Running head: Multilevel SEM with Pseudobalanced Groups

To appear in: Structural Equation Modeling, 2000

The Accuracy of Multilevel Structural Equation Modeling with Pseudobalanced Groups and Small Samples

Joop J. Hox and Cora J.M. Maas

Department of Methodology and Statistics

Utrecht University, the Netherlands

Corresponding author:

J.J. Hox

Department of Methodology and Statistics

Faculty of Social Sciences, Utrecht University

P.O.B. 80140

NL-3508 TC Utrecht

The Netherlands

j.hox@fss.uu.nl

tel. +31 30 2539236

fax +31 30 2535797

Abstract

Social data often have a hierarchical structure, which leads to clustered data. This causes problems in analysis, because the usual assumptions of independently and identically distributed variables are not met. Muthén (1989) describes an estimation method for multilevel factor and path analysis with hierarchical data. For full Maximum Likelihood estimation, his method requires equal group sizes. This article assesses the robustness of the method against unequal groups, small sample sizes at both the individual and the group level, in the presence of either a low or a high intraclass correlation (ICC). The accuracy is assessed by simulation. The within groups part of the model poses no problems. The most important problem in the between groups part of the model, is the occurrence of inadmissible estimates, especially when group level sample size is small (50) while the intracluster correlation is low. This is partly compensated by using large group sizes. When an admissible solution is reached, the factor loadings are generally accurate. However, the residual variances are underestimated, and the standard errors are generally too small. Having more or larger groups or a higher ICC does not effectively compensate for this. Therefore, while the nominal alpha level is 5%, the operating alpha level is about 8% in all simulated conditions with unbalanced groups. The strongest factor is an inadequate sample size at the group level. Imbalance is only a problem for the overall goodness-of-fit test. For balanced data, the chi-square test for goodness-of-fit is accurate. The size of the biases is comparable to the effect of moderate non-normality in ordinary modeling, and in our view, the approximate solution remains a useful analysis tool, provided the group level sample size is at least 100.

The Accuracy of Multilevel Structural Equation Modeling with Pseudobalanced Groups and Small Samples

Social science often studies systems with a hierarchical structure, e.g. the educational system, with a hierarchy of pupils within classes within schools; families, with family members within families; and other social structures where individuals are grouped in organizational or geographical units. Naturally, such systems can be observed at different hierarchical levels, which leads to data that should be regarded as multistage or cluster samples with a number of hierarchical levels. There may be different sets of variables at the separate levels.

Even if the analysis includes only variables at the lowest (individual) level, standard multivariate models are not appropriate. The hierarchical structure of the data creates problems, because the standard assumption of independent and identically distributed observations (i.i.d.) is generally not valid. Multilevel analysis techniques have been developed for the linear regression model (Bryk & Raudenbush, 1992; Goldstein, 1995), and specialized software is now widely available (Bryk, Raudenbush & Congdon, 1996; Rasbash & Woodhouse, 1995). We refer to McArdle and Hamagami (1996) for a comparison between multilevel regression techniques and standard multigroup SEM. The multilevel analysis of structural equation models has been discussed by, among others, Goldstein and McDonald (1988), McDonald (1994), Muthén and Satorra (1989), and Muthén (1989, 1990, 1994). Muthén’s approach is particularly interesting, because he shows that structural equation modeling (SEM) of multilevel data is possible using available SEM software. For an introductory exposition of Muthén's method, see Hox (1995), Kaplan and Elliot (1997) and Li, Duncan, Harmer, Acock, and Stoolmiller (1998). Meanwhile, SEM software has appeared that includes the multilevel extensions (Mplus, see

Muthén & Muthén, 1998; Eqs 6.0, as promised in spring 2000), or acts as a front end for conventional SEM software (Streams, see Gustaffson & Stahl, 1999).

Multilevel Structural Equation Models

In Muthén's multilevel model (Muthén, 1989, 1994), we assume sampling at two levels, with both between group (group level) and within group (individual level) covariation. The starting point is a decomposition of the total scores $\mathbf{Y}_T$ at the individual level, into a between group component $\mathbf{Y}_B$ (the disaggregated group means), and a within group component $\mathbf{Y}_W$, (the individual deviations from the corresponding group means). This leads to additive and uncorrelated scores for the two levels (cf. Cronbach and Webb, 1975; Searle, Casella & McCulloch, 1992). Thus, at the individual score level we have

$$\mathbf{Y}_T = \mathbf{Y}_B + \mathbf{Y}_W \quad (1)$$

and for the sample covariance matrix $\mathbf{S}$ we have

$$\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W \quad (2)$$

Likewise, we distinguish in the population a between group covariance matrix $\mathbf{\Sigma}_B$ and a within group covariance matrix $\mathbf{\Sigma}_W$. Muthén (1989, 1990) formulates structural equation models for $\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_W$, and discusses maximum likelihood procedures to estimate these. In the special case of balanced groups, estimation is straightforward (Muthén, 1989). If we have *G* balanced groups, with *G* equal group sizes *n* and a total sample size *N*=*nG*, we define two sample covariance matrices: the pooled within covariance matrix $\mathbf{S}_{PW}$ and the scaled between covariance matrix $\mathbf{S}^*_B$. These are given by:

$$\mathbf{S}_{PW} = \frac{\sum_{g}^{G}\sum_{i}^{n}\left(Y_{gi} - \overline{Y}_g\right)\left(Y_{gi} - \overline{Y}_g\right)'}{N-G} \qquad (3)$$

and

$$\mathbf{S}^*_B = \frac{\sum_{g}^{G} n\left(\overline{Y} - \overline{Y}_g\right)\left(\overline{Y} - \overline{Y}_g\right)'}{G-1} \qquad (4)$$

