Sample Sizes for Two-Group Second Order Latent Growth Curve Models

Linda Wänström
Stockholm University

Second order latent growth curve models (Duncan & Duncan, 1996; McArdle, 1988) can be used to study group differences in change in latent constructs. We give exact formulas for the covariance matrix of the parameter estimates and an algebraic expression for the estimation of slope differences. Formulas for calculations of the required sample size are presented, illustrated and discussed. They are checked by Monte Carlo simulations in Mplus and also by Satorra and Saris' (1985) power approximation techniques for small and medium effect sizes (Cohen, 1988). Results are similar across methods. Not surprisingly, sample sizes decrease with effect sizes, indicator reliabilities, number of indicators, frequency of observation, and duration of study. The relative importance of these factors is also discussed, alone and in combination. The use of the sample size formula is illustrated in a hypothetical example.

Key words: sample size; latent growth curve models; reliability.

Acknowledgement: The work on this paper was supported by the Bank of Sweden Tercentenary Foundation, grant K2000-5063.

1. Introduction

Individual changes in a population are often more interesting than changes of the population as a whole. One may want to look at the individual changes in one or several variables, the variation around this mean change, as well as differences between groups of people or the interplay between variables. Longitudinal studies differ from cross-sectional studies foremost because units are followed over time so that the correlation within units over time can be explored. Many statistical methods[1] are available for longitudinal data e.g. dependent *t* tests, repeated measures ANOVA, multivariate ANOVA, autoregressive models and random effect models. The choice of the most suitable technique depends on the research question, the structures of the data, and the underlying theory (e.g. Curran & Bollen, 2001; Diggle et al., 2002; Ferrer & McArdle, 2003).

This paper focuses on a longitudinal approach called latent growth curve (LGC) modeling. It may be a good choice when there are relatively few

[1] See e.g. Diggle, Heagerty, Liang, and Zeger (2002) for descriptions of longitudinal models.

measurements on a large number of individuals (Sayer & Cumsille, 2002). In LGC modeling each individual is characterized by his or her personal development, which in turn is described by a number of individual latent factors plus a residual term. The development of each individual might be around a straight line, $g_t = \pi_1 + \pi_2 x_t$, where there are two factors, an "intercept": π_1 and a "slope": π_2. Thus, in a population there are as many lines as there are individuals. The population mean of the intercepts and slopes are $E(\pi_1, \pi_2) = (\alpha_1, \alpha_2)$. The individual's variation around his or her curve is described by residual terms ζ_t. We will examine the simple case in which they are normal with variance $\sigma^2_{\eta t}$,: $\eta_t = g_t + \zeta_t$. Often, the exact level, η_t, of the individual at time *t*, cannot be observed directly but only through indicators having a certain distribution. In a simple case the observed indicators, $y_{tk} = \eta_t + \varepsilon_{tk}$, are unbiased, independent normal and have variances $\sigma^2_{\varepsilon tk}$. This simple case is illustrated in Figure 1 with 3 indicators and 4 equidistant measurement occasions.

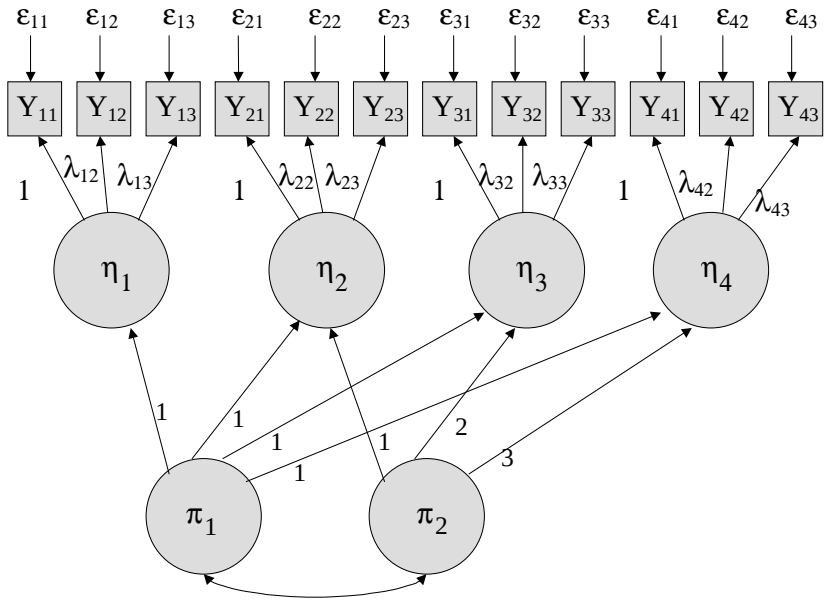

Figure 1. Path diagram of an LGC model with indicators. Circles indicate latent variables, squares indicate observed variables, one-headed arrows indicate regression coefficients, and double headed arrows indicate covariances.

The LGC model has been widely used in the social, developmental, and educational literature to study change. It makes it possible to examine changes in longitudinal data, to model different error structures and to use various types of structural models. For example, Lance, Vandenberg, and Self

(2000) used an LGC model to study change in employee attachment for new bank employees. Stoel, Peetsma, and Roelveld (2003) modeled changes in language ability, school investment, and self-confidence for students[2], whereas Hong and Ho (2005) used students' tests in reading, math, and science obtained in 1988, 1990, and 1992 to study their achievements[3]. The LGC model may also be used to compare the growth of multiple populations, such as males and females, children with married or divorced parents, or randomized groups in intervention settings analyses (e.g. Muthén & Curran, 1997).

In this paper we will examine the sample sizes required to detect differences in latent factors between groups. In many research situations, the length of a confidence interval may be more important than the power of a test. Our methods are easily adapted to both cases. We give an algebraic expression for the variance of the estimated slope in a group and for slope differences that extends work by Raudenbush and Liu (2001). This expression is used to investigate required sample sizes. In addition, sample sizes are examined using power approximation techniques developed by Satorra and Saris (1985). A sample of the results are also checked against simulations. Effects of the number of indicators and their reliabilities are investigated in situations with different frequencies of observation and study duration. Results from the formula are illustrated numerically in tables, graphs and in an example.

