Statistica Neerlandica (2004) Vol. 58, nr. 2, pp. 127–137

Robustness issues in multilevel regression analysis

Cora J. M. Maas* and Joop J. Hox

Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, P.O. Box 80140, NL-3508 TC Utrecht, the Netherlands

A multilevel problem concerns a population with a hierarchical structure. A sample from such a population can be described as a multistage sample. First, a sample of higher level units is drawn (e.g. schools or organizations), and next a sample of the sub-units from the available units (e.g. pupils in schools or employees in organizations). In such samples, the individual observations are in general not completely independent. Multilevel analysis software accounts for this dependence and in recent years these programs have been widely accepted. Two problems that occur in the practice of multilevel modeling will be discussed. The first problem is the choice of the sample sizes at the different levels. What are sufficient sample sizes for accurate estimation? The second problem is the normality assumption of the level-2 error distribution. When one wants to conduct tests of significance, the errors need to be normally distributed. What happens when this is not the case? In this paper, simulation studies are used to answer both questions. With respect to the first question, the results show that a small sample size at level two (meaning a sample of 50 or less) leads to biased estimates of the second-level standard errors. The answer to the second question is that only the standard errors for the random effects at the second level are highly inaccurate if the distributional assumptions concerning the level-2 errors are not fulfilled. Robust standard errors turn out to be more reliable than the asymptotic standard errors based on maximum likelihood.

Key Words: multilevel modeling, sample size, cluster sampling, maximum likelihood, (robust) standard errors, sandwich estimate, Huber/White correction.

1 Introduction

Social research questions often relate to hierarchical systems. For instance, in educational research the achievements of the students are frequently modeled as the result of a combination of individual characteristics, such as intelligence and behavior, and school characteristics, such as the number of students in a group and

*c.maas@fss.uu.nl

the expertise of the teachers. In the educational system, the students constitute the lower level and the schools the higher level. Other examples are organizational research with individuals nested within organizations and longitudinal research with repeated observations nested within individuals. Standard multivariate models are not appropriate for the analysis of such hierarchical systems, even if the analysis includes only variables at the lowest (individual) level, because the individual observations are in general not independent. This results in a violation of the standard assumption of independent and identically distributed (i.i.d.) errors. The consequences of using uni-level analysis methods on multilevel data are well-known: the parameter estimates are unbiased but inefficient, and the standard errors are negatively biased, which results in spuriously 'significant' effects (cf. De Leeuw and Kreft, 1986; Snijders and Bosker, 1999; Hox, 1998, 2002). Multilevel analysis techniques have been developed for regression models (Bryk and Raudenbush, 1992; Goldstein, 1995), and specialized software is widely available (e.g., Raudenbush *et al.*, 2000; Rasbash *et al.*, 2000).

The maximum likelihood estimation methods used commonly in multilevel analysis are asymptotic, which translates to the assumption that the sample size is large. This raises questions about the accuracy of the various estimation methods with relatively small sample sizes. This especially concerns the higher level(s), because the sample size at the highest level (the sample of groups) is by definition smaller than the sample size at the lowest level. Simulations by Van der Leeden and Busing (1994) and Van der Leeden *et al.* (1997) suggest that when assumptions of normality and large samples are not met, the standard errors have a downward bias. In addition, the group level variance components tend to be underestimated. Simulation studies by Busing (1993) and Van der Leeden and Busing (1994) indicate that for highly accurate group level variance estimates many groups (more than 100) are needed (cf. Afshartous, 1995). In contrast, Browne and Draper (2000) report that with as few as six to twelve groups, Restricted ML (RML) estimation provides reasonable variance estimates and, with 48 groups, both RML and Full Information ML (FML) estimation produce reasonable variance estimates. The simulations by Van der Leeden *et al.* (1997) show that the standard errors of the variance components are generally estimated too small, with RML again being 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, in general, a large number of groups appears more important than a large number of individuals per group. Although the results of the available simulation studies are not completely in agreement with each other, they all conclude that the regression coefficients are estimated without bias while their standard errors tend to be biased downward with small sample sizes at the group level. Variance components are more susceptible to bias; they tend to be estimated too small with standard errors that may also be strongly biased downward with small sample sizes at the group level (cf. VERBEEK, 2000). The various reports diverge as to the conclusion at precisely which point the group sample size becomes 'too small'.

