Sufficient Sample Sizes for Multilevel Modeling

Cora J. M. Maas and Joop J. Hox

Utrecht University, The Netherlands

Abstract. An important problem in multilevel modeling is what constitutes a sufficient sample size for accurate estimation. In multilevel analysis, the major restriction is often the higher-level sample size. In this paper, a simulation study is used to determine the influence of different sample sizes at the group level on the accuracy of the estimates (regression coefficients and variances) and their standard errors. In addition, the influence of other factors, such as the lowest-level sample size and different variance distributions between the levels (different intraclass correlations), is examined. The results show that only a small sample size at level two (meaning a sample of 50 or less) leads to biased estimates of the second-level standard errors. In all of the other simulated conditions the estimates of the regression coefficients, the variance components, and the standard errors are unbiased and accurate.

Keywords: multilevel modeling, hierarchical linear model, sample size, cluster sampling

Introduction

Social and organizational research often studies the relationship between individuals and the groups or organizations they belong to. The general concept is that individuals interact with their social contexts, and therefore the individual persons are influenced by the social groups or contexts to which they belong, and the properties of those groups are in turn influenced by the individuals who make up that group. Generally, the individuals and the social groups are conceptualized as a hierarchical system of individuals and groups, with individuals and groups defined at separate levels of this system.

Such systems can be observed at different levels, and as a result produce data with variables observed at several distinct hierarchical levels. This leads to research and analysis problems that focus on the interaction of variables that describe the individuals and variables that describe the groups. This kind of research is now generally referred to as *multilevel research.* Several analysis strategies exist for analyzing multilevel data; for an overview, we refer to Klein and Kozlowski (2000). One important class of analysis methods is the hierarchical linear regression model, or multilevel regression model. As Cohen and Cohen (1983) show, the ordinary multiple regression model is highly versatile. Using dummy coding for categorical variables, it can be used for analysis-of-variance (ANOVA) models as well as for the more usual multiple regression models. This versatility carries over to multilevel regression analysis, which is essentially a multilevel extension of multiple regression analysis. In addition to organizational research, where the levels are defined by the individuals and the distinct levels in the organizations they belong to (cf. Bryk & Raudenbush, 1989), multilevel regression analysis can be applied to longitudinal data, where the levels are defined by the measurement occasions nested within individuals (Raudenbush, 1989; Snijders, 1996). For a general introduction to multilevel modeling of hierarchical data, we refer to Snijders and Bosker (1999), Heck and Thomas (2000), Raudenbush and Bryk (2002), and Hox (2002).

The maximum likelihood (ML) estimation methods used commonly in multilevel analysis are asymptotic, which translates to the assumption that the sample size must be sufficiently large. This raises questions about the acceptable lower limit to the sample size, and the accuracy of the estimates and the associated standard errors with relatively small sample sizes. In multilevel studies, the main problem is usually the sample size at the group level, because the group-level sample size is always smaller than the individual-level sample size. Increasing the number of groups may be difficult for two distinct reasons. Firstly, there are costs involved. Increasing the number of individuals in the sample means collecting data on more individuals within the sampled organizations. Increasing the number of organizations in the sample means bringing a new organization into the study. The latter is typically more expensive than the former. Secondly, we may already have all existing organizations in our study. If the object is to study how

Methodology 2005; Vol. 1(3):86–92
DOI 10.1027/1614-1881.1.3.86

organizational behavior in Switzerland is affected by characteristics of the different cantons, 26 is the absolute limit to the canton-level sample size.

The few simulation studies that have been carried out to date, which are reviewed later in this paper, suggest that the group-level sample size is generally more important than the total sample size, with large individual-level sample sizes partially compensating for a small number of groups. However, these studies use different approaches, and there is no strong evidence to guide researchers in their multilevel design decisions. There is a need for more insight in the problem of multilevel sample-size requirements. In this study, we use simulation to examine the accuracy of the parameter estimates and the corresponding standard errors, at different sample sizes for the number of groups and the number of individuals. The simulation design will be explained in more detail after the next section, which describes the multilevel regression model and reviews currently available simulation studies.

