Multilevel Analysis of Grouped and Longitudinal Data

Joop J. Hox

Utrecht University

Second draft, to appear in: T.D. Little, K.U. Schnabel, & J. Baumert (Eds.). *Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples*. Hillsdale, NJ: Erlbaum (in preparation).

1. The multilevel regression model for grouped data

Social and behavioral research often concerns problems that have a hierarchical structure, with individuals nested within groups. In multilevel analysis, such data structures are viewed as a multistage sample from a hierarchical population. For example, in educational research we may have a sample of schools, and within each school a sample of pupils. This results in a data set consisting of pupil data (e.g., SES, intelligence, school career) and school data (e.g., school size, denomination, but also aggregated pupil variables such as mean SES). In this chapter, the generic term *multilevel* is used to refer to analysis models for hierarchically structured data, with variables defined at all levels of the hierarchy. Typically, the research problem includes hypotheses of relationships between variables defined at different levels of the hierarchy.

A well-known multilevel model is the hierarchical linear regression model, which is essentially an extension of the familiar multiple regression model. It is known in the literature under a variety of names, such as 'hierarchical linear model' (Raudenbush & Bryk, 1986; Bryk & Raudenbush, 1992), 'variance component model' (Longford, 1989), and 'random coefficient model' (De Leeuw & Kreft, 1986; Longford, 1993). This model has become so popular that 'multilevel modeling' has become almost synonymous with 'applying a multilevel regression model.' However, since we also have multilevel extensions of other models, such as factor analysis or structural equation models (SEM), I reserve the term 'multilevel model' for the general case, and refer to specific classes of models as 'multilevel regression analysis' and 'multilevel SEM.'

The *multilevel regression model* is a hierarchical linear regression model, with a dependent variable defined at the lowest (usual the individual) level, and explanatory variables at all existing levels. Using dummy coding for categorical variables, it can be used for analysis of variance (ANOVA) models. The model has been extended to include dependent variables that are binary, categorical, or other non-normal data, and generalized to include multivariate response models and cross-classified data (cf. Bryk & Raudenbush, 1992; Longford, 1993; and especially Goldstein, 1995).

1.1 The basic multilevel regression model

In most applications, the first (lowest) level consists of individuals, the second level of groups of individuals, and higher levels of sets of groups. Conceptually, the model can be viewed as a hierarchical system of regression equations. For example, assume that we have collected data in J schools, with a different number of pupils N_j in each school. On the pupil level, we have the dependent variable Y_{ij} (e.g., school career) and the explanatory variable X_{ij} (e.g., pupil SES). We can set up a regression equation to predict the dependent variable Y from the explanatory variable X:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}. \qquad (1)$$

In equation (1), X_{ij} and Y_{ij} are the scores of pupil i in school j, β_{0j} is the regression intercept, β_{1j} the regression slope, and e_{ij} the residual error term. The multilevel regression model depicted in equation (1) specifies that the different schools are characterized by different regression equations; each school has its own intercept β_{0j} and slope β_{1j}.

The regression coefficients β_{0j} and β_{1j} vary across the schools, and we may model this variation with explanatory variables at the school level. Assume that we have one school level explanatory variable Z_j (e.g., school size). Then, the model for the ß's becomes:

$$\beta_{0j} = \beta_{00} + \beta_{01}Z_j + u_{0j}, \tag{2}$$
$$\beta_{1j} = \beta_{10} + \beta_{11}Z_j + u_{1j}. \tag{3}$$

In equation (2), β_{00} and β_{01} are the intercept and slope of the regression equation used to predict β_{0j} from Z_j, and u_{0j} is the residual error term in the equation for β_{0j}. Thus, if β_{01} is positive and significant, we conclude that school career outcome in large schools is higher than in small schools. Similarly, in equation (3), β_{10} and β_{11} are the intercept and slope to predict β_{1j} from Z_j, and u_{1j} is the residual error term in the equation for β_{1j}. Thus, if β_{11} is positive and significant, we conclude that the effect of X_{ij}, pupil SES in our example, is stronger in large schools. In this example, we have an interaction effect of Z_j (school size) and X_{ij} (pupil SES) on Y_{ij} (school career). This becomes more clear if we write the model as a single equation, by substituting equations (2) and (3) into equation (1):

$$Y_{ij} = \left[\beta_{00} + \beta_{10}X_{ij} + \beta_{01}Z_j + \beta_{11}Z_jX_{ij}\right] + \left[u_{1j}X_{ij} + u_{0j} + e_{ij}\right]. \tag{4}$$

In the multilevel regression model in equation (4), which is special case of the general mixed linear model (Harville, 1977), we can distinguish two parts. The fixed part contains the regression coefficients beta and their associated variables: $[\beta_{00} + \beta_{10}X_{ij} + \beta_{01}Z_j + \beta_{11}Z_jX_{ij}]$. The regression coefficients no longer carry a subscript *j* for schools; in the combined equation they refer to the average value of the regression across all schools. The random part contains the residual error terms: $[u_{0j} + u_{1j}X_{ij} + e_{ij}]$. Note that these do carry subscripts for the schools; the residual error terms represent the deviation of the schools' regression coefficients from their overall mean. The interaction term Z_jX_{ij} is sometimes referred to as a *cross-level interaction*, because it involves explanatory variables from different levels. The individual level errors e_{ij} are assumed to be independent and to have a normal distribution with mean zero and variance σ_e^2. The school level errors u_j are assumed to be independent and to have a multivariate normal distribution with mean vector zero and covariance matrix Σ. Since the school level errors are the schools' deviations from the overall regression, this is equivalent to assuming that the regression coefficients β_j follow a multivariate normal distribution.

