The concept of explained proportion of variance or modeled proportion of variance is reviewed in the situation of the random effects hierarchical two-level model. It is argued that the proportional reduction in (estimated) variance components is not an attractive parameter to represent the joint importance of the explanatory (independent) variables for modeling the dependent variable. It is preferable instead to work with the proportional reduction in mean squared prediction error for predicting individual values (for the modeled variance at level 1) and the proportional reduction in mean squared prediction error for predicting group averages (for the modeled variance at level 2). It is shown that when predictors are added, the proportion of modeled variance defined in this way cannot go down in the population if the model is correctly specified, but can go down in a sample; the latter situation then points to the possibility of misspecification. This provides a diagnostic means for identifying misspecification.

Modeled Variance in Two-Level Models

TOM A. B. SNIJDERS
University of Groningen, Netherlands
ROEL J. BOSKER
University of Twente, Netherlands

The concept of explained variance is well-known in multiple regression analysis: it gives an answer to the question, how much of the variability of the dependent variable is accounted for by the linear regression on the explanatory variables. This concept, however, is somewhat complicated in the hierarchical random effects model, which is so fruitful in multilevel research (see, among many others, Aitkin and Longford 1986; Raudenbush and Bryk 1986; Goldstein 1987; Bryk and Raudenbush 1992). One way to approach the concept of explained variance is to transfer its customary treatment, well-known from multiple linear regression, in a straightforward way to multilevel models: treat proportional reductions in the estimated variance components as analogues of R^2 values. There are several variance components in hierarchical random effect models (e.g., in a random intercept two-level model, there is a random effect

AUTHORS' NOTE: *We would like to thank Stephen W. Raudenbush for his comments on an earlier draft of this article.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 22, No. 3, February 1994 342-363

for either level). So this approach to defining explained variance leads to several R^2 values, one for each variance component. Raudenbush and Bryk (1986, p. 9) and Bryk and Raudenbush (1992, pp. 65, 70) follow this approach; it is also proposed by Prosser, Rasbash, and Goldstein (1991, p. 13). Some practitioners, however, have run into problems with this definition of R^2: it sometimes happens that adding explanatory variables increases rather than decreases some of the variance components. This leads, under the stated definition of R^2, to negative values for the contribution of this explanatory variable to R^2 and sometimes even to negative values for R^2. Examples will be provided in the sequel. Negative values for R^2 clearly are undesirable and do not correspond with its intuitive interpretation.

This article explores the concept of explained variance for hierarchical random effects two-level models. As we dislike the term explained variance because of the often ill-founded connotation of causal explanation, we use the more neutral term of *modeled variance*. It is well known that, in multiple regression, the concept of R^2 makes sense only in the case that the predictor variables are random variables (i.e., the design is observational rather than experimental). The reason is that the variance accounted for by the regression model depends not only on the regression coefficients and the residual variance, but also on the variances and covariances of the predictor variables. In other words, although the regression model is conditional on the values of the predictor variables, the value of R^2 is population dependent. Correspondingly, we assume throughout the article that the predictor variables are random variables over some population, so that we can take expectations with respect to them.

The article consists of four parts. In the first place, we desire to understand better why variance components might go up when predictor variables (used in this article as a synonym of explanatory variables; we shall not use the term independent variables) are added. To this end, we consider some simple two-level models where explicit expressions are available for estimated variance components, and we indicate circumstances under which estimated variance components should indeed go up in the course of a forward model selection process. Second, we propose for two-level random effects models a definition for R^2, or proportion of modeled variance, at the lower level; and similarly for the higher level. Third, we investigate whether the so

defined R^2-type parameters can go down when predictors are added to the model. As the last part of the article, we give formulae extending these definitions to random slope models, which have more than two variance components.

HOW IS RESIDUAL VARIANCE AFFECTED BY PREDICTORS?

Let us first try to get an understanding of the way in which predictor variables influence the estimated variance components at the two levels by considering a simple model with equal group sizes. We consider a two-level model with a random group main effect but no other random effects:

$$Y_{ij} = X_{ij}\beta + U_{0j} + E_{ij}, \tag{1}$$

where i denotes the level-1 unit (also called individual i) and j the level-2 unit (also designated as group j), X_{ij} is a row vector of predictor variables, and β is a column vector of regression coefficients; U_{0j} is the random effect of group j, having mean 0 and variance τ^2, whereas E_{ij} is the residual at level 1 having mean 0 and variance σ^2. The random effects U_{0j} and E_{ij} are assumed to be independent. A normality assumption will be made for U_{0j} and E_{ij} when discussing parameter estimation, but not for the definition of analogues of R^2. The predictor variables X_{ij} are random; they are assumed to be independent of U_{0j} and E_{ij}; it is further assumed that the distribution of the vector X_{ij} is the same for all i and j, and that for identical group sizes, the group means $\overline{X}_{.j}$ have the same distribution for all j. No specific distributional forms such as normality are assumed for the X_{ij}. The number of groups (level-2 units) is N; the group size is equal to n for all $j = 1, \ldots, N$.

The assumptions with respect to X_{ij} are so general that some of the predictor variables (components of the vector X_{ij}) might be variables at the level of the individuals whereas others might be group variables (i.e., not dependent on i). Also, the X_{ij} do not need to be independent within the groups. As an example, it is allowed that, for a given substantive explanatory variable, the within-group deviation score (with group means necessarily equal to 0) as well as the group mean

are contained in the vector X_{ij}. Thus it is possible to specify model (1) in such a way that within-group regression coefficients and between-group regression coefficients are allowed to be different, but also to specify the model in such a way that within-group and between-group regressions are required to be the same for some of the substantive explanatory variables.

