order to include them as exogenous variables in multilevel factor analysis.

MEASUREMENT BIAS AT LEVEL 2

With two-level SEM, the covariance matrix is modelled as the sum of the within level (Level 1) and the between level (Level 2) covariance matrices (Muthén, 1990; Rabe - Hesketh, Skrondal & Pickles, 2004):

$$\boldsymbol{\Sigma}_{TOTAL} = \boldsymbol{\Sigma}_{BETWEEN} + \boldsymbol{\Sigma}_{WITHIN}. \quad (1)$$

For example, consider data concerning the closeness of teacher-child relations, obtained using a 5-item questionnaire, completed by teachers for several of their pupils. The (pooled,

within class) differences between children are modelled by $\mathbf{\Sigma}_{WITHIN}$. Teachers also differ in the general closeness of their relations with children. The differences between teachers are modeled by $\mathbf{\Sigma}_{BETWEEN}$. At the within and between levels, distinct measurement models can be used to describe the covariances between the item scores. In the present study we employ the linear factor model as the measurement model (Mellenbergh, 1994).

Jak, Oort, and Dolan (in press) described a five-step procedure to investigate measurement bias in multilevel data. In this procedure, Step 1 involves testing the necessity of applying multilevel analysis, Step 2 consists of establishing a measurement model at Level 1, Step 3 involves testing for measurement bias at Level 1, Step 4 refers to testing for cluster bias, and Step 5 concerns explaining the cluster bias with observed Level 2 variables. The present study focusses on Steps 4 and 5 of this procedure.

Testing for cluster bias (Step 4) can be seen as a global test for measurement bias with respect to all possible Level 2 violators. In case of cluster bias, one or more indicators measure different constructs in different clusters. In the closeness example, if there is cluster bias, this means that there are other cluster level variables than closeness causing differences between the teachers' scores. The test for cluster bias involves testing whether the factor loadings are equal across levels, and whether the residual variance at Level 2 is zero (Jak, Oort & Dolan, 2013). If there is no cluster bias, for p observed variables and k common factors, the following model holds:

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

and

$$\mathbf{\Sigma}_{WITHIN} = \mathbf{\Lambda}\,\mathbf{\Phi}_{WITHIN}\,\mathbf{\Lambda}' + \mathbf{\Theta}_{WITHIN}. \qquad (2)$$

In this model, $\mathbf{\Phi}_{BETWEEN}$ and $\mathbf{\Phi}_{WITHIN}$ are k by k covariance matrices, $\mathbf{\Theta}_{WITHIN}$ is a p by p (diagonal) matrix with residual variances, and $\mathbf{\Lambda}$ is a p by k matrix with factor loadings at the between- and within-level, respectively. Cluster bias appears as residual variance at the between level, and can be modeled by estimating a (diagonal) p by p matrix with residual variance at the between level ($\mathbf{\Theta}_{BETWEEN}$), so that $\mathbf{\Sigma}_{BETWEEN} = \mathbf{\Lambda}\,\mathbf{\Phi}_{BETWEEN}\,\mathbf{\Lambda}' + \mathbf{\Theta}_{BETWEEN}$. The test for cluster bias involves testing whether $\mathbf{\Theta}_{BETWEEN}$ is zero.

Equality of the factor loadings across levels implies that the common factors have the same interpretation across levels (Muthén, 1990; Rabe-Hesketh, Skrondal & Pickles, 2004). In the closeness example, it means that the factor at the within level can be interpreted as child-level closeness, and the factor at the between level as teacher-level closeness.

If there are no factors other than closeness influencing the teachers' scores, then there is no residual variance in the Level 2 common factor model; the closeness factor explains all

variance and covariance at the between level. In that case, it would not be necessary to include other variables in the model, which possibly cause measurement bias at the between level. If there is residual variance however, this cluster bias can possibly be explained by measured between level variables.

Testing for measurement bias with respect to specific between level variables (the final step of the five-step procedure) requires the availability of measured between level variables that can be added to the model. Measurement bias with respect to such variables can be investigated using RFA (Oort, 1992), which is statistically equivalent to MIMIC modeling (Muthén, 1989). In an RFA model, the variables that possibly violate measurement invariance, the violators, are correlated with the common factors, whereas in MIMIC the common factors are regressed on the violators. In both models, measurement bias is represented as a direct effect of the violator on the indicators. Testing for measurement bias is only appropriate if there is variance in the indicators that is not explained by the common factor. So, if there is no cluster bias, i.e., if the residual level 2 variance is zero, investigating measurement bias with respect to possible between level violators would be superfluous.

The test for cluster bias thus serves as a global test of measurement invariance at the between level. However, the cluster bias test is subject to Type 1 errors (false positives) and Type 2 errors (false negatives). If the power of the overall cluster bias test is sufficient, or at least larger than the power of the RFA test, then not detecting cluster bias will render the RFA test unnecessary. In that case, the RFA test will not detect bias that the test for cluster bias would not detect. However, if the power of the test for cluster bias is smaller than the power of the RFA test, it is possible that a researcher will not detect cluster bias with the global cluster bias test, but will detect measurement bias with respect to particular Level 2 variables with the RFA test. This raises the important question concerning the informativeness of the global test compared to the RFA test. Specifically, if the cluster bias test is highly informative, the RFA test could be superfluous, while if the cluster bias test is not really informative, the cluster bias test itself would be of no use. In this paper we use simulated data to compare the power and false positive rates of the test for cluster bias and the RFA test in several conditions, varying the size of the bias and the sample size at both levels.

METHOD

To compare the performance of the test for cluster bias and the RFA test, we generated 500 datasets for each of 18 conditions, according to a factorial design with the following three factors:

- Bias effect size (none, small, large)
- Between level sample size (50 clusters, 100 clusters)
- Within level sample size (2, 5, 25 observations per cluster)

In all conditions, the population model was a two-level, one-factor model with 5 indicators, with one observed covariate (violator) at the between level. Population values are given in Figure 1. In the population, factor loadings are equal across levels, and there is no residual variance at the between level. With these population values, 50% of the total variance is residual variance and 10% of the variance of an unbiased indicator is at the between level (ICC = .10).

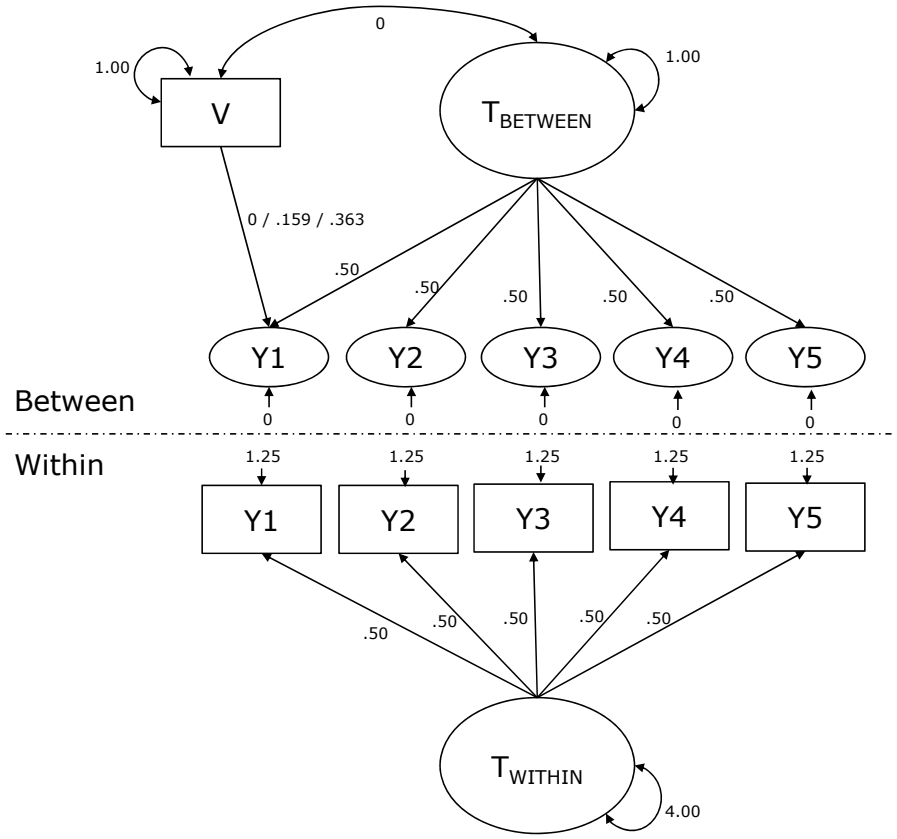


Figure 1. Two-level measurement model with population parameter values. In conditions with 0, 1, and 5 % bias, the corresponding values for the effect of V on Y1 were 0, .159, and .363 respectively.

