Using Two-Level Factor Analysis to Test for Cluster Bias in Ordinal Data

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

DOI
10.1080/00273171.2014.947353
Publication date
2014
Document Version
Final published version
Published in
Multivariate Behavioral Research

Routledge
Taylor & Francis Group

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: http://www.tandfonline.com/loi/hmbr20

Using Two-Level Factor Analysis to Test for Cluster Bias in Ordinal Data

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

To cite this article: Suzanne Jak, Frans J. Oort & Conor V. Dolan (2014) Using Two-Level Factor Analysis to Test for Cluster Bias in Ordinal Data, Multivariate Behavioral Research, 49:6, 544-553, DOI: 10.1080/00273171.2014.947353

To link to this article: https://doi.org/10.1080/00273171.2014.947353

Multivariate Behavioral Research, 49:544–553, 2014

ISSN: 0027-3171 print / 1532-7906 online
DOI: 10.1080/00273171.2014.947353

Routledge
Taylor & Francis Group

Using Two-Level Factor Analysis to Test for Cluster Bias in Ordinal Data

Suzanne Jak
Department of Social Sciences–Methodology and Statistics, Utrecht University

Frans J. Oort
Department of Education, University of Amsterdam

Conor V. Dolan
Department of Biological Psychology, VU University

The test for cluster bias is a test of measurement invariance across clusters in 2-level data. This article 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 nonnormality of the data gives untrustworthy results, whereas the unscaled LRT and the Wald test have acceptable false positive rates and perform well in terms of empirical power rate if the amount of cluster bias is large. The test for cluster bias is illustrated with data from research on teacher-student relations.

If data have a two-level structure, as is the case with data from students in school classes, it may be 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 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, as implicated by the definition of Mellenbergh (1989). In the present study, the factor model is the measurement model of interest (Mellenbergh, 1994). In this case, measurement invariance is often called *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). Following the terminology of Widaman and Reise (1997), we distinguish the following forms of invariance: *configural invariance,* comprising equal factor structure 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. If one treats the clustering variable as fixed, multigroup factor analysis is an obvious choice to investigate measurement bias. If the clusters are treated as a sample from a population of clusters, random effects modeling is suitable. To test strong factorial invariance across clusters in multilevel data, the test for cluster bias can be used (Jak, Oort, & Dolan, 2013). 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 multilevel

Correspondence concerning this article should be addressed to Suzanne Jak, Utrecht University, Padualaan 14, 3584 CH, Utrecht, The Netherlands. E-mail: S.Jak@uu.nl

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hmbr.

structural equation modeling enables the investigation of the possible causes of cluster bias by testing the significance of the direct effects of potential causes on the observed indicators (assuming the potential causes have been measured). Statistical methods to investigate measurement bias across clusters in continuous data have been developed (B. Muthén, 1990; Rabe-Hesketh, Skrondal, & Pickles, 2004) and have been found to perform well with continuous item responses (Jak et al., 2013). As in educational and psychological testing, item responses are often ordinal, for example, 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 categorical data as well. The purpose of this article therefore is 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 ORDINAL DATA WITH THE TWO-LEVEL FACTOR MODEL

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 $\{1, 2, \ldots, C_p\}$ if a certain threshold on the underlying variable y_{pij}^* is passed (see Lord & Novick, 1968; B. Muthén, 1984; Olsson, 1979; Christoffersson, 1975). For example, given a variable with five response options, there are four threshold parameters τ, such that

$$y_{pij} = \begin{cases} 1 & \text{if} \quad y_{pij}* \leq \tau_1 \\ 2 & \text{if} \quad \tau_1 < y_{pij}* \leq \tau_2 \\ 3 & \text{if} \quad \tau_2 < y_{pij}* \leq \tau_3 \\ 4 & \text{if} \quad \tau_3 < y_{pij}* \leq \tau_4 \\ 5 & \text{if} \quad y_{pij}* > \tau_4. \end{cases} \tag{1}$$

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 ($\boldsymbol{\eta}_{ij}$):

$$\boldsymbol{y}_{ij}* = \boldsymbol{\mu}_j + \boldsymbol{\eta}_{ij}. \tag{2}$$

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 $\boldsymbol{\eta}_{ij}$ ($\boldsymbol{\Sigma}_{\text{WITHIN}}$):

$$\boldsymbol{\Sigma}_{\text{TOTAL}} = \boldsymbol{\Sigma}_{\text{BETWEEN}} + \boldsymbol{\Sigma}_{\text{WITHIN}}. \tag{3}$$