2. Previous work

2.1. Latent growth curve (LGC) models

LGC models were developed by Meredith and Tisak (1990), although they originate from work by Rao (1958) and Tucker (1958), and they have been further developed by others, including McArdle (1988) and Willett and Sayer (1994). They are latent variable models and can be thought of as confirmatory factor analysis models where longitudinal behavior, e.g. growth or decline, is captured and tested through latent variables as well as latent "errors" at each time point (Meredith & Tisak, 1990; Muthén & Curran, 1997). More thorough descriptions of LGC models can be found, for example, in Duncan, Duncan, Strycker, Li, & Alpert (1999), Curran and Muthén (1999), McArdle (1988), McArdle and Bell (2000), Meredith and Tisak (1990), Muthén and Curran (1997), Raykov (2000), and Willet and Sayer (1994).

As mentioned previously, the construct of interest can often not be measured directly, but instead through indicators (see Figure 1). Second order LGC

[2] They used the longitudinal PRIMA cohort study.
[3] They used the NELS data.

models (Duncan & Duncan, 1996; McArdle, 1988) can be used to study systematic change in latent constructs, and this provides advantages over first order models (e.g. Sayer & Cumsille, 2002). First, the error term related to each occasion can be partitioned into measurement error and error related to a unique factor (indicator residual), and to error related to deviations from the growth curve (occasion residual). Second, assumptions of factorial invariance over time and across groups can be explicitly tested. Descriptions of second order LGC models can be found, for example, in Sayer and Cumsille (2002).

LGC modeling is an approach to growth curve modeling, and most LGC models are also multilevel models (for descriptions of multilevel models, see e.g. Bryk & Raudenbush, 1992; Goldstein, 2003; Longford, 1993) under certain restrictions. These models have fixed effects, e.g. the average intercept and the average slope corresponding to the means of the latent factors, and random effects, e.g. the deviations from the average intercept and slope, and deviations from the individual growth trajectories. They are therefore also called mixed models or random effect models. For descriptions of similarities and differences between LGC modeling and multilevel modeling, see e.g. Raudenbush (2002), Hox (2000), Hox and Stoel (2005), MacCallum (2000), and Stoel, van Der Wittenboer and Hox (2003).

Because LGC models are latent variable models, parameters can be estimated through Structural Equation Modeling (SEM) (Bollen, 1989; Jöreskog & Sörbom, 1979, 1993; Loehlin, 1992) methods (e.g. Willet & Sayer, 1994). Regular SEM software, such as LISREL, Mx, Calis, Amos, EQS or Mplus, as well as standard statistical packages such as S+, R, or Stata (GLM or GLAMM) or multilevel packages such as MLwiN or HLM can be used in estimation of parameters.

2.2 Sample size determination

When planning a longitudinal study it is desirable to know the number of participants needed. This might be particularly important if the planned intervention is costly. Several authors have investigated sample size and power for detection of group differences in longitudinal models (e.g. Liu and Liang, 1997; Liu, Shih, & Gehan, 2002; Rochon, 1991; 1998; Shieh, 2003). In the multilevel literature, Raudenbush and Liu (2000; 2001) presented closed form formulas for variances used in power calculations of group differences in means and trends in longitudinal models based on calculating the noncentrality parameter of the *F* distribution. Effects of duration of study, frequency of observation, and effect size, among other factors, could be evaluated using these formulas. In the medicine literature, closed form formulas have also been presented for related models (e.g. Liu and Liang, 1997; Liu, Boyett, and Xiong (2000); Liu, Shih, and Gehan, 2002; Rochon, 1991).

Common approaches for studying power and sample sizes for LGC models include Satorra and Saris (1985) power approximation technique, and Monte Carlo simulations. Muthén and Curran (1997) and Curran and Muthén (1999) showed how to use Satorra and Saris' (1985) technique to investigate the power to detect non-zero differences in mean slopes as well as interactions in LGC models. Hertzog, Ghisletta, Lindenberger, and Oertzen (2006) used the same approximation techniques to study power to detect covariances between slopes of parallell change processes. Muthén and Muthén (2002) showed how Monte Carlo simulations could be used to obtain sample sizes for LGC and SEM models. Fan and Fan (2005) and Fan (2003) used simulations to study power to detect linear growth in single group studies and group differences in intercepts and slopes respectively. Duncan, Duncan, Strycker, and Li (2002) provide a discussion of power estimation techniques for multiple groups within LGC- and other models.

Power and sample sizes have also been evaluated for other latent variable models. For example, MacCallum and colleagues (MacCallum, Browne, & Cai, 2006; MacCallum, Browne & Sugawara, 1996; MacCallum & Hong, 1997) used the RMSEA- and other indices encountered in SEM analyses to estimate the noncentrality parameter needed to compute power. Hancock (2001) investigated power and sample sizes for detecting differences in single latent constructs.

Previous studies on the power of LGC and related models have found positive effects of sample size, effect size, and frequency of observation on precision and power (Fan & Fan, 2005; Muthén & Curran, 1997; Fan, 2003; Raudenbush & Liu, 2001). A long duration of the study is also positive (Muthén & Curran, 1997; Raudenbush & Liu, 2001), whereas missing data are negative (Muthén & Muthén, 2002). LGC methods have also been found better than univariate ANCOVA (Muthén & Curran, 1997; Curran & Muthén, 1999), ANOVA and dependent *t* tests, and highly more powerful than MANOVA (Fan & Fan, 2005; Fan, 2003). However, Fan (2003) suggested that ANOVA may be more powerful in detecting differences in intercepts.

