QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No.2, 2013, pp. 19 – 28

SAMPLE SIZE AND STRUCTURE FOR MULTILEVEL MODELLING: MONTE CARLO INVESTIGATION FOR THE BALANCED DESIGN

Edyta Łaszkiewicz
Department of Spatial Econometrics, University of Lodz
e-mail: elaszkiewicz@uni.lodz.pl

Abstract: The aim of the study is to examine the robustness of the estimates and standard errors in the case of different structure of the sample and its size. The two-level model with a random intercept, slope and fixed effects, estimated using maximum likelihood, was taken into account. We used Monte Carlo simulation, performed on a sample of the equipotent groups.

Keywords: multilevel model, Monte Carlo, sample size

INTRODUCTION

Sufficient sample is one of the most important problem in the multilevel modelling (see e.g. Mass and Hox [2004, 2005] or Snijders [2005] to mention just a few). The most basic design conditions like a number of groups at each level of the analysis and its size determine the ability to obtain accurate (unbiased) estimates of the regression coefficients, standard errors and power of tests[1]. Additionally, Busing [1993] found out the insufficient sample size (10 to 50 groups with 5 or 10 individuals) might be responsible for the model nonconvergence. Despite the asymptotic properties of the multilevel models estimators (like REML or IGLS), due to which larger sample guarantees the bias reduction, in the centre of interest is the downward limit of the sample [Mass and Hox 2005]. Accordingly, the adequate (sufficient) sample size can be define as such the minimum sample, which guarantees the unbiasedness (or more precisely: acceptable low size of the bias). Such definition is consistent with Snijders and Bosker [1993], who use the

[1]Other factors like the estimation method, proportion of singletons, value of the intraclass correlation, collinearity or model complexity, which also might affect the estimates, are not wider describe as they are not take into consideration in this study.

term ‘conditionally optimal’ to characterise the sample size which allows to yield the minimal standard errors for the particular parameters or other constraints. Although the literature about the sufficient sample size is large, there is still no consensus how it should looks like, what is the result of i.e. using different simulation conditions and/or simulation designs. Let review only the guidelines for 2-level models estimated using the balanced sample. We start from the recommendations for the unbiased parameter and standard errors estimates, then concentrate on the suggestions based on the maximization the power of the tests.

Kreft [1996] recommended ‘30/30’ rule which means minimum 30 observations per group and minimum 30 units at each level of the analysis to unbiased estimate all parameters and their standard errors. As pointed by Mass and Hox [2005], such number of groups gives unbiased results except the standard error estimates of the random effects at the level-2. Accordingly, Hox [1998] recommended minimum 20 observations for 50 groups if the cross-level interaction is tested. Although both the number of groups and the number of observations per group are important to obtain the unbiased results, the sensitivity of the fixed and random effects (and their standard errors) estimates to above is different. When the accuracy of the variance components estimates is influenced strongly by the number of groups, fixed effects estimates are less susceptible to the data sparseness. Similar conclusions were drawn by Newsom and Nishishiba [2002] and Clarke and Wheaton [2007], who confirmed that the unbiased estimates of the fixed effects might be received even for the small sample. As the variance components estimates are often in the main centre of the interest in the multilevel models, additional suggestions dealing with the random effects were concerned in detail. Mok [1995] noticed that 5 groups at the second level gives a notably bias of the variance estimates, while Clarke and Wheaton [2007] suggested at least 10 observation per group for at least 100 groups is needed to obtain the unbiased estimate of the intercept variance. If the slope variance is estimated they recommended at least 200 groups with minimum 20 observation per group. Although for the accurate estimates of the variance components (often underestimated) at least 100 units is needed, in practise such sample would be hard to obtain [see Mass and Hox 2004]. According to all of the mentioned guidelines, rather than the large number of observations per unit, the large number of groups seems to be more important to receive the accurate estimates.