Any structural equation model can be fitted to the within and between level covariance matrices. We consider a two-level factor model for p observed variables and k common factors:

$$\begin{aligned} \boldsymbol{\Sigma}_{\text{BETWEEN}} &= \boldsymbol{\Lambda}_{\text{BETWEEN}} \boldsymbol{\Phi}_{\text{BETWEEN}} \boldsymbol{\Lambda}_{\text{BETWEEN}}{}' \\ &\quad + \boldsymbol{\Theta}_{\text{BETWEEN}}, \\ \boldsymbol{\Sigma}_{\text{WITHIN}} &= \boldsymbol{\Lambda}_{\text{WITHIN}} \boldsymbol{\Phi}_{\text{WITHIN}} \boldsymbol{\Lambda}_{\text{WITHIN}}{}' + \boldsymbol{\Theta}_{\text{WITHIN}}, \end{aligned} \tag{4}$$

where $\boldsymbol{\Phi}_{\text{BETWEEN}}$ and $\boldsymbol{\Phi}_{\text{WITHIN}}$ are k by k covariance matrices of common factors, $\boldsymbol{\Theta}_{\text{BETWEEN}}$ 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 Expectation–Maximization algorithm using adaptive numerical quadrature (denoted by robust maximum likelihood [MLR] estimation in Mplus; L. K. 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 et al. (2013), strong factorial invariance across clusters implies

$$\begin{aligned} \boldsymbol{\Sigma}_{\text{BETWEEN}} &= \boldsymbol{\Lambda} \boldsymbol{\Phi}_{\text{BETWEEN}} \boldsymbol{\Lambda}', \quad \text{and} \\ \boldsymbol{\Sigma}_{\text{WITHIN}} &= \boldsymbol{\Lambda} \boldsymbol{\Phi}_{\text{WITHIN}} \boldsymbol{\Lambda}' + \boldsymbol{\Theta}_{\text{WITHIN}}. \end{aligned} \tag{5}$$

If there is no cluster bias, the factor loadings are equal across levels, and there is no residual variance at the between level. If the factor loadings are not equal across levels, the common factors do not have the same interpretation across levels (B. Muthén, 1990; Rabe-Hesketh et al., 2004). The test for cluster bias addresses strong factorial invariance across clusters by testing the significance of cluster level residual variance in a factor model with equal factor loadings across levels. See Table 1 for an overview of the implications in two-level models at the various levels of factorial invariance in a multigroup model. The cluster bias model can test whether strong invariance holds (by setting $\boldsymbol{\Lambda}_{\text{WITHIN}} = \boldsymbol{\Lambda}_{\text{BETWEEN}}$ and testing $\boldsymbol{\Theta}_{\text{BETWEEN}} = 0$) but cannot differentiate between violations of weak and strong factorial invariance. If $\boldsymbol{\Theta}_{\text{BETWEEN}} \neq 0$, this can also be caused by a difference in factor loadings across clusters, which is a violation of weak factorial invariance (Jak et al., 2013). Also, there is no additional test for strict factorial invariance across clusters. In the current simulation study we focus on testing intercept differences in situations where factor loadings are equal across clusters. Hereafter, when we use the term "test for cluster bias," we mean a test of strong factorial invariance across clusters in a two-level model.

TABLE 1
Comparison of the Restrictions in a Multigroup Model and the Implications in a Two-Level Model With Different Levels of Factorial Invariance

Level of Factorial Invariance	Restrictions in Multigroup Model	Implications in Two-Level Model
Configural	pattern($\boldsymbol{\Lambda}_g$) = pattern($\boldsymbol{\Lambda}$)	—
Weak	$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$	$\boldsymbol{\Lambda}_{WITHIN} = \boldsymbol{\Lambda}_{BETWEEN}$
Strong	$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$, $\boldsymbol{\nu}_g = \boldsymbol{\nu}$	$\boldsymbol{\Lambda}_{WITHIN} = \boldsymbol{\Lambda}_{BETWEEN}$, $\boldsymbol{\Theta}_{BETWEEN} = 0$
Strict	$\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$, $\boldsymbol{\nu}_g = \boldsymbol{\nu}$, $\boldsymbol{\Theta}_g = \boldsymbol{\Theta}$	—

Note. ν is a p-dimensional vector of intercepts. Subscript g is used for group/cluster.