All the LGC studies mentioned above focused on first order models, i.e. models where the measurement errors and the individual residuals are non-distinguishable. Second order LGC models are used more and more in the social sciences. These models are often preferred to first order models, whenever applicable, due to their variance decomposition properties and possibilities of factorial invariance testing (e.g. Sayer & Cumsille, 2002). A researcher with limited resources benefits from knowing whether to focus on getting more participants, using more indicators, using fewer but more reliable indicators, or measuring the participants on more occasions.

3. The model

3.1 One group

Consider a group with n individuals. Each person is observed at T occasions. At each occasion K indicators are measured. Let y_{itk} denote the observed value of individual i, of indicator k at occasion t. Define x_t to be the elapsed time from start to the t:th measurement occasion (x_1=0). Raudenbush and Liu (2001) have investigated the effect of study duration and frequency of observation. Following their definitions, we define f, the frequency of observation, to be the number of observations per time unit: $f = (t - 1) / x_t$ for equally spaced occasions. The total study time (the duration) is $D=x_T$, which in this case equals $D = (T – 1)/f$. If the measurements are made every second year, the frequency will be $f = 1/2$, and $x_t = 0, 2, 4, \ldots, 2T\text{-}2$ for $t = 1, 2, 3, 4, \ldots, T$, and $D = 2T – 2$.

The LGC model[4] consists of a measurement part and a structural part. The structural part describes the development of the latent variables and the measurement part the relation between the measurements and the latent variables. Every individual is characterized by J latent factors, $\Pi=(\pi_1, \ldots, \pi_J)'$, which are assumed to come from a normal distribution with mean vector $\alpha_p=(\alpha_1, \ldots, \alpha_J)'$ and covariance matrix Σ_π. The structural part[5] relates an individual's true latent levels at each time to his latent factors

$$\eta_h = \alpha_h + B_h\Pi + \zeta_h, \tag{1}$$

where η_h is a column vector with the individual's latent level at the T measurement occasions, α_h is an intercept vector, B_h is a matrix with TxJ factor loadings and ζ_h is a random residual vector with mean 0 and covariance matrix Σ_η.

Often Π and η_h are stacked above each other. In that case one can write the structural part as

$$\begin{pmatrix}\eta_h\\ \Pi\end{pmatrix} = \begin{pmatrix}\alpha_h\\ \alpha_p\end{pmatrix} + \begin{pmatrix}0 & B_h\\ 0 & 0\end{pmatrix}\begin{pmatrix}\eta_h\\ \Pi\end{pmatrix} + \begin{pmatrix}\zeta_h\\ \zeta_p\end{pmatrix}$$

This expression may be written (see e.g. Sayer and Cumsille (2002))

[4] We will focus on the second order model in which indicators of latent variables are observed at each occasion.

[5] See e.g. Jöreskog and Sörbom (1979) for a general SEM model with a measurement and a structural part.

$$\eta = \alpha + B\eta + \zeta \qquad (2)$$

where α is a $T+J$ vector of factor means, B is a $(T+J)\times(T+J)$ matrix of factor loadings and ζ is a $T+J$ vector of latent residuals which are normal with zero mean and covariance matrix Ψ. The column vector η now contains the $T+J$ latent factors.

In order to describe the model in Figure 1 we have $\alpha = (0,0,0,0,\alpha_1,\alpha_2)'$

$$B_h = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{pmatrix}', \qquad \Sigma_\pi = \begin{pmatrix} \sigma^2_{\pi_1} & \sigma_{\pi_1\pi_2} \\ \sigma_{\pi_1\pi_2} & \sigma^2_{\pi_2} \end{pmatrix} \quad and \quad \Sigma_\eta = diag(\sigma^2_{\eta_t})$$

The measurement part of the model relates the observations to the latent variables

$$Y = \tau + \Lambda\eta + \varepsilon, \qquad (3)$$

where Y is the column vector with the KT observed measurements, the column vector τ contains the corresponding KT measurement intercepts and Λ is a $KT\times(T+J)$ matrix containing the factor loadings λ. The vector ε contains KT normal error terms with zero mean and covariance matrix Θ. One factor loading for each latent factor, η_t, can be fixed to 1 to set the scale. The corresponding intercept, τ, can be fixed to 0 for identification purposes. In the model corresponding to Figure 1, $\tau = (0, \tau_{12}, \tau_{13}, 0, \tau_{22}, \tau_{23}, 0, \tau_{32}, \tau_{33}, 0, \tau_{42}, \tau_{43})'$ and a typical row of Λ is $(0, \lambda_{22}, 0,0,0,0)$.

3.2 Analysis of the one group model

In our case, (3) can be simplified to

$$Y = \tau + \Lambda_h\eta_h + \varepsilon, \qquad (4)$$

since we cannot measure the basic latent factors directly, only the individual's latent level at certain time points. The matrix Λ_h now contains KTxT elements. Inserting (1) into (4) we get

$$Y = \tau + \Lambda_h(\alpha_h + B_h\Pi + \varsigma_h) + \varepsilon. \qquad (5)$$

The expected value is thus $\tau+\Lambda_h\alpha_h+\Lambda_hB_h\alpha_p$ and the variance is $V = \Lambda_hB_h\Sigma_\Pi B_h'\Lambda_h' + \Lambda\Sigma_\eta\Lambda_h' + \Theta$. We cannot estimate both α_h and α_p separately. To solve this we will set $\alpha_h = 0$. The expected value is thus $\tau+\Lambda_hB_h\alpha_p$.

For all components to be estimable the total number of parameters in τ and α_p must be less than the dimension, *TK*. Some components in τ can be set to 0. We may then write the expected value $\Lambda_0(\alpha_p,\tau_0)'$ with $\Lambda_0=(\Lambda_h B_h, I)$. In the example of Figure 1, the number of independent parameters are $2+T(K-1)$.