1.2 Estimation and significance testing in the multilevel regression model

The parameters (regression coefficients and variance components) of the multilevel regression model are commonly estimated using Maximum Likelihood (ML) methods. Asymptotic standard errors are available for hypothesis testing. The usual significance test in maximum likelihood estimation is the so-called *Wald test*: the parameter estimate is divided by its standard error, and the result is referred to the standard normal distribution. Bryk and Raudenbush (1992) argue, that for the fixed effects it is better to refer this ratio to a Student distribution with *J-q-1* degrees of freedom (*J* = number of groups; *q* = number of fixed parameters), and to use a chi-square test for the random effects. For a discussion of the issues involved in choosing between such tests, see Hox (1998).

The likelihood function can be used to test the significance of the difference between two nested models. Most multilevel programs output a value that is called the deviance (computed as: deviance = minus two times the log likelihood). If a smaller model is a subset of a larger model, which means that it is obtained by either dropping parameters or imposing constraints on the larger model, the difference between the two deviances can be tested against a chi-square distribution. The degrees of freedom for this test is the difference in the number of parameters estimated. This test can be used instead of the Wald test, for multivariate tests of groups of parameters, and for tests of the variances in the random part where it is thought to be more accurate than the Wald test (Goldstein, 1995).

Two different Maximum Likelihood functions are currently used in the available software: Full ML (FML) and Restricted ML (RML). FML includes the fixed parameters in the likelihood function, RML does not. Most software offers a choice between the two methods. Since RML does not include the fixed parameters in the likelihood function, a deviance test based on RML can only be used to test for differences in the random part.

1.3 Example of multilevel regression analysis of grouped data

The multilevel regression model is most appropriate for data structures that have many groups, because it is more flexible and more parsimonious than analysis of variance-type models. For instance, assume that we study school careers in 50 schools. In each school, we take one class, and measure the pupils' achievement score, their SES and sex, the class size, and how experienced their teacher is. We have a total of 979 pupils from 50 classes, with an average class size just under 20. Note that by taking one class per school we have collapsed the school and the class level: it is impossible to distinguish between school and class effects. Table 1 presents the results of a sequence of multilevel regression models, with the achievement score as the dependent variable.

Table 1. Results analysis of school achievement example. Parameters (standard errors).

Explanatory variable:	intercept only	+ pupil variables	+ teacher variables	+ random slope	+ cross-level interaction
Fixed part:					
intercept	49.6 (.73)	16.7 (.91)	11.3 (4.70)	14.8 (4.52)	49.7 (.47)
pupil SES		9.0 (.17)	9.0 (.17)	9.0 (.35)	8.9 (.17)
pupil sex		-.8 (.31)	-.8 (.30)	-.6 (.28)	-.6 (.28)
class size			-.1 (.22)	-.1 (.20)	-.1 (.20)
teacher exp.			.5 (.07)	.2 (.07)	.5 (.07)
p.ses*t.exp					.3 (.03)
Random part:					
σ_e^2	88.8 (4.12)	21.9 (1.01)	21.9 (1.01)	17.9 (.85)	17.9 (.85)
σ_{interc}^2	21.7 (5.27)	20.0 (4.01)	9.2 (2.07)	49.7 (13.64)	8.9 (1.96)
σ_{ses}^2				4.9 (1.24)	.3 (.3)
$\sigma_{interc/ses}$				-14.4 (3.96)	.9 (.55)
Deviance:	5730	7258	5942	5908	5798

In Table 1, we present several different models. The first model, the Intercept Only model, serves as a baseline; it tells us that the total variance is divided into 88.8 at the pupil level, and 21.7 at the school level. This gives us the intraclass correlation, which is the proportion of variance at the group level. In the school data the intraclass correlation is (21.7/110.5=) 0.20. In other words: 20% of the variance is at the school level. Some of this variation is explained by pupil and school characteristics. Pupil SES turns out to have a regression coefficient with significant variance across schools, which is partly explained by the interaction with the teachers' experience. For a more detailed introduction to multilevel regression, I refer to Bryk and Raudenbush (1992) and Hox (1995).

2. The multilevel regression model for longitudinal data

In the previous section, the individuals are considered to be the lowest level of the hierarchy. In longitudinal research, we have a series of repeated measures for each individual. One way to model such data is to view the series of repeated measures as a separate level below the individual level. The individual level becomes the second level, and it is possible to add a third and higher levels for possible group structures. Multilevel models for longitudinal data are discussed by, amongst others, Bryk and Raudenbush (1987, 1992) and Goldstein (1987, 1995), for an introduction see Snijders (1996).

2.1 The basic model for longitudinal data

The multilevel regression model for longitudinal data is a straightforward application of the multilevel regression model described above. It can also be written as a sequence of models for each level. At the lowest, the repeated measures level, we have:

$$Y_{tj} = \beta_{0i}+\beta_{1i}T_{ti}+\beta_{2i}X_{ti}+e_{ti}. \tag{5}$$

In equation (5), Y_{ti} is the dependent variable of individual i measured at time point t, T is the time variable that indicates the time point, and X_{ti} is a *time varying covariate*. For example, Y_{ti} could be a reading score of a pupil, T_{ti} the age at the time the reading score is measured, and X_{ti} the experience of the teacher the pupil has at time t. Pupil characteristics, such as gender, are *time invariant covariates*, which enter the equation at the second level:

$$\beta_{0i} = \beta_{00}+ \beta_{01}Z_i+u_{0i}, \tag{6}$$
$$\beta_{1i} = \beta_{10}+ \beta_{11}Z_i+u_{1i}, \tag{7}$$
$$\beta_{2i} = \beta_{20}+ \beta_{21}Z_i+u_{2i}. \tag{8}$$