The Multilevel Regression Model

Assume that we have data from J groups, with a different number of respondents n_j in each group. On the respondent level, we have the outcome of respondent i in group j, variable Y_{ij}. We have one explanatory variable X_{ij} on the respondent level, and one group-level explanatory variable Z_j. To model these data, we have a separate regression model in each group as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}. \quad (1)$$

The variation of the regression coefficients β_j is modeled by a group-level regression model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j}, \quad (2)$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + u_{1j}, \quad (3)$$

The individual-level residuals e_{ij} are assumed to have a normal distribution with mean zero and variance σ^2_e. The group-level residuals u_{0j} and u_{1j} are assumed to have a multivariate normal distribution with expectation zero, and to be independent from the residual errors e_{ij}. The variance of the residual errors u_{0j} is specified as σ^2_e and the variance of the residual errors u_{0j} and u_{1j} is specified as σ^2_{u0} and σ^2_{u1}.

This model can be written as one single regression model by substituting Equations 2 and 3 into Equation 1. Substitution and rearranging terms gives

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}X_{ij}Z_j + u_{0j} + u_{1j}X_{ij} + e_{ij} \quad (4)$$

The segment $\gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}X_{ij}Z_j$ in Equation 4 contains all the fixed coefficients; it is the fixed (or deterministic) part of the model. The segment $u_{0j} + u_{1j}X_{ij} + e_{ij}$ in Equation 4 contains all the random error terms; it is the random (or stochastic) part of the model. The term $X_{ij}Z_j$ is an interaction term that appears in the model because of modeling the varying regression slope β_{1j} of the respondent-level variable X_{ij} with the group level variable Z_j.

Even if the analysis includes only variables at the lowest (individual) level, standard multivariate models are not appropriate. Multilevel models are needed because grouped data violate the assumption of independence of all observations. The amount of dependence can be expressed as the intraclass correlation (ICC) ρ. In the multilevel model, the ICC is estimated by specifying an empty model, as follows:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}. \quad (5)$$

This model does not explain any variance in Y. It only decomposes the variance of Y into two independent components: σ^2_e, which is the variance of the lowest-level errors e_{ij}, and σ^2_{u0}, which is the variance of the highest-level errors u_{0j}. Using this model, the ICC ρ is given by the equation

$$\rho = \sigma^2_{u0}/(\sigma^2_{u0} + \sigma^2_e). \quad (6)$$

In addition to the sample sizes at the separate levels, the size of the ICC also may affect the accuracy of the estimates (Goldstein, 1995). Therefore, in our simulation, we have varied not only the sample size at the individual and the group level, but also the ICC. In general, what is at issue in multilevel modeling is not so much the ICC, but the *design effect*, which indicates how much the standard errors are underestimated in a complex sample (Kish, 1965) compared to a simple random sample. In cluster samples, the design effect is approximately equal to 1 + (average cluster size − 1) × ICC. In the multilevel context, Muthén and Satorra (1995) view a design effect of two as small. 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.

So far, three major sample properties have been discussed that may affect the estimates: number of groups, number of level-one units, and the ICC. In addition, the estimation method may be important. Multilevel modeling mostly uses ML estimation. Two ML functions are common in multilevel modeling: full ML (FML) and restricted ML (RML) (for a description of these, see Hox, 2002). The difference between FML and RML is that RML maximizes a likelihood function that is invariant for the fixed effects (Goldstein, 1995). Since RML takes the uncertainty in the fixed parameters into account when estimating the random parameters, it should in theory lead to better estimates of the variance

components, especially when the number of groups is small (Raudenbush & Bryk, 2002).