If V has full rank, the vector $(\alpha_p,\tau_0)'$ can now be estimated, for each individual, with the well-known generalized least square estimate $(\alpha_p^*,\tau_0^*)' = (\Lambda_0'V^{-1}\Lambda_0)^{-1}\Lambda_0'V^{-1}Y$ with variance $(\Lambda_0'V^{-1}\Lambda_0)^{-1}$ assuming that all parameters are estimable. The estimate based on *n* individuals is the average of the *n* individual estimates and its variance is $(\Lambda_0'V^{-1}\Lambda_0)^{-1}/n$. If we know the covariance matrix we can now use standard statistical techniques to decide what sample sizes are needed to obtain a certain precision or power. If the covariance matrix is unknown the formulas hold approximately if the sample sizes are large enough to make good variance estimates. However, the required sample size is then expressed in terms of the unknown parameters and to use the formula in practise one must guess their value. This means that it is not so important to get exact values for small sample sizes since the size will be uncertain anyway.

3.3 Two or more groups

In several studies there may be two or more groups. For example, there may be one group of males and one group of females or there may be one or several treatment and control groups. We enter a superscript, *g*, to denote group. Equation (2) becomes[6]

$$\eta^{(g)} = \alpha^{(g)} + B^{(g)}\eta^{(g)} + \zeta^{(g)} \tag{6}$$

and equation (3) becomes

$$Y^{(g)} = \tau^{(g)} + \Lambda^{(g)}\eta^{(g)} + \varepsilon^{(g)} \tag{7}$$

where the parameters may differ between groups. If we have two independent groups with no common parameters it is easily seen that the best estimate of the difference between them is the difference of the corresponding group parameters with variance

$(\Lambda_0^{(1)\prime}V^{(1)-1}\Lambda_0^{(1)})^{-1}/n^{(1)} + (\Lambda_0^{(2)\prime}V^{(2)-1}\Lambda_0^{(2)})^{-1}/n^{(2)}$.

Muthén and Curran (1997) and Curran and Muthén (1999) presented a way of testing the equality of slope parameters by introducing an "added curve factor due to treatment" (Muthén & Curran, 1997; p. 378). A model with an

[6] This is based on the general latent variable multiple population formulation (see e.g. Bollen, 1989; Jöreskog & Sörbom, 1979; Muthén & Curran, 1997).

added curve factor is depicted in Figure 2 for the second group (e.g. the treatment group). The first group model is as in Figure 1. All non-zero parameters can be constrained to have the same values in both groups but there may be some extra relations in the second group. In Figures 1 and 2, π_2 is constant between groups but a new growth factor, π_3, with mean α_3 and variance $\sigma^2_{\pi_3}$, is added to the second group that corresponds to $\pi_2^{(2)} - \pi_2^{(1)}$ in our formulation.

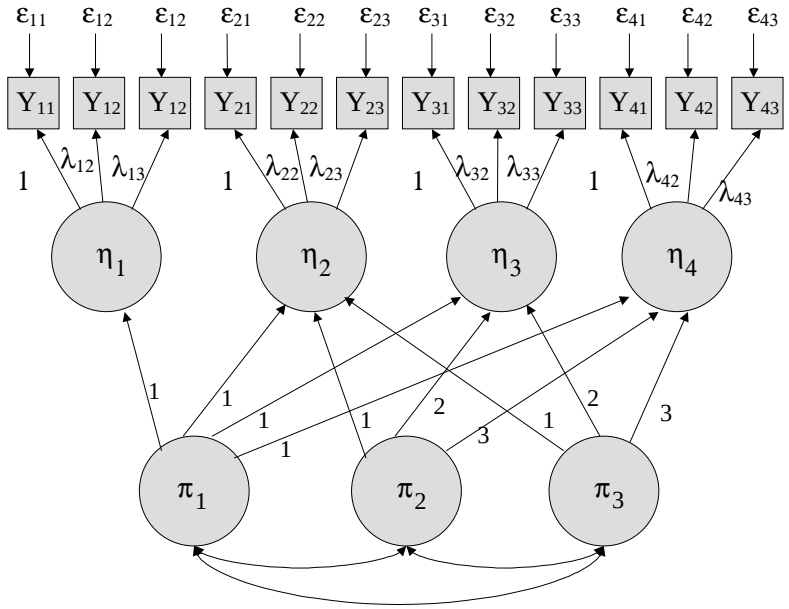

Figure 2. Second order LGC model with an added curve factor (Curran & Muthén, 1999; Muthén & Curran, 1997) for one of the groups.

In many situations it is reasonable to assume that some parameters are common to the groups and that other parameters differ between them. See the Appendix for how to find the variance of least squares estimates when some parameters are equal across groups, using matrix forms.

4. Variance formulas and sample size

4.1 Different group means

From now on we restrict ourselves to the simple case in Figure 1 with two groups and two factors, intercept and slope. We have already expressed the estimates and variances in matrix terms. In this special model they can be

expressed in ordinary algebraic expressions. The parameters of interest are the slopes, $\pi_2^{(1)}$ and $\pi_2^{(2)}$ with means, $\alpha_2^{(1)}$ and $\alpha_2^{(2)}$, in the two groups.

