Statistica Neerlandica (2010) Vol. 64, nr. 2, pp. 157–170
doi:10.1111/j.1467-9574.2009.00445.x

The effect of estimation method and sample size in multilevel structural equation modeling

Joop J. Hox*, Cora J. M. Maas and Matthieu J. S. Brinkhuis

Department of Methodology and Statistics, Utrecht University, Utrecht The Netherlands

Multilevel structural equation modeling (multilevel SEM) has become an established method to analyze multilevel multivariate data. The first useful estimation method was the pseudobalanced method. This method is approximate because it assumes that all groups have the same size, and ignores unbalance when it exists. In addition, full information maximum likelihood (ML) estimation is now available, which is often combined with robust chi-squares and standard errors to accommodate unmodeled heterogeneity (MLR). In addition, diagonally weighted least squares (DWLS) methods have become available as estimation methods. This article compares the pseudobalanced estimation method, ML(R), and two DWLS methods by simulating a multilevel factor model with unbalanced data. The simulations included different sample sizes at the individual and group levels and different intraclass correlation (ICC). The within-group part of the model posed no problems. In the between part of the model, the different ICC sizes had no effect. There is a clear interaction effect between number of groups and estimation method. ML reaches unbiasedness fastest, then the two DWLS methods, then MLR, and then the pseudobalanced method (which needs more than 200 groups). We conclude that both ML(R) and DWLS are genuine improvements on the pseudobalanced approximation. With small sample sizes, the robust methods are not recommended.

Keywords and Phrases: Two-level structural equation modeling, estimation method, simulation.

Multilevel structural equation modeling (SEM) has become an established method to analyze multilevel multivariate data. A variety of approaches to multilevel analysis of structural equation models has been proposed by, among others, Goldstein and McDonald (1988), Muthén and Satorra (1989), Muthén (1989, 1994), McDonald (1994), and Raudenbush and Sampson (1999). The first generally useful estimation method was the pseudobalanced method (Muthén, 1989, 1994). The advantage of the pseudobalanced method is that it is simple to implement in existing software, so most major SEM packages have included this method. When SEM software does not

*j.hox@uu.nl
After the online publication of this article, Cora Maas passed away on 8 February at the age of 45. We will miss her.

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

incorporate multilevel SEM, it can be carried out using a somewhat complicated setup for a multiple group model (Hox, 2002).

The pseudobalanced method is approximate because it assumes that all groups have the same size, and thus ignores unbalance when it exists. Full information maximum likelihood (ML) estimation does not require that groups are balanced, and should produce more accurate estimates. The ML equations and a possible software implementation are described by Muthén (1990), and more recently by Mehta and Neale (2005). A recent development is using ML estimation with robust chi-squares and standard errors (MLR). This produces the same parameter estimates, but the chi-square for the model test and the standard errors for the parameters are calculated differently. MLR is assumed to be robust against moderate violations of assumptions, including unmodeled heterogeneity.

Recently, a limited information diagonally weighted least squares (DWLS) estimation procedure has become available for multilevel SEM (Asparouhov and Muthén, 2007). In this procedure, ML methods are used to estimate the means and the within groups and between groups covariance matrices separately, after which DWLS is used to estimate the parameters of the multilevel SEM.

This study uses simulation to examine the accuracy of pseudobalanced estimation, full ML, and DWLS with unbalanced data and varying sample sizes. In addition, the differences between asymptotic normal theory and robust standard errors and chi-squares is studied. The next section describes these estimation methods in more detail, ending with the questions that stimulated this simulation study.

1 Multilevel structural equation models

In multilevel SEM, we assume sampling at two levels, with both between group (group level) and within group (individual level) covariation. More than two levels are possible, which leads to extensions of the methods described next.

1.1 Pseudobalanced estimation

The starting point for the pseudobalanced approach is Cronbach and Webb's (1975) decomposition of the total scores at the individual level $\mathbf{Y}_T$ into a between group component $\mathbf{Y}_B$, which are the disaggregated group means, and a within group component $\mathbf{Y}_W$, which are the individual deviations from the corresponding group means. This leads to additive and orthogonal scores for the two levels.