Some simulations address the effect of violation of the assumption of normally distributed residual errors. In general, the effect of violation of the assumption of normal errors resembles the effect of small sample sizes: both with small sample sizes and with non-normal errors, the regression coefficients and their standard errors show little or no bias, but variance components and their standard errors may be severely biased. We will present an investigation of the effect of non-normal errors, and also investigate the efficacy of robust standard errors. One method of obtaining better tests and confidence intervals when distributional assumptions do not hold is to correct the asymptotic standard errors. One well-known correction method to produce robust standard errors is the so-called Huber/White or sandwich estimator (HUBER, 1967; WHITE, 1982), which is available in several of the available multilevel analysis programs (e.g., RAUDENBUSH *et al.*, 2000; RASBASH *et al.*, 2000).

In this paper simulation studies are used to determine the influence of different sample sizes on the accuracy of the estimates (regression coefficients and variances) and to determine the consequences of the violation of the assumption of normally distributed errors at the second level of the multilevel regression model. A recent simulation study on multilevel structural equation modeling (HOX and MAAS, 2001) suggests that the size of the intraclass correlation (ICC) also affects the accuracy of the estimates, therefore this factor is also included in the simulation design. The research questions are: (1) what group level sample size can be considered adequate in general and more specific in the situation that the assumption of normally distributed residuals is not met, and (2) how well does the sandwich estimator perform when the assumption of normally distributed residuals is not met.

2 The sandwich estimator

One method of obtaining better tests and confidence intervals is to correct the asymptotic standard errors, using the so-called Huber/White or sandwich estimator (HUBER, 1967; WHITE, 1982). In the maximum likelihood approach, the usual estimator of the sampling variances and covariances is the inverse of the Information matrix (Hessian matrix, cf. ELIASON, 1993). Using matrix notation, the asymptotic variance-covariance matrix of the estimated regression coefficients can be written as follows:

$$\mathbf{V}_A(\hat{\beta}) = \mathbf{H}^{-1} \tag{1}$$

where $\mathbf{V}_A$ is the asymptotic covariance matrix of the regression coefficients, and $\mathbf{H}$ is the Hessian matrix. The Huber/White estimator is given as:

$$\mathbf{V}_R(\hat{\beta}) = \mathbf{H}^{-1}\mathbf{C}\mathbf{H}^{-1} \tag{2}$$

where $\mathbf{V}_R$ is the robust covariance matrix of the regression coefficients, and $\mathbf{C}$ is a correction matrix. The correction matrix, which is 'sandwiched' between the two $\mathbf{H}^{-1}$ terms, is based on the observed raw residuals. Details of the Huber/White correction for the multilevel model are given by GOLDSTEIN (1995) and RAUDENBUSH and BRYK (2002). If the residuals follow a normal distribution, $\mathbf{V}_A$ and $\mathbf{V}_R$ are both consistent estimators of the covariances of the regression coefficients, but the model-based asymptotic covariance matrix, $\mathbf{V}_A$, is more efficient and the model-based standard errors are in general smaller. However, when the residuals do not follow a normal distribution, the model based asymptotic covariance matrix is both inaccurate and inconsistent, while the observed residuals based sandwich estimator $\mathbf{V}_R$ is still a consistent estimator of the covariances of the regression coefficients. This makes inference based on the robust standard errors less dependent on the assumption of normality, at the cost of sacrificing some statistical power.