In a first order model, the minimum variance, unbiased estimate of the mean for one group and its variance are

$$\hat{\pi}_1 + \bar{x}\hat{\pi}_2 = \bar{Y} = \sum_{t=1}^{T} Y_t / T, \qquad Var(\bar{Y}) = \frac{1}{n}\left(\frac{\sigma_Y^2}{T} + \sigma_{\pi_1}^2 + \bar{x}^2 \sigma_{\pi_2}^2 + 2\bar{x}\sigma_{\pi_1\pi_2} \right) \tag{8}$$

where σ_Y^2 is the common residual variance of *Y* at each occasion, *T* is the number of measurements, and $\sigma_{\pi_1}^2$ and $\sigma_{\pi_2}^2$ are factor variances. The estimate and variance of the slope are

$$\hat{\pi}_2 = \frac{\sum_{t=1}^{T} (x_t - \bar{x})(Y_t - \bar{Y})}{\sum_{t=1}^{T} (x_t - \bar{x})^2}, \qquad Var(\hat{\pi}_2) = \frac{1}{n}\left(\frac{\sigma_Y^2}{\sum_{t=1}^{T} (x_t - \bar{x})^2} + \sigma_{\pi_2}^2 \right) \tag{9}$$

where x_t is the elapsed time since the start of the study to the *t*:th measurement occasions. The first term inside the brackets is the conditional variance of the least squares estimate of the individual slope. In cases of equally spaced measurements at times 0, 1, …, *T* - 1, this can also be written $12(T-1)\sigma_Y^2 / D^2 T(T+1)$ (Friedman, Furberg & DeMets, 1985; see also Raudenbush & Liu, 2001) where *T* is the number of occasions and *D* is the duration. We can also write this as $(12 f^2 \sigma_Y^2 / (T^3 - T)$ where *f* is the frequency of observation. The second term inside the bracket of (9) is variance across individuals. Finally, for two groups with no common parameters the estimate of the mean slope difference is just the difference $\hat{\pi}_3 = \hat{\pi}_2^{(1)} - \hat{\pi}_2^{(2)}$. Its variance is the sum of the two group variances.

The formulas above can be extended to incorporate a second order model with several measurements on each occasion. Assume that we have *K* unbiased and independent indicators with equal variances per factor $\eta_t = g_t + \zeta_t$. The mean of the indicators for individual *i* at occasion *t* has the following conditional variance

$$\hat{\eta}_{ti} = \sum_{k=1}^{K} Y_{kti} / K, \qquad Var(\hat{\eta}_{ti} \mid \pi_1, \pi_2) = \frac{\sigma_\varepsilon^2}{K} + \sigma_{\eta_t}^2 \tag{10}$$

where the first variance term is the measurement variance and the second term is the residual variance around the individual line. Assuming equal

residual variances at each occasion, we can substitute $\sigma_Y^2 = \sigma_\eta^2 + \sigma_\varepsilon^2 / K$ into (8) and (9) above.

4.2 Equal intercepts in groups

As mentioned previously, some factors may be equal across groups. E.g. when the assignment to groups is random and done at start (x_1=0), it is reasonable to assume that the starting values (intercepts) are equal. In that case the difference between the two group means is also an estimate of the difference in slope means. Combining the two estimates in an optimal way we get

$$\hat{\pi}_3 = \frac{\left(\frac{\bar{Y}^{(2)} - \bar{Y}^{(1)}}{\bar{x}}\right)\left(Var\left(\hat{\pi}_2^{(2)} - \hat{\pi}_2^{(1)}\right) - Cov\right) + \left(\hat{\pi}_2^{(2)} - \hat{\pi}_2^{(1)}\right)\left(Var\left(\frac{\bar{Y}^{(2)} - \bar{Y}^{(1)}}{\bar{x}}\right) - Cov\right)}{Var\left(\hat{\pi}_2^{(2)} - \hat{\pi}_2^{(1)}\right) + Var\left(\frac{\bar{Y}^{(2)} - \bar{Y}^{(1)}}{\bar{x}}\right) - 2Cov} \quad (11)$$

where $Cov = Cov((\hat{\pi}_2^{(2)} - \hat{\pi}_2^{(1)}), (\bar{Y}^{(2)} - \bar{Y}^{(1)}) / \bar{x})$. Assuming equal variances and sample sizes in the groups, the variance of this new estimator can be shown to be

$$Var(\hat{\pi}_3) = \frac{2}{n}\left(V_2 + \sigma_{\pi_2}^2 - \frac{\left(V_2 \bar{x} - \sigma_{\pi_1\pi_2}\right)^2}{V_1 + \sigma_{\pi_1}^2 + V_2 \bar{x}^2}\right) \quad (12)$$

where

$$V_1 = \frac{\sigma_\eta^2 + \sigma_\varepsilon^2 / K}{T}, \quad (13)$$

$$V_2 = \frac{\sigma_\eta^2 + \sigma_\varepsilon^2 / K}{(T^3 - T)/12 f^2}, \quad (14)$$

and σ_η^2 is the common residual variance of the first order factors, σ_ε^2 is the common residual variance of the indicators, $\sigma_{\pi_1}^2$ and $\sigma_{\pi_2}^2$ are the variances of the latent intercept and slope, and $\sigma_{\pi_1\pi_2}$ is their covariance. If $\sigma_{\pi_1\pi_2} = 0$, formula (12) can also be written

$$Var(\hat{\pi}_3) = \frac{2}{n}\left(\frac{1}{\frac{1}{V_2} + \frac{\bar{x}^2}{V_1 + \sigma_{\pi_1}^2}} + \sigma_{\pi_2}^2 \right). \quad (15)$$

4.3 Relation between variance and sample size

The standard sample size formula for tests under normality with known variance (see e.g. Desu & Raghavarao, 1990) is

$$n = \frac{\left(\lambda_{\alpha/2} + \lambda_{\beta}\right)^2 2\sigma^2}{e.s.^2} \quad (16)$$

where n is the sample size per group, α is the significance level, 1-β the power, $\lambda_{\alpha/2}$ and λ_{β} quantiles of the normal curve, $2\sigma^2/n$ the variance of the estimate and *e.s.* the effect size. We can use formula (16) for LGC models, once $2\sigma^2$ has been determined. Several authors have presented formulas for correlated observations where (16) is adjusted by a factor containing ρ, the correlation among the measurements (see e.g. Liu & Liang, 1997; Diggle et al., 2002). For LGC models, the numerators of (12) or (15) can be used for $2\sigma^2$ in (16).