In the population we can also distinguish the between-group covariance matrix $\mathbf{\Sigma}_B$ and the within-group covariance matrix $\mathbf{\Sigma}_W$. In the special case of balanced groups, estimation turns out to be straightforward (Muthén, 1989). In the case of G balanced groups, with all G group sizes (GS) equal to n, and total sample size $N=nG$, we can define two sample covariance matrices: the pooled within covariance matrix $\mathbf{S}_{PW}$ and the scaled between covariance matrix $\mathbf{S}_B^*$. As Muthén (1989, 1990) shows, $\mathbf{S}_{PW}$ is the ML estimator of $\mathbf{\Sigma}_W$, and $\mathbf{S}_B^*$ is the ML estimator of the composite

$\Sigma_W + c\Sigma_B$, with scaling parameter c equal to the common GS n. In the more general unbalanced case, $\mathbf{S}_{PW}$ is still the ML estimator of Σ_W, but $\mathbf{S}^*_B$ now estimates a different between-groups matrix for each set of groups with common GS n. Thus, ML estimation for unbalanced groups using this approach implies a separate between-group model for each distinct GS, with different scaling parameters c_d for each. MUTHÉN (1989) proposed a simplified method, which uses one single $\mathbf{S}^*_B$ with an *ad hoc* estimator $c*$, close to the average sample size, for the scaling parameter c. The result is a limited information ML solution, which MCDONALD (1994) called a pseudobalanced solution, but it has also become known as Muthén's ML (MUML) solution. We will refer to it as the pseudobalanced solution. MUTHÉN (1989) claims that this estimator is unbiased and consistent. The simulation studies referred to in the introduction confirm this, but also show that in the unbalanced case the standard errors and chi-square model tests are not as accurate as would be desired. YUAN and HAYASHI (2005) show analytically that pseudobalanced standard errors and chi-square tests only lead to correct inferences when the between-level sample size goes to infinity and the coefficient of variation of the GS goes to zero. Thus, both simulations and analytical work agree that larger sample sizes do not improve the accuracy with seriously unbalanced data.

1.2 ML and MLR

The pseudobalanced approach follows the conventional notion that structural equation models are constructed for the covariance matrix with added mean vector. The multilevel full ML approach defines the model and the likelihood in terms of the individual or raw data. ARBUCKLE (1996) presents this method in the context of SEM estimation with incomplete data. The SEM likelihood function for raw data is given by

$$F = \sum_{i=1}^{N} \log |\Sigma_i| + \sum_{i=1}^{N} \log \left(\mathbf{x}_i - \mu_i\right)' \Sigma_i^{-1} \left(\mathbf{x}_i - \mu_i\right), \qquad (1)$$

where the subscript i refers to the observed cases, $\mathbf{x}_i$ to the variables observed for case i, and $\boldsymbol{\mu}_i$ and Σ_i contain the population means and covariances of the variables observed for case i. MEHTA and NEALE (2005) show that for multilevel data, with individuals nested within groups, the ML fit function given by Equation (1) applies, with clusters as observations, and individuals within clusters as variables. Thus, their approach incorporates multilevel analysis in general SEM, allowing for intercept and slope variation across groups.

A recent development is to use robust standard errors and chi-squares for significance testing when violations of the assumptions of the asymptotic tests are suspected. Several corrections have been proposed for the chi-square model test, the most often used are the SATORRA–BENTLER (1994) and the YUAN–BENTLER (1998) corrections. The robust standard errors are generally Huber–White sandwich estimators (HUBER, 1967; WHITE, 1982), using the observed residual variances to correct the asymptotic standard errors. The robust chi-square tests and standard errors are generally more accurate than the asymptotic tests when data are non-normal and when the model is mis-specified (Chou, Bentler and Satorra, 1991; Curran, West and Finch, 1996). With multilevel data, robust chi-squares and standard errors are assumed to offer some protection against unmodeled heterogeneity, which may result from mis-specifying the group-level model, or by omitting a level. The present simulation study includes ML estimation of parameter estimates both with asymptotic and with robust standard errors and chi-square. The software employed (Mplus, cf. Muthén and Muthén, 1998–2007) uses the Yuan–Bentler (1998) robust chi-square and sandwich standard errors. The robust ML approach is denoted in Mplus and in this article as MLR. It should be stressed that MLR results in the same parameter estimates as ML; only the standard errors and chi-square tests are computed differently.