Jak et al. (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 use simulated data 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. Population values for the parameter estimates are proportional to the values used by Jak et al. (2013). All parameter values were rescaled so the residuals at the within level follow a standard normal distribution, which matches the implementation of multilevel categorical data analysis using the probit link function in Mplus (L. K. Muthén & Muthén, 2007). 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 nonzero 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 (within + between) variance was common variance and 50% was residual variance (calculated using population values from Figure 1 as $(\theta_{WITHIN} + \theta_{BETWEEN}) / (\lambda_{WITHIN}{}^2 * \varphi_{WITHIN} + \theta_{WITHIN} + \lambda_{BETWEEN}{}^2 * \varphi_{BETWEEN} + \theta_{BETWEEN}) = (1 + 0) / (.45^2 * 4 + 1 + .45^2 * 1 + 0) = 1 / 2.01 = .50$). 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 ($b = .142$), or large, contributing 5% of the total variance ($b = .324$). We considered conditions with 100 clusters of 25 (total sample size is 100*25 = 2,500) and conditions with 50 clusters of 25 (total sample size is 50*25 = 1,250). We categorized the continuous normal data into two or five 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 (two or five), and frequency distribution (symmetrical or asymmetrical) yielded 3*2*2*2 = 24 conditions. We generated 500 samples per condition.

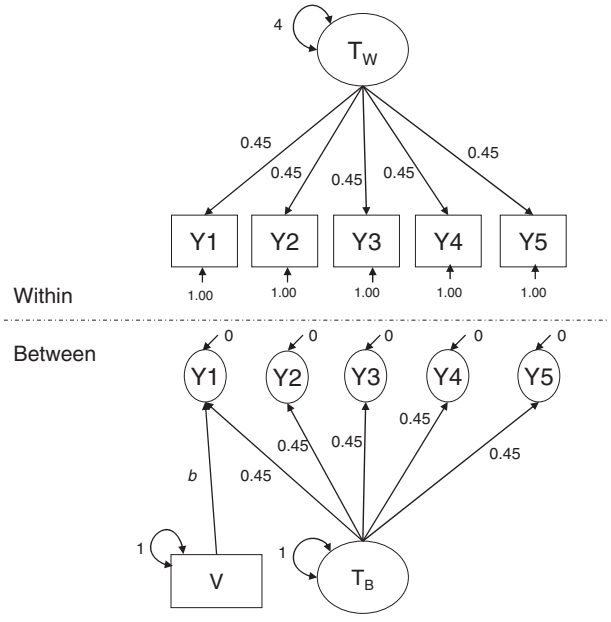


FIGURE 1 Two-level measurement model with population parameter values. In conditions with 0, 1, and 5% bias, the corresponding values for b were 0, .142, and .324, respectively.

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 model:

$$\mu_{ij} = \nu_i + \lambda_i T_j + b\, V_j, \quad (6)$$

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, ν_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 Σ_{WITHIN} that is calculated as $\Sigma_{WITHIN} = \mathbf{\Lambda\Phi}_{WITHIN}\ \mathbf{\Lambda}' + \mathbf{\Theta}_{WITHIN}$ (see Equation 1). 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. For example, a symmetric continuous response variable with large (5%) bias has a variance of 2.12 ($2.01 + b^2 = 2.01 + .324^2 = 2.12$). Greater variance leads to bigger tails in the continuous distribution and slightly more scores in the extreme categories of the categorical distribution, leading to population proportions of .11, .20, .39, .20, .11.

Analysis

We used robust maximum likelihood (MLR) estimation with the probit link function in Mplus (L. K. Muthén & Muthén, 2007) to fit the models to the generated data sets. MLR estimation of the parameters with multilevel ordinal data 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 (LRT), the likelihood ratio test with a correction factor (scaled LRT; Satorra & Bentler, 2001), and of the univariate Wald test.

