Cluster bias: Testing measurement invariance in multilevel data

Jak, S.

Publication date
2013

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.

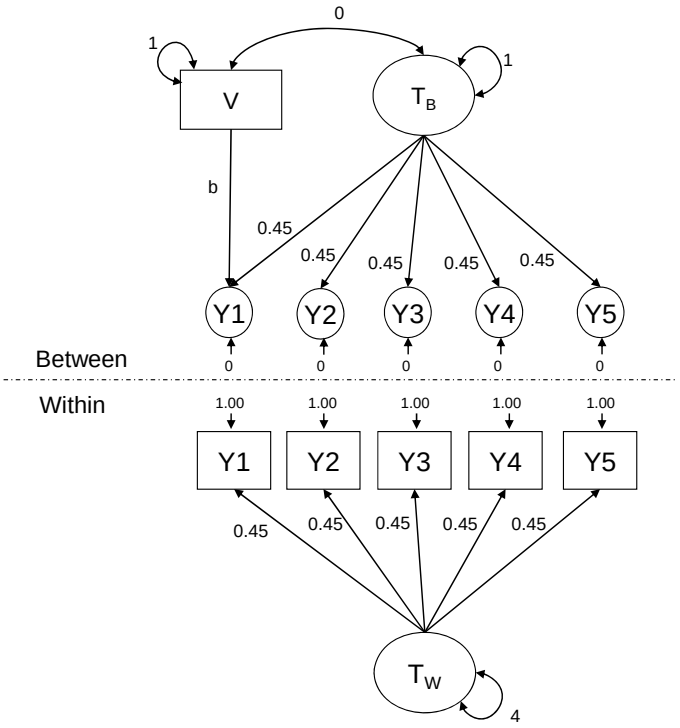

Figure 1. Two-level measurement model with population parameter values. In conditions with 0, 1, and 5 % bias, the corresponding values for *b* were 0, .142, and .324 respectively.

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.