Review of Existing Research on Sample Size

There are some simulation studies on this topic, which mostly investigate the accuracy of the estimates for the fixed and random parameters with small sample sizes at either the individual or the group level. Comparatively less research investigates the accuracy of the standard errors. Most simulation research that addresses the accuracy of the significance test or the coverage of the confidence interval is based on asymptotic reasoning: the standard error is used in conjunction with the standard normal distribution to generate a *p*-value or confidence interval. Other approaches exist that may be more valid in small samples, but are not available in all multilevel software; such alternative approaches are taken up in the discussion.

Testing variances using their standard errors is not optimal because it assumes normality, and because the null hypothesis that a variance is zero is a test on the boundary of the permissible parameter space (variances cannot be negative), where standard likelihood theory is no longer valid. A variety of alternative approaches have been proposed (cf. Berkhof & Snijders, 2001, for a review). Since testing variances using their asymptotic standard errors is still widely used, we incorporate this procedure in our review and simulation, but discuss the issues separately for the fixed and the random part.

Accuracy of Regression Coefficients and Their Standard Errors

The estimates for the regression coefficients appear generally unbiased, for ordinary least squares (OLS) and generalized least squares (GLS) as well as ML estimation (Van der Leeden & Busing, 1994; Van der Leeden, Busing, & Meijer, 1997). OLS estimates are less efficient; Kreft (1996), reanalyzing results from Kim (1990), finds OLS estimates that are about 90% efficient. Simulations by Van der Leeden and Busing (1994) and Van der Leeden et al. (1997) suggest that even when assumptions of normality and large samples are not met, ML-based standard errors for the fixed parameters have only a small downward bias. In general, a large number of groups appears more important than a large number of individuals per group.

Accuracy of Variance Components and Their Standard Errors

Estimates of the lowest-level variance are generally very accurate. The group-level variance components are sometimes underestimated. Simulation studies by Busing (1993) and Van der Leeden and Busing (1994) show that for accurate group-level variance estimates many groups (more than 100) are needed (cf. Afshartous, 1995).

The simulations by Van der Leeden et al. (1997) show that the standard errors used to test the variance components are generally estimated too small, with RML again more accurate than FML. Symmetric confidence intervals around the estimated value also do not perform well. Browne and Draper (2000) report similar results. Typically, with 24–30 groups, Browne and Draper report an operating alpha level of about 9%, and with 48–50 groups about 8%. Again, a large number of groups appears more important than a large number of individuals per group.

Simulation Design

The Simulation Model and Procedure

We use a simple two-level model, with one explanatory variable at the individual level and one explanatory variable at the group level, conforming to Equation 4, which is repeated here:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}X_{ij}Z_j + u_{0j} + u_{1j}X_{ij} + e_{ij} \quad \text{(4, repeated)}$$

Three conditions are varied in the simulation: (1) number of groups (NG: three conditions, NG = 30, 50, 100), (2) group size (GS: three conditions, GS = 5, 30, 50), and (3) intraclass correlation (ICC: three conditions, ICC = 0.1, 0.2, 0.3).[1]

The number of groups is chosen so that the highest number should be sufficient given the simulations by Van der Leeden et al. (1997). In practice, 50 groups is a frequently occurring number in organizational and school research, and 30 is the smallest acceptable number according to Kreft and De Leeuw (1998). Similarly, the group sizes are chosen so that the highest number should be sufficient. A group size of 30 is normal in educational research, and a group size of 5 is normal in family research and in longitudinal research, where the measurement occasions form the lowest level. The ICCs

1 Corrections for clustering based on the design effect (Kish, 1965) assume equal group sizes; multilevel analysis does not. We carried out some preliminary simulations to assess whether having balanced or unbalanced groups has any influence on multilevel ML estimates, with a view to including this in the simulation design. However, despite extreme unbalance, there was no discernible effect of unbalance on the multilevel estimates or their standard errors.

span the customary range of ICC coefficients (Gulliford, Ukoumunne, & Chinn, 1999).