Likelihood Ratio Test

The LRT can be used to test the difference in fit between two nested models. The test statistic is calculated as −2 times the difference in the log-likelihood of two models. The difference in −2 times the log-likelihoods is asymptotically chi-square distributed with degrees of freedom equal to the difference in the number of parameters of the two models. The distribution is central chi-square if the more parsimonious model is true and noncentral chi-square otherwise.

Scaled LRT

Mplus gives correction factors to compute a scaled LRT when using MLR estimation. The scaling factors are estimated on the basis of the multivariate kurtosis in the observed data and are used to correct the LRT for nonnormality. The scaled LRT statistic can be calculated using the log-likelihood of the null model (L_0) and the alternative model (L_1), the number of parameters in the null model (p_0) and alternative model (p_1), and the scaling factors in the null and alternative model (c_0 and c_1, respectively) using the formula

$$\text{Scaled LRT statistic} = -2 * (L_0 - L_1)/\text{cd}, \text{ where}$$

$$\text{cd} = (p_0 * c_0 - p_1 * c_1)/(p_0 - p_1) \quad (7)$$

(see http://www.statmodel.com/chidiff.shtml).

Wald Test

With the Wald test we test whether the parameter equals a hypothesized value. If this value is zero, the Wald statistic is the parameter estimate divided by its standard error. With MLR estimation, Mplus provides standard errors based on a sandwich estimator (Huber–White sandwich estimators; Huber, 1967; White, 1982), which are robust to nonnormality. The robust standard errors are the only standard errors available in Mplus and therefore are used in this simulation study.

Expectations

As in our simulation study all continuous variables underlying the discrete item responses were normally distributed, we expect the unscaled LRT to perform better than the scaled LRT in line with the findings by Hox, Maas, and Brinkhuis (2010) with continuous data. In two-level data the largest sample size is at the within level, therefore the LRT has more power to detect misfit at the within level than at the between level. Ryu and West (2009) presented some level-specific test statistics for two-level data. These measures are unfortunately not suitable to test for cluster bias because the level-specific measures require a saturated model at one of the two levels, whereas the test for cluster bias requires equality constraints on the factor structure across levels.

As in this study we only detect misfit (due to nonzero residual variance) at the between level, we expected that the

TABLE 2
Proportions of True Positives and Problems for All Conditions, With $\alpha = .05$ and $\alpha = .10$

Condition				Power $\alpha = .05$			Power $\alpha = .10$			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 cluster bias; Cat. = number of response categories; LRT = likelihood ratio test; 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. Boldface indicates power larger than 80.

power of the LRT, the scaled LRT, and the Wald test increases with the number of clusters and with the size of the bias.

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), and
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 LRTs are associated with a significant difference in the likelihoods of Model 0 and Model 1 in conditions with bias, given the level of significance. As in Model 1 there is one variance parameter estimated more than in Model 0: the test has one degree of freedom. The false positives of the LRTs are indicated by a significant difference in fit between Model 0 and Model 2 (also a test with one degree of freedom). Two types of false positive results can be distinguished. In conditions without cluster bias, a significant difference between Model 0 and Model 2 indicates a false positive result. The proportion of these false positives across simulated data sets is expected to equal the alpha level. In conditions with bias in Item 1, a significant difference in fit between Model 0 and Model 2 indicates a false positive test with a misspecified model of which the proportion across simulated data sets is expected to be higher than the alpha level.

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, that is, by testing the significance of the Level 2 residual variance of Item 2, whereas there is cluster bias in Item 1.

All tests are univariate, that is, we test the presence of nonzero residual variance one at a time. We chose univariate tests because with discrete multilevel data it is advisable to keep models as restricted as possible to avoid estimation problems. Each freely estimated residual variance at the between level requires one dimension of integration, and with too many dimensions the estimation process breaks down (Asparouhov & Muthén, 2007).

