Standard Errors and Sample Sizes for Two-Level Research

Tom A. B. Snijders; Roel J. Bosker

Journal of Educational Statistics is currently published by American Educational Research Association.

Standard Errors and Sample Sizes for Two-Level Research

Tom A. B. Snijders
University of Groningen
Roel J. Bosker
University of Twente

The hierarchical linear model approach to a two-level design is considered, some variables at the lower level having fixed and others having random regression coefficients. An approximation is derived to the covariance matrix of the estimators of the fixed regression coefficients (for variables at the lower and the higher level) under the assumption that the sample sizes at either level are large enough. This covariance matrix is expressed as a function of parameters occurring in the model. If a research planner can make a reasonable guess as to these parameters, this approximation can be used as a guide to the choice of sample sizes at either level.

Multilevel and, in particular, two-level designs are used frequently in educational and social research. Hierarchical linear models incorporating both random and fixed effects provide a useful statistical paradigm for situations where nesting is an obvious and direct consequence of multistage sampling as well as situations with nested sources of random variability. (See Bryk & Raudenbush, 1992, for a general introduction; see Raudenbush, 1988, for a review.) Some applications of such models follow.

1. Nesting of microunits within macrounits—for instance, students within schools (e.g., Aitkin & Longford, 1986). The sampling design can be, but does not need to be, a multistage sample.
2. Multivariate analysis, with randomly sampled measurements on a latent construct where the measurements are nested within randomly sampled units (e.g., Goldstein, 1987, p. 61).
3. Assessment of change in a repeated measurements design, with the repeated measures assumed to be random samples within units, the samples being ordered in time (e.g., Bryk & Raudenbush, 1987).

Correspondence concerning this article should be sent to Tom A. B. Snijders, VSM-FPPSW, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands; e-mail T.A.B.SNIJDERS@PPSW.RUG.NL.

4. Meta-analysis, with units nested within the studies that are reviewed (e.g., Raudenbush & Bryk, 1985).
5. Logical extensions of the situations in Points 1 to 4, like repeated measurements on students nested within schools, or students within cohorts within schools (Bosker, Guldemond, Hofman, & Hofman, 1989; Goldstein, 1987; Raudenbush, 1989).

When a researcher is designing a multistage sampling scheme for the first of these situations, important decisions must be made with respect to the sample sizes at the various levels. In most of the other applications, there are, of course, severe restrictions on the sample size at the microlevel—for example, one cannot take any number of tests on pupils in repeated measurement designs. For a two-level design in educational research, the question might be phrased like this: Should one investigate many schools with few students per school or few schools with many students per school? Until recently, this design question was not at stake, because applications of multilevel models in educational research were restricted to secondary analyses of large scale survey data—for example, Coleman, Hoffer, and Kilgore's High School and Beyond data cited in Raudenbush and Bryk (1986), Inner London Education Authority (ILEA) data by Aitkin and Longford (1986) and Goldstein (1987), the data of a Dutch census cohort by Kreft (1987) and by Bosker et al. (1989), and the Scottish School Leavers Survey data by Willms (1987). One might ask questions, however, about the power of the tests used in these designs to obtain an impression of the validity of the results presented. Now that multilevel software is widely available (VARCL3, VARCL9 by Longford, 1988; HLM2.0 by Bryk, Raudenbush, Congdon, & Seltzer, 1988; ML3 by Prosser, Rasbash, & Goldstein, 1991), one can foresee that many educational and social researchers will test their hypotheses within a multilevel framework. This situation leads to the specific multilevel design question: How should researchers choose sample sizes at the macro- and microlevel in order to ensure a desired level of power given a relevant (hypothesized) effect size and a chosen significance level α? Until now, little has been known about this subject for multilevel designs. Raudenbush, reviewing educational applications of hierarchical linear models, concludes the following:

> Interval estimates and hypothesis tests for HLM-type models rely on large sample properties of maximum likelihood estimates. Little is known about the small sample behavior of the estimates. Standard errors of macroeffects are least well understood when the number of groups is small. Standard errors of microeffects are even more problematic. Investigators are just beginning to understand the effect of errors of variance component estimation on standard errors of shrinkage estimators. (1988, p. 111)

Goldstein suggests the use of pilot studies or even simulation studies to optimize the design of a multilevel study (1987, p. 87). Although useful, this

strategy gives only ad hoc and nongeneralizable rules for the design of multilevel studies. In this article, we study expressions for large-sample standard errors and power of macroeffects (i.e., fixed regression coefficients) in designs with macrolevel and macro–micro interaction (cross-level) effects on microlevel outcomes; in subsequent research, the small-sample quality of these large-sample approximations is being investigated. The question of standard errors for microeffects (i.e., posterior means) is not treated here.

This article presents some approximate formulas for the standard errors of estimated regression coefficients in two-level designs. The approximations involved are large sample approximations in which it is assumed that the sample size at either level is sufficiently large. (In further research, how large this sample size should be to give satisfactory accuracy will be investigated; see the discussion section.) These standard errors are functions, among others, of the sample sizes at the macro- and microlevel. If reasonable estimates or guesses are available for the parameter values that are necessary to compute these formulas (notably, covariance matrices of predictor variables and variance components), then it is possible to determine sample sizes at the macro- and microlevel that are approximately optimal. Optimality of sample sizes means here that the design yields minimal standard errors for the parameters of most interest, given budgetary or other constraints. (This optimality is conditional, of course, on the correctness of the estimated parameter values used in these computations. In view of the asymptotic nature of the expressions for the standard errors, calculated approximate optima should not be trusted if the associated micro- or macrolevel sample size is less than 10.) Using the well-known approximate formula relating effect size and standard errors to statistical power, the formulas for standard errors can be used to calculate approximately optimal sample sizes for attaining a desired level of power.

The Two-Level Linear Model

A two-level structure is assumed throughout this article. For the sake of concreteness, units at the two levels are called *students* and *schools*, respectively; the reader may, of course, substitute any other words relevant to his or her field. In order to avoid complex discussion, it is assumed that the sample size at the first (micro) level—that is, the number of students sampled per school—is constant over schools; it is denoted by n. The number of schools sampled is N; hence the total sample size is nN. The fact that the number of students per school is assumed constant does not constitute a serious problem for the intended application: We wish to arrive at approximately optimal sample sizes at either level; the approximate optimality will not be seriously threatened by minor fluctuations of sample sizes per school due to nonresponse, some small group sizes, and so forth. The

dependent variable is denoted Y; there are assumed to be K explanatory variables at the student level. The observations of the dependent variable for the n students in the jth school are arranged in the random vector $\mathbf{Y}_j$ for which the usual linear model is assumed:

$$\mathbf{Y}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \mathbf{R}_j \qquad j = 1, \ldots, N \tag{1}$$

where $\mathbf{X}_j$ is the design matrix for school j and $\boldsymbol{\beta}_j$ its vector of regression coefficients. For the residual vector $\mathbf{R}_j$, we assume homoscedasticity, absence of correlation, and zero mean:

$$E(\mathbf{R}_j) = \mathbf{0}, \quad \text{Cov}(\mathbf{R}_j) = \sigma^2 \mathbf{I}, \tag{2}$$

where $\mathbf{I}$ is the identity matrix and σ^2 the residual variance at the student level. The student explanatory variables are supposed to be ordered in two subsets of sizes K_1 and K_2 ($K_1 + K_2 = K$), so that the first subset contains the variables with regression coefficients β_{jk} that are constant across schools:

$$\beta_{jk} = \beta_{0k}, \qquad k = 1, \ldots, K_1,$$

while the second subset contains the variables having random regression coefficients β_{jk} ($k = K_1 + 1, \ldots, K$). These random regression coefficients are assumed to satisfy, in their turn, a linear model incorporating explanatory variables at the school level. (This model can be with or without a constant term; this is expressed by the vector $\mathbf{e}$ in (8) below.) It will be convenient to formulate this in a model for the entire $\boldsymbol{\beta}_j$ vector:

$$\boldsymbol{\beta}_j = \mathbf{W}_j \boldsymbol{\gamma} + \mathbf{U}_j. \tag{3}$$

Here, $\mathbf{W}_j$ is the design matrix for school j, $\boldsymbol{\gamma}$ is the vector of regression coefficients at the school level, and $\mathbf{U}_j$ is the residual vector at the school level. The moment assumptions for $\mathbf{U}_j$ are

$$E(\mathbf{U}_j) = \mathbf{0}, \quad \text{Cov}(\mathbf{U}_j) = \boldsymbol{\tau}. \tag{4}$$