There are $3 \times 3 \times 3 = 27$ conditions. For each condition, we generated 1,000 simulated data sets, assuming normally distributed residuals. The multilevel regression model, like its single-level counterpart, assumes that the explanatory variables are fixed. Therefore, a set of X and Z values are generated from a standard normal distribution to fulfill the requirements of the simulation condition with the smallest total sample size. In the condition with the larger sample sizes, these values are repeated. This ensures that in all simulated conditions the joint distribution of X and Z are identical. The regression coefficients are specified as follows: 1.00 for the intercept, and 0.3 (a medium effect size; cf. Cohen, 1988) for all regression slopes. The residual variance σ_e^2 at the lowest level is 0.5. The residual variance σ_{u0}^2 follows from the ICC and σ_e^2, given Equation 6. Busing (1993) shows that the effects for the intercept variance σ_{00} and the slope variance σ_{11} are similar; hence, we chose to set the value of σ_{11} equal to σ_{00}. To simplify the simulation model, the covariance between the two u-terms is assumed equal to zero.

Two ML functions are common in multilevel estimation: FML and RML. We use RML, since this is almost always at least as good as FML, and sometimes better, especially in estimating variance components (Browne, 1998). The software MLwiN (Rasbash et al., 2000) was used for both simulation and estimation.

Variables and Analysis

The accuracy of the parameter estimates (factor loadings and residual variances) is indicated by the percentage relative bias. Let $\hat{\theta}$ be the estimate of the population parameter θ; then the percentage relative bias is given by $100 \times \hat{\theta}/\theta$. The accuracy of the standard errors is investigated by analyzing the observed coverage of the 95% confidence interval. Since there are 27,000 simulated conditions, the power is huge. (Given 1,000 simulations in each condition, the standard error for the coverage probabilities is about 0.007.) As a result, at the standard significance level of alpha $= 0.05$, extremely small bias or coverage errors become significant. Therefore, our criterion for significance is an alpha $= 0.01$ for the main effects of the simulated conditions. The interactions are tested blockwise (two-way, three-way), with a blockwise Bonferroni correction. Even at this stricter 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.

Results

Convergence and Inadmissible Solutions

The estimation procedure converged in all 27,000 simulated data sets. The estimation procedure in MLwiN can and sometimes does lead to negative variance estimates. Such solutions are inadmissible, and the common procedure is to constrain such estimates to the boundary value of zero. However, all 27,000 simulated data sets produced only admissible solutions.

Parameter Estimates

The fixed parameter estimates, the intercept and regression slopes, have a negligible bias. The average bias is smaller than 0.05%. The largest bias was found in the condition with the smallest sample sizes in combination with the highest ICC: there the percentage relative bias was 0.3%. This is of course extremely small. Moreover, there are no statistically significant differences in bias across the simulated conditions.

The estimates of the random parameters, the variance components, also have a negligible bias. The average bias is smaller than 0.05%. The largest bias was found in the condition with the smallest sample sizes in combination with the highest ICC: there the percentage relative bias was again 0.3%. Also, there are no statistically significant differences in bias across the simulated conditions.

Standard Errors

To assess the accuracy of the standard errors, for each parameter in each simulated data set the 95% confidence interval was established using the asymptotic standard normal distribution (cf. Goldstein, 1995). For each parameter a noncoverage indicator variable was set up that is equal to zero if the true value is in the confidence interval, and equal to one if the true value is outside the confidence interval. The effect of the number of groups on the noncoverage is presented in Table 1, the effect of the group size on noncoverage is presented in Table 2, and the effect of the ICC on noncoverage is presented in Table 3. Logistic regression was used to assess the effect of the different simulated conditions on the noncoverage, which is reported in Tables 1,2,3 as a p-value for each outcome (last column in the tables).