By substitution, we get the single equation model:

$$\begin{aligned} Y_{tj} = {} & \beta_{00}+\beta_{10}T_{ti}+\beta_{20}X_{ti}+\beta_{01}Z_i+\beta_{11}Z_iT_{ti}+\beta_{21}Z_iX_{ti}+ \\ & u_{1i}T_{ti}+u_{ij}X_{ti}+u_{0i}+e_{ti} \end{aligned} \tag{9}$$

In longitudinal research, we sometimes have repeated measurements of individuals, who are all measured together on a small number of fixed occasions. This is typically the case with panel research. If we simply want to test the null hypothesis that the means are equal for all occasions, we can specify the repeated measures as observations on a multivariate response vector and use

Multivariate Analysis of Variance (MANOVA). Multilevel regression offers an essentially equivalent analysis. We use a multivariate response model (Goldstein, 1995) with dummy variables indicating the different occasions. The standard multilevel model in equation (9) assumes that the residual errors e_{ti} are independent and have constant variance over time, while MANOVA assumes an unconstrained covariance matrix for the e_{ti}. Bryk and Raudenbush (1992: 132) argue that uncorrelated errors may be appropriate in short time series. However, the multilevel regression model can be extended to include an unconstrained covariance matrix at the lowest level (Goldstein, 1995; Maas & Snijders, 1997). Multilevel analysis is useful for fixed occasion data when there are missing observations, due to absence of individuals at specific occasions, or due to panel attrition. Since multilevel models do not assume equal numbers of identical occasions for all individuals, such missing data poses no special problems, while MANOVA handles missing data by deleting all individuals with incomplete data.

The number of occasions and their spacing may also vary across individuals. This is often the case in growth curve models, where an individual's development is a function of the *time* or *age* at the different occasions. The most important advantage of multilevel analysis for such research problems is its flexibility. In the multilevel regression model, the development over time is modeled by a linear regression equation, with possibly different regression coefficients for different individuals. Thus, each individual gets their own growth curve, specified by individual regression coefficients that may depend on individual attributes. Quadratic and higher functions can be used to model nonlinear dependencies on time, and both time varying and person level covariates can be added to the model. For a more detailed discussion of such models, I refer to the chapter by MacCallum and Kim (this volume).

As was mentioned above, multilevel regression models for growth curves commonly assume residual errors that are uncorrelated over time. Especially for growth curves with closely spaced observations, this assumption is implausible. Models that are more complex are possible for the residual errors e_{ti}. For instance, we can specify an autocorrelation structure for the residuals, or we can model the variance of the residuals as a function of time or age. Some such models are discussed by Gibbons et al. (1993) and Goldstein (1995), and the program MixReg (Hedeker & Gibbons, 1996) has built-in options for correlated errors.

2.2 Example of multilevel analysis of longitudinal data

The example data have been generated by Rogosa and Saner (1995), with 200 individuals measured at five equidistant time points, and at the person level one time invariant covariate. The model for these data is given by:

$$Y_{tj} = \beta_{00}+\beta_{10}T_{ti}+\beta_{01}Z_i+\beta_{11}Z_iT_{ti}+u_{1i}T_{ti}+u_{0i}+e_{ti} \quad (10)$$

The time points T are coded as t=0,1,2,3,4, and the covariate Z is centered around its overall mean. As a result, the intercept can be interpreted as the expected value at the first occasion, for individuals with an average value of Z. Using time points t=1,2,3,4,5 and raw scores for Z would be completely equivalent, but slightly more difficult to interpret.

Table 2 presents the results of a multilevel analysis of these longitudinal data. Model (1) is a model that contains only an intercept term; this model serves as a null model. The intercept-only model estimates the repeated measures variance as 83, and the person level variance as 42. Thus, the intraclass correlation or the proportion variance at the person level is estimated as (42/(42+83))=0.34. Approximately two-thirds of the variance of the measures is variance over time, and one-third is variance between individuals. In model (2), the time variable is added as a predictor with varying coefficients for different persons. The model predicts a value of 44 at the first occasion, which increases by 5 on each succeeding occasion. The variance components for the intercept and the regression slope for the time variable are both significant. The significant intercept variance means that individuals have different initial states, and the significant slope variance means that individuals also have different growth rates. There is a correlation between the initial status and the growth rate; individuals who start high tend to grow at a slower rate. In model (3), the covariate *Z* is added to the model, both as a direct effect and as an interaction with the time variable. Part of the variation of the intercept and slope coefficients can be modeled by the covariate. Note that the correlation between the intercept and slope is different; in model (2), it is an ordinary correlation, but in model (3), it is a partial correlation, conditional on the covariate *Z*.

Table 2. Multilevel results Rogosa/Saner data (ML estimates, parameters and se's)

Model	(1)	(2)	(3)
interc	54.0 (.55)	44.0 (.56)	44.0 (.51)
time		5.0 (.17)	5.0 (.14)
Z			1.5 (.14)
Z*time			.6 (.07)
σ^2_e	84.4 (4.22)	11.9 (.69)	11.9 (.69)
σ^2_{int}	42.7 (6.01)	54.8 (6.21)	45.3 (5.26)
σ^2_{time}		4.4 (.56)	2.7 (.40)
$\sigma_{int*time}$		-3.8 (1.39)	-7.7 (1.25)
r_{i*t}		-.24	-.69
deviance	7525	6251	6032