The order of the student explanatory variables implies the following matrix partitionings:

$$\boldsymbol{\beta}_j = \begin{pmatrix} \boldsymbol{\beta}_j^{(1)} \\ \boldsymbol{\beta}_j^{(2)} \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{pmatrix},$$

$$\mathbf{X}_j = (\mathbf{X}_{1j} \quad \mathbf{X}_{2j}), \quad \mathbf{U}_j = \begin{pmatrix} \mathbf{0} \\ \mathbf{U}_{2j} \end{pmatrix},$$

$$\mathbf{W}_j = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{2j} \end{pmatrix}, \quad \boldsymbol{\tau} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\tau}_2 \end{pmatrix},$$

implying indeed that $\boldsymbol{\beta}_j^{(1)} = \boldsymbol{\gamma}_1$ independently of j. Relations 1 and 3 can now be combined into

$$\mathbf{Y}_j = \mathbf{X}_{1j}\,\boldsymbol{\gamma}_1 + \mathbf{X}_{2j}\,\boldsymbol{\beta}_j^{(2)} + \mathbf{R}_j, \tag{5a}$$

$$\boldsymbol{\beta}_j^{(2)} = \mathbf{W}_{2j}\,\boldsymbol{\gamma}_2 + \mathbf{U}_{2j}, \tag{5b}$$

or equivalently into

$$\mathbf{Y}_j = \mathbf{X}_{1j}\,\boldsymbol{\gamma}_1 + \mathbf{X}_{2j}\,\mathbf{W}_{2j}\,\boldsymbol{\gamma}_2 + \mathbf{X}_{2j}\,\mathbf{U}_{2j} + \mathbf{R}_j. \tag{5c}$$

It is assumed that all student variables with random regression coefficients (those in $\mathbf{X}_{2j}$) are school deviation scores, except possibly for the first variable, which is a constant (giving the random intercept). In other words, all columns of $\mathbf{X}_{2j}$ have a zero mean, for all j, except possibly for the first column, which then has all its entries equal to unity, for all j. (This centering assumption is used in Formula 24a below; without this assumption, fourth between-school moments of the explanatory variables would be needed in the expression for the standard errors. A less restrictive assumption that could also be used is independence of the school means of the $\mathbf{X}_{1j}$ and the $\mathbf{X}_{2j}$ variables.)

The covariance matrix of $\mathbf{Y}_j$ is given by

$$\mathrm{Cov}(\mathbf{Y}_j) = \sigma^2\,\mathbf{I} + \mathbf{X}_j\,\boldsymbol{\tau}\mathbf{X}_j' = \sigma^2\,\mathbf{I} + \mathbf{X}_{2j}\,\boldsymbol{\tau}_2\,\mathbf{X}_{2j}'. \tag{6}$$

This demonstrates the well-known decomposition into student-level and school-level variance or covariance components.

It will be assumed below (again in order to keep the discussion as simple as possible in this situation) that, for every regression coefficient β_{2jk} (i.e., for the variable regression coefficients at the school level), the same explanatory variables at the school level are used. Formulas where this assumption is not made can be derived along the same lines as those presented in this article. This more general approach is not followed here for the following reasons: Notation and formulas become considerably more complex; the purpose of the formulas in this article is design rather than inference, so that it is not required here to achieve the most parsimonious model; the full covariance matrix of the explanatory variables at school level is necessary anyway to compute the formulas, so the greater generality would not lead to parsimony in the required parameters, and the approach followed in the derivations is a large-sample approach of such a kind that noninclusion of unimportant effects will not lead to smaller standard errors for important effects.

Let $\mathbf{W}_{3j}$ denote the (column) vector of these explanatory variables for

school j. Then the design matrix $\mathbf{W}_{2j}$ in (5) has the form

$$\mathbf{W}_{2j} = \begin{pmatrix} \mathbf{W}'_{3j} & \mathbf{0} & \mathbf{0} & . & . & \mathbf{0} \\ \mathbf{0} & \mathbf{W}'_{3j} & \mathbf{0} & . & . & \mathbf{0} \\ . & & . & . & . & . \\ \mathbf{0} & . & . & . & . & \mathbf{W}'_{3j} \end{pmatrix} = \mathbf{I} \otimes \mathbf{W}'_{3j}, \qquad (7)$$

where $\otimes$ denotes the Kronecker product (e.g., Rao, 1973, p. 29). It will usually be convenient to let the first variables in $\mathbf{X}_{2j}$ and also in $\mathbf{W}_{3j}$ be constant (representing the regression intercepts at the student and the school level, respectively) and equal to unity.

The n rows of $\mathbf{X}_j$ are identically distributed; the ith row, where i refers to a student within school j, is partitioned as

$$(\mathbf{X}'_{1ji} \quad \mathbf{X}'_{2ji}).$$

Because the variables in $\mathbf{X}_{2j}$ are assumed to be within-school deviation scores, except possibly for the first one, the within-school mean of $\mathbf{X}_{2ji}$ is constant, with all elements zero except possibly the first one. This constant mean vector is denoted

$$\mathbf{e} = \mathrm{E}(\mathbf{X}_{2ji}) = 0 \quad \text{or} \quad (1, \quad 0, \quad 0, \ldots, 0)'. \qquad (8)$$

The within-school mean of $\mathbf{X}_{1ji}$ is denoted

$$\mathbf{M}_{1j} = \mathrm{E}(\mathbf{X}_{1ji}) \qquad \text{(within-school } j\text{)}. \qquad (9)$$

The within-school covariance matrix of $\mathbf{X}_{ji}$ is denoted by

$$\mathbf{\Sigma}^W = \begin{pmatrix} \mathbf{\Sigma}^W_{11} & \mathbf{\Sigma}^W_{12} \\ \mathbf{\Sigma}^W_{21} & \mathbf{\Sigma}^W_{22} \end{pmatrix} = \mathrm{Cov}\begin{pmatrix} \mathbf{X}_{1ji} \\ \mathbf{X}_{2ji} \end{pmatrix} \qquad \text{(within schools)}. \qquad (10)$$

This within-school covariance matrix is assumed to be independent of the school j (homoscedasticity). The mean over schools of $\mathbf{M}_{1j}$ is denoted

$$\mathbf{\mu}_1 = \mathrm{E}(\mathbf{M}_{1j}), \qquad (11)$$

and the between-schools covariance matrix of $\mathbf{X}_{1ji}$, which is the covariance matrix of the school means $\mathbf{M}_{1j}$, is denoted $\mathbf{\Sigma}^B_{11}$. In addition to $\mathbf{M}_{1j}$, the school variables in the row vector $\mathbf{W}_{3j}$ also play a role. The corresponding mean value and joint (between-schools) covariance matrices are denoted

$$\mathbf{\mu}_3 = \mathrm{E}(\mathbf{W}_{3j}), \quad \mathbf{\Sigma}^B = \begin{pmatrix} \mathbf{\Sigma}^B_{11} & \mathbf{\Sigma}^B_{13} \\ \mathbf{\Sigma}^B_{31} & \mathbf{\Sigma}^B_{33} \end{pmatrix} = \mathrm{Cov}\begin{pmatrix} \mathbf{M}_{1j} \\ \mathbf{W}_{3j} \end{pmatrix} \qquad \text{(between schools)}. \qquad (12)$$

Estimators

This article focuses on estimating the vector of regression parameters $\mathbf{\gamma}$. Assuming for the moment that the variance σ^2 and the covariance matrix $\mathbf{\tau}$ are known, the usual estimator for $\mathbf{\gamma}$ within the paradigm of hierarchical

linear models is the generalized least squares (GLS) estimator for the model defined by (1)–(4), which is also the maximum likelihood estimator if multivariate normality is assumed. Note that

$$E(\mathbf{Y}_j) = \mathbf{X}_j \, \mathbf{W}_j \, \boldsymbol{\gamma},$$

while the covariance matrix is given by (6). The generalized least squares estimator is the value for $\boldsymbol{\gamma}$ that minimizes

$$\sum_{j=1}^{N} (\mathbf{Y}_j - \mathbf{X}_j \, \mathbf{W}_j \, \boldsymbol{\gamma})'(\sigma^2 \mathbf{I} + \mathbf{X}_j \, \boldsymbol{\tau} \mathbf{X}_j')^{-1}(\mathbf{Y}_j - \mathbf{X}_j \, \mathbf{W}_j \, \boldsymbol{\gamma}).$$

Standard theory (e.g., de Leeuw & Kreft, 1986, pp. 70–71; Rao, 1973) shows that the solution is

$$\hat{\boldsymbol{\gamma}}_{\text{GLS}} = \mathbf{V}^{-1} \sum_{j=1}^{N} \mathbf{W}_j' \, \mathbf{X}_j'(\sigma^2 \mathbf{I} + \mathbf{X}_j \, \boldsymbol{\tau} \mathbf{X}_j')^{-1} \, \mathbf{Y}_j \tag{13}$$

with

$$\mathbf{V} = \sum_{j=1}^{N} \mathbf{W}_j' \, \mathbf{X}_j'(\sigma^2 \mathbf{I} + \mathbf{X}_j \, \boldsymbol{\tau} \mathbf{X}_j')^{-1} \mathbf{X}_j \, \mathbf{W}_j. \tag{14}$$

This estimator is also the empirical Bayes (EB) estimator given by Efron and Morris (1975) and Raudenbush and Bryk (1986).