We first consider unrestricted maximum likelihood estimation. Under the assumption of normality for the random residuals U_{0j} and E_{ij}, the log likelihood is given by

$$-N(n-1)\log(\sigma) - \tfrac{1}{2}N\log(\sigma^2 + n\tau^2) - \tfrac{1}{2}\sigma^{-2}SS_w(\beta) - \tfrac{1}{2}(\sigma^2 + n\tau^2)^{-1}SS_B(\beta), \quad (2)$$

where $SS_w(\beta)$ and $SS_B(\beta)$ denote the residual within- and between-groups sums of squares:

$$SS_w(\beta) = \tilde{\sum}_{i,j} \{Y_{ij} - \overline{Y}_{.j} - (X_{ij} - \overline{X}_{.j})\beta\}^2, \quad (3)$$

$$SS_B(\beta) = \tilde{\sum}_{j} n(\overline{Y}_{.j} - \overline{X}_{.j}\beta)^2. \quad (4)$$

To obtain the unrestricted (full-information) maximum likelihood (ML) estimate, the likelihood is maximized as a function of β, σ^2, and τ^2. It is seen from (2) that with regard to the vector β of regression coefficients, this means that a weighted sum of the within-group and between-group residual sums of squares is minimized. This minimization can be considered as two separate minimizations, one for the within-group and one for the between-group sums of squares, if and only if for every predictor variable that has within- as well as between-group variability, the within-group regression coefficient is allowed to be different from the between-group regression coefficient; in other words, when for all variables with a positive between-group variance, the group means are also included as predictor variables in the regression model. It follows from differentiation with respect to σ^2 and τ^2 that the ML estimates for the variance components are given by

$$\hat{\sigma}^2 = SS_w(\hat{\beta})/(N(n-1)) + \min\{0,[SS_B(\hat{\beta}) - SS_w(\hat{\beta})/(n-1)]/Nn\}, \quad (5)$$

$$\hat{\tau}^2 = \max\{0,[SS_B(\hat{\beta})/N - \hat{\sigma}^2]/n\} = \max\{0,[SS_B(\hat{\beta}) - SS_w(\hat{\beta})/(n-1)]/Nn\}. \quad (6)$$

These expressions are well-known; for the compound symmetry model (i.e., the case without the X_{ij}s), they were derived by Herbach (1959) and are given, for example, in Searle (1971, p. 419).

What will happen to these estimates when a researcher has fitted a model of this kind and then adds another predictor? Most predictors have both within-group and between-group variability. The two extreme possibilities, where this is not so, are a group-level variable and a within-group deviation variable. By definition, the former has zero within- and the latter zero between-group variance. Let us consider these two extreme possibilities in turn. To simplify the discussion, we suppose, for the time being, that within-group and between-group regression coefficients for all level-1 predictor variables are allowed to be different, so that the within-groups and the between-groups residual sums of squares, SS_w and SS_B, vary independently as functions of the regression coefficients β. In other words, for every predictor variable that has a positive between-group as well as within-group variance, its group mean is included in the list of predictor variables.

When the main effect of a group variable (i.e., a level-2 variable) is added, the term $(X_{ij} - \overline{X}_{.j})$ in SS_w is not affected. This means that SS_w, and hence $\hat{\sigma}^2$, does not change. Unless the new predictor is completely collinear with the predictors already used, it will be possible, however, to obtain a smaller value for the between-group residual variance SS_B, and hence the estimate, $\hat{\tau}^2$, for the group-level variance component will be diminished.

Now suppose that a within-group deviation variable is added: that is, a variable with within-group means $\overline{X}_{.j}$ equal to 0. Then the between-group sum of squares is not affected, and only the within-group sum of squares SS_w will become smaller. This means, as can be seen from equations (5) and (6), that $\hat{\sigma}^2$ will become smaller *while* $\hat{\tau}^2$ *becomes larger*. Another way to explain this is as follows. Unexplained within-group variability is represented completely by σ^2. Unexplained between-group variability, however, is constituted by variation ascribed to τ^2 as well as variation ascribed to σ^2:

$$\text{var}(\overline{Y}_{.j}|X_{1j}, \ldots, X_{nj}) = \tau + \frac{\sigma^2}{n}. \tag{7}$$

Adding a predictor that models a part of the within-group variability must decrease the estimate of σ^2; if this predictor does not model part of between-group variability then, because unexplained between-group variability remains the same, the decrease of σ^2 must be balanced by an increase of the estimate of τ^2.

The preceding discussion was formulated for the unrestricted maximum likelihood estimator. However, the explanation based on (7) for the fact that adding a within-group deviation variable must increase the estimate of τ^2 is a direct consequence of the model assumptions and does not depend on the estimator employed. Other estimators, such as the restricted maximum likelihood estimator (REML, also called MLR for maximum likelihood-random; see Dempster, Rubin, and Tsutakawa 1981) present a similar behavior.

In designs with differing group sizes n_j, the formulas are more complicated and explicit expressions for the ML estimates of σ^2 and τ^2 cannot be given. Also in cases where, for some variables, within-group regression coefficients are restricted to be identical to between-group regression coefficients, the situation is more complicated. However, in both of these situations, provided that no equality assumptions between within-group and between-group regression coefficients are made that are strongly at odds with the data, the general conclusions of the preceding analysis still are valid to the effect that adding a group-level variable will decrease the estimate for τ^2 but hardly affect the estimate for σ^2, whereas adding a within-group deviation variable will decrease the estimate for σ^2 and at the same time increase the estimate for τ^2.

These theoretical insights are confirmed by practical experience. In data analysis, it is usual to have explanatory variables that have a positive within- as well as a positive between-group variance; it is also common to have different group sizes and, for some variables, the constraint that within-group and between-group regression coefficients are identical (i.e., the group mean is not included in the list of explanatory variables). In this general situation it is not uncommon to see estimated variance components going up when explanatory variables are added. Especially if the between-group variance of an explanatory variable is small compared to within-group variance

divided by group size, it is not unusual to observe an increased estimate for τ^2 when the variable is added to the model.