It would be nice to have clear-cut values for the amount of variance explained by the various effects. Bryk and Raudenbush (1992) suggest to use the residual error variance of the intercept-only model as a benchmark, and to examine how much this goes down when explanatory variables are added to the model. In Table 2, this strategy leads to inconsistencies, because in model (2) the residual error variance for the intercept actually goes up when the time variable is added to the model! The reason is that in multilevel models with random coefficients the notion of 'amount of variance explained at a specific level' is not a simple concept. As Snijders and Bosker (1994) explain in detail, the problem arises because the statistical model behind multilevel models is a hierarchical sampling model: groups are sampled at the higher level, and at the lower level individuals are sampled within groups. This sampling process creates some variability between the groups, even if there are in fact no real group differences. In time series, the lowest level is a series of measurements, which in many cases are (almost) the same for all individuals in the sample. Thus, the variability between persons in the time series variable is in fact much lower than the hierarchical sampling model assumes. Snijders and Bosker (1994) describe procedures to correct the problem. A simple approximation is to use the occasion level error variance from model (1), and the person level error variance of model (2) that includes the time variable. Then, we observe that the error variance at the repeated measures level goes down from 84.4 to 11.9, which means that the time variable explains about 86% of the variance between the occasions. To see how much variance the person level variable *Z* explains, we regard the intercept variance of 54.8 in model (2) as the error variance, and observe that in model (3) this variance goes down to 45.3. Thus, at the person level, the covariate *Z* explains about 17% of the initial variation between the persons. Likewise, we can calculate that *Z* explains about 39% of the initial variance of the time slopes. These values are indications, procedures that are more precise are given by Snijders and Bosker (1994).

As the example makes clear, applying multilevel regression models to longitudinal data is straightforward, especially if we restrict ourselves to a single dependent variable, and to a linear or polynomial function of the time variable. More complicated models are possible, such as nonlinear models or multivariate models for several dependent variables; for these I refer to the chapter by MacCallum and Kim in this book, and to Goldstein (1995).

3 Multilevel structural equation models

Structural Equation Models (SEM) for multilevel data are described, amongst others, by Goldstein and McDonald (Goldstein & McDonald, 1988; McDonald & Goldstein, 1989), Muthén and Satorra (Muthén, 1989; Muthén & Satorra, 1989), Longford and Muthén (Longford & Muthén, 1992; Longford, 1993) and Lee and Poon (1992). Non-technical introductions are given by Muthén (1994) and McDonald (1994) and in Hox (1995). This section describes an approximation proposed by Muthén (1989, 1994), that makes it possible to use standard SEM software to analyze multilevel structural equation models.

3.1 The basic multilevel structural equations model

Multilevel structural models assume that we have a population of individuals that are divided into groups. Assume that we have multivariate data from *N* individuals in *G* groups. For each individual, we replace the observed *Total* scores $\mathbf{y}_T$ by their components: the between-groups component $\mathbf{y}_B$, which are the disaggregated group means, and the within-groups component $\mathbf{y}_W$, which are the individual deviations from the group mean. These two components have the attractive property that they are orthogonal and additive:

$$\mathbf{y}_T = \mathbf{y}_B + \mathbf{y}_W. \qquad (11)$$

If we decompose the population data into between-groups variables and within-groups variables, we can distinguish three population covariance matrices: the *total* covariance matrix Σ_B, the *between-groups* covariance matrix Σ_B, and the *within-groups* covariance matrix Σ_W. Just like the group means and the individual deviation scores themselves, the covariance matrices Σ_B and Σ_W are orthogonal and sum to the total covariance matrix Σ_T:

$$\Sigma_T = \Sigma_B + \Sigma_W. \qquad (12)$$

Muthén (1989) shows that an unbiased estimate of the population within-groups covariance matrix Σ_W is given by the pooled within groups-covariance matrix $\mathbf{S}_{PW}$:

$$\mathbf{S}_{\mathbf{PW}} = \frac{\sum_{1}^{G}\sum_{1}^{n_g} (\mathbf{y}_{ig} - \overline{\mathbf{y}}_{g})(\mathbf{y}_{ig} - \overline{\mathbf{y}}_{ig})'}{N-G} \tag{13}$$

This corresponds to the conventional equation for the covariance matrix of the individual deviation scores, with N-G in the denominator instead of the usual N-1. Thus, we can model the population within group structure by constructing and testing a structural model for $\mathbf{S}_{PW}$.

The between groups covariance matrix for the disaggregated group means $\mathbf{S}_{B,}$ calculated in the sample, is given by:

$$\mathbf{S}_{\mathbf{B}} = \frac{\sum_{1}^{G} n_g (\overline{\mathbf{y}}_{\mathbf{g}} - \overline{\mathbf{y}})(\overline{\mathbf{y}}_{\mathbf{g}} - \overline{\mathbf{y}})'}{\mathbf{G}} \tag{15}$$

This corresponds to the conventional equation for the covariance matrix of the disaggregated group means, with G in the denominator instead of the usual N-1. Unfortunately, $\mathbf{S}_B$ is not a simple estimator of the population between groups covariance matrix Σ_B. Instead, $\mathbf{S}_B$ is an estimator of the sum of two matrices:

$$\mathbf{S}_B = \Sigma_W + c\Sigma_B \tag{15}$$

where *c* is a scaling factor equal to the common group size (Muthén, 1989, 1994). Thus, to model the between groups structure, we must specify for $\mathbf{S}_B$ two models: one for the within-groups structure and one for the between-groups structure. Muthén (1989, 1994) proposes to use the multigroup option of standard SEM software to analyze these models. There are two groups, with covariance matrices $\mathbf{S}_{PW}$ and $\mathbf{S}_B$, based on N-G and G observations. The model for Σ_W must be specified for both $\mathbf{S}_{PW}$ and $\mathbf{S}_B$, with equality restrictions between both 'groups.' The model for Σ_B is added to the model for $\mathbf{S}_B$, with the scale factor *c* built into the model.