1.3 Two-step DWLS

Asparouhov and Muthén (2007) describe a limited information WLS approach to multilevel SEM. This approach is a two-step method. In the first step, univariate ML methods are used to estimate the vector of means $\boldsymbol{\mu}$ at the between-group level, and the diagonal elements of $\boldsymbol{\Sigma}_W$ and $\boldsymbol{\Sigma}_B$. Next, the off-diagonal elements of $\boldsymbol{\Sigma}_W$ and $\boldsymbol{\Sigma}_B$ are estimated using bivariate ML methods. Finally, the asymptotic covariance matrix for these estimates is obtained. In the second step, the multilevel model parameters are estimated for both levels using WLS.

It should be noted that standard WLS uses a weight matrix based on the sampling covariances of all estimated parameters. This is a square matrix, which for the unrestricted model has dimensionality $q \times q$, with q depending on the number of estimated parameters. So, q equals $(p \times (p+1))/2$, where p is the number of estimated parameters. As q is an exponential function of the number of parameters p, the asymptotic covariance matrix tends to be very large. Especially for the between part of the model, the number of elements in this matrix can easily become larger than the number of groups (NG). Unless the NG is extremely large, it is preferable to use only the diagonal of this weight matrix (cf. Muthén, Du Toit, and Spisic, 1997), which results in DWLS estimation. In Mplus, choosing the diagonal weight matrix always implies using a robust chi-square, with WLSM using a mean-corrected (first order) and WLSMV, a mean-and-variance-corrected (second order) correction. In the simulation described next, the smaller sample sizes are too small to permit full WLS estimation; hence, we use WLSM and WLSMV only. WLSM and WLSMV are both DWLS methods that lead to the same estimates and standard errors, but to different goodness-of-fit (chi-square) tests.

1.4 Accuracy, estimation method, and sample size in multilevel SEM

Several studies have shown that the pseudobalanced approach results in unbiased and accurate parameter estimates (Muthén, 1990; Hox, 1993; McDonald, 1994).

Hox and Maas (2001) present a simulation that investigates the accuracy of the model test and standard errors. Their study finds that with unbalanced data a group-level sample size of 100 is required for sufficient accuracy of the model test and confidence intervals (CIs) for the parameters. With the group-level sample size set to 50, the parameter estimates are estimated accurately, but the standard errors are too small, leading to 95% CIs that are actually lower than 90%. Even with the larger sample size, the pseudobalanced significance tests are approximate, with an operating alpha level around 8% instead of the nominal alpha of 5%. Yuan and Hayashi (2005) show analytically that pseudobalanced standard errors are always biased downwards when the groups are unbalanced, and that this bias is independent of the between-level sample size.

Full ML estimation in multilevel SEM should lead to more accurate chi-squares and standard errors for unbalanced multilevel data than the pseudobalanced method (cf. Liang and Bentler, 2004). The two-step DWLS approach offers the opportunity to obtain efficient estimates when full ML estimation would need to use numerical integration, which is the case when categorical data are analyzed. Although we do not investigate categorical data, we include two-step DWLS in our simulation to gauge the general accuracy of this method. In addition, we examine for the ML method both asymptotic (ML) and robust (MLR) standard errors and chi-squares. Given that we simulate multivariate normal data, the asymptotic (normal theory) standard errors should be more accurate than the robust standard errors. We include the robust methods in our simulation, because these methods are increasingly the methods chosen by default in the available software, so it is of interest to assess how much accuracy is lost if they are employed with normal data, where robust estimation is not necessary.

In this simulation study, we examine the accuracy of pseudobalanced estimation, ML, MLR, WLSM, and WLSMV with different sample sizes at the individual and group levels. As with balanced data pseudobalanced estimation and ML are equivalent, the simulation includes only unbalanced data. The study also varies the ICC. With respect to the estimation methods, we expect ML estimates to be more accurate than pseudobalanced estimates by a considerable margin. As ML is asymptotically efficient, we expect ML estimates to be more accurate than the two-step DWLS estimates, but only with a small margin. We expect the robust standard errors and chi-squares to be less accurate than the asymptotic standard errors and chi-squares.