Results

In Table 2 and Table 3, the true and false positive rates in all conditions are shown for three tests: the LRT, the 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 produced 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 indicates a problem with this test. The last three columns of Table 2 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

TABLE 3
Proportions of False Positives and Problems for All Conditions, With $\alpha = .05$ and $\alpha = .10$

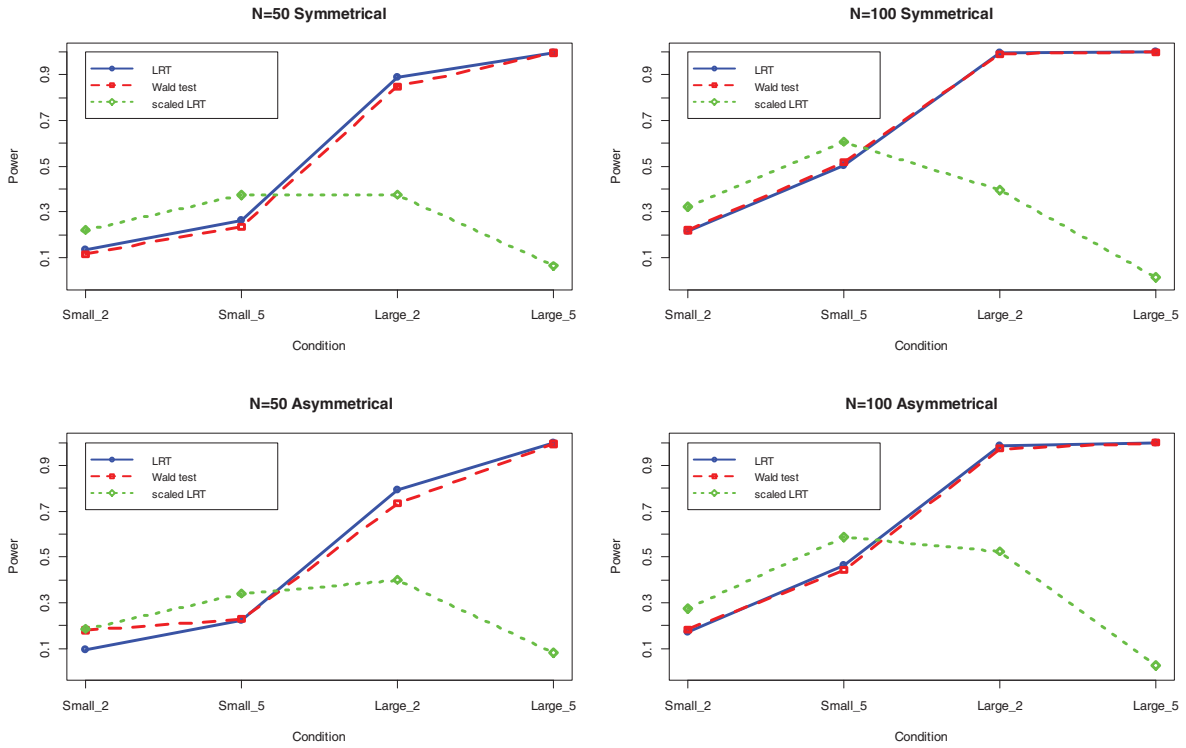
Condition				False Positives $\alpha = .05$			False Positives $\alpha = .10$			Problems		
	N	Size	Cat.	LRT	Scaled LRT	Wald	LRT	Scaled LRT	Wald	Negative LRT	Negative Scaled LRT	Incorrect*SE*s
Symmetrical	50	None	2	.026	.078	.022	.044	.122	.078	.548	.490	.092
			5	.018	.120	.034	.050	.144	.122	.624	.574	.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 cluster bias; Cat. = number of response categories; LRT = likelihood ratio test; 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.

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. This is in line with simulation results of testing latent interactions by Cham, West, Ma, and Aiken (2012), who found inflated Type I error rates for the scaled LRT in some conditions. As the scaled LRT therefore is of limited use in testing for cluster bias, we focus on 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 bias was small or absent. In these situations, the log-likelihood of the least constrained model might not have been at its true maximum, yielding a –2 log-likelihood that is higher than the –2 log-likelihood of the constrained model. These computational problems are known to be present when the multilevel model is complex (Grilli & Rampichini, 2007). Problems with the Wald test concerned untrustworthy standard errors (indicated by an Mplus warning about nonpositive definiteness of the first-order derivative product matrix). These problems, although 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 2) in all conditions where 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. In conditions with small bias, the power with five response options was around twice as large as the power in the conditions with two response options. A larger number of clusters also contributed positively to the statistical power. Figure 2b shows the false positive rates for the three tests with alpha = .05. Whereas the LRT and the Wald test yield around 5% false positive rates in all conditions, the scaled LRT always yields around 10% false positive rate. As expected, 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 were higher if the size of cluster bias was large, particularly with the Wald test. So, the larger the misspecification of the model, the larger the false positive rates. The highest false positive rates were found in the symmetrical condition with large bias and 100 clusters. Asymmetry of

(a) Power to detect cluster bias in Item 1 (biased item).