In practice, of course, σ^2 and $\boldsymbol{\tau}$ are unknown and have to be estimated from the data. This is usually done by maximum likelihood methods under the normality assumption, which can be implemented by a variety of algorithms: iterated generalized least squares (Goldstein, 1986), Fisher's scoring method (Longford, 1987), or the EM algorithm (Raudenbush, 1988). The resulting version of $\hat{\boldsymbol{\gamma}}_{\text{GLS}}$ where estimates of σ^2 and $\boldsymbol{\tau}$ are used will be denoted $\hat{\boldsymbol{\gamma}}$. Its covariance matrix can be approximated in the usual way as the inverse of the observed information matrix. These calculations can be carried out rather easily for given data $\mathbf{Y}_j$ ($j = 1, \ldots, N$) if Fisher's scoring method is used. We are interested now, however, in an expression for $\text{Cov}(\hat{\boldsymbol{\gamma}})$ that can be computed before the data are collected, using estimates or guesses for the relevant unknown parameters. This makes the approach via the observed information matrix less suitable.

We assume that N is sufficiently large to ensure that the (co)variance parameters σ^2 and $\boldsymbol{\tau}$ are estimated so precisely that the precision of the estimator for $\boldsymbol{\gamma}$ is hardly affected by the fact that estimated instead of true values of σ^2 and $\boldsymbol{\tau}$ are used. Hence $\text{Cov}(\hat{\boldsymbol{\gamma}})$ may be approximated by $\text{Cov}(\hat{\boldsymbol{\gamma}}_{\text{GLS}})$. (Asymptotically as N tends to infinity, the limits of N times these covariance matrices are identical.) This implies that the usual expression for the covariance matrix of the GLS estimator may be used as an approximation:

$$\text{Cov}(\hat{\boldsymbol{\gamma}}) \approx \text{Cov}(\hat{\boldsymbol{\gamma}}_{\text{GLS}}) = \mathbf{V}^{-1}, \tag{15}$$

where $\mathbf{V}$ is given by (14).

Covariance Matrix of the Estimators

The specification of $\mathbf{V}$ implied by the model assumptions can be exploited in order to turn Expressions 14 and 15 into something more useful. Recall from (5) that $\boldsymbol{\gamma}$ is partitioned into the vector $\boldsymbol{\gamma}_1$, containing the student level regression coefficients that are constant across schools, and the vector $\boldsymbol{\gamma}_2$, containing the regression coefficients for the fixed part $\mathbf{W}_{2j}\,\boldsymbol{\gamma}_2$ of the varying regression coefficients $\boldsymbol{\beta}_j^{(2)}$. Correspondingly, $\mathbf{V}$ is partitioned into

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}.$$

(The following derivation can be skipped by readers who are interested only in the results; they can resume reading with Formula 31.) In order to find a useful reexpression of (14), we define

$$\mathbf{T}_j = \mathbf{X}_j\,\boldsymbol{\tau}\mathbf{X}_j' + \sigma^2\mathbf{I} = \mathbf{X}_{2j}\,\boldsymbol{\tau}_2\mathbf{X}_{2j}' + \sigma^2\mathbf{I} \tag{16}$$

so that

$$\mathbf{V} = \sum_{j=1}^{N} \mathbf{W}_j'\,\mathbf{X}_j'\,\mathbf{T}_j^{-1}\,\mathbf{X}_j\,\mathbf{W}_j.$$

Using the partitioning $\mathbf{X}_j = (\mathbf{X}_{1j}\,\mathbf{X}_{2j})$, $\mathbf{W}_j = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{2j} \end{pmatrix}$ and defining

$$\mathbf{U}_{stj} = \mathbf{X}_{sj}'\,\mathbf{T}_j^{-1}\,\mathbf{X}_{tj} \qquad \text{for } s,t = 1,2,$$

this can be expressed as

$$\mathbf{V} = \sum_{j=1}^{N} \begin{pmatrix} \mathbf{U}_{11j} & \mathbf{U}_{12j}\,\mathbf{W}_{2j} \\ \mathbf{W}_{2j}'\mathbf{U}_{12j}' & \mathbf{W}_{2j}'\mathbf{U}_{22j}\,\mathbf{W}_{2j} \end{pmatrix}. \tag{17}$$

We found it convenient to use the well-known matrix identity

$$\begin{aligned}(\sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\tau}\mathbf{X}')^{-1} = \sigma^{-2}\mathbf{I} &- \sigma^{-2}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &+ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \boldsymbol{\tau})^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\end{aligned} \tag{18}$$

(e.g., de Leeuw & Kreft, 1986, Formula 15; Rao, 1973, p. 33) for $\mathbf{X} = \mathbf{X}_{2j}$. This yields

$$\begin{aligned}\mathbf{U}_{11j} = \sigma^{-2}\mathbf{X}_{1j}'\mathbf{X}_{1j} &- \sigma^{-2}\mathbf{X}_{1j}'\mathbf{X}_{2j}(\mathbf{X}_{2j}'\mathbf{X}_{2j})^{-1}\mathbf{X}_{2j}'\mathbf{X}_{1j} \\ &+ \mathbf{X}_{1j}'\mathbf{X}_{2j}(\mathbf{X}_{2j}'\mathbf{X}_{2j})^{-1}\mathbf{C}_{2j}^{-1}(\mathbf{X}_{2j}'\mathbf{X}_{2j})^{-1}\mathbf{X}_{2j}'\mathbf{X}_{1j},\end{aligned} \tag{19}$$

where

$$\mathbf{C}_{2j} = \sigma^2(\mathbf{X}_{2j}'\mathbf{X}_{2j})^{-1} + \boldsymbol{\tau}_2.$$

It also follows from (18) that

$$\mathbf{T}_j^{-1}\mathbf{X}_{2j} = \mathbf{X}_{2j}(\mathbf{X}_{2j}'\mathbf{X}_{2j})^{-1}\mathbf{C}_{2j}^{-1}$$

so that

$$U_{12j} = X'_{1j} X_{2j} (X'_{2j} X_{2j})^{-1} C_{2j}^{-1} \tag{20}$$

$$U_{22j} = C_{2j}^{-1}. \tag{21}$$

Approximation for the Covariance Matrix

The approximation for the covariance matrix is obtained by replacing the sums of squares and cross-products in the matrix V with their expected values in a judicious way. First this is done with respect to the distribution of student-level explanatory variables within schools: Terms of the form $X'_{sj} X_{tj}$ are replaced by their expected values. The resulting approximate expression for V is replaced by its expectation with respect to its distribution over schools. The main difficulties with this approximation procedure may be expected to lie at the within-school level: It is a valid first-order asymptotic approximation for $n \to \infty$, which we wish to apply for intermediate values of n (10 and higher). It should be noted that, for the design purpose that is treated in this article, we aim at formulas that are not overwhelmingly complex and that give reasonable approximations without necessarily being exact. We are now conducting simulation studies to investigate the validity of these approximations, and initial results are encouraging. In the last two sections of this article, we discuss the practical validity of this approximation.

Expectations within schools are taken for the student variables in matrices X_{1j} and X_{2j}. Note that

$$X'_{sj} X_{tj} = \sum_{i=1}^{n} X_{sji} X'_{tji} \qquad s, t = 1, 2.$$

This is approximated by its expected value:

$$X'_{sj} X_{tj} \approx n M_{sj} M'_{tj} + n \Sigma_{st}^{W},$$

where M_{2j} is understood to equal e given in (8). For $s = t = 2$, this implies

$$n^{-1} X'_{2j} X_{2j} \approx E_{22},$$

where

$$E_{22} = ee' + \Sigma_{22}^{W}, \tag{22}$$

independently of j. It follows that

$$C_{2j} \approx \tau_2 + n^{-1} \sigma^2 E_{22}^{-1}. \tag{23}$$

Replacing all factors $X'_{tj} X_{sj}$ in (19)–(21) by their within-schools expected values now results in

$$U_{11j} \approx n\sigma^{-2}(M_{1j} M'_{1j} + \Sigma_{11}^{W}) - n\sigma^{-2}(M_{1j} e' + \Sigma_{12}^{W}) E_{22}^{-1}(eM'_{1j} + \Sigma_{21}^{W})$$
$$+ (M_{1j} e' + \Sigma_{12}^{W}) E_{22}^{-1} F E_{22}^{-1}(eM'_{1j} + \Sigma_{21}^{W}) \tag{24a}$$

245

$$\mathbf{U}_{12j} \approx (\mathbf{M}_{1j}\,\mathbf{e}' + \boldsymbol{\Sigma}_{12}^{W})\mathbf{E}_{22}^{-1}\,\mathbf{F} \tag{24b}$$

$$\mathbf{U}_{22j} \approx \mathbf{F}, \tag{24c}$$

where

$$\mathbf{F} = (\boldsymbol{\tau}_2 + n^{-1}\sigma^2\,\mathbf{E}_{22}^{-1})^{-1}. \tag{25}$$