To illustrate this, data from a study by Vermeulen and Bosker (1992) on the effects of part-time teaching in primary schools is used. The dependent variable is an arithmetic test score; the sample consists of 718 third-grade pupils in 42 schools. An intelligence test score is used as predictor variable. Group sizes range from 1 to 33 with an average of 20. We balanced this design to 33 schools with 10 pupils in each school by deleting schools with less than 10 pupils from the sample and randomly sampling 10 pupils from each of the remaining schools. In Table 1, the results of the analyses (calculated using ML3 by Prosser et al. 1991) are summarized, both for the balanced and for the entire data set. From Table 1 we see that in the balanced as well as in the unbalanced case, $\hat{\tau}^2$ increases as a within-group deviation variable is added as an explanatory variable to the model. Furthermore, the estimates of the variance components in the balanced case behave exactly as predicted. In the unbalanced case, $\hat{\sigma}^2$ increases slightly when adding the group variable to the model. When R^2 is defined as the proportional reduction in residual variance, as discussed earlier, then R^2 on the group level is negative for model C, while for the entire data set R^2 on the pupil level is negative for model B. Estimating σ^2 and τ^2 using restricted maximum likelihood results in slightly different parameter estimates. The pattern, however, remains the same.

We would like to stress that the possibility, discussed above, of an increase of residual variance estimates when predictor variables are added is not a consequence of misspecification. The model without and the model with the extra predictor variable could both be valid statistical models for the observations at hand (although the latter model would be better in the sense of having a greater explanatory power). To illustrate this, we now turn from the data and the estimators to the population. Consider the following two different models, both of which we assume to be valid for one given population (X_{ij} is random and independent of U_{0j} and E_{ij}, and all group sizes are equal to n):

$$Y_{ij} = \beta_0 + X_{ij}\beta + U_{0j} + E_{ij}, \tag{8}$$

$$Y_{ij} = \beta_0 + \tilde{U}_{0j} + \tilde{E}_{ij} \tag{9}$$

TABLE 1: Modeling Variance by Within- and Between-Group Variables

	$\hat{\sigma}^2$	$\hat{\tau}^2$
I. Balanced Design		
A. $Y_{ij} = \beta_0 + U_{0j} + E_{ij}$	8.694	2.271
B. $Y_{ij} = \beta_0 + \beta_1 \overline{X}_{.j} + U_{0j} + E_{ij}$	8.694	0.819
C. $Y_{ij} = \beta_0 + \beta_2(X_{ij} - \overline{X}_{.j}) + U_{0j} + E_{ij}$	6.973	2.443
II. Unbalanced Design		
A. $Y_{ij} = \beta_0 + U_{0j} + E_{ij}$	7.653	2.798
B. $Y_{ij} = \beta_0 + \beta_1 \overline{X}_{.j} + U_{0j} + E_{ij}$	7.685	2.038
C. $Y_{ij} = \beta_0 + \beta_2(X_{ij} - \overline{X}_{.j}) + U_{0j} + E_{ij}$	6.668	2.891

where X_{ij} is a vector of within-group deviation variables: $\overline{X}_{.j} = 0$ for all j, and where the variances of the random residual variables are denoted τ^2, σ^2, $\tilde{\tau}^2$, and $\tilde{\sigma}^2$, respectively. The assumption that (8) and (9) both are valid models for the same joint distribution of the variables X_{ij} implies that the residuals, $\tilde{U}_{0j}$ and $\tilde{E}_{ij}$ in model (9), have to incorporate the effects modeled in (8) by $X_{ij}\beta$. The variance components τ^2, σ^2, $\tilde{\tau}^2$, and $\tilde{\sigma}^2$ are identified by the variances and the within-group covariances. Recall that the X_{ij} are random vectors. Because of the assumption that $\overline{X}_{.j} = 0$ for all groups j, it cannot be assumed that the X_{ij} are independent within groups. Instead, we assume that their distribution is permutation-symmetric within each group (i.e., the joint distribution of X_{1j} to X_{nj} is invariant under permutations of the indexes 1 to n), and independent and homoscedastic across groups. Denoting the covariance matrix of X_{ij} by Σ_X, it can be shown that the cross-covariance matrix of X_{ij} and $X_{i'j}$ for $i \neq i'$ (i.e., the covariance matrix between the predictor variables for two different pupils in the same school) is equal to $-(n-1)^{-1}\Sigma_X$. This implies that if (8) and (9) both are valid statistical models for the joint distribution of the Y_{ij}, then

$$\text{var}(Y_{ij}) = \beta'\Sigma_X\beta + \tau^2 + \sigma^2 = \tilde{\tau}^2 + \tilde{\sigma}^2$$

$$\text{cov}(Y_{ij}, Y_{i'j}) = -(n-1)^{-1}\beta'\Sigma_X\beta + \tau^2 = \tilde{\tau}^2,$$

which implies that the variance components of model (8) are related to those in model (9) by

$$\tau^2 = \tilde{\tau}^2 + (n-1)^{-1}\beta'\Sigma_X\beta,$$

$$\sigma^2 = \tilde{\sigma}^2 - \{n/(n-1)\}\beta'\Sigma_X\beta.$$

It is seen that the within-group deviation variable as a predictor contributes a negative amount to the residual variance at level 1, which is balanced by an extra positive contribution to the residual variance at level 2. It is also seen that model (8) can only be represented in the form (9) if the intercept variance is large enough:

$$\tau^2 \geq (n-1)^{-1}\beta'\Sigma_X\beta.$$

It can be concluded that the correspondence between the variance components τ^2, σ^2 and $\tilde{\tau}^2$, $\tilde{\sigma}^2$ in model (8) and (9), respectively, contains the design variable n and is not very appealing conceptually: the predictors X_{ij} do not model any variation at the level of group means, so the models are formally identical at the aggregated level of group means, but nevertheless the variance components τ^2 and $\tilde{\tau}^2$ at the group level are different. This demonstrates that the partitioning of variance between the random level-1 residual E_{ij} and the level-2 intercept U_{0j} can depend in a conceptually rather undesirable way on the predictors. This is an argument against using the intercept variance τ^2 as the fundamental parameter for comparison of level-2 modeled variance for different models. In the next section we propose, therefore, a different parameter.