BIAS EFFECT SIZE

Bias was introduced in the first indicator, by including a direct effect of the violator on this indicator. Small sized bias was defined as a direct effect of .159, which amounts to 1% of the total variance of the indicator being caused by the violator. Large sized bias was

defined as a direct effect at .363, which amounts to 5% of the total variance being caused by the violator.

Between level sample size

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

Within level sample size

A within level sample size of 25 corresponds to the typical size of a school class (e.g. Elffers, 2012; Thoonen, Sleegers, Peetsma & Oort, 2010). Group sizes of 5 are common in data from organisational research where it is a typical size of a working team (e.g. Jackson & Joshi, 2004; Koman & Wolff, 2007). Cluster sizes of 2 correspond to a typical cluster size in data from family research (e.g. Duncan, Alpert & Duncan, 1998; Voorpostel & Blieszner, 2008).

Data generation

We generated continuous multivariate normally distributed data using the same procedure as Jak, Oort, & Dolan (2013), using the R-package 'mvtnorm' (Genz et al., 2012).

Likelihood ratio test and Wald test

The likelihood ratio test (LRT) and the Wald test were used to test the significance of parameters. The likelihood ratio equals the difference in -2 log likelihoods of a model with and without the parameter(s) of interest. The difference between the -2 log likelihoods follows a chi-square distribution with degrees of freedom equal to the difference in numbers of parameters between the two models, assuming the parameter of interest is zero. If the chi-square test is significant, given the chosen alpha level, then the hypothesis of the parameter(s) of interest being zero is rejected. In the present study, we used robust maximum likelihood estimation (MLR) in Mplus (Muthén & Muthén, 2007) to obtain parameter estimates. The differences in -2 log likelihoods of models that are estimated with MLR theoretically need a correction to approximate the chi-square distribution (Satorra & Bentler, 2001). However, simulation studies have showed, that conducting the LRT with this correction often leads to untrustworthy results and that the corrected LRT does not perform better than the uncorrected LRT (Cham, West, Ma & Aiken, 2013; Jak, Oort, Dolan, 2013). In this study we therefore apply the uncorrected LRT.

The Wald test is based on the parameter estimate divided by its standard error, and tests the hypothesis that the parameter is zero. The Wald test is asymptotically equivalent to the LRT (Engle, 1983).

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

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

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

TESTING FOR LEVEL 2 BIAS WITH THE CLUSTER BIAS TEST AND THE RFA TEST

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

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

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

estimated for all indicators, allowing for possible cluster bias in the indicators that is not explained by the violator (Case C).

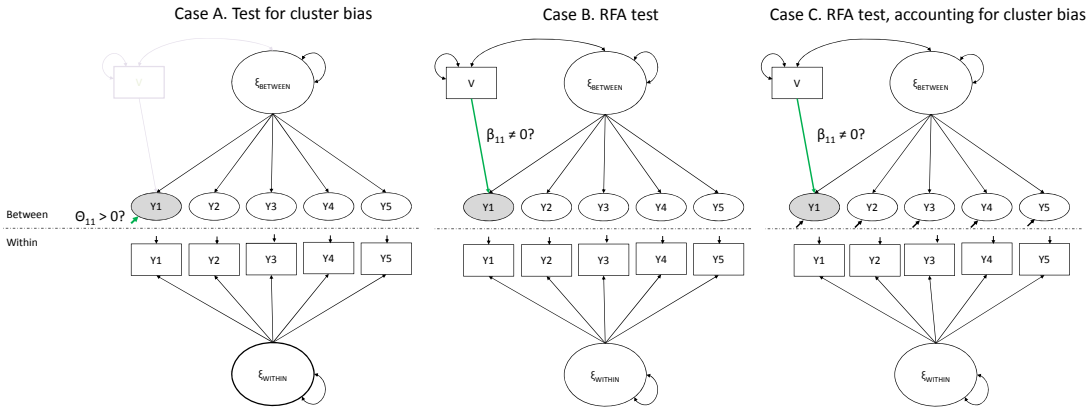


Figure 2. Three models that were used to investigate the power of three tests (corresponding to Case A, Case B and Case C).

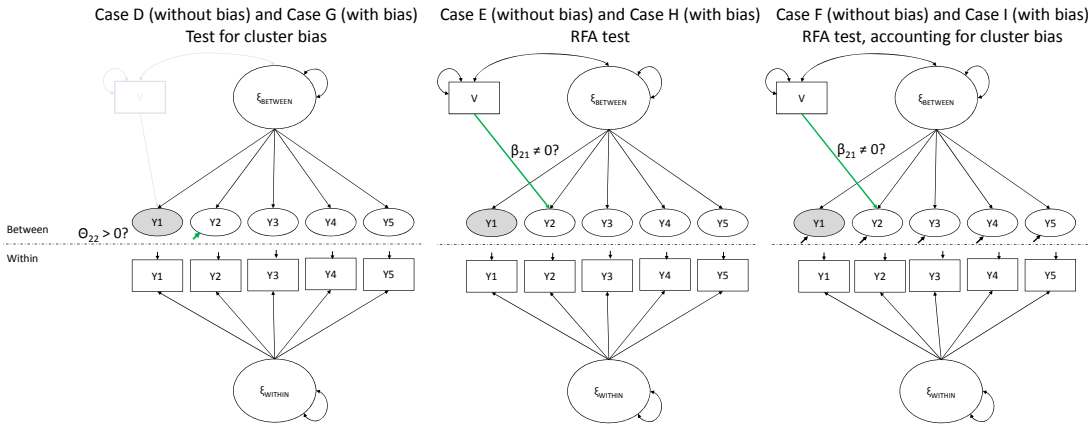


Figure 3. The models that were used to investigate the false positive rates (without bias) and false positive rates with a misspecified model (with bias in Indicator 1) (Case D through I).

We investigated the false positive rates of all tests in three ways (see Figure 3). Firstly, we tested for bias in the conditions where no bias was introduced (Cases D, E and F). Secondly, we tested for bias in Indicator 2 (an unbiased indicator), in conditions where the bias was in Indicator 1 (Cases G, H and I). So in these cases we investigate the false positive rates with a misspecified model.

In the conditions with bias in Indicator 1, we investigated a third type of false positives (Case J, see Figure 4). In these cases we accounted for the bias by letting the violator have a direct effect on this indicator. We then tested the residual variance in Indicator 1. As the violator is the only cause of the cluster bias, significance of the residual variance represents a false positive result.

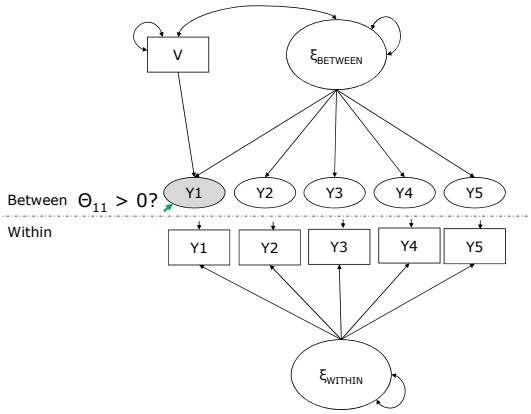


Figure 4. The model that was used to evaluate the false positive rates of the cluster bias test after accounting for the bias in the RFA model (Case J).