(b) False positive rates of testing cluster bias in Item 2 (unbiased item), while the cluster bias is in Item 1 (in the conditions with bias).

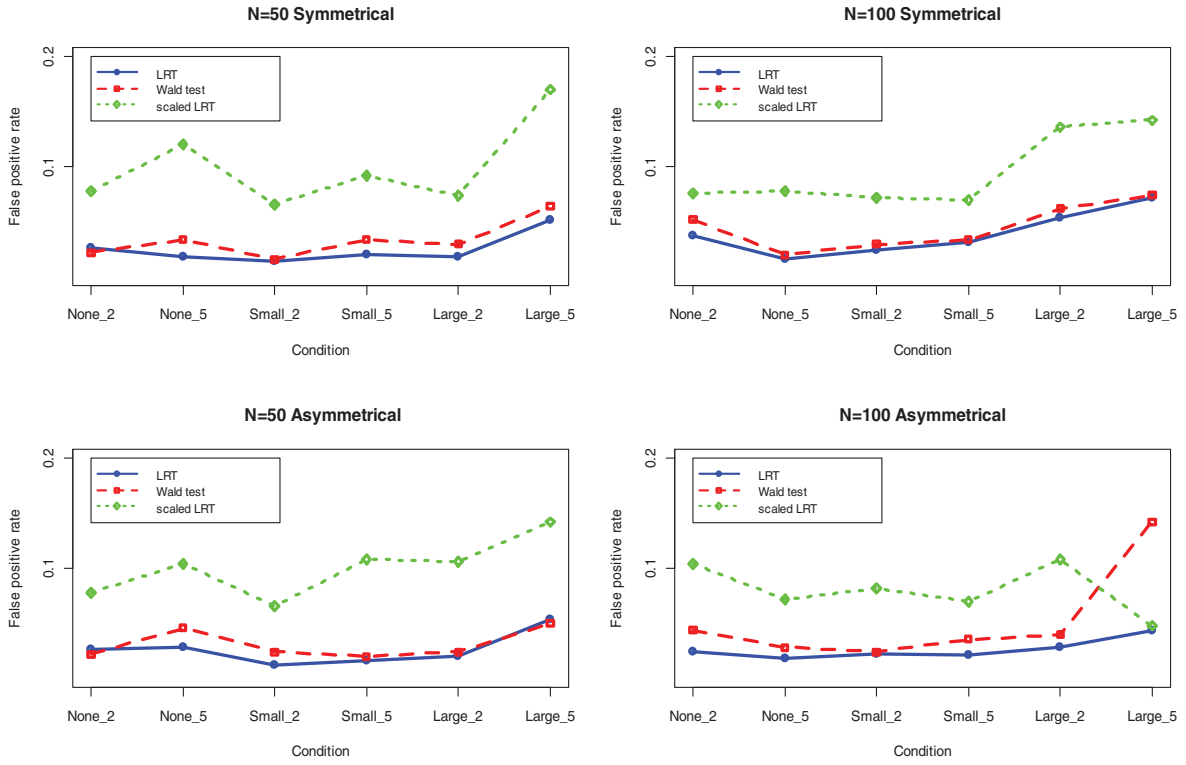


FIGURE 2 A comparison of the power and false positive rate of the LRT, the Wald test, and the scaled LRT with alpha = .05 in different conditions. *Note.* LRT = Likelihood ratio test; None_2: Condition without bias and 2 response options; None_5: Condition without bias and 5 response options; Small_2: 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.

TABLE 4
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[a]	Proportion Bias Total[b]
0. Invariance	25168.90	1.232					
1. $\theta_{\text{BETWEEN},11}$	25035.44	1.208	287.66	133.46	5.933	.584	.333
2. $\theta_{\text{BETWEEN},22}$	25148.30	1.218	26.29	20.61	2.966	.190	.091
3. $\theta_{\text{BETWEEN},33}$	25157.46	1.229	10.08	11.45	2.232	.231	.098
4. $\theta_{\text{BETWEEN},44}$	25142.84	1.212	44.03	26.07	3.798	.348	.166
5. $\theta_{\text{BETWEEN},55}$	25075.82	1.201	387.84	93.08	5.656	.401	.209
6. $\theta_{\text{BETWEEN},66}$	25101.42	1.207	156.24	67.50	4.530	.341	.166

[a]Calculated as residual variance at Level 2/total variance at Level 2. [b]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 overdependent 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.

the response distribution did not substantially affect power or false positive rate.