Sufficient sample size is considered also due to the accuracy of the standard errors estimates but such investigations are in the minority [Mass and Hox 2005]. In the simulation research the most common way to validate standard errors estimates is by checking the accuracy of the significance test or the coverage of the confidence interval (generated by using standard normal distribution and gamma distribution)[2]. Accordingly, Browne and Draper [2000] showed, using IGLS and

[2] Although the assumption about the normality is not optimal, especially if the confidence intervals of the random effects are considered (because of the lack of the confidence

RIGLS estimators, that for at least 48 groups the coverage of the nominal 95% intervals is unbiased (for the fixed effects estimates), when the intervals for the covariance matrix parameters are substantially biased (below 95%). Similarly, Mass and Hox [2005] found out that negative influence of as small as 30 number of groups is small for the standard errors of the fixed effect coefficients (6.0% and 6.4% for the intercept and regression coefficient) and higher for the standard errors of the variance components (around 9% for the level-2 intercept and slope variances). Additionally, in a large (5760 conditions) Monte Carlo experiments Bell et al. [2010] found out that for each type of the predictor variable, treated as the fixed effect, estimated confidence interval coverage is rather constant and higher than for the level-2 estimates, what is consistent with the previous reviewed researches. Finally, according to Snijders [2005] group size is less important for the power of the tests than the number of groups, what is similar to the results for the estimates. The only limitation of the small group size for the power of testing are the random slope variances. As the power of the tests is the result of the standard error size, consistency of the conclusions seems to be natural.

There is no agreement about the negative influence of the data sparseness on the convergence. Although Bell et. al. [2010], Mass and Hox [2004] found out that there is no problem with the model convergence using ML and RIGLS estimator, according to Busing's [1993] findings such problem might occurs if the sample is too small. In practice the generalisation of the presented rules is always limited to the specific cases, e.g. the type of the estimated effect (random, fixed, interaction, cross-level, etc.) or the estimation method.

In the literature, to set the optimal/sufficient sample size, in the multilevel modelling, the simulation method has been chosen more frequently. Another way is to use the approximate formula, relating effect size and standard errors to statistical power of the significance test [Snijders and Bosker 1993]. As was showed by Snijders [2005], the way of computing the sufficient sample size depends on the parameter estimates which the researcher is interested in. Also Moerbeek et al. [2001] presented formulas for calculating the optimal design (the sample size) for the 2-level models with detailed evaluation using *D*-optimality and *L*-optimality criteria. Although the approximate formula seems to be faster in using, its limitation (like the lack of the generalisation) makes Monte Carlo simulation more flexible tool for evaluation the sufficiency of the sample size.

The motivation for this paper is to evaluate by the Monte Carlo simulation the influence of the sample size and its structure on the estimates biasness. The fixed and random parameter estimates and their standard errors are examined in the 2-level model estimated by maximum likelihood (ML). The rest of the paper is divided into the simulation method description and the results discussion.

---

symmetry), in most of the simulation studies such method of evaluation of the standard errors estimates are using [see e.g. Busing 1993, Van der Leeden et al. 1997].

SIMULATION DESIGN

The 2-level model (for the continuous outcome variable Y_{ij}) with two explanatory variables $X_{1,ij}$, $X_{2,ij}$ on the level-1 was examined. The random (or stochastic) part of the model contains: residual error terms at the level-2: $\mu_{0,j} \sim N(0, \sigma^2_{\mu_0})$, $\mu_{1,j} \sim N(0, \sigma^2_{\mu_1})$ and individual-level (level-1) residuals $\varepsilon_{ij} \sim (0,1)$. The fixed (or determinist) part contains β_0, β_1, β_2 coefficients. This model can be written as [Goldstein 2010]:

$$Y_{ij} = \beta_{0,j} + \beta_{1,j} X_{1,ij} + \beta_2 X_{2,ij} + \varepsilon_{ij}, \quad (1)$$
$$\beta_{0,j} = \beta_0 + \mu_{0,j},$$
$$\beta_{1,j} = \beta_1 + \mu_{1,j},$$

where: $i = 1, \dots, M$ and $j = 1, \dots, J$. We assume the structure of the variance-covariance matrix as in the standard multilevel models: $\forall i \neq i' \ \mathrm{cov}(\varepsilon_{ij}, \varepsilon_{i'j}) = 0$, $E(\mu_{1,j}) = E(\mu_{0,j}) = 0$, $j \neq j' \ \mathrm{cov}(\mu_{0,j'}, \mu_{0,j}) = \mathrm{cov}(\mu_{1,j'}, \mu_{1,j}) = 0$, $\mathrm{cov}(\mu_{0,j}, \varepsilon_{ij}) = \mathrm{cov}(\mu_{1,j}, \varepsilon_{ij}) = 0$. The values of the predictors were drawn independently from the normal distribution with variance 1. Model (1) was estimated via ML.