In equation (3), $\mathbf{S}_{PW}$ is the covariance matrix of the individual deviation scores, with denominator *N-G* instead of *N-1*. In equation 4, $\mathbf{S}^*_B$ is *n* times the covariance matrix of the group means. This is equal to the covariance matrix of the disaggregated means, with denominator *G-1* instead of *N-1*. Muthén (1989, 1990) has shown that $\mathbf{S}_{PW}$ is the maximum likelihood estimator of $\mathbf{\Sigma}_W$, with sample size N-G, and $\mathbf{S}^*_B$ is the maximum likelihood estimator of the composite $\mathbf{\Sigma}_W$ + c$\mathbf{\Sigma}_B$, with sample size *G*, and *c* equal to the common group size *n*:

$$\mathbf{S}_{PW} = \hat{\Sigma}_W \tag{5}$$

and

$$\mathbf{S}^*_B = \hat{\Sigma}_W + c\hat{\Sigma}_B \tag{6}$$

Equations 3 through 6 suggest using the multi-group option of conventional SEM software for a simultaneous analysis at both levels. For the within group structure, the same model is specified for $\mathbf{S}_{PW}$ and $\mathbf{S}^*_B$, with equality constraints across both 'groups.' The model for the between group structure includes the constant *C*, equal to the common group size *n*, as a scaling factor. Note that the between group structure is actually a composite of the model for $\mathbf{\Sigma}_W$ and the model for $\mathbf{\Sigma}_B$, with the scaling parameter *C* for the latter.

The unbalanced case, with G groups of unequal sizes, is more complicated. In this case, $\mathbf{S}_{PW}$ is still the maximum likelihood estimator of $\mathbf{\Sigma}_W$, but $\mathbf{S}^*_B$ now estimates a different expression for each set of groups with distinct group size *d*:

$$\mathbf{S}^*_{Bd} = \hat{\Sigma}_W + c_d\hat{\Sigma}_B \tag{7}$$

where equation 7 holds for each distinct set of groups with a common group size equal to n_d, and $C_d=n_d$ (Muthén, 1990, 1994).

Full Information Maximum Likelihood (FIML) estimation for unbalanced groups implies specifying a separate between group model for each distinct group size. These between group models have different scaling parameters C_d for each, and require equality constraints across all other parameters and inclusion of a mean structure (Muthén, 1994, p. 385.). This results in large and complex models, sometimes with groups with a sample size less than the number of elements in the corresponding covariance matrix. This makes full Maximum Likelihood estimation problematic, and therefore Muthén (1989, 1990) proposes ignoring the imbalance, and computing a single $\mathbf{S}^*_B$. The model for $\mathbf{S}^*_B$ includes an ad hoc estimator C^* for the scaling parameter, which is close to the average sample size:

$$C^* = \frac{N^2 - \sum_{g}^{G} n_g^2}{N(G-1)} \tag{8}$$

The result is a Limited Information Maximum Likelihood (LIML) solution, which McDonald (1994) calls a pseudobalanced solution. Muthén (1989, 1990) argues that since $\mathbf{S}^*_B$ is a consistent and unbiased estimator of the composite $\mathbf{\Sigma}_W + c\mathbf{\Sigma}_B$, the pseudobalanced solution should produce a good approximation given large sample sizes.

There has been limited research into the robustness of this pseudobalanced solution. Muthén (1990) presents a MIMIC (Multiple Indicator Multiple Independent Causes) model, where the LIML and FIML approaches produce practically identical parameter estimates and chi-square values. McDonald (1994) compares a LIML and a FIML analysis of a path model, for an artificial data set with 50 groups and group sizes randomly drawn from a uniform distribution of 3-100 individuals per group. Despite this extreme imbalance, both approaches produce very similar results. Hox (1993) reports a small simulation study using a factor model with LIML analyses of some artificial data sets. He concludes that LIML estimation performs well if the within group sample size is adequate (at least 200) even with smaller sample sizes at the group level (50 and 100 in his simulation).