3 Method

3.1 *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 (3):

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}X_{ij}Z_j + u_{1j}X_{ij} + u_{0j} + e_{ij} \tag{3}$$

where Y_{ij} is the score of individual i in group j on the dependent variable; X_{ij} is the score of individual i in group j on the independent variable on the individual level; Z_j is the score of group j on the independent variable on the group level; γ_{00} is the general intercept; γ_{10} is the regression coefficient of the direct effect of X_{ij} on Y_{ij}; γ_{01} is the regression coefficient of the effect of Z_j on Y_{ij}; γ_{11} is the regression coefficient of the effect of Z_j on the influence of X_{ij} on Y_{ij}; e_{ij} is the error term on the individual level; u_{0j} is the error term on the group level in the intercept and u_{1j} is the error term on the group level in the effect of Z_j on the influence of X_{ij} on Y_{ij}. The individual-level residuals e_{ij} are assumed to have a normal distribution with mean zero and variance σ_e^2. 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_{00}, and the variance of the residual errors u_{1j} is specified as σ_{11}.

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). The sizes of the conditions are partially based on literature and partially on practical experience.

The size of the highest and lowest number of groups is based on literature (VAN DER LEEDEN *et al.*, 1997; KREFT and DE LEEUW, 1998): 30 groups is mentioned as a minimum, while 100 groups is seen as sufficient. In practice, 50 groups is a frequently occurring number. Similarly, for the highest group size, a size is chosen that should be sufficient on the basis of the literature, while a group size of 30 is normal in educational research, and a group size of five is normal in family research and in longitudinal research. The Intraclass Correlations (ICC's) are totally based on practice. They span the customary level of intraclass correlation coefficients found in multilevel studies.

There are $3 \times 3 \times 3 = 27$ conditions. For each condition, we generated 1000 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 conditions with the larger sample sizes, these values are repeated. This ensures that in all simulated conditions the joint distribution of X and Z are the same. 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_{00} follows from the specification of the ICC and σ_e^2, given formula (4).

$$\mathrm{ICC} = \sigma_{00}/(\sigma_{00} + \sigma_e^2). \tag{4}$$

BUSING (1993) shows that the effects for the intercept variance σ_{00} and the slope variance σ_{11} are similar; hence, we chose to use the same value for σ_{11} as for σ_{00}. To simplify the simulation model, without loss of generality, the covariance between the two u-terms is assumed equal to zero. Given the parameter values, the simulation procedure generates the residual errors e_{ij}, u_{0j} and u_{1j}.

To investigate the influence of non-normally distributed errors we transformed the second level residuals of the first simulation set to a χ_1^2 distribution, rescaled to have the same mean and variance as the generated normal residuals. (We did not investigate non-normality for the first level residuals. Because of the larger sample size at this level, the influence of non-normality will be less than at the second level.) Since a chi-square distribution with one degree of freedom is markedly skewed, we consider this a large deviation of the assumption of having a multivariate normal distribution for the second-level residuals.

Therefore the analysis is carried out twice, once with asymptotic maximum likelihood based standard errors, and once with robust Huber/White standard errors. The software MLwiN (RASBASH *et al.*, 2000) was used for both simulation and estimation. In this program the correction of the sandwich estimation is based on the cross-product matrix of the residuals, taking the multilevel structure of the data into account.

Two maximum likelihood functions are common in multilevel estimation: Full ML (FML) and Restricted ML (RML). We use RML, since this is always at least as good as FML, and sometimes better, especially in estimating variance components (BROWNE, 1998).

3.2 *Analysis*
To indicate the accuracy of the parameter estimates (regression coefficients and residual variances) the percentage relative bias is used. Let $\hat{\theta}$ be the estimate of the population parameter θ, then the percentage relative bias is given by $\hat{\theta}/\theta$. 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; Longford, 1993). For each parameter a non-coverage indicator variable was set up that is equal to zero if its true value is in the confidence interval, and equal to one if its true value is outside the confidence interval. The effect of the different simulated conditions on the non-coverage was analyzed using logistic regression on these indicator variables. Since the total sample size for each analysis is 27 000 simulated conditions, the power is huge. As a result, at the standard significance level of alpha = 0.05, extremely small effects become significant. Therefore, our criterion for significance is alpha = 0.001 for the main effects of the simulated conditions.