MODELED VARIANCE AT EITHER LEVEL

In multiple linear regression, the customary R^2 measure can be introduced in several ways—for instance, as the maximal squared correlation coefficient between the dependent variable and some linear combination of the predictor variables or as the proportional reduction in the residual variance parameter due to the joint predictor variables. The latter description is convenient as a utilitarian definition because it is easily calculated from computer output; from a conceptual point of view, however, this definition is not the most appealing one. In the context of multilevel modeling, we need some reflection about what constitutes a suitable definition because it is not a priori clear that the definitions used in the multiple linear regression framework can be automatically carried over to the multilevel framework

and still make sense; nor is it certain that the various possible definitions will coincide here, as they do for multiple linear regression.

In our view, the most appealing principle to define measures of modeled (or explained) variation is the principle of proportional reduction of prediction error. This is one of the definitions of R^2 in multiple linear regression and can be described as follows: A population of (X_i, Y_i) values is given, with a known joint probability distribution; β is the value for the vector v for which the expected squared error $E(Y_i - X_i v)^2$ is minimal. If the value of X_i is unknown, then the best predictor for Y_i is its expectation E(Y), with mean squared prediction error $\text{var}(Y_i)$; if X_i is given, the linear predictor of Y_i with minimum squared error is the regression value $X_i\beta$, with mean squared prediction error $E(Y_i - X_i\beta)^2$. The proportional reduction of the mean squared error of prediction is defined as

$$\frac{\text{var}(Y_i) - \text{var}(Y_i - X_i\beta)}{\text{var}(Y_i)} = 1 - \frac{\text{var}(Y_i - X_i\beta)}{\text{var}(Y_i)};$$

this formula expresses one of the equivalent ways to define R^2.

The same principle can be used to define modeled proportion of variance in multilevel random effect models. In the case of two-level models, however, a basic question is, What is predicted? An individual value Y_{ij} at the lowest level, or an aggregated value $\overline{Y}_{.j}$ at a higher level? On the basis of the distinction between the two levels, two concepts of modeled proportion of variance in a two-level model can be defined. To introduce the basic idea, first consider a two-level random effects model with a random intercept and some predictor variables with fixed effects but no other random effects:

$$Y_{ij} = X_{ij}\beta + U_{0j} + E_{ij}, \tag{10}$$

where the random residuals U_{0j} and E_{ij} are uncorrelated and have expectation zero; and where the X variables are random and not correlated with the U and the E variables. We assume that (10) is a model with intercept: The first variable in X is identically equal to unity. Because we wish to discuss the definition of modeled proportion of variance as a population parameter, we assume, temporarily, that the vector β of regression coefficients is known.

For the level-1 modeled proportion of variance, we consider the prediction of Y_{ij} for a randomly drawn level-1 unit i within a randomly drawn level-2 unit j. If the values of the predictors X_{ij} are unknown, then the best predictor for Y_{ij} is its expectation, $\mu\beta$ where $\mu = EX_{ij}$; the associated mean squared prediction error is $\text{var}(Y_{ij})$. If the value of the predictor vector X_{ij} for the given unit is known, then the best linear predictor for Y_{ij} is the regression value $X_{ij}\beta$; the associated mean squared prediction error is $\text{var}(Y_{ij} - X_{ij}\beta) = \sigma^2 + \tau^2$. The level-1 modeled proportion of variance is defined as the proportional reduction in mean squared prediction error:

$$R_1^2 = 1 - \frac{\text{var}(Y_{ij} - X_{ij}\beta)}{\text{var}(Y_{ij})}. \tag{11}$$

How can this parameter R_1^2 be estimated? For unbalanced data, the sample variance is not necessarily the best estimator for $\text{var}(Y_{ij})$. An estimator that is more in line with the random effects two-level model is $\hat{\sigma}_0^2 + \hat{\tau}_0^2$, where $\hat{\sigma}_0^2$ and $\hat{\tau}_0^2$ are defined as the estimators for the two-level model with a random intercept but without any predictors:

$$Y_{ij} = \beta_0 + U_{0j} + E_{ij}. \tag{12}$$

Because the sample variance and $\hat{\sigma}_0^2 + \hat{\tau}_0^2$ are two estimators for the same parameter, their outcomes will not be very different; in any case, they should only rarely be significantly different. Using the complete (unbalanced) data from the earlier mentioned study of Vermeulen and Bosker (1992) once again, $\hat{\sigma}_0^2 + \hat{\tau}_0^2$ amounts to 10.45, whereas the sample variance is 10.14.

Then the most straightforward way to estimate R_1^2 is to consider $\hat{\sigma}^2 + \hat{\tau}^2$ for the reference model (12) as well as for model (10), and compute 1 minus the ratio of these values. In other words, R_1^2 is just the proportional reduction in the value of $\hat{\sigma}^2 + \hat{\tau}^2$ due to including the X variables in the model. For a sequence of nested models, the contributions to the estimated value of (11) due to adding new predictors can be considered to be the contribution of these predictors to the modeled variance at level 1.

To illustrate this, we once again use the data from the first (balanced) example, and estimate the proportional reduction of prediction

error for a model where within- and between-groups regression coefficients might be different. From Table 2 we see that $\hat{\sigma}^2 + \hat{\tau}^2$ for model (A) amounts to 10.965, and for model (D) to 7.964. R_1^2 is thus estimated to be $1 - (7.964/10.965) = 0.274$.