2 Method

2.1 The simulation model

We use a simple confirmatory factor model with six variables, two factors in the within part, and one factor in the between part. Figures 1 and 2 present the path diagram for the between and within parts, with the population parameter values.

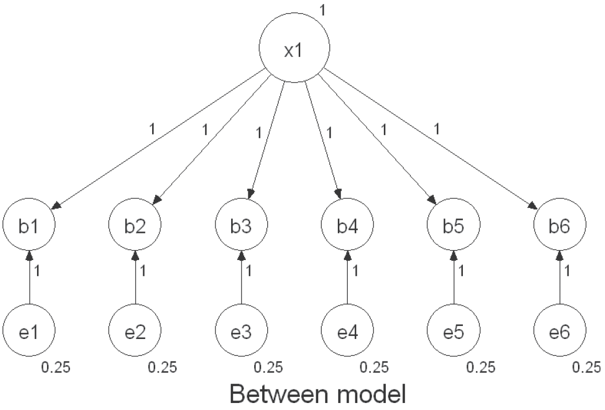

Fig. 1. Path diagram for the between model.

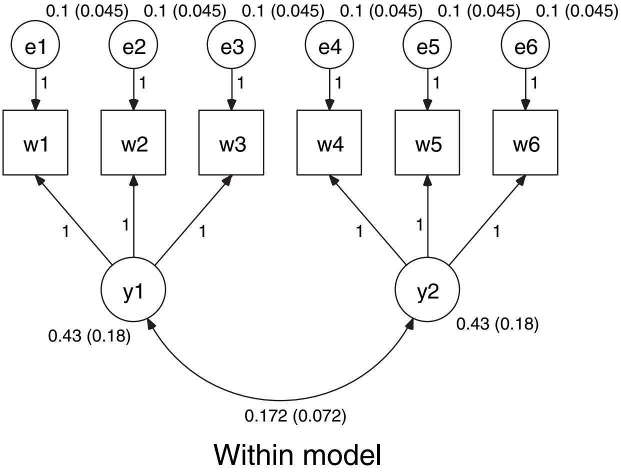

Fig. 2. Path diagram for the within model. Unbracketed parameter values correspond to a population model with a low intraclass correlation (ICC) and values within brackets to a population model with a high ICC.

2.2 Simulation procedure

Four conditions are varied in the simulation:

(i) method of estimation/type of standard error or chi-square, five conditions: pseudobalanced method, ML, MLR, and two-step DWLS (WLSM and WLSMV for the mean and the mean-and-variance-corrected chi-squares);
(ii) NG, three conditions: NG=50–100–200;
(iii) average GS, three conditions: GS=5–10–25; and
(iv) ICC low versus high, two conditions: ICC=0.05 [low]– 0.15 [high].

The sample sizes (NG) at the group level are 50, 100, and 200, respectively. These sample sizes are chosen so that the highest number conforms to Boomsma's (1983) recommended lower limit for achieving good ML estimates with normal data, derived from his robustness studies for single-level data. The lower values have been chosen because, in multilevel modeling, obtaining data from as many as 200 groups can be difficult, and many studies have far less than 200 groups.

To maximize the effect of imbalance, the GS where chosen to be quite different. To create unbalanced data, we employ two distinct GS, with exactly half the groups being small and the other half being large. For the three average GS, the unbalanced sample sizes are as follows: for GS = 5: 3/7; for GS = 10: 5/15; for GS = 25: 13/37. Thus, the large GS is about thrice as large as the small GS.

Several model parameters of the within part had to be modified to set the ICC at a low or a high level. Figure 2 presents the values for the residual variances, factor variances, and the covariance that lead to a low ICC of 0.05. The values within brackets lead to a high ICC of 0.15. At both the within and between levels the model was identified by fixing the factor variances to the specified values. Thus, the estimated parameters are the factor loadings, the factor covariance, and the residual variances.