4 Results

4.1 *Normal distributed level-2 errors*

4.1.1 *Parameter estimates*
Both the fixed parameter estimates (the intercept and regression slopes) and the random parameters (the variance components), have a negligible bias: less 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%.

4.1.2 *Standard errors*
The coverage of both fixed and random effects is significantly affected by the number of groups and by the group size. Coverage is not sensitive to the Intraclass Correlation. The effect of the number of groups on the coverage is presented in Table 1, and the effect of the group size on coverage is presented in Table 2.

There are no effects of the number of groups on the standard errors of the fixed regression coefficients. The effect of the number of groups on the standard errors of the random variance components are shown in Table 1. With 30 groups, the coverage rate for the second-level intercept variance is 91.0%, and the coverage rate

Table 1. Coverage of the 95% confidence interval by number of groups (0.9225 < C.I. < 0.9747, * = significant at 0.001).

		E0	U0	U1
Number of groups	30	0.9428	0.9104*	0.9120*
	50	0.9438	0.9261	0.9282
	100	0.9514	0.9404	0.9426

Table 2. Coverage of the 95% confidence interval by group size (0.9265 < C.I. < 0.9735, * = significant at 0.001).

	E0	U0	U1
Group size			
5	0.9403	0.9259*	0.9213*
30	0.9489	0.9250*	0.9337
50	0.9488	0.9261*	0.9278

for the second-level slope variance is 91.2%. The amount of coverage here implies that the standard errors for the second-level variance components are estimated about 15% too small.

Table 2 shows that in the case of a group size of five the coverage rate for the second-level slope variance is 92.1%. This amount implies that the standard errors are estimated about 3.1% too small.

4.2 *Non-normal distributed level-2 errors*

4.2.1 *Parameter estimates*

For the 27 conditions the mean relative bias is calculated. The percentage relative bias for the fixed and the random parameters is the same for the ML- and the robust estimations, because we investigate the parameter estimates and not their standard errors. Only the "worst" condition, only 30 groups with five individuals and an ICC of 0.1 shows a statistically significant bias. However, from a practical perspective this significant effect is totally irrelevant (variance estimated as 0.492 instead of 0.50).

4.2.2 *Standard errors*

Table 3 shows the coverage of the 95% confidence interval for the fixed effects, without a breakdown in conditions. There are two statistically significant effects, both refer to the level-1 regression coefficient.

The coverage of the fixed effects is significantly affected by the Number of Groups and by the Group Size. With respect to the Number of Groups, the results are as expected: more groups lead to a closer approximation of the nominal coverage (see Table 4). With respect to the Group Size, this is not the case. Having larger groups does not improve the situation.

Table 3. Coverage of the 95% confidence interval for the main fixed effects (0.9260 < C.I. < 0.9740; * = sign., α = 0.001).

	ML-estimation	Robust estimation
Intercept	0.9322	0.9291
X	0.9262	0.9229*
Z	0.9458	0.9402
XZ	0.9484	0.9365

Table 4. Coverage of the 95% confidence interval for the fixed effects by number of groups and group size (0.9265 < C.I. < 0.9735; first the value for the ML-estimation; second for the robust estimation, * = significant at 0.05).

	Intercept	X	Z
Number of groups			
30	0.9271/0.9214*	0.9171*/0.9120*	0.9397/0.9306
50	0.9302 /0.9279	0.9246/0.9214*	0.9498/0.9439
100	0.9392/0.9379	0.9370/0.9353	0.9480/0.9462
Group size			
5	0.9422/0.9390	0.9378/0.9328	0.9543/0.9440
30	0.9266/0.9236*	0.9247*/0.9221*	0.9414/0.9353
50	0.9278/0.9247*	0.9162*/0.9139*	0.9417/0.9413

Table 5 shows the coverage of the random effects, without a breakdown in conditions. Only the ML standard errors for the lowest level parameter are correct. At this level, the robust standard error gives an overcorrection. Both the ML and the robust estimated standard errors of the second level variances are incorrect. The robust estimators are, however, better than the ML estimators.