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 (Koomen, Verschueren, & Pianta, 2007; Pianta, 2001). The scale comprises six items. Dependency refers to overdependent and clingy child behavior. The dependency items are given in the note to Table 3. Data of 1,493 students were gathered from 659 primary school teachers (182 men, 477 women) from 92 regular elementary schools. We use teacher as the clustering variable. Each teacher reported on 2 or 3 students. One hundred eighty-two male teachers reported on 242 boys and 227 girls; 477 female teachers reported on 463 boys and 561 girls. The children were in Grades 1 through 6. Responses were given on a 5-point scale ranging from 1 (*definitely does not apply*) to 5 (*definitely does apply*).

Statistical Analysis

Results from a two-level model are reported (students nested in teachers) because school-level variance was very low (intraclass correlations .02–.06). In earlier research, treating the responses as continuous outcomes, a one-factor model was found to fit the item responses adequately (Spilt, Koomen, & Jak, 2012). We use a one-factor model with cluster invariance restrictions (see Equation 1) 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 a familywise alpha of .05. We used the one-sided test because we were testing the significance of a variance, which cannot have values below zero (Stoel, Garre, Dolan, & van den Wittenboer, 2006).

Results

Table 4 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 LRT, the chi-square value associated with the scaled LRT, and the Wald statistic. For the Wald statistic, we test against a critical value of 2.39 (i.e., the z value associated with a one-sided alpha level of .05) divided by the number of tests to be performed (six). For the LRTs, the critical value was 6.96 (i.e., the chi-square value associated with a one-sided alpha level of .05/6). Table 4 shows that all chi-square values were larger than this critical value, so, according to the LRTs, 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 residual variance caused by between-level factors other than the common factor 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, that is, 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.

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 sufficient 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 size of the bias.

In the data from the illustration, the clusters were smaller than in the simulation study (average cluster size was around 3 in the illustration and 25 in the simulation). Jak and Oort (under review) conducted a simulation study to evaluate the test for cluster bias with varying cluster sizes with continuous data. They found that false positive rates were not affected by cluster size, but the power increased with increasing cluster size. Keeping the overall sample size constant, the power to detect small bias was .50 with 50 clusters of 25, decreasing to a power of .20 with 250 clusters of size 5, and a power of .17 with 625 clusters of size 2. Although these results were found in conditions with continuous data, the same pattern may be expected with ordinal data. In our illustration, the employed tests thus had very low power. Still we detected significant residual variance in all indicators.

In this article we used MLR estimation. An alternative estimator, suitable for more complex models, is the multilevel version of diagonally weighted least squares (denoted by WLSM in Mplus; Asparouhov & Muthén, 2007). In comparison with MLR, WLSM 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 can also be investigated using item response models for measurement bias (Fox & Verhagen, 2010; Verhagen & Fox, 2012). Verhagen and Fox (2012) demonstrated the use of Bayesian methods to test invariance hypothesis in the random item effects modeling framework. An advantage of their method over the test for cluster bias is that the investigation of differences in factor loadings across clusters is more straightforward.

In conclusion, this study showed that cluster bias can be tested in ordinal data using ordinal factor analysis. We prefer and advise the use of 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.

FUNDING

C. V. Dolan is supported by the European Research Council Genetics of Mental Illness (ERC-230374).

REFERENCES

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.

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.

Engle, R. F. (1983). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In M. D. Intriligator & Z. Griliches (Eds.), *Handbook of econometrics* (pp. 796–801). Amsterdam, The Netherlands: Elsevier.

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, UK: Routledge Academic.

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

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

Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 221–233). Berkeley: University of California Press.

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

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*(2), 265–282.

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

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

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.

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

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

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

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

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

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org/

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

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, *16*(4), 583–601.

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

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

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.

Spilt, J. L., Koomen, H. M., & 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*(3), 363–378.

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.

Verhagen, A. J., & Fox, J.-P. (2012). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*, 383–401.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

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.