Table 1 shows that the effect of the number of groups on the standard errors of the fixed regression coefficients is small. With 30 groups, the noncoverage rate is 6.0% for the regression coefficient for X and 6.4% for the intercept, while the nominal noncoverage rate is 5%. We regard this difference as unimportant. The effect of the number of groups on the standard errors of the variance

Table 1. Noncoverage of the 95% Confidence Interval by Number of Groups

Parameter	Number of groups			*p*-value[a]
	30	50	100	
U0	.089	.074	.060	.0000
U1	.088	.072	.057	.0000
E	.058	.056	.049	.0101
INT	.064	.057	.053	.0057
X	.060	.057	.050	.0058
Z	.052	.051	.050	.9205
XZ	.056	.052	.050	.2187

[a] *p*-value for the effect of number of groups on mean parameter estimate.

Table 2. Noncoverage of the 95% Confidence Interval by Group Size

Parameter	Group size			*p*-value[a]
	5	30	50	
U0	.074	.075	.074	.9419
U1	.078	.066	.072	.0080
E	.061	.051	.051	.0055
INT	.062	.056	.055	.0846
X	.055	.054	.057	.6086
Z	.055	.048	.049	.0782
XZ	.057	.051	.050	.1169

[a] *p*-value for the effect of group size on mean parameter estimate.

Table 3. Noncoverage of the 95% Confidence Interval by Intraclass Correlation

Parameter	Intraclass correlation			*p*-value[a]
	0.1	0.2	0.3	
U0	.073	.073	.077	.4973
U1	.071	.073	.073	.8871
E	.056	.055	.053	.6212
INT	.059	.059	.055	.4238
X	.056	.052	.059	.2014
Z	.049	.049	.055	.1175
XZ	.051	.055	.052	.3931

[a] *p*-value for the effect of intraclass correlation on mean parameter estimate.

components is definitely larger. With 30 groups, the noncoverage rate for the second-level intercept variance is 8.9% (U0), and the noncoverage rate for the second-level slope variance is 8.8% (U1). Although the coverage is not grotesquely wrong, the 95% confidence interval is clearly too short. The amount of noncoverage here implies that the standard errors for the second-level variance components are estimated about 15% too small. With 50 groups, the effects are smaller, but still not negligible. The noncoverage rates of 7.4% and 7.2% imply that the standard errors are estimated about 9% too small.

Table 2 shows the same pattern: the coverage of the 95% confidence interval is good for the regression coefficients, and less accurate for the variances. The coverage rates improve when the group size increases, but the group size has a smaller effect on the coverage rates than the number of groups. The noncoverage of the second-level variances does not improve when the group size increases.

Table 3 again shows that the coverage rates for the regression coefficients are better than for the variances. The difference in ICC has no effect on the coverage rate.

The interaction effects of these simulated characteristics are presented in Table 4. Tested with a blockwise Bonferroni correction, none of the interactions were statistically significant.

Given that we use asymptotic (large sample) methods, the absence of bias in the parameter estimates and the small bias in the standard errors is encouraging. To investigate the limits of our results, we carried out one additional simulation, with only 10 groups of group size five, varying the ICC from 0.1 to 0.3. This was inspired by a statement in Snijders and Bosker (1999, p. 44) that multilevel modeling becomes attractive when the number of groups is larger than 10. This is indeed a very small group-level sample size, but given our simulation results not impossibly small. The results of this small simulation are presented in Table 5.

The regression coefficients and lowest-level variance components are again estimated without bias. However, the group-level variance components were estimated much too large, with a bias up to 25%. The standard errors were too small, both for regression coefficients and for variances. The noncoverage rates for the regression coefficients ranged between 5.7% and 9.7%, and for the second-level variances they ranged between 16.3% and 30.4%. In our view, this indicates that having only 10 groups is not enough. Although the ML-standard errors of the regression coefficients are still within bounds of reason, the standard errors of the second-level variances are clearly unacceptable.