Now we turn to the level-2 modeled proportion of variance. It is natural to define this as the proportional reduction in mean squared prediction error for the prediction of $\overline{Y}_{.j}$ for a randomly drawn level-2 unit j. If the values of the predictors X_{ij} for the set of level-1 units i within level-2 unit j are completely unknown, then the best predictor for $\overline{Y}_{.j}$ is its expectation, which is again $\mu\beta$ where $\mu = EX_{ij}$; the associated mean squared prediction error is $\text{var}(\overline{Y}_{.j})$. If the values of the predictors X_{ij} for all i in this particular group j are known, then the best linear predictor for $\overline{Y}_{.j}$ is the regression value $\overline{X}_{.j}\beta$; the associated mean square prediction error is $\text{var}(\overline{Y}_{.j} - \overline{X}_{.j}\beta) = \sigma^2/n_j + \tau^2$. The level-2 modeled proportion of variance is now defined as the proportional reduction in mean squared prediction error for $\overline{Y}_{.j}$:

$$R_2^2 = 1 - \frac{\text{var}(\overline{Y}_{.j} - \overline{X}_{.j}\beta)}{\text{var}(\overline{Y}_{.j})}. \tag{13}$$

It must be noted that this parameter is similar to the value of R^2 (defined in the classical way) in the aggregated regression analysis, where $\overline{Y}_{.j}$ is regressed on $\overline{X}_{.j}$. This is apparent from definition (13). (The similarity, however, is conditional on the correct model specification; if a two-level model is estimated with incorrectly assumed equalities of within-group and between-group regression coefficients, then the estimated value for R_2^2 will be smaller than the R^2 estimated in the aggregated regression analysis, because then a nonoptimal value for β is used.)

To estimate the level-2 modeled proportion of variance, we follow a similar approach as for estimating R_1^2: For balanced data, we estimate R_2^2 as the proportional reduction in the value of $\hat{\sigma}^2/n + \hat{\tau}^2$.

In the example given earlier, for model (a) the value of $\hat{\sigma}^2/n + \hat{\tau}^2$ is $8.694/10 + 2.271 = 3.140$, whereas for model (b) this amounts to $6.973/10 + 0.991 = 1.688$. R_2^2 is thus estimated at $1 - (1.688/3.140) = 0.462$.

For unbalanced data, it must be noted that the mean squared error for predicting a group mean naturally depends on the group size; one

TABLE 2: Estimating R_1^2, the Level-1 Modeled Variance (balanced data)

	$\hat{\sigma}^2$	$\hat{\tau}^2$
A. $Y_{ij} = \beta_0 + U_{0j} + E_{ij}$	8.694	2.271
D. $Y_{ij} = \beta_0 + \beta_1(X_{ij} - \bar{X}_{.j}) + \beta_2\bar{X}_{.j} + U_{0j} + E_{ij}$	6.973	0.991

could use as the value for n either a value deemed a priori to be representative, or the harmonic mean, defined by $\{(1/N)\sum_j(1/n_j)\}^{-1}$.

BEHAVIOR OF R_1^2 *AND* R_2^2 *WHEN PREDICTORS ARE ADDED*

The starting point of this article was a criticism of the use of proportional reductions in the variance components σ^2 and τ^2 as definitions of modeled fractions of variance: it was pointed out that the variance components can increase when predictors are added. In particular, $\hat{\tau}^2$ increases when a successful within-group deviation variable is added. What about our definitions of R_1^2 and R_2^2, based on proportional reduction of prediction error? Is it possible that adding predictor variables leads to smaller values of R_1^2 and R_2^2? Can we even be sure at all that these quantities are positive?

It turns out that a distinction must be made between the population parameters R_1^2 and R_2^2 and their estimates from data. For the population parameters, we have the following properties.

Proposition 1. Suppose that

$$Y_{ij} = X_{ij}\beta + U_{0j} + E_{ij}, \tag{14}$$

where the correlations between the X variables on the one hand, and the U and E variables on the other, are zero. Let $X_{ij}^{(k)}$ be any subvector containing k elements of the random vector X_{ij}, and let $\beta^{(k)}$ be any fixed k vector. Then

$$\mathrm{Var}(Y_{ij} - X_{ij}^{(k)}\beta^{(k)}) \geq \mathrm{Var}(Y_{ij} - X_{ij}\beta). \tag{15}$$

This implies that the value of R_1^2 when computed for all predictors in the vector X is not smaller than the value of R_1^2 when computed for only the predictors in the vector $X^{(k)}$; in particular, $R_1^2 \geq 0$.

Proof. It holds that

$$\mathrm{Var}(Y_{ij} - X_{ij}^{(k)}\beta^{(k)}) = \mathrm{Var}(Y_{ij} - X_{ij}\beta + X_{ij}\delta) \tag{16}$$

where

$$X_{ij}\delta = X_{ij}\beta - X_{ij}^{(k)}\beta^{(k)}.$$

Such a vector δ exists because $X_{ij}^{(k)}$ is a linear function of X_{ij}. The model for Y_{ij} implies that $Y_{ij} - X_{ij}\beta = U_{0j} + E_{ij}$ and that this latter random variable is uncorrelated with X_{ij}. Therefore, (16) is equal to

$$\begin{aligned}\mathrm{Var}(U_{0j} + E_{ij} + X_{ij}\delta) &= \mathrm{Var}(U_{0j} + E_{ij}) + \mathrm{Var}(X_{ij}\delta) \\ &\geq \mathrm{Var}(U_{0j} + E_{ij}) = \mathrm{var}(Y_{ij} - X_{ij}\beta).\end{aligned}$$

This proves (15). It follows immediately from applying (15) to the definition of R_1^2 that this parameter for the entire vectorX_{ij} is not smaller than for the subvector $X_{ij}^{(k)}$. By taking $X_{ij}^{(k)}$ as the vector with the constant variable as its only element, the last result implies that $R_1^2 \geq 0$. QED.

Note that normality assumptions or assumptions on the covariance structure of the residuals $U_{0j} + E_{ij}$ are not needed for this proposition; the conditions only refer to the correct specification of the regression part $X_{ij}\beta$ and to the zero correlation between the explanatory variables and the residuals.

The interpretation of this proposition is that population values of R_1^2 in correctly specified models become smaller when predictor variables are deleted. In a similar way, it can be proved that population values of R_2^2 in correctly specified models become smaller when predictor variables are deleted, provided that the variables U_{0j} and E_{ij} on one hand are uncorrelated with all the X_{ij} variables on the other hand.