The studies by Muthén (1990), McDonald (1994), and Hox (1993) all evaluate the accuracy of the parameter estimates (regression coefficients and variances) of the pseudobalanced solution by comparing them to either FIML estimates or known population values. The accuracy of the standard errors and the goodness-of-fit test is not assessed. Here, we investigate the accuracy of the pseudobalanced solution, with respect to the parameter estimates, their standard errors, and the overall chi-square model test. We compare balanced and unbalanced data, with varying number of groups, varying group sizes, and a varying proportion of group level variance. Since the balanced data actually produce a Full Information Maximum Likelihood solution, the parameter estimates are unbiased with asymptotically correct standard errors. Thus, for balanced data, we expect standard errors and chi-squares close to their nominal values. For unbalanced data, the pseudobalanced solution is unbiased and consistent, but it does not take into account the full variability of the data. Hence, we expect the unbalanced data to produce unbiased parameter estimates, with standard errors and chi-squares that deviate to an unknown degree from their nominal values. Simulation studies on the multilevel regression model (for a review see Hox, 1998), suggest that the number of groups is generally more important than the total sample size, especially for estimates of between group parameters. We expect this to hold in multilevel SEM as well, with large group sizes partially compensating for a small number of groups. A simulation study by Muthén, Wisnicky and Nelson (1991) suggests that the intraclass correlation (ICC), which indicates the amount of variance at the group level, also affects the accuracy of the estimates. In general, what is at issue in multilevel modeling is not so much the intraclass correlation, but the *design effect*, which indicates how much the standard errors are underestimated (Kish, 1965). In cluster samples, the design effect is approximately equal to 1+(average cluster size-1)*ICC. If the design effect is smaller than two, using single level analysis on multilevel data does not seem to lead to overly misleading results (Muthén & Satorra, 1995). In our simulation setup, we have chosen values for the ICC and group sizes that make the design effect larger than two in all simulated conditions.

Method

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. Figure 1 presents the path diagram for the within and the between part, with the population parameter values.

--- Figure 1 about here ---

Simulation Procedure

Four conditions are varied in the simulation: (1) Balanced vs. Unbalanced groups (Balance: 2 conditions), (2) Number of Groups (NG: 3 conditions, NG = 50-100-200), (3) (average) Group Size (GS: 3 conditions, GS = 10-20-50), and (4) IntraClass Correlation low vs. high (ICC: 2 conditions, created by giving between factor X1 a variance of 0.25 or .5). In the unbalanced condition, we employ two distinct group sizes, with exactly half the groups being small and the other half being large. For the three group sizes the unbalanced sample sizes are, for GS=10: 5/15, for GS=20: 10/30, for GS=50: 25/75. Thus, in the imbalanced condition, the large group size is three times as large as the small group size.

The number of groups (50-100-200) is chosen so that the highest number conforms to Boomsma's (1983) recommended lower limit for achieving good Maximum Likelihood estimates with normal 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 group sizes were chosen to be as different as possible. The factor variances we use lead to intraclass correlations that would, in practice, be considered high. Ignoring the effect of the error variances, the systematic variances of the factors lead to an average ICC of 0.20 in the Low, and 0.33 in the High condition. In educational research, most ICC's are below 0.20. However, in family research, or when group characteristics such as sociometric status are studied, ICC's above 0.33 do occur. The relatively high ICC values in our simulation are chosen to obtain better insight into the effect of ICC variations in combination with other factors, especially imbalance.

There are 2x3x3x2=36 conditions. For each condition, we generate 1000 data sets, assuming normally distributed latent variables. The simulated data are generated for the within and between model separately, using standard procedures. Subsequently, the within group and between group data are added following equation (1). Finally, a dedicated program (Hox, 1999) computes for each simulated data set the pooled within group and scaled between group covariance matrices following equations (3) and (4). For each simulation condition, the total data set consists of a stacked file of 1000 within + between covariance matrices according to the specifications. These are then analyzed using LISREL 8.14 (Jöreskog & Sörbom, 1993), which produces a file of 1000 sets of parameter estimates, standard errors and fit statistics.

The model used to generate data is presented in the appendix, with the implied within and between population covariance matrices. Fixing the factor variances at the population values identified the model. Thus, the estimated parameters are the factor loadings and the residual variances.

Variables and Analysis

The percentage relative bias is used to indicate the accuracy of the parameter estimates (factor loadings and residual variances). Let $\hat{\theta}$ be the population parameter θ, then the percentage relative bias is given by $[[\hat{\theta}-\theta]/\theta]\times 100\%$. The relative bias for the standard errors is computed by comparing the estimated values with the asymptotic value estimated in the population. In addition, we present the observed coverage of the 95% confidence interval. The accuracy of the chi-square model test is indicated by computing the relative bias, comparing the estimated value with the expected value, which is equal to the degrees of freedom (*df*=18). In addition, we present the standard deviation of the observed chi-squares compared to their expected value, which is equal to $(2\text{x}df)^{0.5} = 6$, and the percentages of models that are rejected at a significance level of 5 percent. Since the total sample size for each analysis is 36000 simulated conditions, the power is huge, and at the standard significance level of alpha=0.05, extremely small effects become significant. Hence, our criterion for significance is an alpha = 0.001 for the main effects. The interactions are tested blockwise (2-way, 3-way, 4-way), with a Bonferroni correction added for separate interaction effects. Even at this strict level of significance, some of the statistically significant biases correspond to differences in parameter estimates that do not show up before the third decimal place. These small effects are discussed in the text, but not included in the various tables.