The reasoning given above assumes *balanced* data, where all groups have the same group size *n*. For unbalanced data, we proceed as if the group sizes were equal, and calculate the scaling factor as a combination of the observed group sizes given by:

$$C = \frac{N^2 - \sum_{1}^{G} n_g^2}{N(G-1)} \tag{16}$$

This *pseudobalanced* solution (McDonald, 1994), which Muthén (1989, 1994) calls the MUML estimator (cf. Gustafsson & Munck, this volume), is not a full maximum likelihood solution. However, if the samples are reasonably large, and the group sizes are not extremely different, the pseudobalanced estimates are close enough to the full maximum likelihood estimates to be useful in their own right. Comparisons of pseudobalanced estimates with full maximum likelihood estimates or with known population values have been made by Muthén (1990, 1994), Hox (1993), and McDonald (1994). They all conclude that the pseudobalanced estimates are generally accurate.

The multilevel part of the covariance structure model outlined above is simpler than that of the multilevel regression model. It is comparable to the multilevel regression model with random variation of the intercepts. There is no provision for randomly varying slopes (factor loadings and path coefficients). Chou, Bentler and Pentz (this volume) show that, provided that the sample sizes of the groups are large, it is possible to analyze more general models by employing a two-step approach.

Multilevel covariance structure analysis models the population covariance matrices Σ_B and Σ_W, by specifying by separate structural models for the between-groups and within-groups covariances. To apply structural equation models to multilevel data, it is helpful to have a program that calculates the required matrices and the scaling factor. Public domain programs for this are available from Muthén (BW) and Hox (SPLIT2). Gustaffson has developed a program called STREAMS (Gustaffson & Stahl, 1997, see also Gustafsson & Munck in this book), which not only calculates the required matrices, but also writes the required program setups for a variety of SEM packages.

3.2 Example of multilevel factor analysis

The example data are taken from Van Peet (1992). They are the scores on six intelligence tests of 269 children from 49 families. The six intelligence test are: word list, cards, matrices, figures, animals, and occupations. The data have a multilevel structure, with children nested within families. Assuming that intelligence is strongly influenced by shared genetic and environmental influences, we expect strong between family effects.

To begin, the individual scores on the six measures are decomposed into the disaggregated group means and the individual deviations from these group means. The intraclass correlations (the proportion of variance at the family level) of the six tests vary from 0.18 to 0.36, which suggests considerable family level variance. The model specification strategy starts with seeking a within model, and then goes on with the between model, for details see Hox (1995). The principle of utilizing the simplest model that fits well, leads to a model with two factors on the individual level, and one general factor on the family level. The path diagram is given in Figure 1 below.

The model in fits fairly well (χ^2=28.0, df=17, p=.05; GFI=.97; TLI=0.93). Figure 1 actually shows the model for the between groups matrix, which contains a model for both Σ_B and Σ_W. The lower half is the model for Σ_W, and the upper half is the model for Σ_B. The model for Σ_W is also specified for the pooled within groups matrix, with equality restrictions for all corresponding parameters. The fixed regression coefficients of 2.34 for the between group variables, that are represented by the factors *wlb* to *anb*, are the scaling constants, set equal to the square root of *c*. The one-factor model for the between group variables looks quite ordinary, until we realize that the factor loadings are actually regression coefficients going from one latent variable to another. This does not affect their interpretation as factor loadings, it only complicates the setups.