For estimates of R_1^2 and R_2^2, the situation is different: it cannot be proved, in general, that these estimates become smaller when predictor variables are deleted. The important point is that, for the monotonicity property of the population values of R_1^2 and R_2^2, it had to be assumed in Proposition 1 that the larger of the two models being compared was correctly specified. This implies the following: when it is observed that an estimated value for R_1^2 or R_2^2 becomes considerably smaller by the addition of a predictor variable, or considerably larger by the deletion of a predictor variable, this suggests that the larger model is

misspecified. In this sense, changes in R_1^2 or R_2^2 serve as a diagnostic for possible misspecification. Because the proposition does not require the assumption of normal distributions or of a specific structure for the random part of the model, this possibility of misspecification refers to the fixed part of the model (i.e., the specification of the explanatory variables having fixed regression coefficients).

An important type of misspecification in two-level models is the restriction that a certain variable has the same within-group as between-group regression coefficients, whereas in the population these coefficients are different. We have simulated some examples where this anomalous behavior of R_1^2 and R_2^2 was indeed observed.

Example. Data were simulated with $N = 30$ groups of size, $n = 5$ with one predictor variable according to the model,

$$Y_{ij} = X_{ij} - 2\bar{X}_{.j} + U_{0j} + E_{ij},$$

where the random variables X_{ij}, U_{0j}, and E_{ij} all are independent and normally distributed with $\mathrm{Var}(X_{ij}) = 1, \sigma^2 = 1, \tau^2 = 0.1$. When the model,

$$Y_{ij} = \beta_0 + U_{0j} + E_{ij}, \tag{17}$$

was fitted, the following estimates were obtained:

$$\hat{\sigma}^2 = 2.228, \hat{\tau}^2 = 0.088;$$

fitting the model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + U_{0j} + E_{ij} \tag{18}$$

yielded the estimates

$$\hat{\sigma}^2 = 1.026, \hat{\tau}^2 = 1.077, \hat{\beta}_1 = 1.046.$$

Model (18) is incorrectly specified because it is based on the assumption of equal within-groups and between-groups regression coefficients for X. We see that the value for the mean squared prediction error at level 2 is for model (18), given by $\hat{\tau}^2 + \hat{\sigma}^2/n = 1.282$; this is larger than for the intercept-only model (17), which has 0.534. Correspondingly, the estimated value for R_2^2 is negative:

$$R_2^2 = 1 - 1.282/0.534 = -1.401.$$

Example. Another data set was simulated with the same parameters as the previous one, except that the difference between the between-group and the within-group regression coefficients was even stronger, namely, by using the regression model

$$Y_{ij} = -2X_{ij} + 3\overline{X}_{.j} + U_{0j} + E_{ij}.$$

Here the estimates obtained for the intercept-only model (17) were

$$\hat{\sigma}^2 = 5.457, \hat{\tau}^2 = 4.047,$$

whereas the estimates obtained for the incorrect model (18) were

$$\hat{\sigma}^2 = 0.976, \hat{\tau}^2 = 9.702, \hat{\beta}_1 = -1.978.$$

In this case, both $\hat{\sigma}^2 + \hat{\tau}^2$ and $\hat{\sigma}^2/n + \hat{\tau}^2$ are larger for the incorrect larger model; the estimates for the proportions of modeled variance are negative,

$$\hat{R}_1^2 = -0.124 \text{ and } \hat{R}_2^2 = -0.926.$$

The fact that estimated values for R_1^2 and R_2^2 can be negative might seem an undesirable feature of these statistics. Keeping in mind, however, that the population values for correctly specified models are necessarily nonnegative and that these population values decrease when relevant predictor variables are deleted from a well-specified model, this property of the estimators for R_1^2 and R_2^2 turns into an advantage because now we have an interesting new diagnostic for misspecification. For well-specified models, the estimated R_1^2 and R_2^2 will behave nicely just like their population counterparts; if estimated R_1^2 and R_2^2 are strongly negative or decrease by a significant amount when predictor variables are added, then the specification of the fixed part of the model must be doubted.

PROPORTION OF MODELED VARIANCE FOR MODELS WITH SEVERAL RANDOM EFFECTS

The idea of using the proportional reduction in the prediction error for Y_{ij} and $\overline{Y}_{.j}$, respectively, as the definitions of modeled variance at

either level can be immediately extended to two-level models with one or more random regression coefficients. This model is formulated as

$$Y_{ij} = X_{ij}\beta + \sum_h Z_{hij}U_{hj} + E_{ij}, \tag{19}$$

where h ranges from 0 to H, and Z_0 is identically equal to unity so that U_{0j} is the random intercept as in (10). The Z_h are level-1 variables with random regression coefficients; there might be an overlap between the X variables and the Z variables. It is assumed that the U_{hj} are random variables with means zero and (co)variances $\text{cov}(U_{hj}, U_{kj}) = \tau_{hk}$, and that they are uncorrelated with the E_{ij}; further, the X and Z variables are assumed to be statistically independent of the U and the E variables.

One warning is in order before we embark on the extension of R_1^2 and R_2^2 to model (19). The random effects part of the model $\sum_h Z_{hij}U_{hj}$ is of no help to predict Y_{ij} or $\overline{Y}_{.j}$ because the U_{hj} are unknown. Therefore, although it is necessary for a consistent treatment of R_1^2 and R_2^2 to define these parameters and their estimators also for model (19), in practical data analysis, the addition to model (10) of random slopes so that model (19) is obtained will not lead to important changes in estimated values for R_1^2 or R_2^2. (And if it does, one should again be suspicious of misspecification.)