In all 36000 simulations, the estimation algorithm converged. About 7.3% of simulations have at least one negative variance estimate, all of these very large. These cases also have at least one loading that is much larger than the population value. All inadmissible estimates occurred in the between groups part of the model. Since the inadmissible solutions produce extreme outliers in the parameter estimates, we will analyze only cases with admissible solutions. The distribution of inadmissible solutions across the simulation conditions is analyzed below, using logistic regression with the analysis strategy given above. The relative bias is analyzed using MANOVA procedures with the set of parameters (loadings, variances) as multivariate outcomes. Since the effects on separate loadings or variances within a set were in all cases very similar, we report the average (marginal) effect of the simulated conditions. Preliminary analyses indicated that the within part of the model, where the smallest effective sample size is 450 (N-G = 500 – 50 = 450, cf. equation 3), and imbalance is not an issue, shows only small departures from the nominal values in all conditions. Therefore, we will discuss the results for the within part only briefly, and concentrate on the between part.

Results

Inadmissible Solutions

The occurrence of inadmissible solutions depends above all on the number of groups (NG). NG has values 50/100/200, and the corresponding percentages of inadmissible solutions are 17%/5%/0%. The Intraclass Correlation (ICC) also has an effect. When the ICC is low, the percentage of inadmissible solutions is 13%, and when the ICC is high, that percentage drops to 2%. The Group Size (GS) has a smaller effect: GS has values 10/20/50, and the corresponding percentages of inadmissible solutions are 9%/7%/6%. Unbalance has the smallest effect: balanced data have 6% inadmissible solutions, and unbalanced data 8%. All these effects are significant at $p<.001$. The largest effects are those of the number of groups and the ICC. The one significant two-way interaction also involves the number of groups and the intraclass correlation. Table 1 presents the number of inadmissible solutions for these conditions and their interaction.

--- Table 1 about here ---

Clearly, the worst case concerning inadmissible solutions is to have a small number of groups together with a low intraclass correlation.

Parameter Estimates

In the within part of the model, the factor loadings have an overall relative bias of 0.0%, with no differences across the conditions. The relative bias of the error variances is -0.2%, with extremely small differences across the conditions. The largest bias is -0.7%, which occurs when the number of groups is 50 and the group size is 10.

In the between part of the model, the relative bias is somewhat higher. Overall, the relative bias of the factor loadings is 2.6%. Thus, a loading of 0.40 is typically estimated as 0.41. All main effects of Number of Groups, Group Size, ICC and Balance are significant, but the differences are generally small. The largest effects are, in order of magnitude, the effect of ICC, Number of Groups, and their interaction. The relative bias of the residual variances is higher; overall, it is -14.7%. The largest effects are again the effect of ICC, Number of Groups and their interaction. The relative biases for these conditions are presented in Table 2.

--- Table 2 about here ---

Table 2 presents the percentage bias relative to the nominal value. Since the relative bias differs very little across the factor loadings or variances, the results reported in Table 2 and further are the marginal (average) bias for all loadings or variances. Thus, the entry of 13.0 for the combination of a low ICC and a small number of groups means that the typical estimates of the factor loadings are 13% too high. This means that a loading of 0.3 would typically be estimated as 0.34, and a loading of 0.5 would typically be estimated as 0.57. As Table 2 shows, the bias of the factor loadings is generally very small, except in the combination of small number of groups and low ICC. The effects of Group Size and Balance on the factor loadings (not in the table) are negligible. In small groups, and with unbalanced data, the factor loadings are typically estimated about 1% too high. This difference would only show up in the third decimal place, and these effects are therefore not included in the table.

The error variances are moderately underestimated. When the Number of Groups is small, and the ICC is low, they are typically estimated at about half their true value. The other simulated conditions produce a much smaller bias. As with the factor loadings, the effects of Group Size and Balance on the variances (not in the table) are much smaller, about 4% bias for small groups or unbalanced data.

Standard Errors

In the within part of the model, the standard errors of the factor loadings have a relative bias of 0.3%, with minor differences for varying number of groups and group sizes. The relative bias of the standard error of the error variances is 0.6%, with again minor differences for varying number of groups and group sizes.

In the between part of the model, the standard errors show a clear bias. Overall, the relative bias of the standard errors for the factor loadings is -11.9%. Thus, a standard error of 0.15 is typically estimated as 0.13. All main effects of Number of Groups, Group Size, ICC and Balance are significant. The interaction between ICC and Number of Groups is also significant. The largest effects are, in order of magnitude, the effect of ICC, Number of

Groups, and their interaction. The relative bias of the standard errors for the residual variances is a very small -0.1. There are some small effects of, in order, NG, ICC, their interaction, followed by Balance and Group size. Table 3 presents the relative bias of the standard errors for the loadings and variances for the different levels of the four conditions, and Table 4 presents the relative bias for the interaction of ICC and Number of Groups.

--- Tables 3 and 4 about here ---

To assess the impact of the combination of biased estimates and biased standard errors, we computed the 95% confidence interval for each parameter, and counted how many times these intervals covered the true population value. The coverage of the factor loadings is affected by, in order of magnitude, Balance, ICC, and Number of Groups. There was no significant interaction. The coverage of the variances is affected by Balance, Number of Groups, and their interaction, in that order. Table 5 presents the coverage of the 95% confidence intervals for the loadings and variances for the different levels of the four conditions, and Table 6 the effect of the interaction of Balance and Number of Groups on the coverage for the variances.

--- Tables 5 and 6 about here ---

As Table 5 shows, in general the coverage of the 95% confidence intervals is too low. The coverage for the variances in the unbalanced procedure with a small number of groups improves when the number of groups increases from 50 to 100, but does not improve further if the number of groups rises to 200.