As mentiond previously, the length of a confidence interval may be more important than the power of a test. One may want to find a 95% confidence interval for the mean group difference that is shorter than a given value *h*. Equation (16) easily translates into the required group sample sizes for a certain length of a confidence interval. All we do is to substitute $\lambda_{\alpha/2} + \lambda_{\beta}$ by $2\lambda_{\alpha/2}$ and e.s. by *h*.

4.4 Reliability effects

As mentioned previously, one aim of this paper is to investigate sample sizes for second order LGC models with varying number of indicators and indicator reliabilities. Since the reliability often varies with time[7] we will use the reliability at t = 1. The indicator residual variance (error variance) and the indicator reliability are related to each other, and knowing one, in addition to knowing the total factor variance, Var(η_1), we can compute the other. The error variance[8] can be obtained from the reliability R_k

[7] Previous studies have also focused on reliability at time 1, see e.g. Hertzog et al. (2006).
[8] This formula follows from $R_k = \lambda_k^2 Var(\eta_1) / \left(\lambda_k^2 Var(\eta_1) + \sigma_{\varepsilon k}^2\right)$ which is related to the definition of reliability from classical test theory (e.g. McDonald, 1999).

$$\sigma_{\varepsilon k}^2 = \frac{\lambda_k^2 Var(\eta_1)(1-R_k)}{R_k} \quad (17)$$

where λ_k^2 is the *k*th indicator loading. For simplicity, we will assume that all *K* reliabilities are equal and we may drop the subscript *k*. The error variance, σ_ε^2, corresponding to *R*, can then be used in (13) and (14).

4.5 Standardized effect size

Equation (16) entails the raw effect size, e.g. the true difference in slope means across groups, however it might be beneficial to know the standardized effect size as well in order to compare models. Cohen (1988) provided guidelines for effect sizes in the social sciences. According to him, a standardized effect size, *d*= .2, can be considered small, whereas *d* = .5 and *d* = .8 are medium and large, respectively. Consider formula (16). It is easy to see that the sample sizes under large, medium, and small effect sizes relate to each other as $.8^{-2} = 1.5625$ to $.5^{-2} = 4$ to $.2^{-2} = 25$, since we assume that the variance is known or well estimated. For example, we can go from small (*d* = .2) to large (*d* = .8) effect size by dividing our obtained sample size by 25/1.5625=16.

Effect size can be standardized in different ways for longitudinal data. Following Curran & Muthén (1999) and Muthén & Curran (1997) we will define the standardized effect size *d* as the difference in means between the groups at the last time point divided by the standard deviation at that time point.

$$d = \frac{\alpha_3 D}{\sqrt{\sigma_{\eta_t}^2 + \sigma_{\pi_1}^2 + D^2\sigma_{\pi_2}^2 + 2D\sigma_{\pi_1\pi_2}}} \quad (18)$$

where α_3 is the mean of π_3, and *D* is the duration of the study.

Other standardizations are also possible. For example, Fan (2003) considered both the standardized mean difference at the beginning and at the end of the study, and Raudenbush and Liu (2001) used an effect size of the group slope difference divided by the population standard deviation of the slope, $\alpha_3 \Big/ \sqrt{\sigma_{\pi_2}^2}$. See also Hancock (2001) for a discussion of standards for latent variable means, taking into account reliability of the indicators.

5. Alternative ways to decide sample size

5.1 Satorra and Saris' method

Satorra and Saris (1985) suggested a method for approximating the power of the likelihood ratio test often used in latent variable analyses. The likelihood ratio test statistic is asymptotically χ^2-square distributed under the null hypothesis and asymptotically non-central χ^2-square distributed under the alternative hypothesis. The technique is easy to use with the help of SEM software, and works well for large sample sizes and when the misspecification is small. The procedure includes specifying a model under an alternative hypothesis and computing the implied covariance matrix under this model. This can be done using SEM software by fixing all parameters as in the alternative hypothesis and specifying the sample size to be large. The obtained implied covariance matrix can then be used as an input matrix for analysis under the null hyothesis, i.e. an analysis model is specified with constrained, free, and fixed parameters (as usual in SEM analyses) however with the restriction of fixing the parameter(s) of interest to be as in the null hypothesis. The obtained non-zero χ^2-square value can be used as an approximation of the non-centrality parameter, and power of the test can be computed.

Descriptions of the Satorra-Saris method can be found, for example, also in Duncan et al. (1999; 2002) and in Muthén and Curran (1997). We compared their method to our method in some numeric examples. We used the model proposed by Muthén and Curran (1997) and Curran and Muthén (1999) with an added growth factor, π_3, with mean α_3 and zero variance. First we specified the model under H_1, using a non-zero effect size α_3. The implied means- and covariance matrix was then created using the statistical software Mplus (Muthén & Muthén, 1998-2006) by inputting identity matrices and zero mean vectors, specifying the sample sizes in each group to be large (1000 per group), and fixing all parameters to be equal to the values according to H_1. The resulting implied means- and covariance matrices were then checked for mistakes by estimating α_3 and making sure that all original parameter values were reproduced. The implied means- and covariance matrices were then analyzed according to H_0 (i.e. setting $\alpha_3 = 0$) with a certain sample size. Using the obtained χ^2-value as an approximate non-centrality parameter, the power for a test with 1 degree of freedom (one parameter fixed as in H_0) was computed. The power for various sample sizes was obtained, and the sample size that corresponded to a power of .8 was recorded.