The coverage of the random effects is significantly affected by the Number of Groups, the Group Size and the Intraclass Correlation. The effect of the Number of Groups on the coverage is presented in the first part of Table 6, the effect of the Group Size in the second part, and the effect of the Intraclass Correlation in the third

Table 5. Coverage of the 95% confidence interval for the overall random effects (0.9265 < C.I. < 0.9727; * = sign., α = 0.001).

	ML-estimation	Robust estimation
E0	0.9520	0.9901*
U0	0.6632*	0.8693*
U1	0.6427*	0.8524*

Table 6. Coverage of the 95% confidence interval for the random effects by number of groups and group size (0.9265 < C.I. < 0.9735; first the value for the ML-estimation; second for the robust estimation, * = significant at 0.05).

	E0	U0	U1
Number of groups			
30	0.9487/0.9866*	0.6537*/0.8128*	0.6501*/0.8007*
50	0.9539/0.9903*	0.6701*/0.8734*	0.6471*/0.8506*
100	0.9534/0.9933*	0.6659*/0.9217*	0.6308*/0.9059*
Group size			
5	0.9373/0.9819*	0.7784*/0.9019*	0.7540*/0.8648*
30	0.9630/0.9937*	0.6219*/0.8582*	0.6032*/0.8500*
50	0.9557/0.9947*	0.5893*/0.8478*	0.5708*/0.8423*
ICC			
0.10	0.9520/0.9899*	0.7123*/0.8786*	0.6913*/0.8669*
0.20	0.9521/0.9901*	0.6572*/0.8706*	0.6334*/0.8494*
0.30	0.9519/0.9902*	0.6201*/0.8588*	0.6032*/0.8408*

part. In all three conditions, we observe an overcorrection of the standard error of the lowest level variance by the robust method. At the second level, all effects are significant. As expected, the ML-estimator gives worse results than the robust estimations. As in Table 4, we observe that having larger groups does not improve the situation. Robust standard errors are better than the asymptotic standard errors, but the resulting confidence intervals only begin to approach their nominal coverage at the largest sample of groups ($NG = 100$) used in this simulation.

5 Summary and discussion

In conclusion, with respect to the influence of the sample size in the case of normal distributed errors, there turns out only to be a problem with the standard errors of the second-level variances when the number of groups is substantially lower than 50 and when the group size is lower than 30. With 30 groups, the standard errors are estimated about 15% too small. With a group size of five, the standard errors of the second-level slope variance is estimated 3.1% too small.

These results differ to some extent from the simulation results reported by Busing (1993) and Van der Leeden and Busing (1994). They concluded that, for small sample sizes, the standard errors and corresponding statistical tests are badly biased. However they used a different simulation design. Busing (1993) used much higher intraclass correlations, up to 0.80, which are unlikely to occur in actual data. In addition, the simulated second-level sample sizes were much smaller, starting at a sample of ten groups with five observations each. For these simulated conditions, they reported biased standard statistical tests, especially for the variance components. Snijders and Bosker (1999, p. 44) however, claim that multilevel modeling becomes attractive when the number of groups is larger than ten. To resolve this contradiction, we decided to do one more simulation with only ten groups of group size five. In this simulation, the fixed regression coefficients and variance components were still estimated without bias, except for the second-level variance components when the ICC was low (0.10): there the bias is 25% upwards. The standard errors are now all estimated too small. The non-coverage rates for the fixed effects in this case range between 5.7% and 9.7%, and for the second-level variances they range between 16.3% and 30.4%. Although the standard errors of the fixed effects are still reasonable, the standard errors of the second-level variances are clearly unacceptable. It seems that having as few as ten groups is not enough. If one is interested only in the fixed regression coefficients, it appears less problematic, but we would still advise using bootstrapping or other simulation-based methods to assess the sampling variability instead of ML or robust standard errors.

This leads us to the following rule of thumb: if one is only interested in the fixed effects of the model, ten groups can lead to good estimates. If one is also interested in contextual effects, 30 groups are needed. If one also wants correct estimates of the standard errors, at least 50 groups are needed.