From (8) and the fact that $e_1 \neq 0$ implies $(\boldsymbol{\Sigma}_{22}^{W})_{11} = 0$, it follows that always $\mathbf{E}_{22}^{-1}\,\mathbf{e} = \mathbf{e}$, $\mathbf{e}'\mathbf{E}_{22}^{-1}\,\mathbf{e} = e_1$, and $\boldsymbol{\Sigma}_{12}^{W}\mathbf{E}_{22}^{-1}\,\mathbf{e} = \mathbf{0}$. This yields for (24a, 24b) the equivalent expressions

$$\mathbf{U}_{11j} \approx n\sigma^{-2}\{(1 - e_1)\mathbf{M}_{1j}\,\mathbf{M}_{1j}' + \boldsymbol{\Sigma}_{11}^{W} - \boldsymbol{\Sigma}_{12}^{W}\mathbf{E}_{22}^{-1}\,\boldsymbol{\Sigma}_{21}^{W}\} + (\mathbf{e}'\mathbf{F}\mathbf{e})\mathbf{M}_{1j}\,\mathbf{M}_{1j}'$$
$$+ \mathbf{M}_{1j}\,\mathbf{e}'\mathbf{F}\mathbf{E}_{22}^{-1}\,\boldsymbol{\Sigma}_{21}^{W} + \boldsymbol{\Sigma}_{12}^{W}\mathbf{E}_{22}^{-1}\,\mathbf{F}\mathbf{e}\mathbf{M}_{1j}' + \boldsymbol{\Sigma}_{12}^{W}\mathbf{E}_{22}^{-1}\,\mathbf{F}\mathbf{E}_{22}^{-1}\,\boldsymbol{\Sigma}_{21}^{W} \tag{26a}$$

$$\mathbf{U}_{12j} \approx \mathbf{M}_{1j}\,\mathbf{e}'\mathbf{F} + \boldsymbol{\Sigma}_{12}^{W}\mathbf{E}_{22}^{-1}\,\mathbf{F}. \tag{26b}$$

The second step consists of taking expectations over schools for the school level variables $\mathbf{M}_{1j}$ and $\mathbf{W}_{3j}$. In view of Expression 17 for $\mathbf{V}$, we define

$$\mathbf{V}_{11j} = \mathbf{U}_{11j}$$

$$\mathbf{V}_{12j} = \mathbf{U}_{12j}\,\mathbf{W}_{2j}$$

$$\mathbf{V}_{22j} = \mathbf{W}_{2j}'\,\mathbf{U}_{22j}\,\mathbf{W}_{2j}.$$

It is now assumed that $\mathbf{W}_{2j}$ has the form (7); $\mathbf{V}_{12j}$ and $\mathbf{V}_{22j}$ can then be expressed as

$$\mathbf{V}_{12j} = \mathbf{U}_{12j}(\mathbf{I} \otimes \mathbf{W}_{3j}') = \mathbf{U}_{12j} \otimes \mathbf{W}_{3j}' \approx (\mathbf{M}_{1j}\,\mathbf{e}' + \boldsymbol{\Sigma}_{12}^{W}\mathbf{E}_{22}^{-1})\mathbf{F} \otimes \mathbf{W}_{3j}' \tag{27a}$$

$$\mathbf{V}_{22j} = (\mathbf{I} \otimes \mathbf{W}_{3j})\mathbf{U}_{22j}(\mathbf{I} \otimes \mathbf{W}_{3j}') = \mathbf{U}_{22j} \otimes (\mathbf{W}_{3j}\,\mathbf{W}_{3j}') \approx \mathbf{F} \otimes (\mathbf{W}_{3j}\,\mathbf{W}_{3j}'). \tag{27b}$$

If $\mathbf{W}_{2j}$ does not have the form (7), then more complex expressions can be derived that are analogous to (27).

The expectation of Expression 26a for $\mathbf{U}_{11j}$ involves first and second moments of $\mathbf{M}_{1j}$. These are given by

$$E(\mathbf{M}_{1j}) = \boldsymbol{\mu}_1, \qquad E(\mathbf{M}_{1j}\,\mathbf{M}_{1j}') = \boldsymbol{\mu}_1\,\boldsymbol{\mu}_1' + \boldsymbol{\Sigma}_{11}^{B}.$$

This yields as the expectation of (26a)

$$E(\mathbf{V}_{11j}) \approx \mathbf{V}_{11}^{(n)} = n\sigma^{-2}\{(1 - e_1)(\boldsymbol{\mu}_1\,\boldsymbol{\mu}_1' + \boldsymbol{\Sigma}_{11}^{B}) + \boldsymbol{\Sigma}_{11}^{W} - \boldsymbol{\Sigma}_{12}^{W}\mathbf{E}_{22}^{-1}\,\boldsymbol{\Sigma}_{21}^{W}\}$$
$$+ (\mathbf{e}'\mathbf{F}\mathbf{e})(\boldsymbol{\mu}_1\,\boldsymbol{\mu}_1' + \boldsymbol{\Sigma}_{11}^{B}) + \boldsymbol{\mu}_1\,\mathbf{e}'\mathbf{F}\mathbf{E}_{22}^{-1}\,\boldsymbol{\Sigma}_{21}^{W}$$
$$+ \boldsymbol{\Sigma}_{12}^{W}\mathbf{E}_{22}^{-1}\,\mathbf{F}\mathbf{e}\boldsymbol{\mu}_1' + \boldsymbol{\Sigma}_{12}^{W}\mathbf{E}_{22}^{-1}\,\mathbf{F}\mathbf{E}_{22}^{-1}\,\boldsymbol{\Sigma}_{21}^{W}. \tag{28}$$

As an important special case, let us briefly consider models with an intercept—that is, $e_1 = 1$—and with centered pupil-level variables—that is, $\boldsymbol{\mu}_1 = \mathbf{0}$. In this case, (28) becomes somewhat simpler:

$$E(\mathbf{V}_{11j}) \approx \mathbf{V}_{11}^{(n)} = n\sigma^{-2}\{\boldsymbol{\Sigma}_{11}^{W} - \boldsymbol{\Sigma}_{12}^{W}\mathbf{E}_{22}^{-1}\,\boldsymbol{\Sigma}_{21}^{W}\} + f_{11}\,\boldsymbol{\Sigma}_{11}^{B} + \boldsymbol{\Sigma}_{12}^{W}\mathbf{E}_{22}^{-1}\,\mathbf{F}\mathbf{E}_{22}^{-1}\,\boldsymbol{\Sigma}_{21}^{W}, \tag{28'}$$

where f_{11} denotes the $(1, 1)$ element of matrix $\mathbf{F}$.

Concerning Expression 27a, for $\mathbf{V}_{12j}$, properties of the Kronecker product imply that

$$(\mathbf{M}_{1j}\,\mathbf{e}'\mathbf{A}) \otimes \mathbf{W}'_{3j} = (\mathbf{e}'\mathbf{A}) \otimes (\mathbf{M}_{1j}\,\mathbf{W}'_{3j})$$

for any matrix $\mathbf{A}$. Applying this rule and replacing $\mathbf{M}_{1j}$ and $\mathbf{M}_{1j}\,\mathbf{W}'_{3j}$ by their expected values

$$E(\mathbf{M}_{1j}) = \boldsymbol{\mu}_1, \qquad E(\mathbf{M}_{1j}\,\mathbf{W}'_{3j}) = \boldsymbol{\mu}_1\,\boldsymbol{\mu}'_3 + \boldsymbol{\Sigma}^B_{13}$$

yield

$$E(\mathbf{V}_{12j}) \approx \mathbf{V}^{(n)}_{12} = (\boldsymbol{\Sigma}^W_{12}\mathbf{E}^{-1}_{22}\,\mathbf{F}) \otimes \boldsymbol{\mu}'_3 + (\mathbf{e}'\mathbf{F}) \otimes (\boldsymbol{\mu}_1\,\boldsymbol{\mu}'_3 + \boldsymbol{\Sigma}^B_{13}). \qquad (29)$$

Finally, in Expression 27b for $\mathbf{V}_{22j}$, the factor $\mathbf{W}_{3j}\,\mathbf{W}'_{3j}$ can be straightforwardly replaced by its expected value:

$$E(\mathbf{V}_{22j}) \approx \mathbf{V}^{(n)}_{22} = \mathbf{F} \otimes (\boldsymbol{\mu}_3\,\boldsymbol{\mu}'_3 + \boldsymbol{\Sigma}^B_{33}). \qquad (30)$$

The matrix $\mathbf{V}$, given by (17), can now be approximated by

$$\mathbf{V} \approx N\begin{pmatrix} \mathbf{V}^{(n)}_{11} & \mathbf{V}^{(n)}_{12} \\ \mathbf{V}^{(n)}_{21} & \mathbf{V}^{(n)}_{22} \end{pmatrix}, \qquad (31)$$

where the $\mathbf{V}^{(n)}_{st}$ are defined by (28)–(30) with (22) and (25). The inverse of (31) gives, according to (15), a useful expression for $\text{Cov}(\hat{\boldsymbol{\gamma}})$.