There are $5 \times 3 \times 3 \times 2 = 90$ conditions. For each condition, we generate 1000 data sets, assuming normally distributed latent variables. This results in observed data that meet the assumption of multivariate normality. The 90 simulations where performed using MPlus 5 (Muthén and Muthén, 2007). Software restrictions cause the confounding of estimation method and use of asymptotic versus robust methods in the first condition: in the current version of Mplus the pseudobalanced method produces only asymptotic standard errors, and the DWLS methods produce only robust standard errors and chi-squares, with WLSM and WLSMV differing only in the chi-square. Only in full ML estimation both asymptotic and robust methods are available, and can be compared directly. We will return to software issues briefly in the discussion. For brevity, in section 3, the available combinations of estimation method and type of standard errors or chi-squares are labeled 'estimation method' in the tables.

2.3 Variables and analysis

The percentage relative bias is used to indicate the accuracy of the parameter estimates. Let $\hat{\theta}$ be the estimate of the population parameter θ. Then, the relative bias is given by

$$(\hat{\theta} - \theta)/\theta. \tag{2}$$

The accuracy of the chi-square model test is indicated by the empirical alpha level. In addition, the proportions of parameter estimates falling within their CI are calculated. The relative bias is analyzed using manova procedures with the set of parameters (loadings, variances) as multivariate outcomes.

3 Results

No non-convergent or inadmissible solutions (e.g., negative variance estimates) were encountered across all 90,000 simulated data sets.

3.1 Model fit

Because the fitted models are equal to the population model, the expected value for the chi-square is equal to the degrees of freedom, which is 17, and the expected proportion of significant chi-squares is equal to the alpha level of 0.05. There is an effect on the overall chi-square test of estimation method, GS, and NG. Table 1 presents the empirical alpha level for the different estimation methods and sample sizes. It is clear that the most important factors are the estimation method and the NG. When the other factors are constant, increasing the GS does not improve the accuracy of the model fit test. In this simulation, where the observed data follow a multivariate normal distribution, the normal theory ML chi-square and the WLSMV robust chi-square are the most accurate. With pseudobalanced estimation, the bias of the chi-square test is large, and it actually increases when the GS becomes larger. The accuracy of the robust MLR and WLSM chi-squares is better than the pseudobalanced chi-square but not as good as the asymptotic ML and the robust WLSMV chi-square, even with 200 groups.

3.2 Parameter estimates

The factor loadings and the error variances of the within part of the model have an overall mean bias of −0.001, with negligible differences across the conditions.

Table 1. Empirical alpha level of the chi-square goodness of fit test for different number of groups and group sizes, by estimation method

		Group size			
Estimation method	Number of groups	5	10	25	Overall
Pseudobalanced	50	0.104	0.139	0.151	0.131
	100	0.086	0.126	0.119	0.110
	200	0.086	0.111	0.111	0.103
ML	50	0.071	0.078	0.082	0.077
	100	0.058	0.075	0.066	0.066
	200	0.070	0.055	0.058	0.061
MLR	50	0.087	0.100	0.105	0.097
	100	0.064	0.079	0.070	0.071
	200	0.073	0.059	0.067	0.066
WLSM	50	0.062	0.078	0.095	0.078
	100	0.092	0.092	0.048	0.077
	200	0.085	0.082	0.077	0.081
WLSMV	50	0.030	0.031	0.023	0.028
	100	0.041	0.043	0.037	0.040
	200	0.053	0.038	0.053	0.048

Notes: ML, maximum likelihood; MLR, robust ML; WLSM, weighted least squares, mean-corrected; WLSMV, WLS, mean-and-variance-corrected.

In the between part of the model, the overall mean bias of the factor loadings is −0.008. This bias is somewhat smaller with larger sample sizes, larger NG, and with the pseudobalanced estimation procedure. Across all conditions the differences in bias are negligible. The mean bias of the error variances is somewhat higher; overall, it is −0.023. The bias is smaller with a larger NG and with the pseudobalanced estimation procedure. Across the conditions, the differences in bias are again negligible.

3.3 Standard errors

To investigate the bias of the standard errors of the parameters, the 95% CI for each parameter is computed, and the proportion of intervals that include the true population value is counted.

In the within part of the model, the mean coverage of the 95% CI is 94.1% for the factor loadings and 94.4% for the error variances. The differences across conditions for both the factor loadings and the error variances are very small.

In the between part of the model the mean coverage of the 95% CIs is smaller. For the factor loadings the mean coverage is 93.8%, with small effects of the NG and the GS, and a larger effect of the estimation method. The mean coverage of the variances is only 91.5%. There is small effect of NG, estimation method, and an interaction effect between NG and estimation method. The results for the between part of the model are presented in Table 2.