Three conditions were varied in the simulation: (1) number of groups J={5, 10, 20, 30, 50, 70, 90}, (2) number of observations per group M={5, 10, 20}, (3) values of the parameters (in Table 1). As the value of the intraclass correlation (ICC) influence the results the two different values of ICC were tested. The ICC was calculated as follows: $(\sigma^2_{\mu_0} + \sigma^2_{\mu_1})/(\sigma^2_{\mu_0} + \sigma^2_{\mu_1} + \sigma^2_{\varepsilon})$.

Table 1.Target values of parameters

variant/parameter	β_0	β_1	β_2	$\sigma^2_{\mu_0}$	$\sigma^2_{\mu_1}$	ICC
1	0.60	0.50	0.30	0.50	0.40	0.47
2	0.20	0.30	0.80	0.20	0.30	0.33
3	0.30	0.70	0.80	0.20	0.30	0.33

Source: own calculation

The large variation of the groups number was evaluated because this factor might affects the estimate much more than the group size. For each of the 63 conditions 1000 datasets were simulated using user-written syntax in STATA[3] based on the xtmixed command which allows for the multilevel model estimation.

The accuracy of the estimates was indicated using two measures commonly used in the evaluation of the simulation results:

- Relative bias of an estimator $\widehat{\theta_l}$ for parameter θ_l, defined as:

[3] Monte Carlo simulation syntax is available at: https://sites.google.com/site/elaszkiewicz.

$$B(\widehat{\theta_l}) = \frac{\overline{\widehat{\theta_l}} - \theta_l}{\theta_l} \cdot 100\%, \quad (2)$$

where $\overline{\widehat{\theta_l}}$ is the arithmetic mean calculated from *K*=1000 simulation runs of $\widehat{\theta_{lk}}$. According to Hoogland and Boomsm (1998) unbiased estimates are those for which the relative bias is less than 5%. The relative biases were calculated to evaluate only the parameter estimates.

- Rate of the coverage, calculated as:

$$C\left(se(\widehat{\theta_l})\right) = \frac{\sum C\left(se(\widehat{\theta_{lk}})\right)}{K} \cdot 100\% \quad (3)$$

$$C\left(se(\widehat{\theta_{lk}})\right) = \begin{cases} 1 \; if \; \theta_l \in CI \\ 0 \; if \; \theta_l \notin CI \end{cases},$$

where $se(\widehat{\theta_{lk}})$ is the estimated standard error of the $\widehat{\theta_{lk}}$ at *k*-th run, CI is the 95% confidence interval established separately for the fixed effects as: $\widehat{\theta_{lk}} \pm u_\alpha \cdot se(\widehat{\theta_{lk}})$ and for the random effects as: $\exp(\ln(\widehat{\theta_{lk}}) \pm u_\alpha \cdot \frac{1}{\widehat{\theta_{lk}}} \cdot se(\widehat{\theta_{lk}})$. The indicator was used to check the bias of the standard error estimates.

Additionally, to compare different conditions ANOVA (for the parameter estimates) and logistic regression (for the confidence interval evaluation) were used.

RESULTS AND DISCUSSION

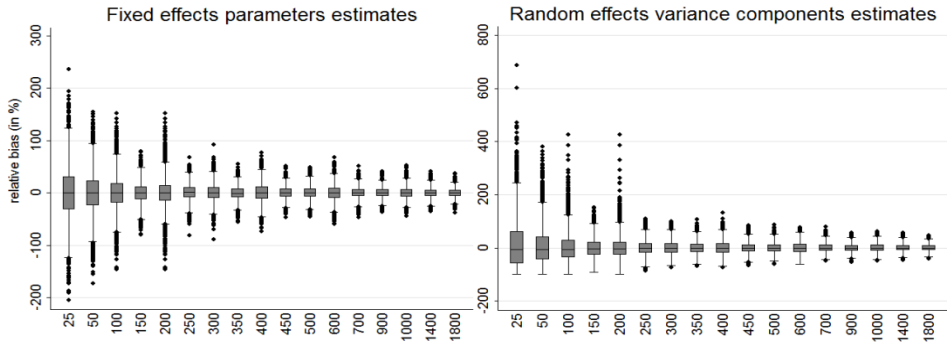
The convergence of model was achieved almost in each case, even for the smallest sample size. However, for the sample of 5 groups with 5 observations per group it was more frequently impossible to estimate standard errors for the random effects variance components due to the singular variance-covariance matrix of the random effects [see e.g. Henderson 1986].