5.2 Monte Carlo Simulations

Both the Satorra-Saris' (1985) technique and our formulas may not work as well for small sample sizes and unknown covariance matrix (e.g. Curran, Bollen, Paxton, Kirby & Chen, 2002; Satorra & Saris, 1985). To check the methods a sample of the analyses were simulated as well. For each simulation, 10,000 samples were generated and analyzed. Power was computed as the proportion of replications that the null hypthesis $\alpha_3 = 0$ was rejected at significance level .05, given that the population value was different from zero. All simulations were checked to make sure that parameter and standard error biases were no larger than 10% (5% for α_3) and that coverage was between .91 and .98 (see Muthén & Muthén, 2002).[9]

6. Numerical illustrations

The formulas in this paper can be used for any desired parameter values. For our illustrations we chose the basic model in Figure 1 with $T = 4$, $D = 3$, and $f = 1$. We then varied the number of indicators, the indicator reliability, the number of occasions, standardized effect size, duration, and frequency of observation. We used the significance level .05 and the power .8 throughout. Cohen (1988) recommended a power of .8 in the social sciences. We used the parameter values $\sigma_\eta^2 = 0.5$, $\sigma_{\pi_1}^2 = 0.5$, and $\sigma_{\pi_2}^2 = 0.1$, providing a commonly seen intercept/slope variance ratio (Muthén & Muthén, 2002). Unless otherwise stated, we have assumed that $\sigma_{\pi_1\pi_2} = 0$ and equal intercepts across groups $\alpha_1^{(1)} = \alpha_1^{(2)}$.

6.1 Comparisons across methods

Table 1 shows the required group sample sizes using the Satorra-Saris (1985) approximation method (top) and the formula (bottom). As shown, the Satorra-Saris sample sizes are slightly lower than those obtained from the formula for $d = .2$, and slightly higher for $d = .5$, but this discrepancy is rounding error due to the fact that Mplus uses only three decimals in computing the implied means- and covariance matrices.

[9] Only fairly large sample sizes were used. Had sample sizes been smaller (e.g. $n = 10$), biases might have been larger.

Table 1. Required Group Sample Sizes for .8 Power to detect Slope Difference

Method			d = .2			d = .5		
Satorra-Saris	R	K =	3	4	5	3	4	5
T = 4, D = 3, f = 1	1		322	322	322	53	53	53
	.9		330	328	326	54	54	53
	.6		367	356	350	60	58	57
	.3		464	431	411	75	70	67
T = 6, D = 3, f = 5/3	1		298	298	298	49	49	49
	.9		305	303	302	50	50	50
	.6		336	327	321	55	54	53
	.3		415	389	372	68	64	61
Sample Size Formula								
T = 4, D = 3, f = 1	1		323	323	323	52	52	52
	.9		330	329	327	53	53	52
	.6		368	357	350	59	57	56
	.3		465	433	412	74	69	66
T = 6, D = 3, f = 5/3	1		299	299	299	48	48	48
	.9		305	304	303	49	49	48
	.6		336	327	322	54	52	52
	.3		416	390	373	67	62	60

Note. $\alpha_1^{(1)} = \alpha_1^{(2)}$, d = standardized effect size, R = reliability, K = number of indicators, T = number of measurement occasions, D = duration of study, and f = frequency of observation.

Table 2 shows simulations for a sample of analyses. The analyses in the simulations are made assuming that the variance-covariances are unknown. The Satorra-Saris sample sizes found from the top of Table 1 were used in the simulations to get approximate power and significance levels. The precision in the table is around .008 for the power curves, and around .004 for the confidence levels. As shown, power is close to .8 for all analyses with d = .2, and the significance level is close to .05. The power and significance levels are slightly inflated for d = .5, however. This is because the test statistic used to evaluate power and significance level in Mplus, the ratio of the parameter estimate to its standard error, is approximately normally distributed in large samples (Muthén & Muthén, 2002), and the sample sizes for d = .5 are fairly small. We can conclude, however, that results from the Satorra-Saris technique and the sample size formula (16) can be seen as practically equivalent, and they are also similar to simulations.

Table 2. Results from Monte Carlo Simulations using Satorra-Saris Sample Sizes from Table 1

			d = .2			d^a = .5		
Power	R	K =	3	4	5	3	4	5
T = 4, D = 3, f = 1	1		.792	-	-	.814	-	-
	.9		.803	.800	.802	.810	.816	.811
	.6		.804	.794	.800	.809	.815	.809
Significance level								
T = 4, D = 3, f = 1	1		.050	-	-	.058	-	-
	.9		.052	.050	.046	.059	.056	.055
	.6		.051	.051	.050	.060	.060	.055

Note. $\alpha_1^{(1)} = \alpha_1^{(2)}$, d = standardized effect size, R = reliability, K = number of indicators, T = number of measurement occasions, D = duration of study, and f = frequency of observation.
[a]The simulations for d = .5 resulted in covariance matrices that were not positive definite for some replications. The largest number of non-positive definite matrices (.0033%) came from the R = .6, K = 3 model.

6.2 Effect size, reliability, and number of indicators

Table 3 and Figure 2 show required group sample sizes for testing $\alpha_3 = 0$ against small (.2) and medium (.5) effect sizes. We do not report large (.8) effect sizes here because of the previously mentioned limitation of the method for small sample sizes. If desired, however, these are easily derived as discussed in section 4.5. We should mention, also, that all sample sizes in Table 3 and Figure 2 would increase by a factor 19/9 had we used the effect size definition of Raudenbusch & Liu (2001). This factor decreases and approaches 1, however, as duration of the study increases.

As expected, effect size has a large effect on required sample size (a factor of 6.25). Also according to expectations, more indicators per occasion do not require as large sample sizes, and this reduction is more noticeable with smaller reliabilities. As expected, models with lower indicator reliabilities require larger sample sizes, and this is more pronounced for models with few indicators. The increase in required sample size is not proportional to the decrease in reliability. The increase is larger at the lower scale of reliabilities.