We test against levels of significance of alpha of 5% and 10%. The test for cluster bias involves testing a variance parameter, which cannot be negative by definition. Therefore, in line with Stoel, Garre, Dolan & van den Wittenboer (2006), we employ one-sided levels of significance of .05 and .10 with the test for cluster bias. Direct effects can be either negative or positive, so with the RFA tests we use two-sided tests. This implies that in order to obtain an alpha level of .05 we used a critical chi-square value of $\chi^2_{crit} = 2.71$ with the test for cluster bias and a critical value of $\chi^2_{crit} = 3.84$ with the RFA tests. With an alpha level of .10, these critical values are $\chi^2_{crit} = 1.64$ for the test for cluster bias and $\chi^2_{crit} = 2.71$ for the RFA tests. Critical values for the Wald-tests are obtained in the same manner (with alpha = .05, $z_{crit} = 1.28$ for the cluster bias test and 1.64 for the RFA test, and with alpha = .10, $z_{crit} = .84$ for the cluster bias test and $z_{crit} = 1.28$ the RFA test).

RESULTS

The results of the LRT and the Wald test are very similar. In the tables and figures we show the outcomes from both the LRT and the Wald test, both at the 5% and 10% level of significance, but below we focus on the LRT with an alpha level of 5%.

POWER

Results of the true positive rates of the three tests are shown in Table 2. Figure 5 shows the results based on the .05 alpha level graphically. With large bias, all bias is detected by all three tests, provided the total sample size is large (100 or 50 clusters with 25 observations per cluster). With smaller samples, the two RFA tests still have adequate power, but the power of the test for cluster bias drops to 69% and 44% with 100 and 50 clusters of 5 observations, respectively, and to even lower levels with 2 observations per cluster.

With small sized bias, the power of all tests is low. The test for cluster bias detects 76% of the bias in the conditions with the largest sample size, and detects less than 10% of the bias in conditions with small sample sizes. The two RFA tests perform better, with acceptable power in conditions with 25 observations per cluster, and in the condition with 100 clusters with 5 observations. With just 50 clusters of 5 observations, the RFA tests detect around 50% of the bias, which drops to around 20% with 50 clusters of 2 observations per cluster. Overall, the power of the RFA tests is considerably larger than the power of the test for cluster bias.

Table 2. Power: Case A, Case B and Case C.

True positive rates of the likelihood ratio test and the Wald test, using the cluster bias model and the RFA model. Based on 500 replications per condition.

			α = .05						**α = .10**					
			Cluster bias		**RFA**		**RFA free Θ**		**Cluster bias**		**RFA**		**RFA free Θ**	
Size bias	**N between**	**N within**	**LRT**	**Wald test**	**LRT**	**Wald test**	**LRT**	**Wald test**	**LRT**	**Wald test**	**LRT**	**Wald test**	**LRT**	**Wald test**
Large	*100*	*25*	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		5	.686	.762	1.00	1.00	1.00	1.00	.794	.880	1.00	1.00	1.00	1.00
		2	.150	.242	.966	.982	.962	.982	.236	.420	.982	.992	.980	.992
	50	*25*	1.00	1.00	1.00	1.00	1.00	.988	1.00	1.00	1.00	1.00	1.00	.990
		5	.438	.532	.994	.996	.988	.992	.598	.736	1.00	1.00	.998	.996
		2	.102	.198	.794	.890	.780	.878	.186	.334	.860	.952	.848	.938
Small	*100*	*25*	.756	.826	1.00	1.00	1.00	1.00	.860	.908	1.00	1.00	1.00	1.00
		5	.108	.174	.808	.888	.792	.884	.194	.308	.878	.942	.862	.942
		2	.074	.134	.418	.558	.404	.550	.126	.252	.520	.702	.512	.690
	50	*25*	.454	.552	.982	.992	.978	.982	.628	.750	.992	.996	.990	.986
		5	.082	.100	.508	.662	.462	.638	.136	.250	.620	.750	.600	.734
		2	.044	.076	.210	.310	.194	.340	.082	.182	.314	.456	.310	.456

Note: The alpha levels are one-sided alpha levels for the test for cluster bias and two-sided alpha levels for the RFA tests.

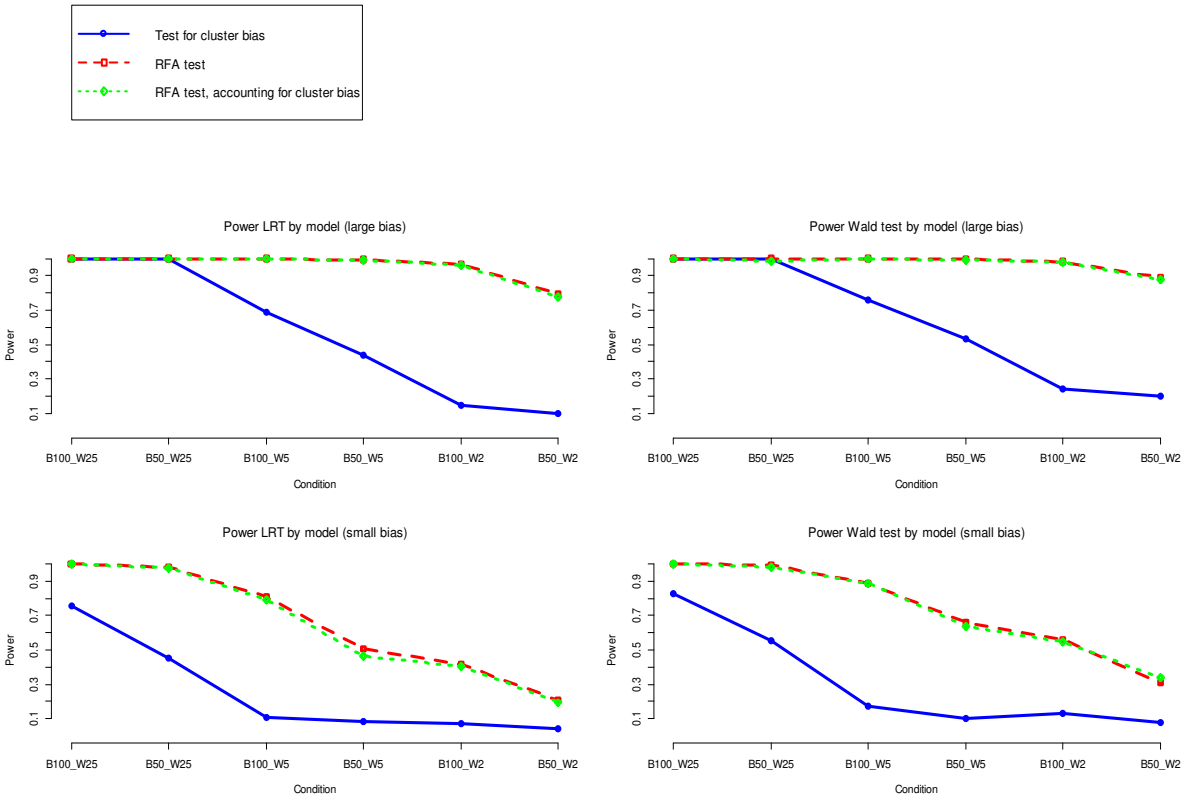


Figure 5. Case A, Case B and Case C (Power).
True positive rates of the test for cluster bias and the two RFA tests with various sample sizes in conditions with large bias (upper part) and small bias (lower part), with the LRT (left part) and the Wald test (right part).
Note: On the X-axis, B100_W25 refers to the condition with 100 clusters with 25 observations per cluster, B50_W25 to the condition with 50 clusters with 25 observations per cluster, and so on.

FALSE POSITIVE RATES

In conditions without bias, the expected false positive rate is the chosen alpha level of significance. Observed false positive rates for all tests are given in Table 3 and Table 4. Figure 6 shows a plot of the false positive rates with an alpha of .05 . The upper two graphs show the false positive rates in the conditions without bias. The false positive rates of the cluster bias test are all under the alpha level, and for the RFA tests we found false positive rates around the chosen alpha level.

We obtained interesting results in conditions, where we introduced the bias in Indicator 1, but we tested bias in Indicator 2. In this case the model is effectively misspecified. When the bias was small, the false positive rates of the test for cluster bias were acceptable, but the two RFA tests identified Indicator as 2 biased in 33% and 21% of the samples in

Table 3. False positives: Case D, Case E and Case F.
False positive rates of the likelihood ratio test and the Wald test, using the cluster bias model and the RFA model. Based on 500 replications per condition.