Parameter estimates

The average relative bias for the fixed effect estimates (0.01%) was lower from the random effect estimates bias, which was 1.07%. Although, there was no significant differences in the relative bias across the fixed parameter estimates, the biases of the $\widehat{\sigma}^2_{\mu_0}$ and $\widehat{\sigma}^2_{\mu_1}$ were significantly different and higher for the first one. Additionally, there was no significant differences between the relative bias of the fixed parameter estimates when three variants of the target values of the parameters were compared. However, the influence of the ICC on the random effect estimates was revealed. For the higher value of the ICC, the lower relative bias of the random effect estimates was achieved. This is consistent with e.g. Newson and Nishishiba [2002], who showed that the ICC value determines the accuracy of the estimates.

The unbiased estimates were achieved for the fixed effect estimates for each of the simulated sample size (Figure 1). Even for the sample as small as 25 observations the relative biases were less than 1% for all of the fixed parameters estimates. In the case of the random effects estimates only for the sample of 25 observations the results were biased. The relative bias for the random intercept variance estimates was 16% and for the random slope variance estimates almost 10%. The relative biases less than 1% for the variance components estimates were achieved for the sample size equal to 100 or higher. Additionally, as the sample size increases, the variance of the parameters estimates has decreased strongly.

Figure 1. Effect of group size on the relative bias of the parameter estimate



Source: own calculation

Table 2. Relative biases (in %) and significance of the group size effect

parameter /group size	5	10	20	*p*-value*
β_0	-0.15	0.10	-0.10	0.89
β_1	-0.22	0.01	0.43	0.27
β_2	-0.20	0.08	0.16	0.22
$\sigma^2_{\mu_0}$	2.70	0.92	0.55	0.00
$\sigma^2_{\mu_1}$	1.97	0.53	-0.25	0.00

* *p*-value for the effect of group size on the relative bias of the parameter estimate

Source: own calculation

Although the unbiased results might be obtained even if 5 observation per unit occurs, the sensitivity of the fixed and random effects estimates for the group size was different (presented in Table 2). Only for the variance components estimates ($\sigma^2_{\mu_0}, \sigma^2_{\mu_1}$) increase of the group size affects significantly the value of the relative bias of the estimates. Such results are similar to Newsom and Nishishiba [2002], Clarke and Wheaton [2007].

According to Table 3, all of the fixed effects estimates are unbiased for the sample consisting of 5 groups. In the opposite, the random effects estimates are biased in such a case. It means that for the unbiased estimates for all of the

parameters the sample of at least 10 groups with 5 observations per group is needed. This is less than in Kreft's '30/30' rule or as Hox [1998] suggested. The differences in the recommendations are the result of not taking into account the unbiasness of the standard errors estimates.

Table3. Relative biases (in %) and significance of the number of groups effect

parameter / nr of groups	5	10	20	30	50	70	90	*p*-value*
β_0	-0.15	-0.01	-0.08	-0.10	0.67	-0.75	0.08	0.79
β_1	0.80	0.23	-0.29	-0.11	-0.10	-0.20	0.18	0.65
β_2	-0.27	0.06	0.04	0.15	0.12	-0.03	0.02	0.89
$\sigma^2_{\mu_0}$	7.90	1.48	0.45	0.14	-0.14	0.26	-0.38	0.00
$\sigma^2_{\mu_1}$	4.73	0.63	0.21	-0.18	-0.46	0.16	0.15	0.00

* *p*-value for the effect of number of groups on the relative bias of the parameter estimate

Source: own calculation

The results showed that unbiasness of the random effects estimates depends more on the number of groups in the sample, than the group size. This conclusion is consistent with Snijders and Bosker[1994].