The coverage of the 95% CI of the loadings is generally better than the coverage of the variances. In Table 2, we see the influence of the estimation method and sample sizes on the operating alpha in detail. The estimation method and the type of standard error have a substanbtial influence. Concerning the factor loadings,

Table 2. Empirical coverage of 95% CI of the loadings and variances in the between part of the model, for different number of groups and group sizes, by estimation method

		Group size		
Estimation method	Number of groups	5	10	25
Pseudo-balanced	50	0.926/0.896*	0.907/0.880	0.917/0.885
	100	0.931/0.916	0.916/0.907	0.920/0.910
	200	0.929/0.921	0.921/0.916	0.922/0.918
ML	50	0.938/0.900	0.933/0.900	0.937/0.900
	100	0.943/0.925	0.940/0.929	0.943/0.930
	200	0.943/0.934	0.946/0.935	0.944/0.938
MLR	50	0.919/0.884	0.911/0.883	0.918/0.883
	100	0.934/0.916	0.928/0.919	0.932/0.918
	200	0.936/0.927	0.943/0.930	0.938/0.933
WLSM and WLSMV	50	0.949/0.900	0.949/0.910	0.950/0.923
	100	0.948/0.933	0.949/0.917	0.948/0.898
	200	0.949/0.921	0.950/0.926	0.948/0.935

Notes: *First number is the coverage of the loadings, and the second is the coverage of the variances. Note that for parameter estimates and corresponding standard errors weighted least squares, mean-corrected (WLSM) and WLS, mean-and-variance-corrected (WLSMV) are equivalent.
ML, maximum likelihood; MLR, robust ML.

when the pseudobalanced estimation method is used, the CIs are much small, and the operating alpha is close to 10%. Increasing the NG of the GS does not improve the coverege of the pseudobalanced estimation. For the ML and WLSM(W) all coverages are good. With the MLR estimation method, increasing the numer of groups leads to better coverages.

Concerning the variances all estimation methods do not perform very well. Only ML estimation leads to acceptable CIs, but only when the NG is 200.

The standard errors are calculated to be the same in WLSM and WLSMW. These methods perform almost equal to the ML method across all sample sizes, and are definitely more accurate than using ML estimation with robust standard errors (MLR).

4 Discussion

It should be noted first that the difference in ICC has no effect on any of the criteria we examined. This is contrary to the results in Hox and Maas (2001) who found that lower ICCs lead to convergence problems. However, in their study the effect of ICC and the amount of systematic variance in the between model was confounded. We conclude that the apparent effect of ICC in Hox and Maas (2001) is actually the effect of having less systematic variance with the low ICC condition. The other results of Hox and Maas (2001) with respect to the pseudobalanced method replicate well in this simulation, and correspond to Yuan and Hayashi's (2005) conclusions.

One general result is that for the within groups model all simulated conditions produced parameter estimates and corresponding standard errors that are accurate, with negligible differences between the estimation methods or the type of standard errors. A general conclusion drawn from these simulation results is that if the interest is only in the within part of the model, for example, when SEM is used with data collected by cluster sampling, analysis of the pooled within covariance matrix only is an accurate and effective approach. Most modern SEM software includes this option.

Differences between simulated conditions appear only in the between groups part of the model, and in the global chi-square test for goodness of fit. The most important factor that determines the accuracy of the results is the estimation method and the type of standard error/chi-square. As expected, pseudobalanced estimation results in chi-squares and standard errors that have a sizeable downward bias. As a result, the empirical alpha level for the chi-square test is unacceptably high, and the CIs are very small. Given that most software now offers the much more accurate full ML method, we do not recommend using the pseudobalanced estimation method.