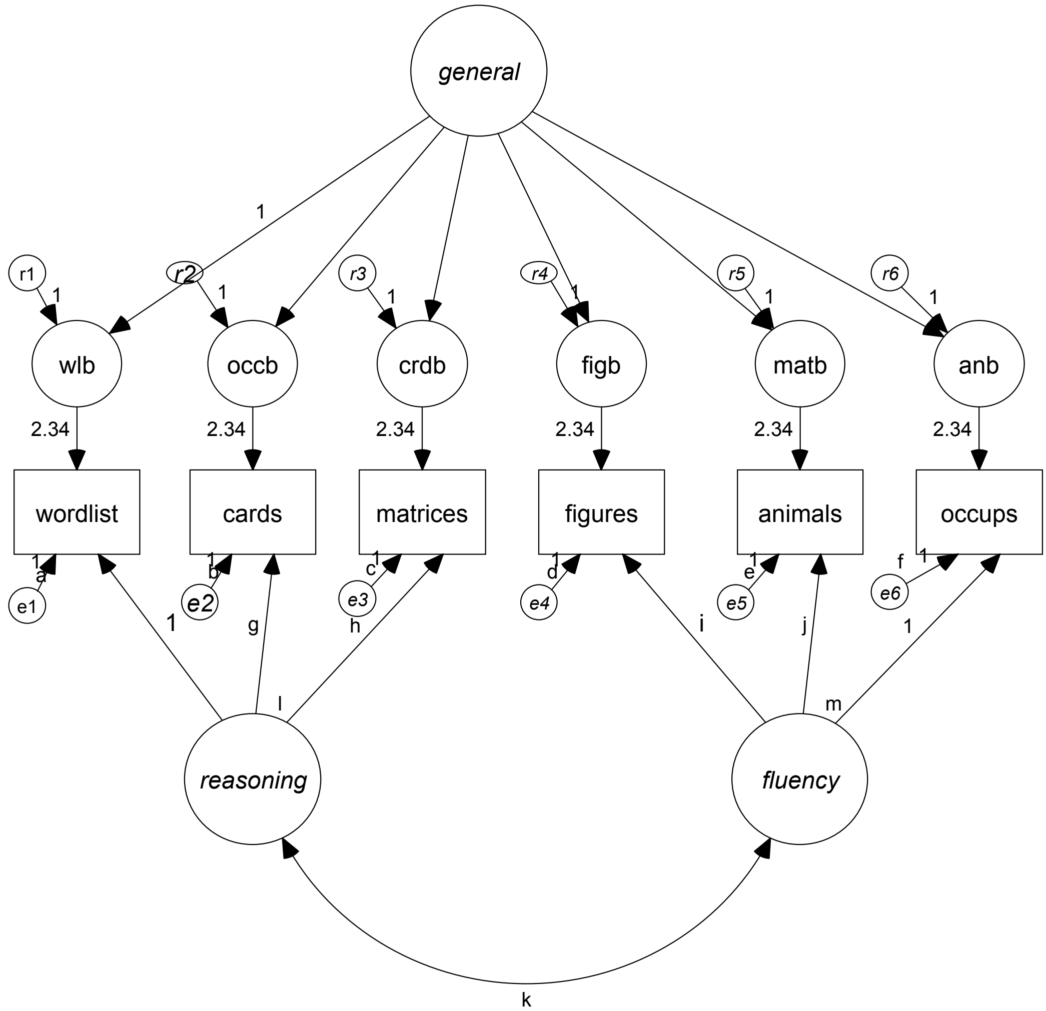


Table 3 presents the standardized parameter estimates for the multilevel factor model.

Table 3 Individual and family model, standardized factor loadings

	Individual		Family
	I	II	I
Word list	.19*	-	.86*
Cards	.33	-	.62
Matrices	.49	-	.95
Figures	-	.29	.51
Animals	-	.45	.91
Occupations	-	.31*	.45

Correlation between individual factors: 0.38; * = fixed;

The model in Table 3 suggests that on the family level, where the effects of the shared genetic and environmental influences are visible, one general (*g*) factor is sufficient to explain the covariances between the intelligence measures. On the individual level, where the effects of individual idiosyncratic influences are visible, we need two factors. The first factor could be interpreted as `reasoning,' and the second as `perceptual fluency.'

Specifying multilevel path models follows the general approach outlined above, but complications may arise because in multilevel path models it is not unusual to have variables that are only defined at the group level, such as 'group size,' or that have virtually no variance at the group level, such as 'sex.' In these cases, the number of variables in the between-groups and the within-groups matrix are not the same. This is not a fundamental problem, and standard SEM software can still be used. However, some SEM software (e.g., AMOS, EQS) handles this with ease, while other programs (e.g., LISREL, LISCOMP) require an ingenious setup to solve this problem.

4. Latent curves and multilevel regression models

An interesting model for longitudinal data with fixed occasions is the latent curve model, also known as the latent growth model. In the latent curve model, the time dimension is incorporated in the specification of the latent variables in the structural model. Consecutive measurements are modeled by a latent variable for the intercept of the growth curve, and a second latent variable for the slope of the curve (c.f. Meredith & Tisak, 1990; Muthén, 1991; Willet & Sayer, 1994, Maccallum & Kim, this volume).

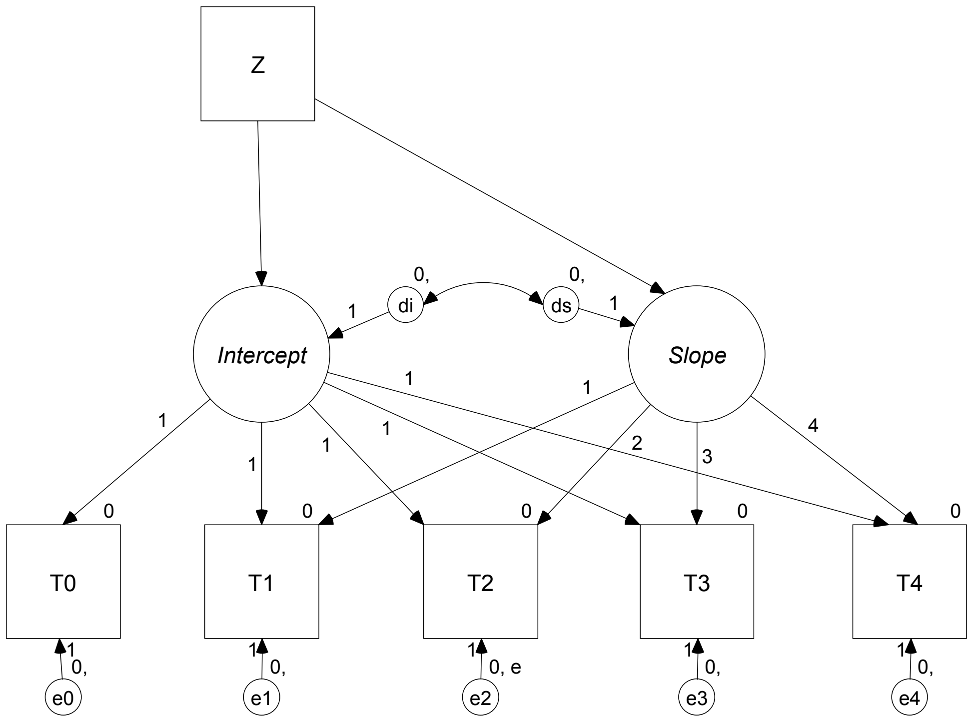
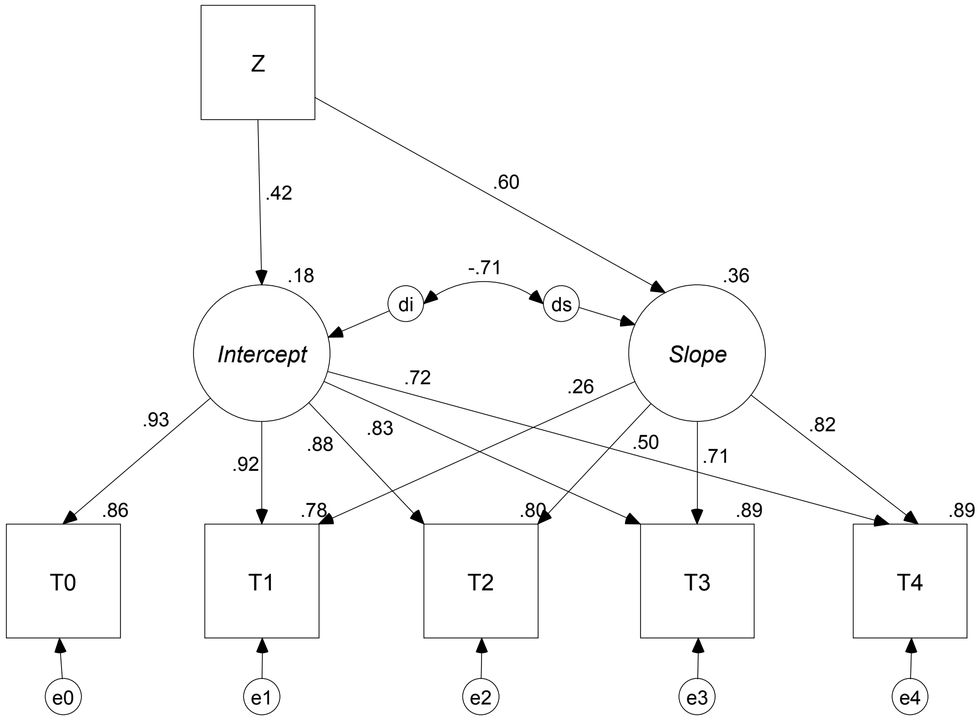