Standard errors

The coverage of the 95% confidence interval (CI) was similar for the fixed effect parameters (93.47%) and for the random effect estimates (94.78%). The results of the logistic regression showed that the rate of the coverage rate for the CI for the random effects depends on the ICC but the fixed effects seems to be not affected by the level of the ICC.

Table4. Influence of the ICC value on the coverage of the 95% confidence interval

parameter / nr of groups*	5	10	20	30	50	70	90
$\sigma^2_{\mu_0}$	0.00	0.00	0.23	0.79	0.95	0.90	0.15
$\sigma^2_{\mu_1}$	0.00	0.16	0.39	0.53	0.47	0.20	0.51

* *p*-value from the logistic regression, where the value of ICC was independent variable

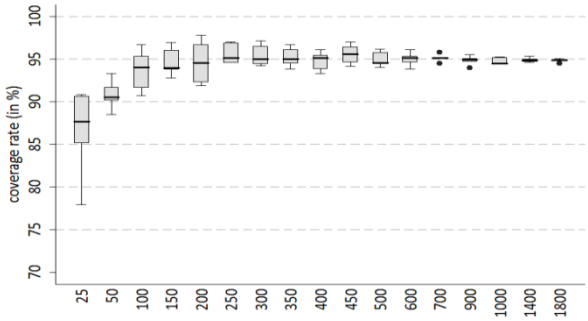
Source: own calculation

Inspired by Mass and Hox [2005], who proved that for the ICC=0.1, 0.2, 0.3, the influence of the ICC value on the coverage rate (for the random effects estimates) occurs only for the extremely small sample size (like 10 groups with 5observations), we checked if the influence of the ICC varies across the number of groups. Our results (presented in Table 4) are more detailed than Mass and Hox [2005] as for each variant of the number of groups separate regression has been done. For at least 10 groups in the sample the impact of the ICC value on the coverage of the 95% CI for random intercepts variance was statistically significant.

In comparison, the CI for random slopes variance was affected by the ICC value only in the sample of 5 groups.

Next, the sample size effect was tested. As expected, with increasing of the sample size results the rate of the coverage grows (Figure 2). For the samples of less than 100 observations at least for one of the parameters the coverage rate of the CI was lower than 90%. It means that for the unbiased standard errors estimates such sample size is the minimum.

Figure 2. Coverage rate of the 95% confidence interval by the group size



Source: own calculation

Although for the smallest sample size there are significant differences in the coverage rate of the CI for the fixed (89.70%) and random (81.52%) effects parameters, the differences decrease as the sample size increase. For the sample of 1800 observations each of the parameter achieved the coverage rate around 95%.

Table 5. Coverage rate of the 95% CI (in %) by group size with significance

parameter /group size	5	10	20	*p*-value*
β_0	93.22	93.15	93.03	0.45
β_1	92.67	93.11	92.90	0.54
β_2	93.91	94.51	94.69	0.00
$\sigma^2_{\mu_0}$	92.31	95.35	95.34	0.00
$\sigma^2_{\mu_1}$	93.80	95.94	95.93	0.00

* *p*-value from the logistic regression, where groups size was independent variable

Source: own calculation

Similar to the relative bias behaviour, the coverage rate increases when the size of the groups grows (Table 5). The positive, significant effect of rising of the group size was noticed for the variance components (boththe intercepts and the slopes) and for the one of the fixed effect parameter. However, difference between coverage rate of the CI for fixed and random parameters was small.Additionally, the significance of the number of groups effect was tested (Table 6). For each of the parameter the coverage rate of the 95% CI depends on the number of groups in

the sample. In the sample of 5 groups 7.09-12.64% of the cases were outside of the 95% CI. As the number of groups increased, the coverage rate also increased.