To investigate the limits of our results, we carried out another simulation. We increased the population values of the residual variances, which decreases the proportion of explained variance in the population model. The results of this simulation were very close to the results reported above. So, the proportion of explained variance does not affect the accuracy of the estimates.

Non-normal distributed residual errors on the second (group) level of a multilevel regression model have an effect on the estimates of the fixed effects. The more groups there are, the better the estimates, but this does not hold for the group size. Having larger groups does not improve the situation. The non-normal distributed level-2 residual errors do have more effect on the estimates of the random effects. The estimates of the variances are unbiased, but the standard errors are not always accurate. At the lowest level, the maximum likelihood standard errors are accurate, while the robust standard errors are overcorrected. The standard errors for the second-level variances are highly inaccurate, although the robust standard errors tend to do better than the maximum likelihood standard errors. With maximum likelihood standard errors, the coverage of the 95% confidence interval for the random effects at the second-level is only 66% and 64%, compared with 87% and 85% for robust standard errors. These results mean that when the group level variances are not normally distributed, neither the maximum likelihood nor the robust estimation of the group level standard errors can be trusted. In the case of robust estimation, this can be compensated for by having a very large number of groups, at the expense of having overcorrected standard errors at the lowest level. An attractive alternative is to use a non-parametric approach, as proposed by Vermunt in this issue of *Statistica Neerlandica.*

Raudenbush and Bryk (2002) suggest that comparing the asymptotic standard errors calculated by the maximum likelihood method to the robust standard errors is a way of appraising the possible effect of model misspecification. Hox (2002) extends this suggestion to model misspecifications including violation of important assumptions. Used in this way, robust standard errors become an indicator for possible misspecification of the model or its assumptions. If the robust standard errors differ considerably from the asymptotic standard errors, this should be interpreted as a warning that certain important assumptions might be violated. Clearly, the recommended action is not to rely simply on the robust standard errors to deal with the misspecification. Our simulation indicates that, unless the number of groups is very large, the robust standard errors are not up to that task. Rather, the reasons for the discrepancy must be diagnosed and resolved.

References

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

Browne, W. J. (1998), Applying MCMC methods to multilevel models, Unpublished Ph.D. Thesis, University of Bath, UK.

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

Bryk, A. S. and S. W. Raudenbush (1992), *Hierarchical linear models*, Sage, Newbury Park, CA.

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

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

Eliason, S. R. (1993), *Maximum likelihood estimation*, Sage, Newbury Park, CA.

Goldstein, H. (1995), *Multilevel statistical models*, Edward Arnold, London.

Hox, J.J. (1998), *Multilevel modeling: when and why*, in: I. Balderjahn, R. Mathar and M. Schader (eds), *Classification, data analysis, and data highways*, Springer Verlag, New York, 147–154.

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, 221–233.

Kreft, I. and J. de Leeuw (1998), *Introducing multilevel modeling*, Sage, Newbury Park, CA.

De Leeuw, J. and I.G.G. Kreft (1986), Random coefficient models for multilevel analysis, *Journal of Educational Statistics* **11**, 57–85.

Longford, N. T. (1993), *Random coefficient models*, Clarendon Press, Oxford.

Rasbash, J., W. Browne, H. Goldstein, M. Yang, I. Plewis, M. Healy, G. Woodhouse, D. Draper, I. Langford and T. Lewis (2000), *A user's guide to MlwiN, Multilevel Models Project*, University of London, London.

Raudenbush, S. W. and A. S. Bryk (2002), *Hierarchical linear models*, (2nd edn), Sage, Thousand Oaks, CA.

Raudenbush, S., A. Bryk, Y. F. Cheong and R. Congdon (2000), *HLM 5, Hierarchical linear and nonlinear modeling*, Scientific Software International, Chicago.

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

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

Van der Leeden, R., F. Busing and E. Meijer (1997), *Applications of bootstrap methods for two-level models*, Unpublished paper, Multilevel Conference, Amsterdam, April 1–2, 1997.

Verbeek, M. (2000), *A guide to modern econometrics*, Wiley, New York.

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

Received: April 2003. Revised: August 2003.