Power and Accuracy in Multilevel Designs:

An Application in Educational Research

Cools,Wilfried (1st Author)
Centre for Methodology of Educational Research
Vesaliusstraat 2, P.O.Box 3762
Leuven 3000, Belgium
E-mail: Wilfried.Cools@ped.kuleuven.be

Van den Noortgate, Wim (2nd Author)
Centre for Methodology of Educational Research
Vesaliusstraat 2
Leuven 3000, Belgium
Wim.VandenNoortgate@kuleuven-kortrijk.be

Onghena,Patrick (3rd Author)
Centre for Methodology of Educational Research
Vesaliusstraat 2
Leuven 3000, Belgium
Patrick.Onghena@ped.kuleuven.be

Various designs can be used to answer specific research questions. Even for a given budget, these designs may differ in the amount of information they provide as quantified by, for example, the accuracy of estimation and/or the power for statistical testing. An appropriate choice of design could therefore save resources. While for simple models it is well understood how to increase the design's efficiency, for more complex models and their corresponding analyses, this relation between the design and its efficiency becomes much less straightforward. This is, for example, the case for multilevel analysis, which is increasingly used as it elegantly takes into account dependencies among observations. These dependencies can arise because of multistage sampling, with sampled observations embedded within clusters, or because repeated measurements were performed, with a sequence of observations embedded within units; additionally also meta-analyses and multivariate analyses can be dealt with (Raudenbush, 1988; Van den Noortgate & Onghena, 2003; Van den Noortgate & Onghena, 2006).

Designing a multilevel study requires —at each of the levels— a choice of the number of units to sample, and possibly of the predictor values, taking into account the available budget and level-dependent costs of sampling units. As such, a differential cost of sampling units at each of the levels causes a trade-off between sampling as many higher level units as possible and sampling as many observations as possible, at least when resources are limited (Cohen, 1998; Mok, 1995). Furthermore, increasing the efficiency at one level may reduce the efficiency at other levels, partly due to the trade-off, requiring efficiency to be determined with respect to a certain effect or set of effects (Raudenbush & Liu, 2000; Snijders & Bosker, 1993). The efficiency of the design for estimating or testing parameters further depends on the population values for the (co)variance parameters.

Analytical studies that address the efficiency of multilevel designs typically make several limiting assumptions that often compromise generalisations to the actual research setting. To complement these studies, a number of simulation studies have been performed. Such numerical studies are nevertheless also difficult to generalize because results are conditional on the specific model and the specific parameter values that were used to generate the data. Therefore, ideally, simulation studies should be set up for each research setting of interest. Unfortunately, writing these macros often presents a too big challenge for behavioral researchers. In this presentation, we discuss power and accuracy for ongoing school effectiveness research, by means of a simulation tool we developed to aid researchers to set up the appropriate macros more easily.

ML-DEs experiment

Body Text The tool that has been developed is named '*MultiLevel Design Efficiency using simulation*' (ML-DEs; Cools, Van den Noortgate, & Onghena, 2006). It is basically a set of scripts in R (R: A Language and Environment for Statistical Computing, 2004) that are to be run sequentially and that allows for setting up macros for simulation and estimation using the special purpose multilevel modeling program MLwiN (Rasbash, Browne, Healy, Cameron, & Charlton, 2005).

Based on the empirical sampling distribution (ESD) of the estimates, the standard error can be approximated by the standard deviation of the estimates, while the bias can be approximated by the difference between the mean parameter estimate and the population value used for simulation. The proportion of replications that leads to a rejection of the null hypothesis approximates the power, or, in case data were generated under the null hypothesis, it approximates the type 1 error probability. Furthermore, the distribution of estimates can be checked for normality or compared with any other distribution. With a growing number of replications these approximations improve.

To set-up an ML-DEs experiment, comparing the efficiency of different multilevel designs, the following sequence is required. First a number of parameters must be specified as input for R, either directly in R-code or using an optional online form to generate these specifications. Second, a first script (R2MLwiN.R) processes these specifications, resulting in several text files that can be executed in MLwiN as macros. The MLwiN macros, when executed, result in several tab-delimited text files for each of the experimental conditions. Each parameter is assigned a text file with the parameter estimates and their estimated standard errors, including some basic statistics and information on convergence. As such they allow for a Wald test for each of the replications, for which the number of rejections of the null hypothesis can be counted. If likelihood ratio tests were requested on any of the random parameters, then additional text files are generated, containing the unique likelihoods for the full and reduced models, for each of the tests. In agreement with Self and Liang (1987), use can be made of a χ^2 mixture to interpret the results for each of the replications, for which the number of rejections of the null hypothesis can be counted. In a third and final step, a second R script (MLwiN2R.R) re-organizes and summarizes these text files, and specifies functions that can be used for visualizing and analyzing the results. For instance, it is possible to plot the ordered set of estimates and their standard errors for each of the conditions, and compare conditions visually, taking into account the whole distribution instead of its summary statistic. For the likelihood ratio tests, for example, *p*-values can be plotted for each of the conditions.