This section is now concluded with a very rough first order approximation that is not of practical use but does give an indication of the asymptotic order of magnitude of $\text{Cov}(\hat{\boldsymbol{\gamma}})$, when both n and N are very large (larger than necessary for the usefulness of (28)–(30)). Inspection of (28)–(30) shows that, if all lower order terms are neglected, the following rough approximation is obtained:

$$\mathbf{V}^{(n)}_{11} \approx n\sigma^{-2}\{(1 - e_1)(\boldsymbol{\mu}_1\,\boldsymbol{\mu}'_1 + \boldsymbol{\Sigma}^B_{11}) + \boldsymbol{\Sigma}^W_{11} - \boldsymbol{\Sigma}^W_{12}\mathbf{E}^{-1}_{22}\boldsymbol{\Sigma}^W_{21}\} \qquad (32)$$

$$\mathbf{V}^{(n)}_{12} \approx (\boldsymbol{\Sigma}^W_{12}\mathbf{E}^{-1}_{22}\,\boldsymbol{\tau}^{-1}_2) \otimes \boldsymbol{\mu}'_3 + (\mathbf{e}'\boldsymbol{\tau}^{-1}_2) \otimes (\boldsymbol{\mu}_1\,\boldsymbol{\mu}'_3 + \boldsymbol{\Sigma}^B_{13}) \qquad (33)$$

$$\mathbf{V}^{(n)}_{22} \approx \boldsymbol{\tau}^{-1}_2 \otimes (\boldsymbol{\mu}_3\,\boldsymbol{\mu}'_3 + \boldsymbol{\Sigma}^B_{33}). \qquad (34)$$

In this approximation, it is assumed that the right-hand sides of (32) and (34) are nonsingular matrices. (If the model has an intercept, so $e_1 = 1$, it implies that $\boldsymbol{\Sigma}^W_{11}$ is nonsingular: The student variables with fixed regression coefficients have positive within-school variances.) Note that Approximation 32 is proportional to n, while (33)–(34) are independent of n. With the inversion formula for partitioned symmetric matrices,

$$\mathbf{V}^{-1} = \begin{pmatrix} \mathbf{V}^{-1}_{11} + \mathbf{Q}\mathbf{P}^{-1}\mathbf{Q}' & -\mathbf{Q}\mathbf{P}^{-1} \\ -\mathbf{P}^{-1}\mathbf{Q}' & \mathbf{P}^{-1} \end{pmatrix},$$

where

$$\mathbf{P} = \mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}^{-1}_{11}\mathbf{V}_{12}, \qquad \mathbf{Q} = \mathbf{V}^{-1}_{11}\mathbf{V}_{12};$$

247

this shows that

$$\text{Cov}(\hat{\gamma}_1) \approx (1/Nn)\sigma^2\{(1 - e_1)(\boldsymbol{\mu}_1\boldsymbol{\mu}_1' + \boldsymbol{\Sigma}_{11}^B) + \boldsymbol{\Sigma}_{11}^W - \boldsymbol{\Sigma}_{12}^W\mathbf{E}_{22}^{-1}\boldsymbol{\Sigma}_{21}^W\}^{-1}, \quad (35)$$

which is of order $1/Nn$, and

$$\text{Cov}(\hat{\gamma}_2) \approx (1/N)(\boldsymbol{\tau}_2^{-1} \otimes (\boldsymbol{\mu}_3\boldsymbol{\mu}_3' + \boldsymbol{\Sigma}_{33}^B))^{-1}, \quad (36)$$

which is of order $1/N$, while $\text{Cov}(\hat{\gamma}_1, \hat{\gamma}_2)$ is of order $1/Nn$, impling that $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are asymptotically (for N and n tending to infinity) uncorrelated. These results are intuitively in accordance with what one might have expected: The reciprocal of total sample size Nn gives the asymptotic order of magnitude of $\text{Cov}(\hat{\gamma}_1)$, referring to the estimated regression coefficients $\hat{\gamma}_1$ for student variables; the reciprocal of the number of schools N gives the asymptotic order of magnitude for $\text{Cov}(\hat{\gamma}_2)$, referring to the estimated regression coefficients $\hat{\gamma}_2$ for school variables (including cross-level interactions). Approximations 35 and 36 are not of practical value, because their validity usually requires values of n in the hundreds; in practice, one has to use (28)–(31).

Determination of Optimal Sample Sizes

The Approximation Formulas 28–31 can be used to derive sample sizes at the micro- and macrolevel that are approximately optimal for achieving a minimal standard error of one of the regression coefficients and for achieving maximal power for testing the significance of this coefficient, when the total budget is given. Because standard errors and powers can be translated into each other by using a normal approximation for the distribution of the estimated coefficient, we will continue this discussion in terms of standard errors only. The approach given below can also be used, of course, to derive optimal sample sizes for minimizing the budget of a study that must achieve a given standard error or a given power.

We first present a very simple special case: nested designs with covariates only at the school level, not at the pupil level. The random intercept is then the only school-level random coefficient ($K_2 = 1$). The covariance matrix $\boldsymbol{\tau}_2$ is now a scalar τ—namely, the variance component at the school level. It is quite easy to demonstrate that the formulas for this case lead to

$$\text{Cov}(\hat{\gamma}_2) \approx (1/N)(\tau + \sigma^2/n)(\boldsymbol{\mu}_3\boldsymbol{\mu}_3' + \boldsymbol{\Sigma}_{33}^B)^{-1}. \quad (37)$$

If there are also no explanatory variables at the school level, then $\boldsymbol{\Sigma}_{33}^B = 0$, and $\boldsymbol{\mu}_3 = 1$, and the model is reduced to the usual compound symmetry model. Then (37) reduces to Formula 10.8 in Cochran (1977, p. 277) for the variance of an estimator in a two-stage sample, if Cochran's finite population factors are set to unity (sampling from a finite population is not taken into account in the present article).

The standard errors of the estimated regression coefficients are the square roots of the diagonal elements of (37). Formula 37 indicates quite straight-

forwardly that, when the total sample size nN is given and one wishes to have a sampling design that is optimal with respect to estimating school effects (contained in γ_2), it is best to take N as large as possible and n equal to 1. (With such a design we cannot estimate σ^2, of course; in practice, we would prefer $n = 2$.) The assumption that nN is given is not very realistic, however, because sampling extra schools is more costly than sampling extra students within a given sample of schools. Assume that interviewing in an extra school costs c times as much as interviewing an extra student in an already sampled school, so that the total variable costs of the project are proportional to $N(c + n)$; suppose that these costs are fixed. Then we can take $N = k/(c + n)$, where k is a constant and $\text{Cov}(\hat{\gamma}_2)$ depends on n via the factor

$$(n + c)(\tau + \sigma^2/n). \tag{38}$$

This factor reaches its minimum when n is equal to

$$n_{\text{opt}} = \sqrt{c\sigma^2/\tau}. \tag{39}$$

Because in our (independent) sampling model finite population corrections are not taken into account, this result corresponds to Formula 10.26 in Cochran (1977, p. 281).

In the more general case, with within-group regressions which may be heterogeneous, it is impossible to derive analytical expressions for optimal sample sizes. The same principles can then be used in a numerical way as follows. Suppose again that budgetary constraints imply a fixed value, denoted k, for $N(c + n)$, where c is a cost parameter. Substituting $N = k/(c + n)$, the standard errors can be expressed as functions of n. For various values of n, and for using the estimated (or guessed) values of the variance components and other parameters, the standard errors of $\hat{\gamma}$ can be calculated from (15), (22), (25), and (28)–(31). (A computer program to perform these calculations can be obtained from us.) Using suitable trial values for n, one can determine the values for n leading to minimal standard errors for the most relevant components of $\hat{\gamma}$ and draw graphs of these standard errors as functions of n. These graphs can then be used for choosing a sensible value for n. Since (31) implies that the standard errors are (in this approximation) proportional to $\sqrt{k}$, the optimum value found for n will not depend on k. In this approximation, a higher budget or a lower desired standard error will lead to sampling more schools with the same number of students per school. The value for N follows, of course, from $N = k/(c + n^*)$, where n^* is the optimal or otherwise sensible value chosen for n.