Summary and Discussion

Summing up, both the regression coefficients and the variance components are all estimated without bias, in all of the simulated conditions. The standard errors of the regression coefficients are also estimated accurately, in all of the simulated conditions. The standard errors of the second-level variances are estimated too small when the number of groups is substantially lower than 100. With 30 groups, the standard errors are estimated about 15% too small, resulting in a noncoverage rate of almost 8.9%, instead of 5%. With 50 groups, the noncoverage drops to about 7.3%. This is clearly different

Table 4. Interaction Effects of the Number of Groups, Group Size, and Intraclass Correlation on the Noncoverage of the 95% Confidence Interval

Number of groups	Group size	ICC	UO	U1	E	INT	X	Z	XZ
30	5	0.1	.103	.116	.069	.064	.060	.056	.058
	5	0.2	.086	.093	.067	.073	.063	.052	.062
	5	0.3	.080	.093	.051	.072	.067	.071	.065
	30	0.1	.088	.064	.047	.060	.065	.052	.046
	30	0.2	.079	.075	.053	.064	.049	.041	.051
	30	0.3	.095	.089	.064	.057	.055	.047	.057
	50	0.1	.084	.082	.070	.065	.062	.037	.048
	50	0.2	.082	.091	.050	.075	.060	.048	.063
	50	0.3	.107	.086	.053	.043	.062	.060	.053
50	5	0.1	.065	.072	.061	.053	.054	.057	.072
	5	0.2	.070	.066	.058	.059	.057	.051	.047
	5	0.3	.080	.084	.068	.061	.057	.040	.048
	30	0.1	.083	.060	.051	.065	.056	.058	.048
	30	0.2	.080	.080	.065	.061	.043	.052	.059
	30	0.3	.082	.065	.052	.053	.070	.040	.054
	50	0.1	.062	.072	.048	.055	.064	.044	.049
	50	0.2	.069	.076	.052	.052	.050	.061	.056
	50	0.3	.073	.071	.052	.054	.059	.055	.039
100	5	0.1	.060	.065	.060	.068	.047	.051	.054
	5	0.2	.060	.059	.062	.048	.041	.053	.062
	5	0.3	.061	.057	.050	.061	.048	.065	.043
	30	0.1	.053	.050	.050	.047	.051	.045	.047
	30	0.2	.059	.058	.043	.053	.051	.039	.046
	30	0.3	.056	.056	.035	.047	.048	.060	.053
	50	0.1	.058	.060	.046	.054	.041	.037	.041
	50	0.2	.071	.059	.043	.047	.058	.047	.053
	50	0.3	.058	.053	.048	.048	.061	.055	.052

Table 5. Noncoverage of the 95% Confidence Interval (10 Groups of Size Five) for Different Intraclass Correlations

Parameter	Intraclass correlation			*p*-value
	0.1	0.2	0.3	
U0	.294	.163	.172	.0000
U1	.304	.215	.203	.0000
E	.095	.084	.073	.2076
INT	.061	.086	.087	.0490
X	.071	.094	.075	.1286
Z	.072	.078	.080	.7829
XZ	.062	.097	.077	.0142

[a] *p*-value for the effect of intraclass correlation on mean parameter estimate.

from the nominal 5%, but in practice probably acceptable.

The results from the various simulations reviewed in this article appear somewhat inconsistent. However, this is probably the result of using different simulation designs and different simulated conditions. Our results are comparable with the simulation results reported by Busing (1993), Van der Leeden and Busing, (1994) and Browne and Draper (2000) when only those conditions are considered that are similar to the conditions in our simulations.

In our study, we investigate the effect of the simulated design characteristics on the accuracy of the parameter estimates and their standard errors, using the standard large-sample approach to constructing the 95% confidence interval. As noted earlier, this is not optimal for testing variances, even if the sample sizes are sufficiently large. Berkhof and Snijders (2001) show that a modified likelihood-ratio test is better for testing variance components. When fixed effects (regression coefficients) are tested in small samples, methods exist that may be more accurate. Approximations based on estimating the number of degrees of freedom for the Wald test (Satterthwaite approximation; cf. Elston, 1998) or on correcting the likelihood-ratio test (Welham & Thompson, 1997) appear to be useful here (Manor & Zucker, 2004). However, these methods are not implemented in widely used multilevel software such as MLwiN or HLM.