First we consider the proportion of modeled variance at level 1. It follows from the model definition that the best linear predictor of Y_{ij} given the X_{ij} values, is $X_{ij}\beta$. Therefore, formula (11) can be retained as the definition of R_1^2. When Y_{ij} is predicted by $X_{ij}\beta$, the associated mean squared prediction error, conditional on the Z_h values, is

$$\text{var}(Y_{ij} - X_{ij}\beta \mid Z_1, \ldots, Z_h) = \sum_{h,k} Z_{hij}Z_{kij}\tau_{hk} + \sigma^2. \tag{20}$$

There is one complication: this is a conditional mean squared prediction error that depends on the covariates Z_h, which are themselves random variables. (When discussing model (10), this complication did not arise because the covariate with a random coefficient in that case was unity, so it remained implicit in the equations.) The unconditional mean squared prediction error is the expected value of (20), given by

$$\begin{aligned} \operatorname{var}(Y_{ij} - X_{ij}\beta) &= \sum_{h,k} E(Z_{hij}Z_{kij})\tau_{hk} + \sigma^2 \\ &= \mu'_Z\tau\,\mu_Z + \operatorname{trace}(\tau(\Sigma_Z^B + \Sigma_Z^W)) + \sigma^2, \end{aligned} \tag{21}$$

where τ is the matrix with elements τ_{hk}, μ_z is the mean of the vector Z, and Σ_Z^B and Σ_Z^W are the between-group and within-group covariance matrices of Z. (Note that $\mu_0 = 1$ and $\tau_{0h} = 0$ for $h = 0, \ldots, H$. For model (10), the case of only a random intercept and no other random group effects, $\tau_{00} = \tau^2$ and (21) reduces to the parameter $\tau^2 + \sigma^2$ that was discussed before.) For model (19), the proportion of modeled variance R_1^2 can be estimated as the proportional reduction in the estimated value of (21) due to the addition of the X_{ij} and/or the Z_{hij} variables.

If the model has only one random slope (i.e., H = 1), then the parameters of the corresponding explanatory variable Z_1 are μ_1, σ_{11}^B, σ_{11}^W, and (21) reduces to

$$\tau_0^2 + 2\mu_1\tau_{01} + \tau_1^2(\mu_1^2 + \sigma_{11}^B + \sigma_{11}^W) + \sigma^2. \tag{21'}$$

For the proportion of modeled variance at level 2, we again consider the prediction of $\overline{Y}_{.j}$ for a randomly drawn level-2 unit j. Suppose first that the number of level-1 units is equal to *n* for each group. If the values of the predictors X_{ij} for the set of level-1 units i within this level-2 unit j are known, then the best linear predictor for $\overline{Y}_{.j}$ is the regression value $\overline{X}_{.j}\beta$. It follows that R_2^2 can again be defined by (13). Conditional on the Z_h values, the mean squared prediction error of $\overline{X}_{.j}\beta$ as a predictor for $\overline{Y}_{.j}$ is

$$\operatorname{var}(\overline{Y}_{.j} - \overline{X}_{.j}\beta \mid Z_1, \ldots, Z_h) = \sum_{h,k} \overline{Z}_{h.j}\overline{Z}_{k.j}\tau_{hk} + \frac{\sigma^2}{n}. \tag{22}$$

The unconditional mean squared prediction error is the expected value of this quantity, given by

$$\begin{aligned} \operatorname{Var}(\overline{Y}_{.j} - \overline{X}_{.j}\beta) &= \sum_{h,k} E(\overline{Z}_{h.j}\overline{Z}_{k.j})\tau_{hk} + \frac{1}{n}\sigma_2 \\ &= \mu'_Z\tau\mu_Z + \operatorname{trace}(\tau(\Sigma_Z^B + \frac{1}{n}\Sigma_Z^W)) + \frac{1}{n}\sigma^2. \end{aligned} \tag{23}$$

The level-2 modeled proportion of variance can be estimated as the proportional reduction in the estimated value of (23) due to the

addition of the X_{ij} and/or the Z_{hij} variables. In case the model has only one random slope, the arguments to reduce (21) to (21′) can be used again, and (23) reduces to

$$\tau_0^2 + 2\mu_1\tau_{01} + \tau_1^2(\mu_1^2 + \sigma_{11}^{B} + \frac{1}{n}\sigma_{11}^{W}) + \frac{1}{n}\sigma^2. \qquad (23')$$

For designs with unequal group sizes n_j, the same approach can be followed as in random intercept models: Use a representative value for n or use the harmonic mean of the group sizes.

The monotonicity properties of R_1^2 and R_2^2, discussed above for the random intercept model, also hold true for the more general model (19). In the proof of Proposition 1, the only assumption is that the fixed regression part $X_{ij}\beta$ is uncorrelated with the residual part $U_{0j} + E_{ij}$. Because in model (19) the X and Z variables are independent of the U and the E variables, the term $\sum Z_{hij}U_{hj}$ in (19) might play the role of U_{0j} in Proposition 1, and the conclusion of the proposition is also valid for this model. Thus the population values of R_1^2 and R_2^2 are nonnegative and increase when predictor variables are added; but their estimates can be negative and can decrease when predictor variables are added, which then points to the possibility of misspecification of the fixed part $X_{ij}\beta$ in model (19).

Example. The data of the first (balanced) example are used again now with a model where the within-group deviation score on the intelligence test is given a random slope:

$$Y_{ij} = \beta_0 + \beta_1(X_{ij} - \overline{X}_{.j}) + \beta_2\overline{X}_{.j} + (X_{ij} - \overline{X}_{.j})U_{1j} + U_{0j} + E_{ij}. \qquad (24)$$

The proportions of modeled variance R_1^2 and R_2^2 will be estimated for model (24); $\text{var}(Y_{ij})$ and $\text{var}(\overline{Y}_{.j})$ (for the denominators of [11] and [13]) are estimated for the reference model that includes a random intercept and no random slopes. Because the variable with random slope is a within-group deviation score, we have $\mu_1 = \sigma_{11}^{B} = 0$. For this data set, $\sigma_{11}^{W} = 8.545$. Estimation of model (24) using ML3 yielded $\hat{\tau}_{00} = 1.029$, $\hat{\tau}_{01} = 0.044$, $\hat{\tau}_{11} = 0.032$, and $\hat{\sigma}^2 = 6.589$. Formula (21′) yields 1.029 + 0.032*8.545 + 6.589 = 7.891. This leads to an estimate for R_1^2 of 1 − (7.891/10.965) = 0.280, which indeed is almost the same as the value

of 0.274 found for the model without a random slope. Formula (23′), for $n = 10$, yields $1.029 + 0.032*8.545/10 + 6.589/10 = 1.715$. Accordingly, the estimate for R_2^2 is $1 - (1.715/3.140) = 0.454$. This is quite close to the value of 0.462 found for the model without the random slope.