			α = .05						**α = .10**					
			Cluster bias		**RFA**		**RFA free Θ**		**Cluster bias**		**RFA**		**RFA free Θ**	
Size bias	**N between**	**N within**	**LRT**	**Wald test**	**LRT**	**Wald test**	**LRT**	**Wald test**	**LRT**	**Wald test**	**LRT**	**Wald test**	**LRT**	**Wald test**
None	*100*	*25*	.038	.068	.044	.086	.042	.086	.086	.130	.090	.218	.074	.218
		5	.034	.080	.056	.134	.048	.130	.084	.160	.102	.216	.094	.214
		2	.024	.072	.032	.088	.024	.092	.064	.148	.068	.186	.064	.192
	50	*25*	.028	.042	.056	.120	.052	.120	.058	.136	.100	.228	.086	.216
		5	.034	.062	.068	.164	.068	.156	.076	.154	.122	.262	.118	.260
		2	.046	.090	.046	.094	.032	.118	.082	.160	.094	.180	.092	.212

Note: The alpha levels are one-sided alpha levels for the test for cluster bias and two-sided alpha levels for the RFA tests.

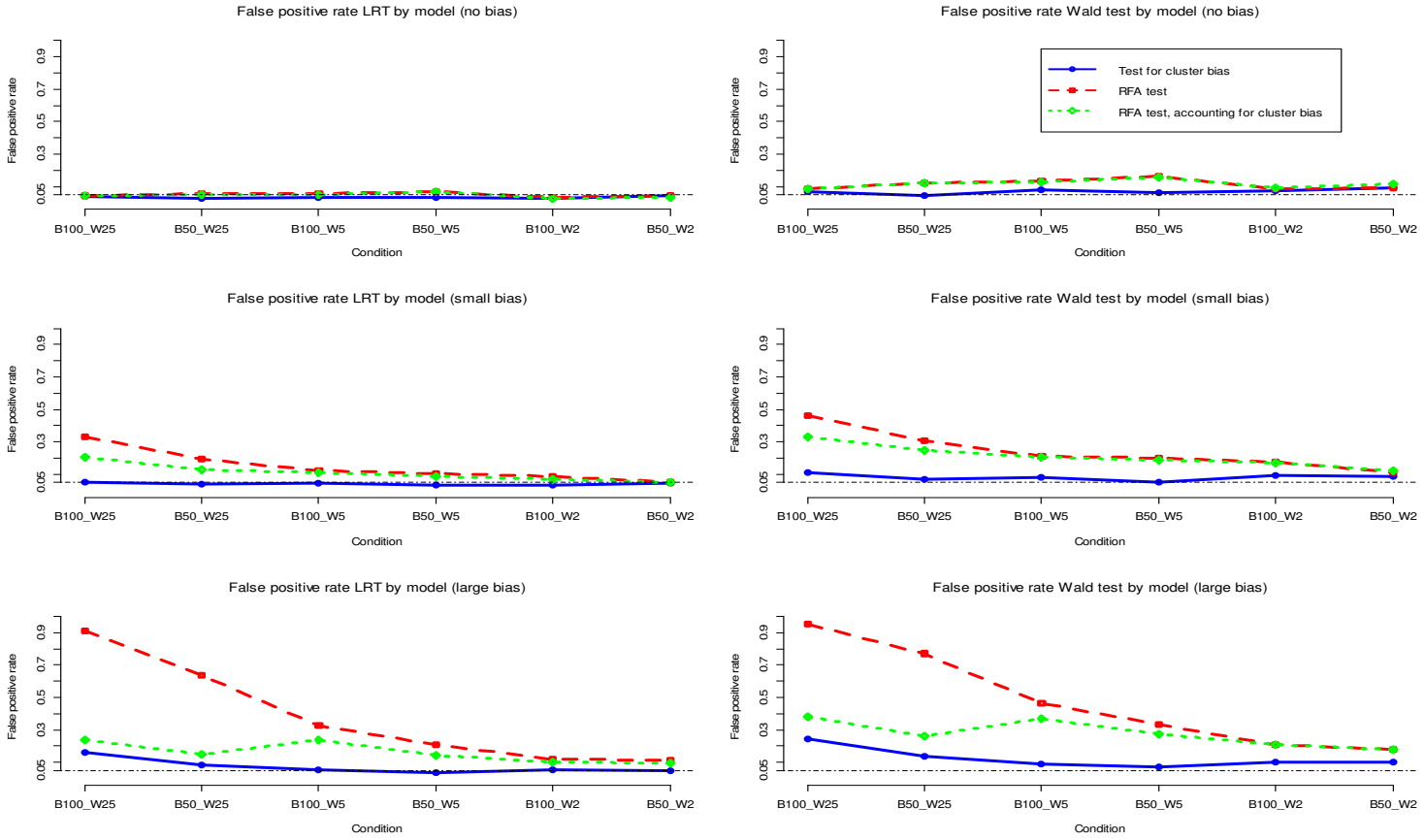


Figure 6. Case D – Case I (False positives) . False positive rates of the test for cluster bias and the two RFA tests with various sample sizes in conditions without bias (upper part), with small bias (middle part) and with large bias (lower part), with the LRT (left part) and the Wald test (right part). In conditions with bias, the bias was in Indicator 1, while we tested bias in Indicator 2.**Note**: On the X-axis, B100_W25 refers to the condition with 100 clusters with 25 observations per cluster, B50_W25 to the condition with 50 clusters with 25 observations per cluster, and so on. **Note:** The nominal alpha level is marked with a straight dotted line.

Table 4. Case G, Case H and Case I: False positives with misspecified model.
False positive rates of the likelihood ratio test and the Wald test in conditions with bias, using the cluster bias model and the RFA model. Based on 500 replications per condition.

			α = .05						**α = .10**					
			Cluster bias		**RFA**		**RFA free Θ**		**Cluster bias**		**RFA**		**RFA free Θ**	
Size bias	**N between**	**N within**	**LRT**	**Wald test**	**LRT**	**Wald test**	**LRT**	**Wald test**	**LRT**	**Wald test**	**LRT**	**Wald test**	**LRT**	**Wald test**
Large	*100*	*25*	.164	.242	.910	.950	.236	.380	.284	.394	.954	.982	.342	.552
		5	.054	.088	.330	.464	.238	.452	.100	.190	.448	.600	.336	.632
		2	.052	.104	.120	.208	.100	.206	.094	.204	.214	.338	.182	.338
	50	*25*	.082	.136	.636	.770	.152	.262	.194	.296	.762	.868	.242	.374
		5	.038	.070	.210	.332	.144	.272	.072	.158	.314	.484	.240	.418
		2	.050	.102	.116	.178	.098	.180	.100	.180	.180	.272	.172	.300
Small	*100*	*25*	.054	.112	.330	.464	.206	.334	.120	.222	.446	.610	.304	.504
		5	.048	.084	.122	.214	.112	.208	.090	.186	.204	.320	.188	.318
		2	.034	.092	.088	.176	.070	.174	.080	.194	.164	.260	.160	.262
	50	*25*	.040	.072	.196	.310	.132	.248	.092	.162	.278	.450	.216	.340
		5	.034	.052	.106	.202	.088	.190	.066	.148	.154	.300	.146	.292
		2	.044	.086	.054	.114	.052	.124	.074	.166	.112	.196	.104	.226

Note: The alpha levels are one-sided alpha levels for the test for cluster bias and two-sided alpha levels for the RFA tests.

conditions with large sample sizes. With smaller sample sizes these percentages drop considerably, and with 50 clusters with 2 observations the false positive rates are around 5% for both tests. With large bias in Indicator 1, the RFA test without estimated residual variance identified Indicator 2 as biased in almost all cases (91%) with large sample size, while the RFA test with residual variance identified 24% of the cases as biased, and the test for cluster bias falsely detected bias in only 16% of the cases. The Wald test shows similar results, but has slightly higher false positive rates overall. In the RFA models, the significant direct effects on Indicator 2 were all negative. The false positive rates of testing cluster bias, while the bias is already accounted for by the violator, are given in Table 5. In all conditions, the false positive rates of the LRT are under the nominal level of significance and the false positive rates of the Wald test fluctuate around the expected alpha level.