Given that our simulation produces data that are multivariate normal, it is no surprise that normal theory asymptotic (ML) standard errors and chi-squares are in general more accurate than the robust standard errors and chi-squares. The robust chi-squares [MLR and WLSM(V)] do not perform well when the NG is small. For the ML estimation method the parameter estimates are identical, and therefore the performance of the asymptotic and robust standard errors can be compared directly. Here, the robust standard errors perform well only with a large NG. This casts some doubt on the routine use of robust standard errors and chi-squares, especially with moderate sample sizes. If the data in fact follow the distributional assumptions, using robust methods is generally less accurate than asymptotic methods. If inspection of the data supports the assumptions, using asymptotic rather than robust standard errors may be preferable. In the present simulation, WLSM and WLSMV produced standard errors for the loadings that are at least as accurate as ML, and appear even better with smaller NG. In fact, WLSM and WLSMV use the same estimation method (two-step DWLS) and the same method to compute robust standard errors. They differ in the way the goodness-of-fit chi-square is computed, and here WLSM is less accurate, whereas WLSMV is almost as accurate as normal theory ML. Further simulations comparing these methods on non-normal data are needed, but based on this simulation the two-step DWLS estimation employed in Mplus' WLSMV method appears promising.

The second factor that determines the accuracy of the statistical tests for the between model is the between-level sample size. When the interest is mostly in the factor loadings, a moderate sample size of 50 groups appears sufficient when the ML or the WLSM or WLSMV method is used. MLR performs well when the NG is increased to 200. Increasing the GS has almost no effect, and in the case of pseudo-balanced estimation even has a negative effect on the accuracy of the tests. It should be noted that the superiority of ML over MLR only holds when the data follow the assumptions. In our simulation, all data are multivariate normal and there is no unmodeled heterogeneity. When the data violate such assumptions, MLR has been found to be more accurate than ML (cf. Maas and Hox, 2004), but it still needs the larger sample sizes to be accurate.

The results of this simulation point out two strategies to increase the accuracy of the statistical tests, in addition to the simple strategy of increasing the group-level sample size, which is not always feasible. First, it is interesting to note that the parameter estimates themselves are accurate in all simulated conditions. This implies that using resampling methods such as the jackknife or the bootstrap should work well, provided the resampling scheme follows the original multilevel sampling scheme. Second, our results indicate that even with 200 groups some tests, such as tests on variances, are still not very accurate. Other studies focusing on multilevel regression also found that the Wald test for variances is not very accurate; for a discussion, see Berkhof and Snijders (2001). Monte Carlo methods could be investigated for more accurate assessment of sampling variability in multilevel SEM.

The present simulation employs Mplus 5.2 for both simulation and estimation. We have included an Appendix with the Mplus setup for one of the simulation runs in the Appendix. Other simulation conditions can be specified by changing certain values in the model simulation section of the setup, as described in the simulation design. Other software (e.g., Lisrel or Eqs, but not Amos) generally also offers both the pseudobalanced approach (generally labeled MUML) and full ML estimation, with a choice of normal theory or robust chi-squares and standard errors. However, at the time of writing Mplus is the only software that includes the two-step DWLS approach. At present, to our knowledge, Mplus and gllamm are the only software packages that are capable of including random slopes in multilevel SEM, using methods described in Mehta and Neale (2005).

Appendix: Simulation setup (Mplus commands)

Table A1. TITLE Simulation run for ML, ICC low, NG=50, GS=10;

```
MONTECARLO:
   NAMES ARE y1-y6;
   NOBSERVATIONS = 250;
   NREPS = 1000;
   SEED = 0;                    ! Comment seed set to ensure complete replicability;
   NCSIZES = 2;
   CSIZES = 25 (3) 25 (7);      ! Produce unbalanced data;
 RESULTS = results01.sav;

MODEL POPULATION:
   %within%
   fw1 BY y1-y3@1;
   fw2 BY y4-y6@1;
   y1-y6@.10;
fw1@.43;                        ! Variances determined by set value of ICC;
   fw2@.43;
   fw1 WITH fw2@.172;
   %between%
   fb1 BY y1-y6@1;
   y1-y6@.25;
   fb1@1

MODEL:
   %within%
   fw1 BY y1-y3*1;
   fw2 BY y4-y6*1;
   y1-y6*.10;
   fw1@.43;
   fw2@.43;
   fw1 WITH fw2*.172;
   %between%
   fb1 BY y1-y6*1;
   y1-y6*.25;
   fb1@1;

ANALYSIS:
   TYPE = TWOLEVEL;
   ESTIMATOR = ML;              ! Estimator and type of SE/Chisquare set here;

OUTPUT:
   TECH9;                       ! Allows monitoring simulations on screen;
```

References

Arbuckle, J. L. (1996), Full information estimation in the presence of incomplete data, in: G. A. Marcoulides and R. E. Schumacker (eds), *Advanced structural equation modeling: issues and techniques*, Lawrence Erlbaum Associates, Mahwah, NJ: pp. 243–277.