DISCUSSION

The concept of proportional reduction in mean squared prediction error leads to R_1^2 and R_2^2 defined by (11) and (13); these are clearly interpretable parameters for the proportion of modeled variance in two-level random coefficient models. For the random intercept model (10), these parameters can be estimated as the proportional reduction in $\hat{\sigma}^2 + \hat{\tau}^2$ and $\hat{\sigma}^2/n + \hat{\tau}^2$, respectively, due to the explanatory variables. For the more general random coefficient model (19), R_1^2 and R_2^2 can be estimated as the proportional reduction in the estimated values for (21) and (23), respectively. The population parameters R_1^2 and R_2^2 have the desirable properties that they are nonnegative and cannot decrease when predictor variables are added to the model, provided that the model for the fixed effects is correctly specified. The indicated estimators, however, do not possess these monotonicity properties for all data sets. When, for a given data set, negative values are estimated for R_1^2 or R_2^2, or when a decrease of the estimate of either of these parameters is observed on the inclusion of an additional predictor in the model, then this must be due either to a chance fluctuation or to misspecification in the model of the set of predictor variables with fixed regression coefficients.

The formulae for estimating R_1^2 and R_2^2 in models with random intercepts only are very easy. Estimating R_1^2 and R_2^2 in models with random slopes is more tedious. The software package HLM (Bryk, Raudenbush, Seltzer, and Congdon 1988), however, provides the necessary estimates because it not only produces estimates of the variance components, but also of the observed residual variances $\text{var}(Y_{ij} - X_{ij}\beta)$. Using these latter estimates, one can calculate the estimate of R_1^2 and R_2^2 straightforwardly.

A further question in model (19) with random regression coefficients is how much the explanatory variables contribute to the prediction of these varying slopes. This question will be treated in a following article.

To conclude this article, we go back briefly to the idea with which we started and which we rejected; namely, the use of proportional reductions in σ^2 and τ^2 as R^2-like measures. The rejection should not be taken absolutely because, even in the context of proportional reduction in prediction error, such measures do have a meaningful interpretation. Consider the general two-level model (19). Parameter σ^2 is the mean squared error for predicting Y_{ij} by $X_{ij}\beta + \sum_h Z_{hij}U_{hj}$; in other words, the mean squared error for the prediction of the individual outcome, when we know not only all covariates but also all peculiarities of group j, expressed in its random effects U_{hj}. In still other words, the proportional reduction in σ^2 is the proportional reduction in mean squared error for predicting Y_{ij} for a random individual in a given group. The only disturbing aspect of this interpretation is that the group is supposed to be known up to its kidneys—after all, the U_{hj} are not directly observable. Another interpretation of σ^2 is that it is $n/(n-1)$ times the mean squared error for predicting $Y_{ij} - \overline{Y}_{.j}$. Hence the proportional reduction in σ^2 is also the proportional reduction in mean squared error for predicting within-group differences.

The proportional reduction in τ^2 can be interpreted in the random intercept model as the proportional reduction in mean squared error for predicting $\overline{Y}_{.j}$ in an infinitely large group: for $n_j \to \infty$, the influence of σ^2 on $\tau^2 + \sigma^2/n_j$ vanishes. So the criticized definitions of R_2 as proportional reductions in the values for σ^2 and for τ^2 can still be interpreted as reductions in mean squared prediction errors, but in rather artificial situations.

REFERENCES

Aitkin, M. and N. T. Longford. 1986. "Statistical Modelling Issues in School Effectiveness Studies (With Discussion)." *Journal of the Royal Statistical Society* 149A:1-43.

Bryk, A. S. and S. W. Raudenbush. 1992. *Hierarchical Linear Models*. Newbury Park, CA: Sage.

Bryk, A. S., S. W. Raudenbush, M. Seltzer, and R. T. Congdon. 1988. *HLM*. Chicago: University of Chicago.

Dempster, A. P., D. B. Rubin, and R. K. Tsutakawa. 1981. "Estimation in Covariance Components Models." *Journal of the American Statistical Association* 76:341-53.

Goldstein, H. 1987. *Multilevel Models in Educational and Social Research*. London: Griffin.

Herbach, L. H. 1959. "Properties of Model II-Type Analysis of Variance Tests: A. Optimum Nature of the F-Test for Model II in the Balanced Case." *Annals of Mathematical Statistics* 30:939-59.

Prosser, R., J. Rasbash, and H. Goldstein. 1991. *ML3: Software for Three-Level Analysis. Users' Guide*. London: University of London, Institute of Education.

Raudenbush, S. W., and A. S. Bryk. 1986. "A Hierarchical Model for Studying School Effects." *Sociology of Education* 59:1-17.

Searle, S. R. 1971. *Linear Models*. New York: Wiley.

Vermeulen, C.J.A.J. and R. J. Bosker. 1992. *De Omvang en Gevolgen van Deeltijdarbeid en Volledige Inzetbaarheid in het Basisonderwijs*. Enschede: University of Twente.

Tom A. B. Snijders is a professor of stochastic models in the social and behavioral sciences at the University of Groningen (the Netherlands). He is one of the scientific directors of the research school ICS (Inter-university Center for Social Science Theory and Methodology). His main research interests are the development of statistical methods for multilevel research and for social networks. Recent publications in these domains, respectively, are "Standard Errors and Sample Sizes for Two-Level Research" (1993, written jointly with Roel J. Bosker), Journal of Educational Statistics*; and "Enumeration and Simulation Methods for 0-1 Matrices With Given Marginals" (1991),* Psychometrika.

Roel J. Bosker is an associate professor in the Department of Education of the University of Twente (the Netherlands). His main research interests are multilevel methods and evaluation research. A recent publication is "Standard Errors and Sample Sizes for Two-Level Research" (1993, written jointly with Tom A. B. Snijders), Journal of Educational Statistics.