Figure 2 shows the path diagram of a simple latent curve model for our panel data, with five time points and one explanatory variable Z. In Figure 1, T0, T1, T2, T3 and T4 are the observations at the five time points. In the latent curve model the expected score at time point zero is modeled by a latent *intercept* factor. The intercept is constant over time, which is modeled by fixing the loadings of all time points on this factor to one. The latent slope factor is the slope of a linear curve, modeled by fixing the loadings of the five time points on this factor to 0, 1, 2, 3 and 4 respectively. Obviously, a quadratic trend would be specified by a third latent variable, with loadings fixed at 0, 1, 4, 9 and 16. The latent curve model includes the means of the variables and factors in the model. All the intercepts of the time points are fixed at zero, which makes the mean of the intercept factor an estimate of the common intercept. The mean of the slope factor is an estimate of the common slope. Individual deviations from the common intercept are modeled by the variance of the intercept factor, and individual deviations in the slope of the curve are modeled by the variance of the slope factor. Both the intercept and slope factor can be modeled by a path model including explanatory variables, in our example the one explanatory variable Z

The latent growth model is a random coefficient model, equivalent to the multilevel regression model with a random component for both the intercept and the slope of the time variable. To make it equivalent to the usual multilevel regression model, we must restrict the error terms *e* to have the same variance, and allow the residuals *d* to correlate. Figure 3 below shows the standardized estimates for our data, and Table 4 shows the unstandardized estimates, next to the estimates obtained previously with the multilevel regression model. Both sets of parameter estimates and standard errors are very close, about as close as we would expect if we had compared estimates produced by two different multilevel or SEM packages (cf. Kreft, de Leeuw & van der Leeden, 1994; Hox, 1995). Since these data are simulated, the population values are known. The last column in Table 4 presents the known population values. All estimates are close to the known population value, none are significantly different.

Table 4. Latent curve model and multilevel results (ML estimates, parameters and se's)

Model	latent curve	multilevel	population value
interc	44.1 (.51)	44.0 (.51)	44
time	4.9 (.14)	5.0 (.14)	5
Z	1.5 (.24)	1.5 (.14)	1.4
Z*time	.61 (.07)	.62 (.07)	.67
σ^2_e	12.4	11.9 (.69)	12
σ^2_{int}	45.1 (5.26)	45.3 (5.26)	47
σ^2_{time}	2.9 (.41)	2.7 (.40)	3.2
$\sigma_{int*time}$	-8.0 (1.28)	-7.7 (1.25)	9.0
r_{i*t}	-.71	70	-.73
	χ^2 33.8 df 13	deviance 6032	



5. Handling missing data

An often-cited advantage of multilevel analysis of longitudinal data is the ability to handle missing data (Bryk & Raudenbush, 1992; Snijders, 1996; Maas & Snijders, 1997). More accurately, this refers to the ability to handle models with varying time points. In a fixed occasions model, observations may be are missing because at some time points respondents were not measured. In MANOVA, the usual treatment of missing time points is to remove the case from the analysis. Multilevel regression models do not assume equal numbers of observations, or even fixed time points, so respondents with missing observations pose no special problems here. However, this advantage of multilevel modeling does not extend to missing observations on the explanatory variables. If explanatory variables are missing, the usual treatment is again to remove the case from the analysis. In structural equation modeling, missing observations can best be handled by direct estimation, where the algorithm estimates the model using all available data on all cases (Arbucle, 1996; Wothke, this volume). At least two SEM programs include direct estimation with missing values: AMOS (Arbucle, 1995) and MX (Neale, 1994).