Since the estimates of the regression coefficients are unbiased, even if the sample is as small as 10 groups of five units, bootstrapping or other simulation-based methods (cf. Goldstein, 2003; Hox, 2002) may also be useful to assess the sampling variability, provided we are interested only in the regression coefficients. Bootstrapping small sample data is discussed by Yung and

Chan (1999). However, these methods are currently only implemented in MLwiN.

Acknowledgments

We gratefully acknowledge the comments of two anonymous reviewers and of Michael Eid on an earlier version.

References

Afshartous, D. (1995, April). *Determination of sample size for multilevel model design.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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

Browne, W. J. (1998). *Applying MCMC methods to multilevel models*. Unpublished doctoral dissertation, University of Bath, UK.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*, 391–420.

Bryk, A. S., & Raudenbush, S. W. (1989). Methodology for cross-level organizational research. *Research in Organizational Behavior*, *7*, 233–273.

Busing, F. (1993). *Distribution characteristics of variance estimates in two-level models*. Unpublished manuscript, Leiden University, the Netherlands.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Elston, D. A. (1998). Estimation of the denominator degrees of freedom of F-distributions for assessing Wald statistics for fixed-effect factors in unbalanced mixed designs. *Biometrics*, *41*, 477–486.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edwards Arnold. New York: Halstead.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Arnold.

Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies. *American Journal of Epidemiology*, *149*, 876–883.

Heck, R. H., & Thomas, S. L. (2000). *An introduction to multilevel modeling techniques*. Mahwah, NJ: Erlbaum.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Kim, K.-S. (1990). *Multilevel data analysis: A comparison of analytical alternatives*. Unpublished doctoral dissertation, University of California at Los Angeles.

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

Klein, K. & Kozlowski, S. W. J. (Eds.). (2000). *Multilevel theory, research, and methods in organizations*. San Francisco: Jossey-Bass.

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Unpublished manuscript, California State University at Los Angeles. Retrieved July 6, 2005 from http://www.calstatela.edu/faculty/ikreft/quarterly.html

Kreft, I. G. G., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage.

Manor, O., & Zucker, D. M. (2004). Small sample inference for the fixed effects in the mixed linear model. *Computational Statistics and Data Analysis*, *46*, 801–817.

Muthén, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 267–316). Oxford, England: Blackwell.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., et al. (2000). *A user's guide to MLwiN*. London: Multilevel Models Project, University of London.

Raudenbush, S. W. (1989). The analysis of longitudinal, multilevel data. *International Journal of Educational Research*, *13*, 721–740.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*. Newbury Park, CA: Sage.

Snijders, T. A. B. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality and Quantity*, *30*, 405–426.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

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 University, the Netherlands.

Van der Leeden, R., Busing, F., & Meijer, E. (1997, April). *Applications of bootstrap methods for two-level models*. Paper presented at the Multilevel Conference, Amsterdam.

Welham, S. J., & Thompson, R. (1997). Likelihood ratio tests for fixed model terms using residual maximum likelihood. *Journal of the Royal Statistical Society, Series B*, *59*, 701–714.

Yung, Y.-F., & Chan, W. (1999). Statistical analysis using bootstrapping: Concepts and implementation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 82–105). Thousand Oaks, CA: Sage.

Address for correspondence

Cora J. M. Maas
Department of Methodology and Statistics
Faculty of Social Sciences, Utrecht University
POB 80140
NL-3508 TC Utrecht
The Netherlands
Tel. +31 30 253 4594
Fax: +31 30 253 5797
E-mail: c.maas@fss.uu.nl