In order to illustrate the logic of the formulas for these designs with cross-level effects, we now work out an example that should help to develop some intuition about the behavior of standard errors in nested designs with heterogeneous regressions. Assume we want to assess the power of a two-level design in which we are primarily interested in the effect on achievement

of a school policy for enhancing educational opportunities for pupils from low socioeconomic status (SES) groups. The specification of the design and of the various variances and covariances can proceed as follows; we use estimates for the variance and covariance components that are in line with findings of school effectiveness research. We assume that predictor variables at the student level are SES and IQ; at the school level, the degree to which the investigated policy is implemented is the only predictor variable. SES is measured as a within-school deviation variable. We suppose that IQ has a fixed, and SES a school-dependent, regression coefficient. The multilevel Regression Model 5c therefore reads

$$\text{ACHIEVEMENT}_{ij} = \gamma_{11} \text{IQ}_{ij} + \gamma_{21} + \gamma_{22} \text{POLICY}_j + \gamma_{23} \text{SES}_{ij}$$
$$+ \gamma_{24} \text{SES}_{ij} * \text{POLICY}_j + U_{21j} + \text{SES}_{ij} * U_{22j} + R_{ij}.$$

The order of predictors in Model 1 and in the within-schools Covariance Matrix 10 is therefore: IQ, constant, SES. The between-schools Covariance Matrix 12 refers to IQ, constant, policy. These covariance matrices, and τ_2, can be specified as follows. It is not a restriction to assume that IQ, SES, and achievement are scaled to have unit variance. The variation in IQ scores is 20% between and 80% within schools. The correlation between IQ score and SES within-school deviation score is assumed to be .30. Further, we assume that the variance of the random school main effects is .10—that is, schools account for 10% of achievement variance—and that the regression coefficient of achievement on SES has a between-school variance of .01. For example, if the average regression coefficient on SES is $\gamma_{23} = .316$, corresponding to 10% explained variance for SES, then 90% of the regression coefficients per school are between .15 and .48. Residual variance is thought to be $\sigma^2 = .5$. The policy variable is also standardized to have zero mean and unit variance. The correlation between the policy variable and mean IQ is estimated at $-.3$, resulting in a covariance of $-.13$. The policy is considered to be effective if the regression coefficient γ_{22} is at least .10, implying that $.10^2/.10 = 10\%$ of the between-school variation in achievement can be attributed to the main effect of the policy implementation. The residual between-school variation in achievement is then $\tau_{11} = .09$. But the policy variable should also affect the within-school relation between achievement and SES. We assumed the variance of the regression coefficients of achievement on SES to be 0.01. The policy is considered to be effective in this respect if it accounts for 25% of this variation, corresponding to $\gamma_{24} = .05$. The residual slope variance is then $\tau_{22} = .0075$. The residual covariance between the random slope and the random intercept is estimated at $-.01$, which amounts to a correlation of $-.39$. (This coefficient is rather hard to specify on the basis of expert knowledge; fortunately, the standard errors of the regression coefficients in this design are not very sensitive to

it.) All the above specifications yield the parameter values

$$\sigma^2 = .5, \quad \boldsymbol{\tau}_2 = \begin{pmatrix} .09 & -.01 \\ -.01 & .0075 \end{pmatrix}, \quad \boldsymbol{\Sigma}^W = \begin{pmatrix} .8 & 0 & .3 \\ 0 & 0 & 0 \\ .3 & 0 & 1.0 \end{pmatrix},$$

$$\boldsymbol{\Sigma}^B = \begin{pmatrix} .2 & 0 & -.13 \\ 0 & 0 & 0 \\ -.13 & 0 & 1.0 \end{pmatrix}, \quad \boldsymbol{\mu}_1 = 0, \quad \boldsymbol{\mu}_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The standard errors of interest are of course those for the school-level predictor POLICY and for the cross-level interaction effect of POLICY * SES. Given these parameters, using a one-tailed significance level of .05, and setting the desired power at .50, these standard errors should not exceed .10/1.645 = .0608 and .05/1.645 = 0.0304, respectively. We specify c and k in $N(c + n) = k$ as 5 and 1000, respectively, and we are now ready to draw the graphs, which give sample sizes and standard errors as functions of n (the number of students per school). The number N of schools in the sample is taken as the largest integer satisfying $N(c + n) \leq k$. First of all, we should be aware of the fact that this budgetary constraint implies that total sample size Nn depends heavily on n. Figure 1 shows the number of schools and the total sample size. As the number of schools slowly decreases, the total sample size tends first rapidly and then rather erratically towards its maximum value ($Nn = 995$). The nonmonotonicity of Nn is a consequence of the integer nature of N. Bearing in mind this relation between n and N, we now will study the behavior of the approximate standard errors calculated from Formulas 28–31. Figure 2 presents the approximate standard errors for the estimated regression coefficients of the school variable POLICY and of the cross-level interaction variable SES * POLICY as a function of n. Because we have assumed in the derivation of our formulas that n is rather large, we will not pay attention to the left part of the graph ($n < 11$). We also have assumed that N is rather large, so the right part of the graph ($n > 91$, corresponding to $N < 11$) should be treated with caution: It gives an underestimate of the standard errors, since $\boldsymbol{\tau}_2$ is assumed to be known. The first conclusion that can be drawn from the figure is that the standard error of the regression coefficient for the school variable POLICY increases as the number of schools declines, although the total sample size increases. This is in accordance with intuition. According to the graph, we have to choose $n \leq 28$ and $N \geq 30$ to keep the standard error below the desired level (.0608). The graph further shows that, in the resulting range $11 \leq n \leq 28$, $30 \leq N \leq 62$, the standard error of the cross-level interaction variable SES * POLICY, also is below the desired level (0.0304). Maximum power for the test of the cross-level interaction effect POLICY * SES can be
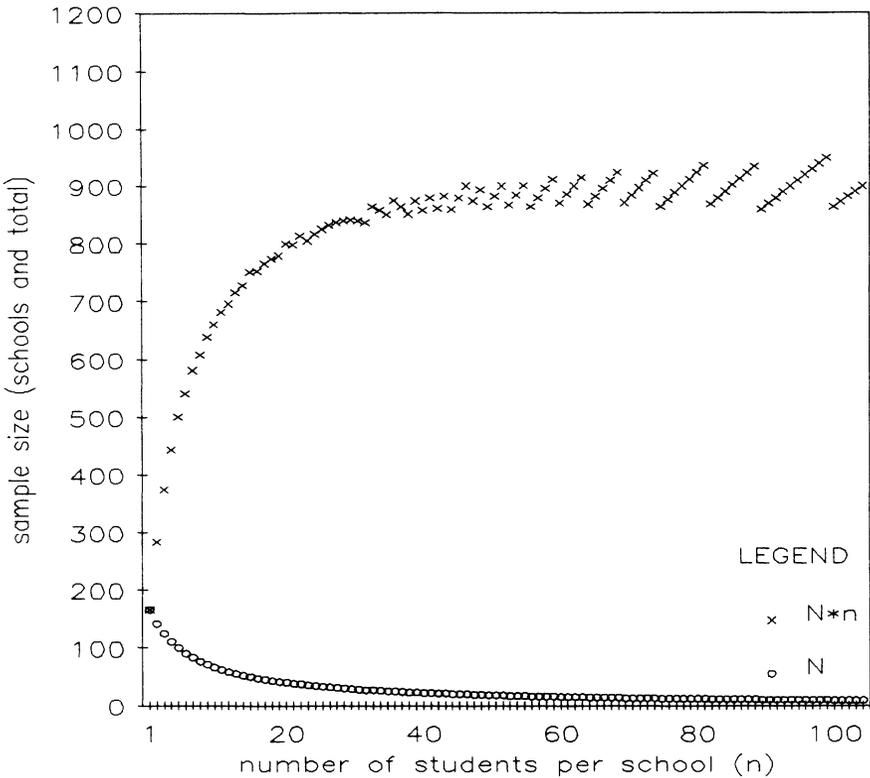
FIGURE 1. *An example of the number of schools sampled and the total sample size as a function of the number of students sampled per school, if there is extra cost associated with sampling a new school*