Little and Rubin (1987, 1989) distinguish between data that are missing completely at random (MCAR) and data that are missing at random (MAR). In both cases, the failure to observe a certain data point is assumed to be independent of the unobserved (missing) value. With MCAR data, the missingness is completely independent of all other variables as well, while with MAR data, the missingness may depend on other variables. For a more detailed treatment of missing data problems see Graham and Hofer (this volume). By way of example, we analyze two data sets with missing data. The first data set is the longitudinal Rogosa/Saner data analyzed presented, with seven percent of the values randomly deleted. This produces an MCAR data set. In the second data set, we have added to this MCAR process a MAR process of panel attrition.

Dependent on the value of the explanatory variable Z, about five percent of the data points T0 to T1 was made missing. Next, if a data point was missing, all following time points were made missing too, to simulate panel attrition. Both data sets are analyzed using standard multivariate MANOVA, multilevel regression analysis, and direct estimation of a latent curve model. The MANOVA results are discussed in the text, the results of the multilevel regression and the latent curve model are in Table 5.

Table 5. Multilevel regression (MR) and latent curve modeling (LCM) on MCAR and MAR data set.

Data	MCAR		MAR	
Model	MR	LCM	MR	LCM
interc	44.2 (.52)	44.2 (.50)	43.7 (.53)	44.1 (.52)
time	4.9 (.14)	4.9 (.14)	4.6 (.15)	4.8 (.15)
Z	1.4 (.24)	1.5 (.24)	2.5 (.20)	1.5 (.25)
Z*time	.61 (.07)	.58 (.07)	.58 (.08)	.55 (.07)
σ^2_e	12.4 (.79)	12.4 (.69)	12.5 (.85)	12.0
σ^2_{int}	40.8 (5.16)	42.0 (5.19)	40.9 (5.32)	42.4 (5.4)
σ^2_{time}	2.5 (.41)	2.6 (.42)	2.5 (.44)	2.7 (.46)
$\sigma_{int*time}$	-6.6 (1.24)	-7.1 (1.27)	-8.8 (1.32)	-7.4
r_{i*t}	-.66	-.68	-.67	.6
% data	73%	93%	64%	84%

Table 5 shows, that if the data are MCAR, all models lead to estimates that are very close to each other and to the known population values that are given in Table 4. This is as expected, because when data are missing completely at random all three methods used are unbiased. However, they are not equally efficient. The last line in Table 5 shows the percentage of data points that are used in the analysis. In the full data set, we have 6 observations for each case, making a total of 1200 data points. In MANOVA, all incomplete cases are removed, which leaves 126 cases or 63% of the original observations. In the multilevel analysis, all cases with a missing observation on the explanatory variable *Z* are removed. For cases with a missing observation on one of the time points, only that specific time point is removed from the data set. This leaves a data set with 866 observations for 186 persons, or 73% of the original number of observations. Finally, the direct estimation method for the latent curve model uses all available data points, or 93% of the original number of cases. However, with these sample sizes and these strong effects, all estimates and standard errors are very similar, and we would draw the same conclusion, irrespective of the method used.

This is not true for the MAR data set. The MANOVA uses 116 or 58% of the original data, multilevel regression 64%, and direct estimation of the latent curve model 84%. The effects for measurement occasion and the covariate Z are again significant in all analyses, including MANOVA. The multilevel regression model estimates the slope for the time as 4.63, and latent curve model estimates the slope as 4.83, both reasonably close to the population value of 5.0. In this example, the multilevel regression performs not quite as well as the latent curve approach, probably because cases with the predictor Z missing had to be deleted here. Simple MANOVA on the observed means does not perform well at all. It is instructive to compare the means presented in Table 6. These are the means at the five time points: in the first column under the casewise deletion scheme that MANOVA uses, in the second column as predicted by the full multilevel regression model, and in the third column as predicted by the latent curve model. Again, the latent curve model performs the best, the multilevel regression somewhat worse, and the MANOVA procedure worst. This is as expected, because the data are missing at random (MAR). Manova on the complete cases, which assumes that data are missing completely at random (MCAR) shows a clear bias. The other two methods, that both only assume MAR, produce very acceptable estimates.

Table 6. Means for different treatments of missing observations, MAR data

time point	population	observed (full deletion)	predicted (multilevel)	predicted (latent curve)
0	44	43.5	43.7	44.1
1	49	46.9	48.4	48.9
2	54	52.5	53.0	53.7
3	59	57.6	57.6	58.5
4	64	62.1	62.2	63.3

6. Concluding comments

For grouped data, both MANOVA and multilevel regression models lead to similar estimates. If there are a large number of groups, more than two nested levels of grouping or explanatory variables at various levels, multilevel modeling is much more flexible than MANOVA. This also holds when the grouping factors are not nested but crossed; Goldstein (1994, 1995) describes extensions of the multilevel model to cross-classified data, which can be modeled using standard multilevel software.

For longitudinal data with fixed occasions, both multilevel regression and latent curve analysis produce very similar estimates. There are, however, differences in the range of applications. In latent curve modeling, each time point is represented by an additional observed variable, while in multilevel regression each time point is an additional observation on the same dependent variable. Multilevel regression analysis is clearly more elegant with a large number of varying measurements. In multilevel regression, it is also simple to add additional levels to the model. On the other hand, latent growth curves can be used in applications where the rate of growth (the slope) is used as a *predictor* in a more complicated path model. In multilevel regression, this is not possible. The last example also suggests that handling missing observations using the direct estimation method is more accurate than the simple multilevel approach.

Multilevel structural equations' modeling is mainly an alternative when we have a complex factor or path model and a large number of groups. Although more than three levels are possible, this is not really attractive; even with two levels the setups can become quite complicated (cf. Hox, 1995; Kaplan & Elliott, 1996). An interesting option is to combine latent growth models with multilevel structural equation models, as in Muthén (1997), The result is a very powerful modeling tool, unfortunately, again with complicated setups.