Model Fit

Since the fitted model equals the population model, the expected value for the chi-square is equal to the degrees of freedom, which is 18. The standard deviation of the chi-square distribution with 18 degrees of freedom is 6. In fitting a correct model, we expect only 5% of the chi-square tests to lead to a rejection at the 5% significance level. Overall, the mean of the chi-squares is 18.8, with a standard deviation of 6.34. The overall bias for the chi-square is 4.0% and there are 6.9% rejected models. Both the size of the chi-square and the probability of the model being rejected are related only to the Balance and the ICC. In the balanced case, the bias is -0.5%, with 4.9% rejections, and in the unbalanced case the bias is 8.6%, with 8.9% rejections. For low ICC, the bias of the chi-square is 2.2%, with 6.1% rejections, and for high ICC it is 5.9%, with 7.7% model rejections.

Summary and Discussion

The conditions that were varied in this simulation have very little impact on the accuracy of the parameter estimates and standard errors of the within-groups part of the model. This is not surprising, since the pooled-within covariance matrix is the maximum likelihood estimator of the population within-groups covariance matrix, the assumptions of the estimation method are met, and the sample size is sufficient. This will generally be the case, since the sample size for the within part of the model is generally much larger than the effective sample size for the between-groups part. The lowest limit is reached when the groups are dyadic couples, as in research on married couples or twins. Most naturally occurring groups will be larger, and consequently the within-groups sample size will generally be much larger than the between groups sample size. In our simulation, the effective within-groups sample sizes range from 450 to 9800 (N-G), which is sufficient for Maximum Likelihood estimation with normal data (cf. Hoogland & Boomsma, 1998). This is useful for researchers who want to estimate SEM models based on data that have been obtained from cluster-sampling or similar sampling schemes. Analyzing the pooled within-cluster covariance matrix instead of the total score covariance matrix proves a simple and effective strategy for controlling the bias stemming from the cluster sampling scheme, a point made earlier by Muthén (1989).

For the between groups part of the model, the number of groups, being the between groups sample size, and the ICC clearly have the largest effect on the accuracy of the estimates and standard errors. In the simulated conditions with 50 groups, the percentage of inadmissible solutions is overall 17%, and 27% when the intraclass correlation is low. A low number of groups is only partially compensated by having large groups, a high ICC, or balanced data. In our analyses, we have omitted the inadmissible solutions. If they are retained, the disturbing effects of these conditions on the parameter estimates become even larger. However, since the inadmissible solutions produce extreme outliers for the parameter estimates, the effects of the various conditions are difficult to interpret if they are included. Since the offending parameter estimates are so far from the admissible values, there is no apparent danger that researchers would attempt to give these a substantive interpretation. However, this also makes the inadmissible solutions completely useless, and their occurrence under specific conditions is a severe limitation on the use of multilevel SEM for applied researchers. Since our ‘Low’ intraclass correlation is relatively high, our results indicate that, if only to avoid inadmissible solutions, researchers should aim at obtaining data from at least 100 groups.

The relative bias of the between groups factor loadings is overall small at +2.6%. Only the combination of the smallest number of groups (50) and a low ICC yields a relative bias (13.0%) higher than the 5% that Hoogland and Boomsma (1998) consider still acceptable. The residual variances are estimated with considerably less accuracy; at least 100 groups or a large ICC are needed here to achieve a relative bias of less than 5% (cf. Table 2).

The between groups standard errors are less accurate than the parameter estimates. Overall, the relative bias is –11.9%, which is larger than the 10% Hoogland and Boomsma (1998) consider acceptable for standard errors. For accurate standard errors, a low ICC and large number of groups (at least 100) is needed. However, since the parameter estimates are also biased, accurately estimated standard errors do not necessarily result in confidence intervals with good coverage. If we require a nominal 95% confidence interval to have an empirical coverage of at least 90% and at most 99%, which corresponds to a bias of about 5%, we need a group level sample of at least 100 groups. With 50 groups, only a low ICC and balanced data will achieve this. Overall, with unbalanced data, even with a large number of groups, the operating alpha level is about 8%, instead of the nominal 5% level. The chi-square model test is accurate only in the balanced case, and in the unbalanced case it shows a bias of 8.6%, which results in a rejection rate of 8.4% instead of the nominal 5%.

There are strong similarities between our results and simulation research in the context of single-level SEM and multilevel regression modeling. The early simulation study by Boomsma (1983) already showed that sample sizes lower than 100 tend to lead to nonconvergence, inadmissible results, and inaccurate estimates. Generally, when the sample size is 200, the Maximum Likelihood method performs rather well (Boomsma, 1983; Chou & Bentler, 1995). Other simulation studies confirm these results. For an overview and meta-analysis of such studies, see Hoogland and Boomsma (1998).

Simulation studies on the accuracy of multilevel regression estimates also show results similar to ours (cf. Afshartous, 1995; Van der Leeden & Busing, 1994; Meijer, Busing & Van der Leeden, 1998; for an overview see Hox, 1998). In general, the regression coefficients corresponding to individual level predictors are estimated accurately, even with small numbers of groups. The variances and regression coefficients at the group level are estimated with less precision. For accurate estimates of the variances at the group level, at least 100 groups appear to be needed (cf. Busing, 1993). Just as in our results for the between groups model, the group level variances in multilevel regression analyses tend to be underestimated, and the standard errors of the parameter estimates that involve the group level are typically too small.