The scripts, the online form, and further information on ML-DEs can be found at the website of the Centre for Methodology of Educational Research at K.U.Leuven: http://ppw.kuleuven.be/cmes/ML-DEs.html.

Example

Following an example borrowed from Snijders and Bosker (1993), assume that a mathematics test is administered to 5 randomly sampled pupils in each of 100 randomly sampled schools, totaling 500 observed test scores. The primary interests could be in the relation between a school's policy and the achievement of its pupils on a mathematics test (β_3) as well as in whether the effect of the socio-economic status (SES) varies over schools ($\sigma^2_{u_2}$). Further, also IQ and the interaction between SES and Policy are taken into account, resulting in the following model:

$$y_{ij} = \beta_0 + \beta_1 * IQ_{1ij} + \beta_2 * SES_{2ij} + \beta_3 * Policy_{3j} + \beta_4 * (SES_{2ij} * Policy_{3j}) + u_{0j} + u_{2j} * SES_{2ij} + e_{0ij} \quad (1)$$

The test score of pupil *i* from school *j* (y_{ij}) is regressed on several predictors of which the values are assumed to be distributed normally. Other distributions for sampling the predictor values are available in ML-DEs. The intercept is assumed to vary randomly over schools, with a mean equal to β_0 and a school-dependent deviation from that mean (u_{0j}). Also the relation between the SES predictor and the achievement on the mathematics test may be different for the 100 schools and therefore is split up in an average relation (β_2) and a group dependent deviation from that relation (u_{2j}). The u's are assumed to be multivariate

normally distributed with zero means and (co)variances ($\sigma^2_{u_0}$, $\sigma^2_{u_2}$ and $\sigma_{u_{02}}$) to be estimated. A possible effect of IQ (β_1) is assumed to be constant over schools. Finally, both the average achievement and the effect of SES are assumed to differ according to the school's policy as indicated by its main and interaction effects (β_3 and β_4).

A possible question could now be whether it is more efficient to sample fewer schools, but more pupils per school. An ML-DEs experiment is set up to compare three conditions, each with a different number of schools to be sampled. Level-dependent costs of sampling would cause these conditions to differ in their number of observations as well. Given a budget and costs of sampling at the various levels, the number of observations budgetted for can be derived. Assuming that sampling an additional school costs as much as sampling 5 pupils in an already sampled school, having 100 schools with 5 pupils in each school (resulting in 500 observations) would require a budget equivalent to observing 1000 pupils in a single school. For the same budget and cost-ratio, the number of pupils that can be observed in each of 55 schools, would be 13, or 715 observations in total. When having only 10 schools sampled, the budget allows for 95 pupils in each school to be observed, or 950 in total. These three conditions can be compared using their resulting sampling distribution. A selected part of the results is shown below, in Tables 1 and 2, and Figures 1 and 2.

Figure 1. Ordered set of valid estimates (β=0) and their confidence intervals, with the upper and lower boundaries indicated by a dot, for the second-level predictor Policy, including vertical lines representing the number of rejections (thick) and number of valid estimations (thin).

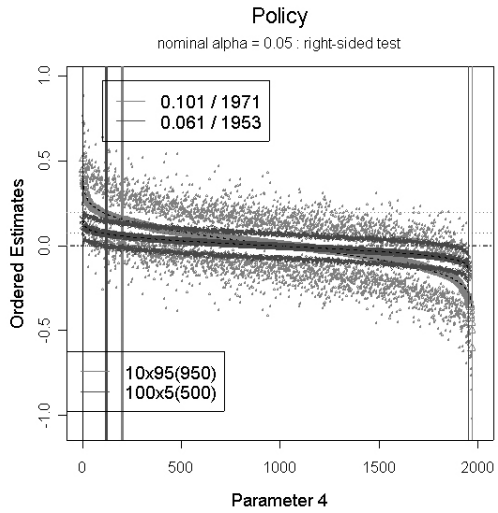
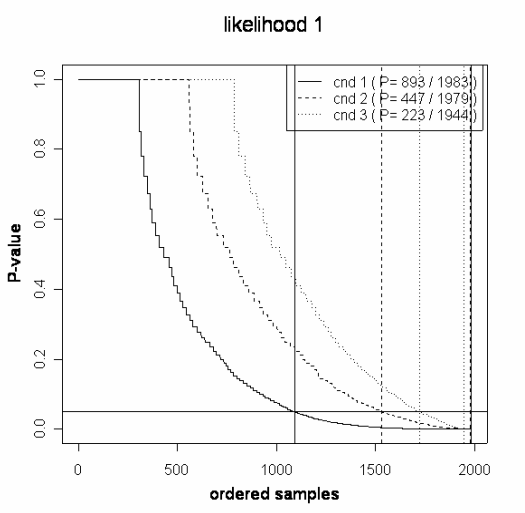