Asparouhov, T. and B. Muthén (2007), Computationally efficient estimation of multilevel high-dimensional latent variable models. *Proceedings of the Joint Statistical Meeting*, August, Salt Lake City, Utah.

Berkhof, J. and T. A. B. Snijders (2001), Variance component testing in multilevel models. *Journal of Educational end Behavioral Statistics* **26**, 133–152.

Boomsma, A. (1983), *On the robustness of LISREL (maximum likelihood estimation) against small sample size and nonnormality*, Sociometric Research Foundation, Amsterdam.

Chou, C. P., P. M. Bentler and A. Satorra (1991), Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: a Monte Carlo study. *British Journal of Mathematical and Statistical Psychology* **44**, 347–357.

Cronbach, L. J. and N. Webb (1975), Between class and within class effects in a reported aptitude × treatment interaction: a reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology* **67**, 717–724.

Curran, P. J., S. G. West and J. F. Finch (1996), The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods* **1**, 16–29.

Goldstein, H. and R. McDonald (1988), A general model for the analysis of multilevel data. *Psychometrika* **53**, 455–467.

Hox, J. J. (1993), Factor analysis of multilevel data: gauging the Muthén model, in: J. H. L. Oud and R. A. W. Van Blokland-Vogelesang (eds), *Advances in longitudinal and multivariate analysis in the behavioral sciences*, ITS, Nijmegen, pp. 141–156.

Hox, J. J. (2002), *Multilevel analysis, techniques and applications*. Lawrence Erlbaum Associates, Mahwah, NJ.

Hox, J. J. and C. J. M. Maas (2001), The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling* **8**, 157–174.

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*, University of California Press, Berkeley, CA, pp. 221–233.

Liang, J. and P. M. Bentler (2004), An EM algorithm for fitting two-level structural equation models. *Psychometrika* **69**, 101–122.

Maas, C. J. M. and J. J. Hox (2004), The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis* **46**, 427–440.

McDonald, R. P. (1994), The bilevel reticular action model for path analysis with latent variables. *Sociological Methods & Research* **22**, 399–413.

Mehta, P. D. and M. C. Neale (2005), People are variables too: multilevel structural equations modeling. *Psychological Methods* **10**, 259–284.

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

Muthén, B. (1990), *Means and covariance structure analysis of hierarchical data*. UCLA Statistics Series, #62, Los Angeles. Available at: http://www.statmodel.com, accessed November 2007.

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

Muthén, L. K. and B. O. Muthén (1998–2007), *Mplus. The comprehensive modeling program for applied researchers. Fifth edition*, Muthén & Muthén, Los Angeles.

Muthén, B. and A. Satorra (1989), Multilevel aspects of varying parameters in structural models, in: R. D. Bock (ed.), *Multilevel analysis of educational data*, Academic Press, San Diego, pp. 87–99.

Muthén, B., S. H. C. Du Toit and D. Spisic (1997), Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Available at: www.statmodel.com, accessed 15 March 2009.

Raudenbush, S. and R. Sampson (1999), Assessing direct and indirect effects in multilevel designs with latent variables. *Sociological Methods and Research* **28**, 123–153.

Satorra, A. and P. M. Bentler (1994), Corrections to test statistics and standard errors in covariance structure analysis, in: A. Von Eye and C. C. Clogg (eds), *Latent variables analysis. applications for developmental research*, Sage, Thousand Oaks, CA, pp. 399–419.

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

Yuan, K.-H. and P. M. Bentler (1998), Robust mean and covariance structure analysis. *British Journal of Mathematical and Statistical Psychology* **51**, 63–88.

Yuan, K.-H. and K. Hayash (2005), On Muthén's maximum likelihood for two-level covariance structure models. *Psychometrika* **70**, 147–167.

Recieived: April 2008. Revised: October 2009.