The approximation of a Full Information Maximum Likelihood solution for unbalanced data by using the average group size in a pseudobalanced solution (cf. equations (3), (4), and (8)) has but a small effect. In our simulations, the operating alpha level for both the overall model test and the tests on the individual parameters is about 8% when we have imbalanced data, which is a reasonable approximation. There are very few and only small interactions of imbalance with the other simulated conditions. It is reasonable to assume that the bias of the significance tests and the confidence intervals depend on the degree of imbalance, which is not varied in our study. We consider the degree of imbalance in our studies rather large, with half of the groups being one-third the size of the other half. In most empirical studies, the degree of imbalance will probably be less. Given our results, we caution against using multilevel SEM when the number of groups is smaller than 100, especially if the intraclass correlation turns out to be low, i.e., under 0.25. Both simulations (Van der Leeden & Busing, 1994; Mok, 1995) and analytic work (Snijders & Bosker, 1993; Cohen, 1998), suggest a trade-off between sample sizes at different levels. The work of Snijders and Bosker merits special attention, because they also consider the cost of obtaining an extra group of size n_g, compared with obtaining n_g more individuals within the available groups. They assume that, in general, obtaining one extra group is more expensive than obtaining the same number of additional individuals within the already available groups. Their article shows that the point is soon reached where statistical reasoning calls for increasing the number of groups, rather than increasing the number of individuals. Our simulations show the same. If the choice is between having a large number of groups, or a large number of individuals, the preferred alternative is to have more groups. In the case where the number of groups is by nature limited, e.g. when the groups are defined by the 50 states of the U.S.A., or the 28 Cantons in Switzerland, other measures become important. Increasing the group sizes helps, as well as increasing the intraclass correlation (difficult to do) and avoiding extreme unbalance. Using a robust chi-square as an alternative for the standard ML chi-square statistic may alleviate the problems (cf. Muthén & Satorra, 1995). A totally different solution might be to use bootstrap methods to correct the point estimates and to obtain empirical standard errors. Meijer, Busing and Van der Leeden (1998) show that there is some promise in that direction. A complication is that the bootstrap sampling procedure should reflect the multilevel model, i.e., we must resample groups, and then resample individuals within groups. Standard SEM software does not do this, and one would need to write a dedicated program to carry it out.

References

Afshartous, D. (1995). Determination of sample size for multilevel model design. Paper, AERA conference, San Francisco, 18-22 april 1995.

Boomsma, A. (1983). *On the Robustness of LISREL (Maximum Likelihood Estimation) against Small Sample Size and Nonnormality.* Amsterdam: Sociometric Research Foundation.

Busing, F. (1993). Distribution characteristics of variance estimates in two-level models. Unpublished manuscript. Leiden: Department of Psychometrics and Research Methodology, Leiden University.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical Linear Models*. Newbury Park, CA: Sage.

Bryk, A.S., Raudenbush, S.W., & Congdon, R.T. (1996). *HLM. Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs*. Chicago: Scientific Software International.

Chou, C.P., & Bentler, P.M. (1995). Estimates and tests in structural equation modeling. In R.H. Hoyle (Ed.), *Structural Equation Modeling: Concepts, Issues, and Applications*. (Pp. 37-55). Newbury Park, CA: Sage.

Cohen, M.P. (1998). Determining sample sizes for surveys with data analyzed by hierarchical linear models. *Journal of Official Statistics*, 14, 267-275.

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

Goldstein, H. (1995). *Multilevel Statistical Model*. London: Griffin.

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

Gustaffson, J.-E., & Stahl, P.E. (1999). *Streams User's Guide, Vs. 2.0*. Mölndal, Sweden: MultivariateWare.

Hoogland, J.J., & Boomsma, A. (1998). Robustness studies in covariance structure modelling: An overview and a meta-analysis. *Sociological Methods & Research*, 26, 3, 329-367.

Hox, J.J. (1993). Factor analysis of multilevel data: Gauging the Muthén model. In J.H.L. Oud & R.A.W.van Blokland-Vogelesang (Eds.). *Advances in longitudinal and multivariate analysis in the behavioral sciences*. (Pp.141-156). Nijmegen: ITS.

Hox, J.J. (1995). *Applied Multilevel Analysis*. Amsterdam, NL: TT-Publikaties.

Hox, J.J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, Data Analysis, and Data Highways*. New York: Springer.

Hox, J.J. (1999). Split2. (Version 2) [Computer Software]. Utrecht, NL: Author. Available from http://www.fss.uu.nl/ms/jh.

Jöreskog, K.G. & Sörbom, D. (1993). *LISREL8: User's Reference Guide.* Chicago: Scientific Software International.

Kaplan, D. & Elliott, P.R. (1997). A didactic example of multilevel structural equation modeling applicable to the study of organizations. *Structural Equation Modeling: A Multidisciplinary Journal,* 4, 1-24.

Kish, L. (1965). *Survey Sampling.* Wiley, New York.