Table 5. Case J. False positive rates of the likelihood ratio test and the Wald test in conditions with bias, using the cluster bias test after accounting for the bias in the RFA model (Figure 4). Based on 500 replications per condition.

			Cluster bias			
			α = .05		**α = .10**	
Size bias	**N between**	**N within**	**LRT**	**Wald test**	**LRT**	**Wald test**
Large	*100*	*25*	.014	.040	.064	.142
		5	.032	.058	.064	.140
		2	.030	.078	.082	.142
	50	*25*	.030	.058	.062	.118
		5	.022	.054	.072	.140
		2	.032	.066	.056	.146
Small	*100*	*25*	.026	.046	.060	.142
		5	.038	.056	.064	.156
		2	.038	.094	.090	.200
	50	*25*	.026	.042	.052	.100
		5	.036	.050	.066	.130
		2	.034	.058	.064	.130

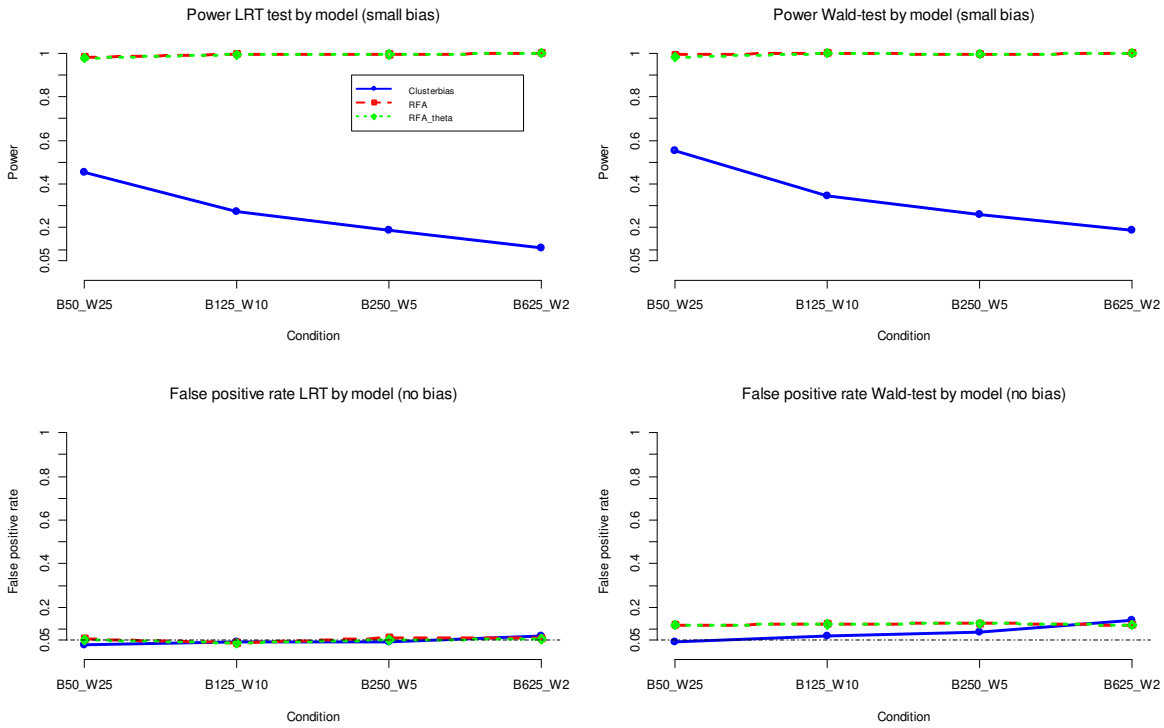


Figure 7. The effect of smaller cluster size. False positive rates (upper part) and power (lower part) of the test for cluster bias and the two RFA tests using the LRT (left part) and the Wald test (right part), with various cluster sizes leading to a total sample size of 1250. **Note**: On the X-axis, B50_W25 refers to the condition with 50 clusters with 25 observations per cluster, B125_W10 to the condition with 125 clusters with 10 observations per cluster, and so on. The nominal alpha level is marked with a straight dotted line.

The results show that the true and false positive rates of all tests vary with the total sample size, irrespective of the sample size at the within or between level. To check whether the within level or between level sample size has a greater effect on the performance of the cluster bias test, we additionally investigated the effect of the within level sample size, relative to the between level sample size. We expected that the false positive rates would increase with smaller cluster sizes, because more random error would be aggregated to the between level. We investigated the false positive rates and the power to detect small bias of the three tests in 4 conditions with a total sample size of 1250 (50 clusters with 25 observations, 125 clusters with 10 observations, 250 clusters with 5 observations and 625 clusters with 2 observations). Figure 7 shows the false positive rates and true positive rates as obtained with the LRT and the Wald test. in the four conditions. With the Wald test, but not with the LRT, there seems to be an upward trend in the false positive rates of the test for cluster bias, but not for the RFA tests. However, the false positive rates of the test for cluster bias are still below the nominal level of significance in all conditions, except for the condition with 625 clusters with 2 observations (where the false positive rate is 7%).

The RFA tests detect almost all bias in all conditions, but the power of the test for cluster bias shows a gradual decrease when cluster size becomes smaller.

DISCUSSION

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

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

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

In practice, researchers can avoid incorrectly identifying indicators as biased, by employing an iterative bias detection procedure. In an iterative procedure, a researcher starts by including a direct effect to the indicator that improves model fit most. The choice of which direct effect to include first can be based on testing direct effects on all indicators one by one, or by inspecting modification indices from the model without any direct effects. In our example, if bias would be tested in the unbiased indicator, while the bias in

Indicator 1 was already accounted for, the factor would not be contaminated, and Indicator 2 would not be marked as biased.

In conclusion, although the test for cluster bias has several advantages, this study showed that including the presumed cause of Level 2 bias in the model to detect measurement bias is a more powerful approach than the test for cluster bias. If a researcher's goal is to investigate measurement bias with respect to (Level 1 and) Level 2 violators, we advise to follow the 5-step approach (Jak et al. in press), and test for Level 2 bias in step 5, while taking cluster bias into account. This study also showed that the power of the test for cluster bias is larger with a smaller number of clusters with a larger size, relative to more clusters with a smaller size.

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, $\mathbf{\Phi}_B$ and $\mathbf{\Phi}_W$ are covariance matrices of the common factors at the between and within level respectively, $\mathbf{\Theta}_B$ and $\mathbf{\Theta}_W$ 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, $\mathbf{\Lambda}_B$ and $\mathbf{\Lambda}_W$ 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.

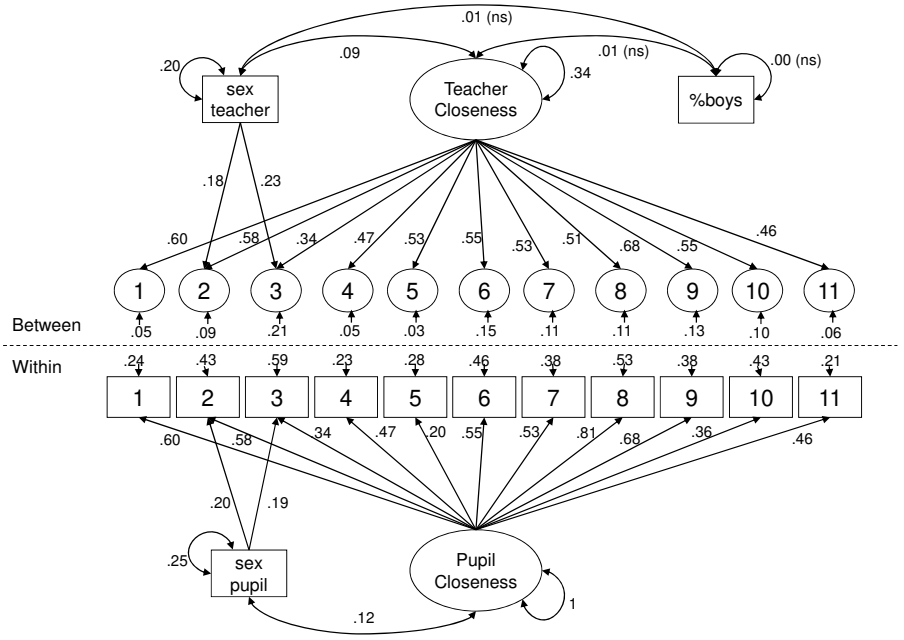
The presence of cluster bias in all closeness items shows that there are other factors than teacher's closeness with pupils that cause differences on the closeness items. Teacher sex could be one explanation for these differences.