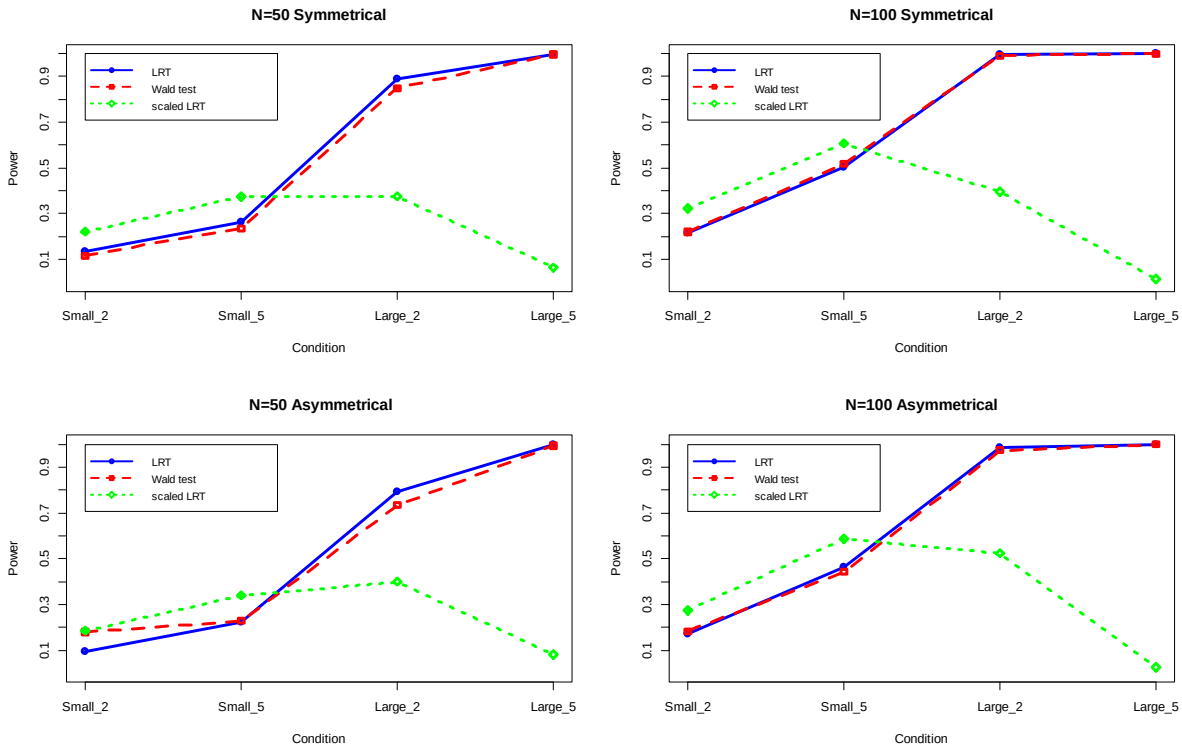

Figure 2a. A comparison of the power of the LRT, the Wald test, and the scaled LRT with alpha = .05 in different conditions.

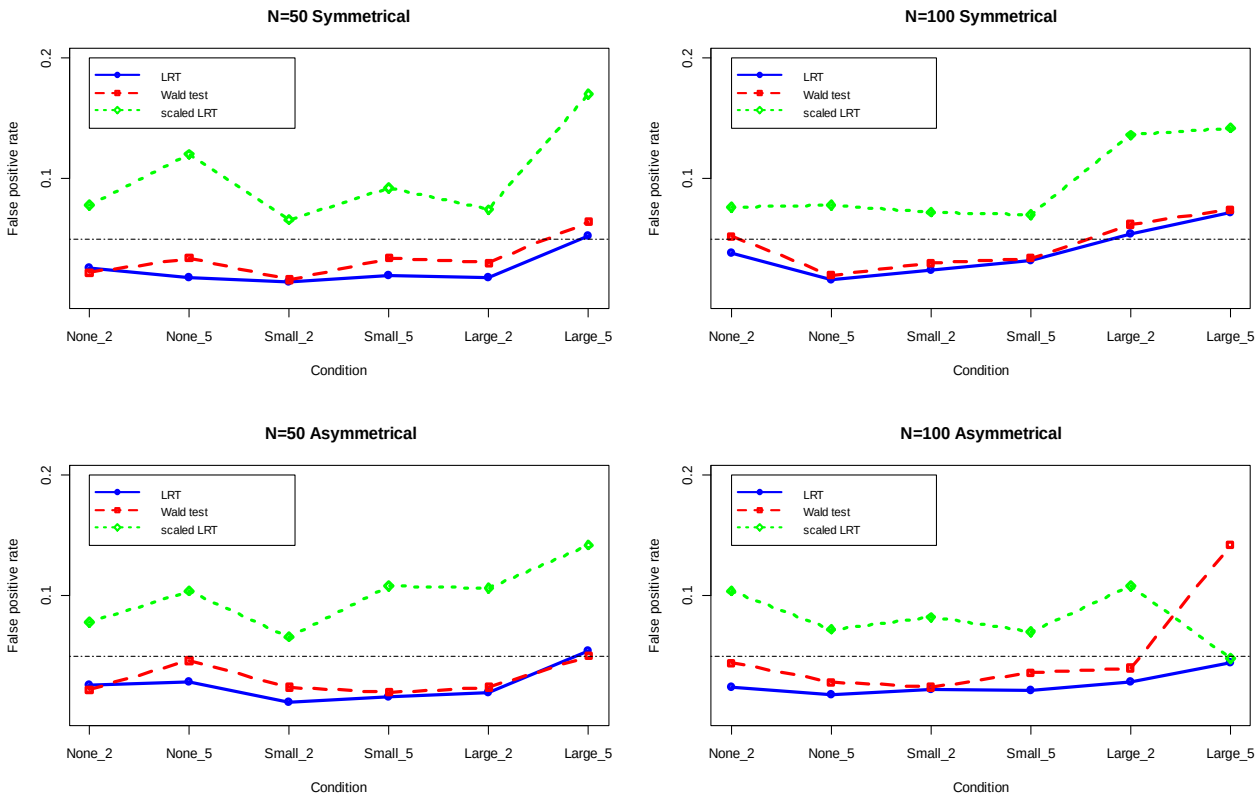

Figure 2b. A comparison of the false positive rate of testing cluster bias in Item 2 (unbiased item), while the cluster bias is in Item 1 (in the conditions with bias) of the LRT, the Wald test, and the scaled LRT with alpha = .05 in different conditions. The straight dotted lines denote the nominal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