Figure 2. Ordered set of valid estimates of the p-value for testing the random slope (SES, $\sigma^2_{u_2}$ = 0.008 and $\sigma_{u_{02}}$ = -0.01), for three conditions: 10, 55 or 100 schools sampled. The nominal alpha (horizontal line) intersects with the number of acceptations vs. rejections (vertical line).



These results indicate that the accuracy for estimating the parameter for Policy is higher when having 100 groups instead of 10, shown by the distribution of estimates under both conditions. The confidence intervals are accordingly quite different, leading to 10.1 percent rejections when having 10 groups compared to 6.1 percent when having 100 groups, with the latter being much closer to the nominal type 1 error probability. For the variance of the effect of SES with population value 0.008, the mixture likelihood ratio test shows that the condition with the smallest number of groups leads to the highest number of correct rejections, or to the highest number of *p*-values smaller than 0.05.

For the presentation, new data from an ongoing large-scale school effectiveness study will be used, similar to the prototypical example that was given in this paper.

Conclusion

Simulation-based studies which explore the efficiency of multilevel designs avoid several stringent assumptions, but remain conditional on the model and designs used in the simulations. ML-DEs is a simulator which assists researchers to set up simulation studies conditional on the specificities of their own

research interest.

Especially because results of different experiments can be combined, this tool provides strong flexibility. Furthermore, the proposed tool can provide a first step into programming macros in MLwiN as it provides structured macros as output that can further be modified to deal with situations that have not been incorporated in the tool itself, and make it a didactical tool for MLwiN macro programming as well.

For now the ML-DEs tool only works for continuous outcomes and strictly hierarchical data, but in the future more complex models will be implemented, including generalized linear mixed models that deal with binary data using PQL and MQL estimation. Also the inclusion of models for data that are not purely hierarchical is aimed at, like cross-classification and multiple memberships. Finally, it may prove worthwhile to include alternative schemes for generating predictor values, including correlations between non-normal predictors.

Table 1 Summary of the results for the fourth coefficient (β3) for three conditions (10/55/100 schools) including the proportion of rejections, the number of valid samples, the bias, and dispersion that inversely reflects precision.

cnd	emp	valid	cfBias	cfDisp
1	0.1010	1971	0.0028	0.1194
2	0.0526	1977	0.0008	0.0492
3	0.0609	1953	-0.0003	0.0451

Table 2 Summary of the results for the likelihood ratio test of the random slope's (SES) variance for three conditions (10/55/100 schools) including the proportion of rejections, the number of valid samples, and the mean and median p-value.

cnd	emp	valid	PMd	PMn
1	0.4503	1983	0.0773	0.2677
2	0.2259	1979	0.2931	0.4386
3	0.1147	1944	0.5537	0.5692

REFERENCES (RÉFERENCES)

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

Cools, W., Van den Noortgate, W., & Onghena, P. (2006). *ML-DEs: MultiLevel Design Efficiency using simulation*. Retrieved December 24, 2006, from Katholieke Universiteit Leuven, Centre for Methodology of Educational Research Web site: http://ppw.kuleuven.be/cmes/ML-DEs.html.

Mok, M. (1995). Sample size requirements for 2-level designs in educational research. *Multilevel Modelling Newsletter, 7*(2), 11-15.

R Development Core Team. (2004). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org.

Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2005). MLwiN (Version 2.02) [Computer software and manual]. Retrieved 31/12/2006 from http://www.cmm.bristol.ac.uk/MLwiN/.

Raudenbush, S. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics, 13*, 85-116.

Raudenbush, S., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*, 199-213.

Self, G., & Liang. K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association, 82*, 605-610.

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

Van den Noortgate, W., & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement*, *63,* 765-790.

Van den Noortgate, W., & Onghena, P. (2006). Analysing repeated measures data in cognitive research: A comment on regression coefficient analyses. *European Journal of Cognitive Psychology, 18,* 937-952.