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

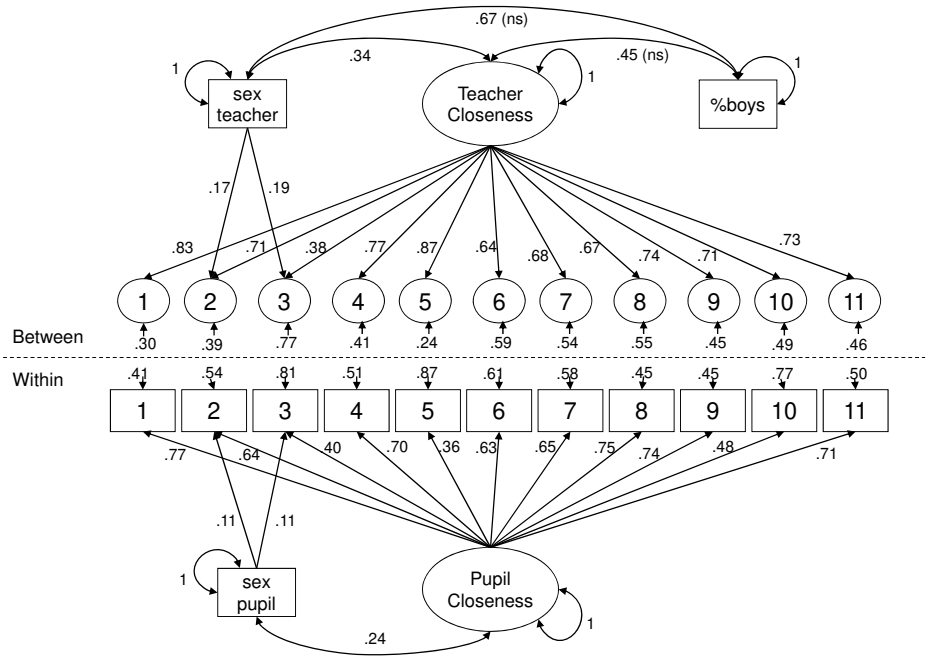


Figure 1. RFA model from Step 5. The upper figure shows the unstandardized parameter estimates, the lower figure shows the standardized parameter estimates (standardized within Level 1 and within Level 2). Non-significant parameter estimates are indicated by '(ns)'.

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.

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.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of educational measurement*, *29*, 67-91.

Asparouhov, T., & Muthén, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. *Proceedings of the 2007 Joint Statistical Meetings, Section on Statistics in Epidemiology* (pp. 2531–2535). Alexandria, VA: American Statistical Association.

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. (2012). Measurement bias detection through factor analysis. *Structural Equation Modeling, 19*, 561-579.

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

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

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley, New York.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological review*, *111*, 1061-1071.

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

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10*, 107-132.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

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

Christoffersson, A. (1975). Factor Analysis of Dichotomized Variables. *Psychometrika, 40*, 5-32.

De Jong, M. G., Steenkamp, J. B. E., & Fox, J. P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT models. *Journal of consumer research, 34*, 260-278.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*, 309-326.

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.

Dudgeon, P. (2003). NIESEM: A computer program for calculating noncentral interval estimates (and power analysis) for structural equation modeling [Computer software].

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

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

Engle, R. F. (1983). Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics. In Intriligator, M. D.; and Griliches, Z.. *Handbook of Econometrics*. Elsevier. pp. 796–801.

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.

Fox, J.-P., & Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In E. Davidov, P. Schmidt,&J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 467–488). London: Routledge Academic.

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

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

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

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129-145

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research*, *18*, 117-144.

Hox, J. (2002). *Multilevel Analysis: Techniques and Applications.* Mahwah, NJ: Erlbaum.

Hox, J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica, 64,* 157-170.

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.

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

Jak, S. & Oort, F.J. (under review). On the power of the test for cluster bias.

Jak, S., Oort, F.J. & Dolan, C.V. (2010). Measurement bias and multidimensionality; an illustration of bias detection in multidimensional measurement models. *Advances in Statistical Analysis, 94,* 129-137.

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.

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

Jak, S., Oort, F.J. & Dolan, C.V. (under review). Using two-level ordinal factor analysis to test for cluster bias in ordinal data.

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.

King-Kallimanis, B. L., Oort, F. J., & Garst, G. J. A. (2010). Using structural equation modelling to detect measurement bias and response shift in longitudinal data. *AStA Advances in Statistical Analysis*, *94*, 139-156.

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

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., Verschueren, K., van Schooten, E., Jak, S., & Pianta, R. C. (2012). 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*, *50*, 215-234.

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 coëfficiënt models.* Oxford: Clarendon Press.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* London: Addison-Wesley.

Lubke, G. H., Dolan, C. V., Kelderman, H. & Mellenbergh, G. J. (2003). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *British Journal of Mathematical and Statistical Psychology, 56,* 231–248.

Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of students' evaluations of college teaching. *American Educational Research Journal, 21*, 341-366.

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

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*, 223-236.

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

Meredith, W. & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44,* 69-77.

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

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

Molenaar, D., Dolan, C. V., & de Boeck, P. (2012). The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. *Psychometrika*, *77*, 455-478.

Molenaar, D., Dolan, C. V., & Verhelst, N. D. (2010). Testing and modelling non-normality within the one-factor model. *British Journal of Mathematical and Statistical Psychology*, *63*, 293-317.

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. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,115*-132.

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

Muthén, B. (1990). *Mean and covariance structure analysis of hierarchical data.* Los Angeles, CA: UCLA statistics series, NO. 62.

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

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*, 431-462.

Muthén, B., Khoo, S.T. & Gustafsson, J.E. (1997). Multilevel latent variable modeling in multiple populations. Unpublished technical report. Retrieved from www.statmodel.com/papers.shtml, March 2, 2011.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus Users Guide*. Fifth Edition. Los Angeles, CA: Muthén & Muthén.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*, 443-460.

Oort, F. J. (1991). Theory of violators: assessing unidimensionality of psychological measures. In: Steyer, R., Wender, K.F., Widaman, K.F. (eds.) Psychometric Methodology, pp. 377–381. Stuttgart: 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.

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, *14*, 587-598.

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

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

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing.* [Computer software]. Vienna, Austria.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika, 69*, 167-190.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics, 19,* 337–350.

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

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modelling. *Structural equation modeling, 16,* 583-601.

Satorra, A., & Bentler, P.M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*, 243-248.

Selig, J. P., Card, N. A., & Little, T. D. (2008). Latent variable structural equation modelling in cross-cultural research: Multigroup and multilevel approaches. In F. J. R. Van de Vijver, D. A. van Hemert, & Y. Poortinga (Eds.), *Individuals and cultures in Multi-level analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Smits, I. A., Dolan, C. V., Vorst, H. C., Wicherts, J. M.,& Timmerman M. E. (2011). Cohort differences in Big Five personality factors over a period of 25 years. *Journal of Personality and Social Psychology, 100(6),* 1124-1138.

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 structures between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229-239.

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

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

Spearman, C. (1928). The sub-structure of the mind. *British Journal of Psychology, 18*, 249-261.

Spilt, J., 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). Statistically based tests for the number of common factors. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

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.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*, 361-370.

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

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.

Verhagen, A. J. & Fox, J.-P. (2012). Bayesian Tests of Measurement Invariance. *British Journal of Mathematical and Statistical Psychology*. DOI: 10.1111/j.2044-8317.2012.02059.

Voorpostel, M., & Blieszner, R. (2008). Intergenerational solidarity and support between adult siblings. *Journal of Marriage and Family*, *70*, 157-167.

Webb, M. L., & Neuharth-Pritchett, S. (2010). Examining factorial validity and measurement invariance of the Student-Teacher Relationships Scale. *Early Childhood Research Quartely, 26,* 205-215.