obtained by choosing $N = 42$ and $n = 19$, whereas maximum power for the test of the school variable POLICY can be obtained by choosing N as large as possible (within the limits of the approximations, this is $N = 62$ with $n = 11$). So in practice, the power in this design can be obtained for both effects that demand interest by taking one class of pupils per school ($n = 25$) and $N = 33$; this is preferable to—for example, two classes per school with $N = 18$. It might be argued, however, that these results are more strongly determined by the restrictions imposed on N and n by $N(c + n) = k$ than by the properties of the standard errors themselves. To show that this is not the case, we take $c = 0$ (i.e., there is no extra cost associated with sampling a new school), in which case $Nn \leq k$. Figure 3 shows for this case the number of schools in the sample and the total sample size as a function of the number of students per school. Since N and n are integers, the total sample size
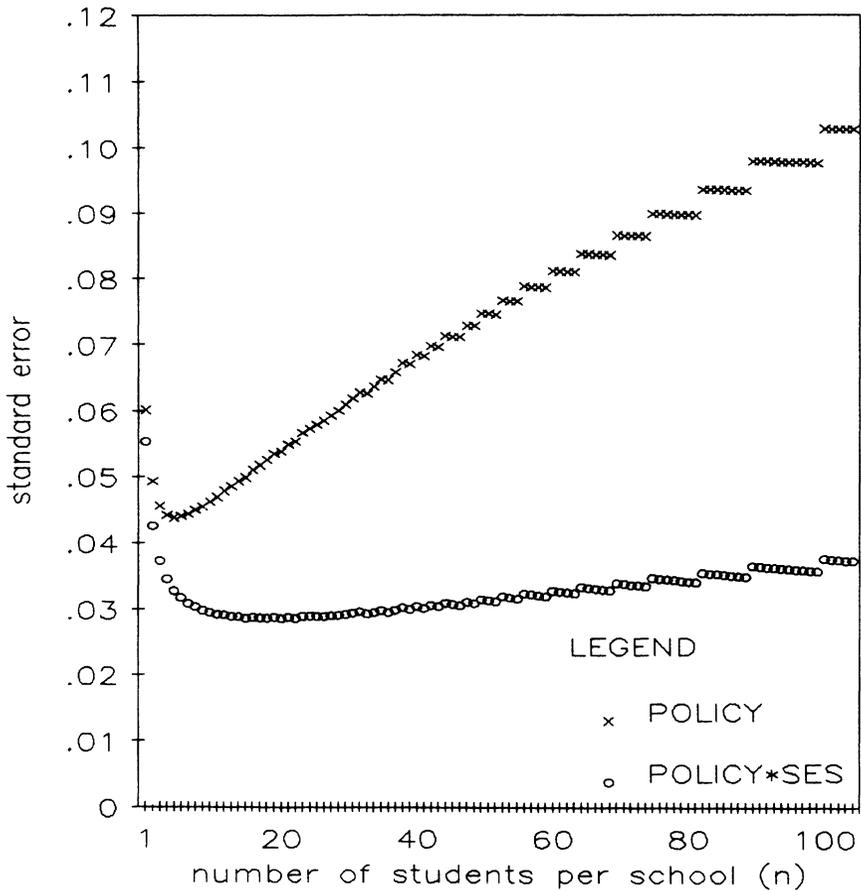
FIGURE 2. *An example of standard errors for regression coefficients of school-level effects and cross-level interaction effects as a function of the number of students sampled per school, if there is extra cost associated with sampling a new school*

fluctuates in this range for n between 910 and 1000. What are the consequences for the standard errors that demand interest once the total sample size is chosen to be (more or less) constant? The results are shown in Figure 4. With Nn constant, it is once again shown that the standard error for the regression coefficient of the school variable deteriorates severely as N decreases. But in this case, the standard error for the regression coefficient of the cross-level effect also is negatively affected by small values for N, and this leads us to the conclusion that N should be taken as large as possible, provided that one remains within the range of validity of the approximations.
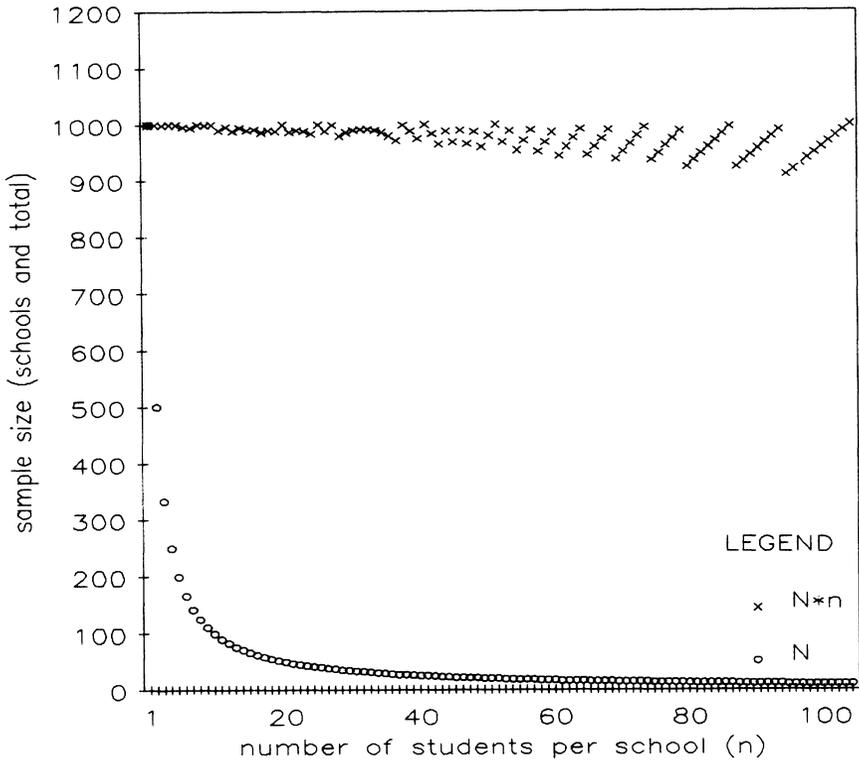
FIGURE 3. *An example of the number of schools sampled and the total sample size as a function of the number of students sampled per school, if there is no extra cost associated with sampling a new school*

Comparison With Standard Errors From the
Inverted Information Matrix

The approximate standard errors given by (28)–(31) can be compared with the standard errors estimated from the inverted, observed information matrix, if one has available a complete data set. We shall present some results where the information matrix is calculated by Longford's (1988) VARCL program. This gives an impression of the precision of our approximations. In the first example, using a random sample from the data collected by Stoel (1982), we have a two-level model with $N = 80$ and $n = 60$. The dependent variable is educational well-being. At the student level, we have gender as a predictor with a fixed regression coefficient and age as a predictor with a variable regression coefficient. Both the random intercept and the random slope are regressed on three group-level variables: the
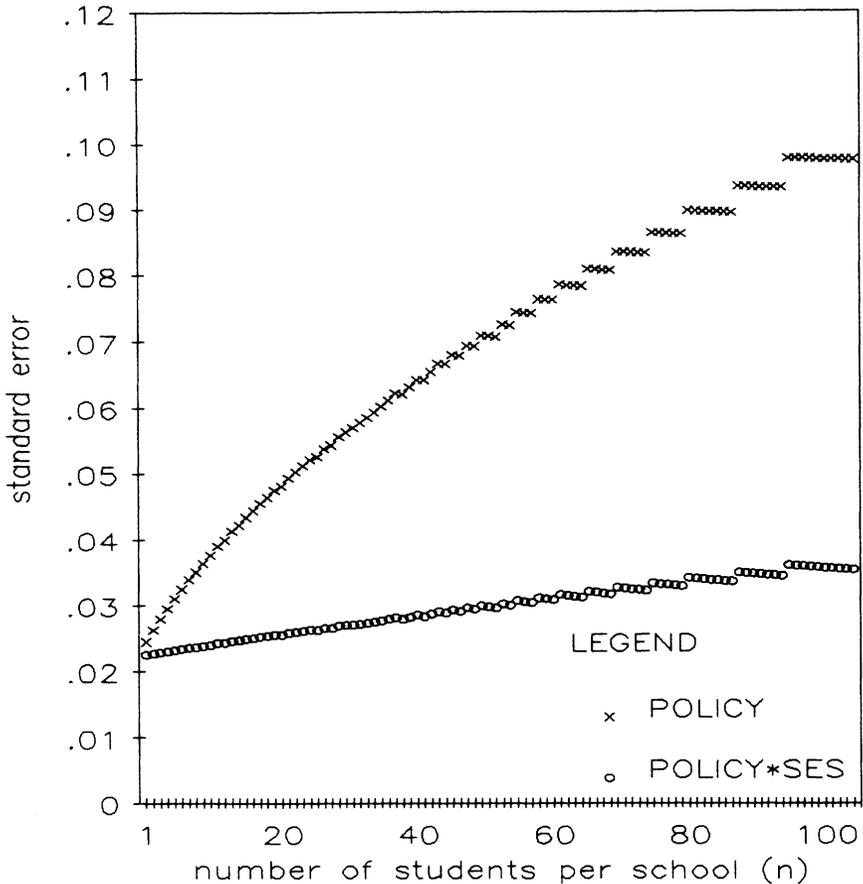
FIGURE 4. *An example of standard errors for regression coefficients of school-level effects and cross-level interaction effects as a function of the number of students sampled per school, if there is no extra cost associated with sampling a new school*

group mean of age (age at student level is centered around these group means), school climate, and percentage of female teachers within the school. (The similarity of several of the standard errors results from the standardization used.) From Table 1, we can conclude that the discrepancy between the VARCL estimates of the standard errors and our approximations is at most 1%. In this example, n was rather large. In school-effects or classroom-effects research, however, data sets often contain only one class of students. For this reason, we give a second example where $N = 40$ and $n = 20$. We use the data collected by Brandsma and Knuver (1988) and Reezigt and Weide (1989). The dependent variable in this case is arithmetic

TABLE 1
Standard errors of regression coefficients (N = 80, n = 40); educational well-being in secondary schools

Type of variable	SE estimate Snijders/Bosker	SE estimate VARCL	Name of variable
1. Fixed student covariate	.01530	.01547	Sex
2. Grand mean	.03116	.03115	
3. First school variable	.03117	.03136	Climate
4. Second school variable	.03149	.03169	Females
5. Third school variable	.03117	.03136	Mean (6)
6. Random student covariate	.01717	.01735	Age
7. (3) * (6) interaction	.01716	.01720	
8. (4) * (6) interaction	.01717	.01729	
9. (5) * (6) interaction	.01879	.01889	

achievement, being predicted by socioeconomic status (student-level variable with fixed regression coefficient) and IQ score (student-level variable with random regression coefficient). Both the random intercept and slope are predicted by two group-level variables: the group mean of the IQ scores (the microlevel IQ scores are centered around these group means) and the variance in the IQ scores within the groups as a measure of the within-classroom heterogeneity of ability levels. In this second example, the discrepancies between the two series of estimated standard errors are small again, 4% at most.