Table 6. Coverage rate of the 95% CI (in %) by number of groups with significance

parameter / nr of groups	5	10	20	30	50	70	90	*p*-value*
β_0	89.29	92.12	93.22	93.91	94.32	94.60	94.48	0.00
β_1	88.01	91.28	93.36	93.92	94.44	94.69	94.56	0.00
β_2	92.91	93.50	94.82	94.69	94.70	94.74	95.22	0.00
$\sigma^2_{\mu_0}$	87.36	93.48	95.94	96.16	96.06	95.70	95.64	0.00
$\sigma^2_{\mu_1}$	92.21	94.79	96.16	96.49	95.98	95.56	95.36	0.00

* *p*-value from the logistic regression, where number of groups was independent variable

Source: own calculation

Surprisingly, for each variant of the number of groups the coverage rate was higher for the variance components estimates. In the case of the 30 groups the non-coverage rate for the fixed effects parameters was 6% (the same as in Mass and Hox [2005]), however for the random effects parameters we obtained 3%, when in Mass and Hox [2005] study the non-coverage rate was 8%. The differences in the results are the effect of the way how the CIs were build. Mass and Hox [2005] used standard normal distribution to establish the CIs for the variance components parameters, when we used Wald-type CIs, which were better performed.

CONCLUSIONS

Our results are consistent with the previous investigation. The unbiased estimates of the fixed effects parameters might be obtained even for the extremely small samples. The structure of the sample (number of groups and group size) do not affect negatively the fixed effects estimates. For the unbiased estimates of the variance components both design conditions are important, but only for too small (less than 10) number of groups results were biased. The evaluation of the standard errors estimates once again proved the major role of the number of groups to guarantee the satisfactory coverage rate of the CIs. Also, we showed that for the random effects the Wald-type confidence interval are better than based on the standard normal distribution. Our results might be generalized but only to the presented conditions. Additional researches are required to examine more advanced multilevel model specification, e.g. cross-classified or multiple membership.

REFERENCES

Bell B. A., Morgan G. B., Kromrey J. D., Ferron, J. M. (2010) The impact of small cluster size on multilevel models: a Monte Carlo examination of two-level models with binary

and continuous predictors, JSM Proceedings, Survey Research Methods Section, pp. 4057-4067.

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

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

Clarke P., Wheaton B. (2007) Addressing data sparseness in contextual population research using cluster analysis to create synthetic neighborhoods, Sociological Methods & Research, 35, pp. 311-351.

Goldstein H. (2010) Multilevel statistical models (4th ed.), New York: Hodder Arnold

Henderson C. R. (1986) Estimation of singular covariance matrices of random effects, Journal of Dairy Science 69.9, pp. 2379-2385.

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

Hox J. J. (1998) Multilevel modeling: When and why, [in:] Balderjahn I., Mathar R., Schader M., Classification, data analysis, and data highways, New York: Springer Verlag, pp. 147-154.

Kreft I. G. G. (1996) Are multilevel techniques necessary? An overview, including simulation studies, Unpublished manuscript, California State University at Los Angeles.

Maas C. J. M., Hox J. J. (2004) Robustness issues in multilevel regression analysis, Statistica Neerlandica, 58, pp. 127-137.

Maas C. J. M., Hox, J. J. (2005) Sufficient sample sizes for multilevel modeling, Methodology, 1, pp. 86-92.

Mass C. J., Hox J. J. (2002) Robustness of multilevel parameter estimates against small sample sizes. Unpublished Paper, Utrecht University.

Moerbeek M., Van Breukelen G. J., Berger M. P. (2001) Optimal experimental designs for multilevel logistic models, Journal of the Royal Statistical Society, Vol.50(1), pp. 17-30.

Mok M. (1995) Sample size requirements for 2-level designs in educational research, Unpublished manuscript, Macquarie University.

Newsom J. T., Nishishiba M. (2002) Nonconvergence and sample bias in hierarchical linear modeling of dyadic data. Unpublished Manuscript, Portland State University.

Snijders T. A. B. (2005) Power and Sample Size in Multilevel Linear Models', [in:] Everitt B.S., Howell D.C. (eds.) Encyclopedia of Statistics in Behavioral Science, Vol. 3, Wiley, pp. 1570–1573.

Snijders T. A. B., Bosker R. J. (1993) Standard Errors and Sample Sizes for Two-Level Research, Journal of Educational Statistics, Vol. 18, No. 3, pp. 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 University.

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