Wei, W., Lu, H., Zhao, H., Chen, C., Dong, Q., & Zhou, X. (2012). Gender differences in children's arithmetic performance are accounted for by gender differences in language abilities. *Psychological science*, *23*, 320-330.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., Baal, G. C. M. van, Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence, 32,* 509-537.

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

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. In M.E. Sobel & M.P. Becker (eds.), *Sociological Methodology 2000*. Washington, D.C.: ASA.

SAMENVATTING (SUMMARY IN DUTCH)

In dit proefschrift worden methoden en procedures voorgesteld die gebruikt kunnen worden voor het toetsen van vraagonzuiverheid (measurement bias) in multilevel data.

Stel dat een onderzoeker geïnteresseerd is in de invloed van motivatie op rekenvaardigheid bij kinderen. Na wekenlang scholen te hebben gebeld, vindt ze 200 leraren en 700 leerlingen bereid aan haar onderzoek mee te werken. De leerlingen vullen een motivatie vragenlijst in met 10 items zoals "Ik denk dat leren rekenen goed voor me is" en "Ik vind rekenen leuk", die gescoord worden op een 7-puntsschaal van 1 (helemaal niet mee eens) tot 7 (helemaal mee eens). De kinderen maken ook een rekenvaardigheidstoets met 60 opgaven die goed of fout gemaakt kunnen worden.

Voordat de onderzoeker een hypothese kan toetsen over de relatie tussen motivatie en rekenvaardigheid, wil zij eerst weten: Zijn deze metingen valide? Resulteren verschillen in motivatie en rekenvaardigheid inderdaad in verschillen in de itemresponsen (Borsboom, Mellenbergh & van Heerden, 2004)? En meten de items dezelfde eigenschappen voor verschillende (groepen) respondenten (Mellenbergh, 1989; Meredith, 1993; Oort, 1992, 1993)? Als de rekenopgaven inderdaad hetzelfde meten voor bijvoorbeeld jongens en meisjes, dan zouden jongens en meisjes met gelijke rekenvaardigheid gemiddeld identieke test scores moeten behalen. Als dit het geval is, zijn de metingen meetinvariant ten opzichte van sekse. Als dit niet het geval is spreken we van vraagonzuiverheid. Om een voorbeeld te geven: een redactiesom zou makkelijker op te lossen kunnen zijn voor meisjes, doordat meisjes gemiddeld gezien beter kunnen lezen dan jongens (Wei et al., 2012). In dat geval zullen meisjes meer goede antwoorden op de som geven dan jongens, terwijl hun rekenvaardigheid gelijk is.

In algemenere zin is vraagonzuiverheid gedefinieerd als een schending van meetinvariantie (Mellenbergh, 1989). Stel dat item X (bijvoorbeeld de redactiesom) ontworpen is om de latente trek T (rekenvaardigheid) te meten. X is meetinvariant ten opzichte van een variabele V (bijvoorbeeld sekse) als de conditionele verdeling van X, gegeven T en V, gelijk is aan de conditionele verdeling van X, gegeven T. Met andere woorden, meetinvariantie geldt als alle invloed van de potentiele schender V op X via T loopt. Zie het figuur op pagina 19 van dit proefschrift voor een grafische weergave van vraagonzuiverheid. De twee meest gebruikte typen modellen voor het onderzoeken van meetinvariantie door middel van structural equation modeling (SEM) zijn multigroep modellen (Sörbom, 1974; Horn & McArdle, 1992; Little, 1997; Widaman & Reise, 1997) en restrictieve factoranalyse (RFA; Oort, 1992, 1998) of equivalente MIMIC (Muthén, 1989) modellen.

Een moeilijkheid is dat we geen directe maat hebben van de latente variabele waar we in geïnteresseerd zijn, zoals rekenvaardigheid of motivatie. We moeten werken met de

geobserveerde itemresponsen. De relatie tussen de geobserveerde itemresponsen en motivatie of rekenvaardigheid kan worden weergegeven in een meetmodel, zoals het lineaire factor model (Mellenbergh, 1994; Spearman, 1904, 1928). In het lineaire factor model wordt de variabele waarin we geïnteresseerd zijn weergegeven als een continue latente factor, die alle gedeelde variantie in de geobserveerde itemresponsen verklaart. Ieder item wordt ook beïnvloed door een unieke factor, die weer bestaat uit een structureel deel (dit deel zorgt voor item-specifieke variantie) en een random deel (de meetfout) (Bollen, 1989).

Het onderzoeken van vraagonzuiverheid dient altijd voorafgegaan te worden door het vinden van een correct meetmodel. **Hoofdstuk 1** van dit proefschrift dient als een introductie over vraagonzuiverheid. Door middel van twee voorbeelden uit een cognitieve vaardigheidstest lieten we zien dat vraagonzuiverheid en multidimensionaliteit nauw aan elkaar verbonden zijn. Een item dat onzuiver is, is multidimensioneel, aangezien het een dimensie meet die niet gemeten diende te worden. Als deze dimensie gerelateerd is aan potentiele schenders van meetinvariantie, (dit zijn vaak variabelen zoals sekse, etniciteit en leeftijd), dan zal dit item onzuiver blijken ten opzichte van deze variabele.

Een andere vraag die de onderzoekster uit het voorbeeld kan stellen is: Worden rekenvaardigheid en motivatie zuiver gemeten in verschillende schoolklassen? Aangezien ze data verzameld heeft van kinderen die gegroepeerd zijn in klassen, hebben de data een multilevel structuur. We kunnen in dit voorbeeld twee niveaus ("levels") onderscheiden: het kindniveau noemen we Niveau 1 en het klasniveau noemen we Niveau 2. Met multilevel SEM kunnen we modellen specificeren op verschillende niveaus van de multilevel data. Kinderen die in dezelfde klas zitten delen kenmerken op klasniveau, zoals de leraar, de samenstelling van de klas en de grootte van de klas. Verschillen in deze kenmerken kunnen leiden tot verschillen in de gemiddelde testscores van kinderen uit verschillende klassen, die niet verklaard worden door de factoren rekenvaardigheid of motivatie. In **Hoofdstuk 2** van dit proefschrift stellen we een toets voor die gebruikt kan worden om te toetsen of metingen onzuiver zijn ten opzichte van schoolklas. Deze toets is algemeen geschikt om onzuiverheid ten opzichte van de clusterende variabele in multilevel data te onderzoeken (bijvoorbeeld bij data van mensen in landen, patiënten in ziekenhuizen, kinderen in families, etc.), vandaar de naam "toets voor clusteronzuiverheid". Clusteronzuiverheid kan onderzocht worden door te toetsen of de factor ladingen op Niveau 1 gelijk zijn aan de factor ladingen op Niveau 2, en of de residuele varianties op Niveau 2 nul zijn. De toets wordt geïllustreerd met data uit onderwijskundig onderzoek. Daarnaast laten we in een simulatie onderzoek zien dat met continue data afkomstig van vijf items, en een groot genoeg aantal clusters, de likelihood ratio test genoeg statistische power heeft om clusteronzuiverheid te ontdekken. Met 50 clusters van 25 observaties per cluster is de power voldoende als de onzuiverheid 3% of meer van de totale variantie van de indicator veroorzaakt. Met slechts 20 clusters met ieder 25 observaties is de power om

clusteronzuiverheid te ontdekken voldoende als de onzuiverheid zorgt voor meer dan 5% van de totale variantie. De proporties vals negatieven waren hoger dan het gekozen significantieniveau in de condities met 100 clusters, maar lager in de condities met 20 clusters.

In **Hoofdstuk 3** breiden we de test voor clusteronzuiverheid uit naar ordinale itemresponsen, met behulp van het ordinale twee-niveau factor model (Grilli & Rampichini, 2007). Op basis van een simulatie onderzoek concluderen we dat clusteronzuiverheid in ordinale data getoetst kan worden met de likelihood ratio test en met de Wald test. Beide tests hebben voldoende power om aanzienlijke hoeveelheden onzuiverheid te detecteren, en hebben acceptabele proporties vals negatieve resultaten. De geschaalde likelihood ratio test, zoals geïmplementeerd in het programma Mplus, wordt niet aangeraden voor het toetsen van clusteronzuiverheid, aangezien in alle condities substantiële aantallen ontoelaatbare resultaten werden gevonden. De voorgestelde toetsen worden geïllustreerd met data over leerkracht-leerling relaties.