Discussion

The examples given provide support for the suitability of Approximation Formulas 28–31 for the purpose of choosing optimal sample sizes at either

TABLE 2
Standard errors of regression coefficients (N = 40, n = 20); arithmetic achievement in primary schools

Type of variable	SE estimate Snijders/Bosker	SE estimate VARCL	Name of variable
1. Fixed student covariate	.02554	.02971	SES
2. Grand mean	.07168	.07332	
3. First school variable	.47006	.48762	Heterogeneity
4. Second school variable	.22951	.23809	Mean (5)
5. Random student covariate	.02916	.02810	IQ
6. (3) * (5) interaction	.19145	.18425	
7. (4) * (5) interaction	.09348	.09130	

level. Let us recall, however, that the researcher needs to have some knowledge beforehand about relevant population parameters: variance components and, in addition, means and covariance matrices (within-school and between-school) of predictor variables. This is usually quite a large number of parameters assumed to be known. The researcher may indeed follow the suggestion by Goldstein (1987) that we quoted earlier: A pilot study may give an impression of these parameter values. The special case leading to Formulas 38–39, as well as some studies conducted by the authors, suggests that for some of the parameters the standard errors may remain close to the minimum for a large range of n values around the optimum. This issue is now being further investigated.

One should bear in mind, however, that the optimality of the sample sizes as derived here is defined here only with respect to the standard errors of the regression coefficients; researchers can also be particularly interested in other parameters—for example, the standard errors of the variance components or the posterior means for some specific schools. A further restriction is imposed by the fact that we assumed n and N to be large in the derivation of our formulas. Our experience to date suggests that we should take $n > 10$ at least. Research to determine the region of validity of the approximations is now in progress. For $2 < n < 11$, better approximations might be needed. As a consequence, the formulas may be of little use for some of the applications mentioned in the introductory section (e.g., repeated measurement designs), where the number of lower level units usually does not exceed $n = 10$. For the special case of school effectiveness research, however, where the researcher has to choose N and n, with n often being at least one class of pupils ($n > 25$), the formulas are clearly of use as has been demonstrated in our examples.

Further research has to deal with two points. As mentioned above, more precise approximations should be derived that are valid for small values of n and N. Secondly, we derived the formulas in order to design two-level studies so as to ensure a satisfactory power of the tests for the regression coefficients. The next step is to use the formulas for this purpose for a number of representative sets of parameter values and try to find out if some general rules can be deduced about optimal sample sizes for some frequently used two-level designs.

References

Aitkin, M., & Longford, N. T. (1986). Statistical modeling issues in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society, Series A, 144*, 419–461.

Bosker, R. J., Guldemond, H., Hofman, R. H., & Hofman, W. H. A. (1989). *Kwaliteit in het voortgezet onderwijs* [Quality in secondary education]. Groningen, The Netherlands: RION.

Brandsma, H. P., & Knuver, J. W. M. (1988). Organisatorische verschillen tussen basisscholen en hun effect op leerlingprestaties [Organizational differences between primary schools and their effects on pupil achievement]. *Tijdschrift voor Onderwijsresearch*, *13*, 201–212.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*, 147–158.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., Congdon, R. T., & Seltzer, M. (1988). *An introduction to HLM: User's guide, version 2.0* [Computer program]. Chicago: University of Chicago.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

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

Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, *74*, 311–319.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, *73*, 43–56.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Griffin.

Kreft, G. G. (1987). *Models and methods for the measurement of school effects*. Unpublished doctoral dissertation, University of Amsterdam, The Netherlands.

Longford, N. T. (1987). A fast scoring algorithm for variance components analysis with hierarchically nested random effects. *Biometrika*, *74*, 817–827.

Longford, N. T. (1988). *VARCL manual*. Princeton, NJ: Educational Testing Service.

Prosser, R., Rasbash, J., & Goldstein, H. (1991). *ML3 software for three-level analysis, user's guide*. London: Institute of Education.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

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

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

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, *10*, 75–98.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*, 1–17.

Reezigt, G. J., & Weide, M. G. (1989, January). *The effects of adaptive instruction in Dutch primary education*. Paper presented at the Second International Conference on School Effectiveness, Rotterdam, The Netherlands.

Stoel, W. G. R. (1982). De grootte van scholen voor voortgezet onderwijs en het welbevinden van de leerlingen [The size of schools for secondary education and the well-being of the pupils]. *Pedagogische Studiën*, *59*, 493–506.

Willms, J. D. (1987). Differences between Scottish Education Authorities in their Examination Attainment. *Oxford Review of Education*, *13*, 211–232.

Authors

TOM A. B. SNIJDERS is Professor, Department of Statistics and Measurement Theory, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands. He specializes in multilevel methods and social network analysis.

ROEL J. BOSKER is Associate Professor, Department of Education, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. He specializes in multilevel methods and evaluation research.

You have printed the following article:

Standard Errors and Sample Sizes for Two-Level Research
Tom A. B. Snijders; Roel J. Bosker
Journal of Educational Statistics, Vol. 18, No. 3. (Autumn, 1993), pp. 237-259.
Stable URL:
http://links.jstor.org/sici?sici=0362-9791%28199323%2918%3A3%3C237%3ASEASSF%3E2.0.CO%3B2-R

This article references the following linked citations. If you are trying to access articles from an off-campus location, you may be required to first logon via your library web site to access JSTOR. Please visit your library's website or contact a librarian to learn about options for remote access to JSTOR.

References

Statistical Modelling Issues in School Effectiveness Studies
M. Aitkin; N. Longford
Journal of the Royal Statistical Society. Series A (General), Vol. 149, No. 1. (1986), pp. 1-43.
Stable URL:
http://links.jstor.org/sici?sici=0035-9238%281986%29149%3A1%3C1%3ASMIISE%3E2.0.CO%3B2-6

Random Coefficient Models for Multilevel Analysis
Jan de Leeuw; Ita Kreft
Journal of Educational Statistics, Vol. 11, No. 1. (Spring, 1986), pp. 57-85.
Stable URL:
http://links.jstor.org/sici?sici=0362-9791%28198621%2911%3A1%3C57%3ARCMFMA%3E2.0.CO%3B2-Z

Data Analysis Using Stein's Estimator and its Generalizations
Bradley Efron; Carl Morris
Journal of the American Statistical Association, Vol. 70, No. 350. (Jun., 1975), pp. 311-319.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197506%2970%3A350%3C311%3ADAUSEA%3E2.0.CO%3B2-1

Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares
H. Goldstein
Biometrika, Vol. 73, No. 1. (Apr., 1986), pp. 43-56.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28198604%2973%3A1%3C43%3AMMLMAU%3E2.0.CO%3B2-5

LINKED CITATIONS

- Page 2 of 2 -

A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models with Nested Random Effects

Nicholas T. Longford
Biometrika, Vol. 74, No. 4. (Dec., 1987), pp. 817-827.
Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28198712%2974%3A4%3C817%3AAFSAFM%3E2.0.CO%3B2-A

Educational Applications of Hierarchical Linear Models: A Review

Stephen W. Raudenbush
Journal of Educational Statistics, Vol. 13, No. 2. (Summer, 1988), pp. 85-116.
Stable URL:
http://links.jstor.org/sici?sici=0362-9791%28198822%2913%3A2%3C85%3AEAOHLM%3E2.0.CO%3B2-T

Empirical Bayes Meta-Analysis

Stephen W. Raudenbush; Anthony S. Bryk
Journal of Educational Statistics, Vol. 10, No. 2. (Summer, 1985), pp. 75-98.
Stable URL:
http://links.jstor.org/sici?sici=0362-9791%28198522%2910%3A2%3C75%3AEBM%3E2.0.CO%3B2-Z

A Hierarchical Model for Studying School Effects

Stephen Raudenbush; Anthony S. Bryk
Sociology of Education, Vol. 59, No. 1. (Jan., 1986), pp. 1-17.
Stable URL:
http://links.jstor.org/sici?sici=0038-0407%28198601%2959%3A1%3C1%3AAHMFSS%3E2.0.CO%3B2-1

Differences between Scottish Education Authorities in Their Examination Attainment

J. Douglas Willms
Oxford Review of Education, Vol. 13, No. 2. (1987), pp. 211-232.
Stable URL:
http://links.jstor.org/sici?sici=0305-4985%281987%2913%3A2%3C211%3ADBSEAI%3E2.0.CO%3B2-0