Li, F., Duncan, T.E., Harmer, P., Acock, A., & Stoolmiller, M. (1998). Analyzing measurement models of latent variables through multilevel confirmatory factor analysis and hierarchical linear modeling approaches. *Structural Equation Modeling: A Multidisciplinary Journal.* 5, 3, 294-306.

McArdle, J.J. & Hamagami, F. (1996). Multilevel models from a multiple group structural equation perspective. Pp. 89-125 in G.A. Marcoulides & R.E. Schumacker (eds.). *Advanced Structural Equation Modeling*. Mahwah, NJ: Erlbaum.

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

Meijer, E., Busing, F.M.T.A., & Van der Leeden, R. (1998). Estimating bootstrap confidence intervals for two-level models. In: J.J. Hox & E.D. de Leeuw (Eds.), *Assumptions, Robustness, and Estimation Methods in Multivaiate Modeling*. Amsterdam, NL: TT-Publikaties.

Mok, M. (1995). Sample size requirements for 2-level designs in educational research. Unpublished manuscript. London: Multilevel Models Project, Institute of Education, University of London. Available from http://www.ioe.ac.uk/multilevel.

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

Muthén, B. (1990). *Means and Covariance Structure Analysis of Hierarchical Data.* Los Angeles: UCLA Statistics series, #62.

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

Muthén, L.K., & Muthén, B.O. (1998). *Mplus. The Comprehensive Modeling Program for Applied Researchers*. Los Angeles: Muthén & Muthén.

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

Muthén, B. & Satorra, A. (1995). Complex sample data in structural equation modeling. Pp. 267-316 in P.V. Marsden (Ed.), *Sociological Methodology, 1995*. Oxford: Blackwell.

Muthén, B., Wisnicky, K.S., & Nelson, G. (1991). *Estimation of covariance structure models with multilevel (clustered) data*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 3-7, 1991.

Rasbash, J., & Woodhouse, G. (1995). *MLn Command Reference*. London: Instutute of Education, University of London.

Searle, S.R., Casella, G., & McCulloch, C.E. (1992). *Variance Components*. New York: Wiley.

Snijders, T.A.B., & Bosker, R.J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*, 237-259.

Van der Leeden, R., & Busing, F. (1994). First iteration versus IGLS/RIGLS estimates in two-level models: A Monte Carlo study with ML3. Unpublished manuscript. Leiden: Department of Psychometrics and Research Methodology, Leiden University.

Appendix

The model for data generation is presented in Table 1 in matrix format. Note that for Y1 and Y2 the variance is 1, and for X1 it is either .25 or .50 depending on the simulation condition for the ICC.

Table 1.

Factor Matrix for Within and Between Model

	Factor loadings within model		Factor loadings between model	All error var.
Variable	Y1	Y2	X1	
1	.3		.5	.25
2	.4		.4	.25
3	.5		.3	.25
4		.3	.5	.25
5		.4	.4	.25
6		.5	.3	.25

The implied population covariance matrices are give in Table 2 and 3:

Table 2.

Implied Covariances, Within Model

Var.	w1	w2	w3	w4	w5	w6
w1	0.34					
w2	0.12	0.41				
w3	0.15	0.20	0.50			
w4	0.00	0.00	0.00	0.34		
w5	0.00	0.00	0.00	0.12	0.41	
w6	0.00	0.00	0.00	0.15	0.20	0.50

Table 3

Implied Covariances, Between, for X1 Variance is 0.25

	b1	b2	b3	b4	b5	b6
b1	0.3120					
b2	0.0500	0.2900				
b3	0.0375	0.0300	0.2730			
b4	0.0625	0.0500	0.0375	0.312		
b5	0.0500	0.0400	0.0300	0.0500	0.2900	
b6	0.0375	0.0300	0.0225	0.0375	0.0300	0.2730

Implied Covariances, Between, for X1 Variance is 0.50

	b1	b2	b3	b4	b5	b6
b1	0.375					
b2	0.100	0.330				
b3	0.075	0.060	0.295			
b4	0.125	0.100	0.075	0.375		
b5	0.100	0.080	0.060	0.100	0.330	
b6	0.075	0.060	0.045	0.075	0.060	0.295

Author Note

We thank Anne Boomsma and Godfried van den Wittenboer, and three anonymous reviewers, for their helpful comments.

Tables

TABLE 1

Percentage of inadmissible solutions, for different Number of Groups (NG) and ICC[a]

NG	50	100	200	Total
ICC				
low	27.4%	9.1%	1.0%	12.5%
high	5.7%	0.0%	0.0%	2.1%
Total	16.6%	4.7%	0.1%	

[a] All data, i.e., data collapsed over different group sizes and balanced/unbalanced conditions

TABLE 2

Relative bias estimates, between model, for different Number of Groups (NG) and ICC[a]

	NG	50	100	200	Total	NG	50	100	200	Total
ICC		bias loadings				ICC	bias variances			
low		13.0	3.0	0.4	5.5	low	-52.9	-16.4	-3.1	-24.1
high		0.1	-0.7	-0.2	-0.3	high	-11.8	-2.8	-1.1	-5.2
Total		6.5	1.2	0.1		Total	-32.3	-9.6	-2.1	

[a] All data, i.e., data collapsed over different group sizes and balanced/unbalanced conditions

TABLE 3

Relative bias standard errors for loadings and variances, main effects[a]

	bias s.e. loadings	bias s.e. variances
Balanced	-11.3	0.2
Unbalanced	-12.4	-0.4
ICC low	-7.1	1.1
ICC high	-16.7	-1.3
NG 50	-15.7	-0.1
NG 100	-11.0	0.5
NG 200	-9.0	0.2
GS 10	-12.7	0.3
GS 20	-11.7	0.0
GS 50	-11.2	0.0

[a]ICC=intraclass correlation; NG=number of groups; GS=group size.