De oorzaak van clusteronzuiverheid is per definitie een variabele op clusterniveau. In het voorbeeld van kinderen in klassen, kan de oorzaak van clusteronzuiverheid liggen in onzuiverheid ten opzichte van de didactische kwaliteiten van de leraar. Om de toets voor clusteronzuiverheid toe te passen, hoeft de werkelijke schender van meetinvariantie (als die er is) niet gemeten te zijn. De toets voor clusteronzuiverheid kan daarom gebruikt worden als een algemene toets voor onzuiverheid ten opzichte van alle mogelijke variabelen op clusterniveau. In **Hoofdstuk 4** vergelijken we de power en de proporties vals positieven van de toets voor clusteronzuiverheid en de RFA-toets. Zoals verwacht heeft de RFA-toets meer power dan de toets voor clusteronzuiverheid. De toets voor clusteronzuiverheid heeft in het algemeen een kleinere hoeveelheid vals positieve resultaten. We concluderen dat het niet vinden van clusteronzuiverheid niet uitsluit dat er significante onzuiverheid ten opzichte van een Niveau 2-variabele gevonden wordt met de RFA-toets.

In het laatste hoofdstuk van dit proefschrift, **Hoofdstuk 5**, stellen we een stapsgewijze aanpak voor om vraagonzuiverheid te onderzoeken ten opzichte van een schender op Niveau 1, een schender op Niveau 2, en de cluster variabele. In deze procedure is de eerste stap het toetsen of multilevel-analyse werkelijk nodig is, Stap 2 is het vinden van een geschikt meetmodel op Niveau 1, Stap 3 is het toetsen van vraagonzuiverheid op Niveau 1, Stap 4 is het toetsen op clusteronzuiverheid, en in Stap 5 wordt de mogelijkheid getoetst dat de clusteronzuiverheid verklaard wordt door Niveau 2-variabelen. De vijf stappen zijn zo ontworpen dat het uiteindelijke model alle bias en inhoudelijke resultaten op beide niveaus laat zien.

Dit proefschrift biedt onderzoekers een methode om op een uitvoerbare en systematische manier vraagonzuiverheid te toetsen in multilevel data.

DANKWOORD

Ik weet nog precies waar ik zat (rechts vooraan, tweede stoel) toen Conor Dolan bij het vak Latente Variabele Modellen uitlegde wat meetinvariantie is. Als ik het overdreven stel, zou ik zeggen dat ik toen op slag verliefd werd op het begrip meetinvariantie. Heel erg blij was ik dan ook, toen ik bij Frans Oort mocht gaan promoveren op het onderwerp, en grotendeels zelf mocht bepalen welke kant ik ermee op wilde. Met het onderwerp was er dus direct een warme band, één van de belangrijkste voorwaarden voor een geslaagd promotietraject. Misschien wel even belangrijk zijn de mensen met wie je werkt, en ook daar ben ik buitengewoon gelukkig mee geweest.

Ten eerste natuurlijk Frans Oort, een betere begeleider en promotor had ik mij niet kunnen wensen. Uit respect voor al zijn kennis en ter onderschrijving van de leraar-leerling relatie heb ik Frans altijd met u aangesproken. Toen hij na enkele maanden vroeg wanneer ik daar eens mee op zou houden antwoordde ik: "als ik gepromoveerd ben". Met het voorbijgaan van de jaren werd de band hechter en, hoewel ik het inmiddels gewend was, vonden collega's het erg vreemd dat ik Frans niet gewoon tutoyeerde. Nu is het zover, en zal ik voortaan gewoon jij proberen te zeggen. Frans, dankjewel voor alles wat ik van je geleerd heb, over SEM en measurement bias, maar ook in bredere zin. Ik ben nog lang niet uitgeleerd, dus ik hoop dat we nog regelmatig samen zullen werken.

Zoals gezegd begon de wens om onderzoek te doen bij de colleges van Conor Dolan, zonder hem had dit proefschrift waarschijnlijk niet bestaan (althans, niet met mij als auteur). Behulpzaam en aardig als hij is, bleef hij tijdens mijn project altijd betrokken bij het onderzoek, wat een rol als mede-promotor vanzelfsprekend maakte. Conor, dankjewel voor je interesse en je snelle en scherpe commentaar op mijn werk. Ook heel veel dank voor het aan mij uitlenen van je spiksplinternieuwe computer. Vers uit de verpakking, voor je zelf ook maar een analyse had gedaan, leende je jouw 8 processoren aan mij uit voor het draaien van de simulaties in hoofdstuk 3. Het lijkt mij tekenend voor jouw onbaatzuchtigheid.

Erik Thoonen, Helma Koomen en stagebedrijf Meurs HRM wil ik bedanken voor het delen van hun data, zodat ik de voorgestelde methoden met echte data kon illustreren. Helma, jij ook erg bedankt voor de fijne samenwerking bij verschillende artikelen buiten dit proefschrift om.

Hoewel ik oorspronkelijk de enige methoden en technieken-aio bij de afdeling pedagogiek en onderwijskunde was, heb ik mij dankzij de vele leuke collega's gelukkig nooit alleen gevoeld. Ik bedank hiervoor Lisette en Maren, kamergenoten van het eerste uur, en mijn latere kamergenoten: Rudy, Annemarie, Harry, Bonne, Hulya, Lisa, Ilona, Annette, Ed en Mathilde. Verder heb ik genoten van SEMmen, vlammetjes, schrijfweek, pubquiz of biertjes met Madelon, Britt, Marloes, Elsje, Debora, Bettina, Jaap, Marjolein, Bellinda en Mariska.

Het schrijven van dit dankwoord dwingt me terug te blikken op de afgelopen vier jaar. Ik kan er daarbij niet omheen dat parallel aan het fijne werk op de UvA, mijn persoonlijke leven zijn dieptepunt bereikte. Mijn lieve vader en moeder zijn overleden tijdens het werken aan dit proefschrift. Hierdoor heb ik bijzonder sterk ervaren wat een fijne mensen ik om mij heen heb. Hoewel onze band losstaat van dit proefschrift, maak ik van de gelegenheid gebruik om zwart op wit mijn dank aan jullie te betuigen.

Ebba, Eva en Frédérique, dank voor jullie jarenlange vriendschap. Sinds we twaalf jaar oud waren zijn wij altijd nauw bij elkaars leven betrokken geweest, en ik ben er van overtuigd dat dit altijd zo zal blijven. Alleen die wetenschap is al genoeg om het leven aan te kunnen. Anke heeft mij onder andere opgebeurd door haar baby Olivier regelmatig in mijn armen te drukken en te beweren dat ze een troostbaby voor mij had gemaakt. Ook Sander is regelmatig koffie komen drinken met baby Maud, waarna het leven altijd lichter leek. Joost en Evelyne wil ik bedanken voor het zijn van hele lieve vrienden. Verder is mijn leven mede leuk gemaakt door de dames van Het Y, want er is niets zuiverender dan lekker waterpoloën en dan doorzakken in de Y-kelder.

Nu onze ouders er niet meer zijn, wordt ons kerngezin gevormd door mijn zus Martine, mijn broers Remco en Wouter, en mijzelf. Gelukkig vinden wij elkaar allemaal fantastisch, en blijken we dat ook in de zwaarste omstandigheden nog te vinden. Ik wil jullie bedanken voor het zijn van de leukste broers en zus van de wereld. Wij, kinderen Jak, zouden het echter moeilijk redden zonder onze wederhelften, Daan, Barbara, Elien en Louise. Jullie wil ik oneindig bedanken voor jullie zorg en toewijding de afgelopen tijd. Ik vergeet alles om mij heen, als ik in de buurt ben van mijn lieve neefjes en nichtjes: Sietse, Mette, Anna, Simon en Jelle, dankjewel dat jullie er zijn. Mijn schoonfamilie wil ik bedanken voor de warmte waarmee ik in de familie verwelkomd ben.

Beste collega Elffers, lieve Louise, van een vage bekende aan de overkant in de catacombe, werd je via New York, G0.04 en de Malediven, afgelopen voorjaar mijn vrouw. Jij bent het happy end van dit verhaal. En we leefden nog lang en gelukkig.