TABLE 4

Relative bias standard errors, between loadings and variances, by NG and ICC[a]

	NG	50	100	200	Total	NG	50	100	200	Total
ICC		bias s.e. loadings				ICC	bias s.e. variances.			
low		-12.7	-7.7	-2.8	-7.1	low	0.2	0.2	0.1	0.1
high		-18.6	-16.3	-15.1	-16.7	high	-2.2	-1.1	0.5	-1.3
Total		-15.7	-11.0	-9.0		Total	-1.0	0.5	0.2	

[a] All data, i.e., data collapsed over different group sizes and balanced/unbalanced conditions.

ICC=intraclass correlation; NG=number of groups; GS=group size.

TABLE 5

Coverage of 95% C.I. for loadings and variances, main effects[a]

	loadings coverage	variances coverage
Balanced	92.6	93.0
Unbalanced	89.1	90.1
ICC low	89.9	91.9[a]
ICC high	91.7	91.8[a]
NG 50	89.5	90.2
NG 100	91.0	92.4
NG 200	91.1	92.9
GS 10	90.9[a]	91.9[a]
GS 20	90.8[a]	91.8[a]
GS 50	90.8[a]	91.8[a]

[a] Main effect not significant, but left in table to show overall coverage.

ICC=intraclass correlation; NG=number of groups; GS=group size.

TABLE 6

Coverage 95% C.I., between variances, by Balance and Number of Groups[a]

NG	50	100	200	Total
Balance	coverage 95% C.I.			
balanced	92.0	93.6	94.5	93.0
unbalanced	88.4	91.2	91.4	90.1
Total	90.2	92.4	92.9	

[a]All data, i.e., data collapsed over different group sizes and ICC conditions

Figure Captions

Figure 1. Path diagram for the within and between simulation model.

Figures

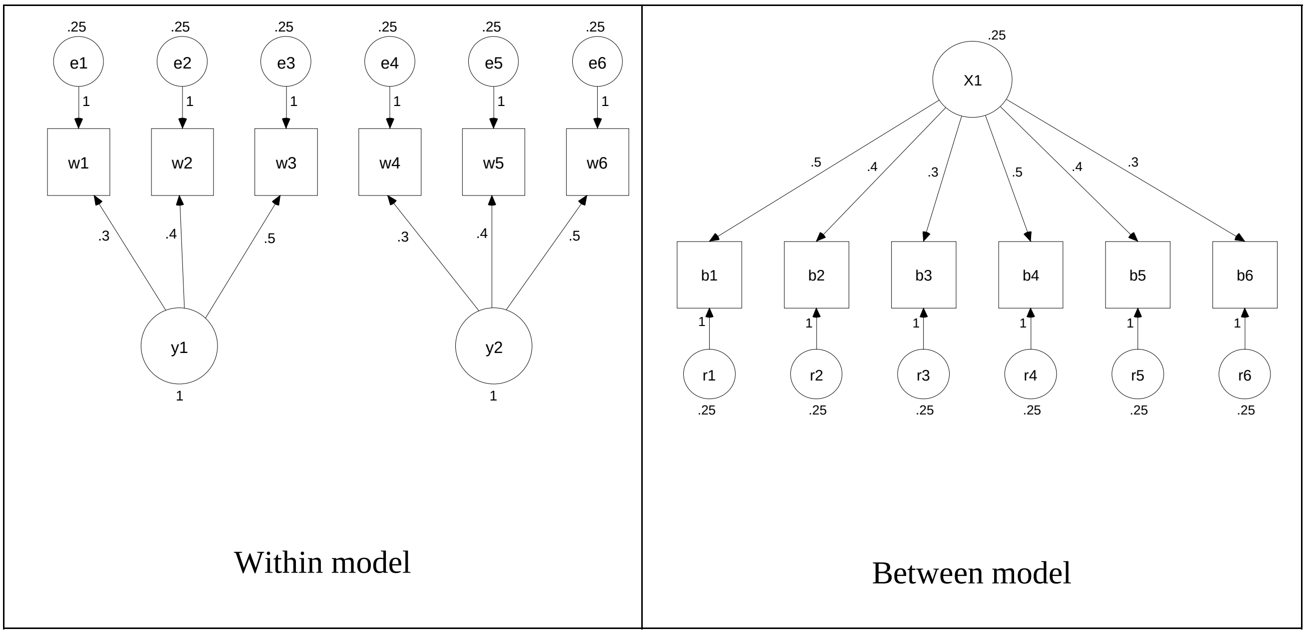

Figure 1. Path diagram for the within and between simulation model.