Table 3. Required Group Sample Sizes for $T = 4$, $D = 3$, $f = 1$, and $\alpha_1^{(1)} = \alpha_1^{(2)}$

d	*R*	*K* =	1	2	3	4	5	6	10	20
0.2	1		323	323	323	323	323	323	323	323
	.9		346	334	330	329	327	327	325	324
	.8		373	349	340	336	333	331	328	325
	.7		406	366	352	345	341	338	332	327
	.6		447	388	368	357	350	346	337	330
	.5		501	418	388	373	363	357	344	333
	.4		577	461	418	396	382	373	354	338
	.3		698	527	465	433	412	399	370	347
	.2		931	651	552	501	469	447	401	363
	.1		1606	999	792	687	622	577	485	409
0.5	1		52	52	52	52	52	52	52	52
	.9		55	53	53	53	52	52	52	52
	.8		60	56	54	54	53	53	52	52
	.7		65	59	56	55	54	54	53	52
	.6		71	62	59	57	56	55	54	53
	.5		80	67	62	60	58	57	55	53
	.4		92	74	67	63	61	60	57	54
	.3		112	84	74	69	66	64	59	56
	.2		149	104	88	80	75	71	64	58
	.1		257	160	127	110	99	92	78	66

Note. $\alpha_1^{(g)}$ is the intercept mean in group g, *d* = standardized effect size, *R* = reliability, *K* = number of indicators, *T* = number of measurement occasions, *D* = duration of study, and *f* = frequency of observation.

Sample size, reliability and number of indicators interact. Let us assume, for simplicity, that $\lambda = 1$. All pairs *(R,K)* such that: $K = (1-R)\left(\sigma_\eta^2 + \sigma_{\pi_1}^2\right)/(R\sigma_\varepsilon^2)$, give the same sample size, which follows from formula (17). We can find *K*, the number of indicators of reliability *R* that are needed to obtain the same sample size as for a model with one indicator with reliability *R** and indicator error variance $\sigma_\varepsilon^2 = (1-R^*)\left(\sigma_\eta^2 + \sigma_{\pi_1}^2\right)/(R^*)$. For example, in a model with $\sigma_\eta^2 = \sigma_{\pi_1}^2 = .5$, say that we want to find the number of indicators with reliability $R = .6$ that will give the same sample size as a model with one indicator with reliability $R^* = .9$. Finding $\sigma_\varepsilon^2 = 1/9$ we obtain $K = 6$. We thus need to use six indicators per measurement occasion of reliability .6 in order to get the same sample size as for a .9 reliability, one-indicator model.

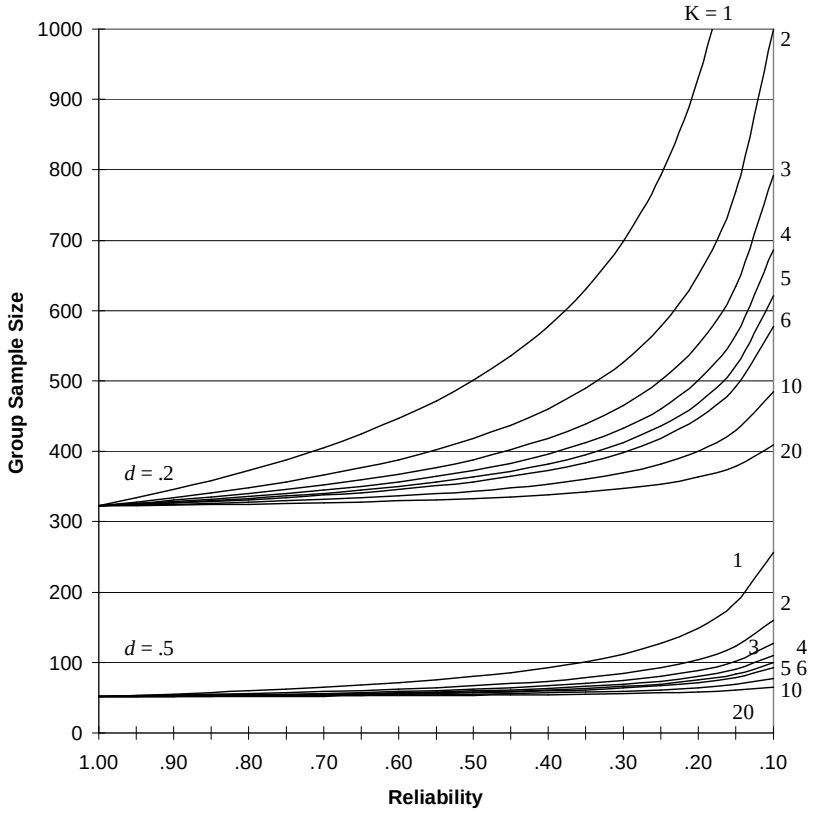

Figure 3. Group sample sizes as a function of reliability, effect size, and number of indicators for models with four measurement occasions.

6.3 Equality of Intercepts

All models mentioned so far have equal starting values, i.e. $\alpha_1^{(1)} = \alpha_1^{(2)}$. This may be reasonable in an intervention study when the intervention is made at time $t = 1$. If the intercepts differ, the group means give no information and formula (15) is simplified to

$$Var(\hat{\pi}_3) = 2\left(V_2 + \sigma^2_{\pi_s}\right)/n. \tag{19}$$

The right half of Table 4 shows required group sample sizes for selected models with unconstrained intercepts. It is seen that substantially larger sample sizes are required when the means of the intercepts cannot be assumed equal across groups.

Table 4. Required Group Sample Sizes for d = 0.2 and D = 3

$T(f)$	R	K =	1	3	5	10	1	3	5	10
			$\alpha_1^{(1)} = \alpha_1^{(2)}$				$\alpha_1^{(1)} \neq \alpha_1^{(2)}$			
2(1/3)	1		341	341	341	341	392	392	392	392
	.9		369	350	347	344	438	408	402	397
	.6		499	396	374	358	668	484	448	420
	.3		859	524	454	399	1356	714	585	489
6(5/3)	1		299	299	299	299	333	333	333	333
	.9		318	305	303	301	366	344	340	337
	.6		401	336	322	311	530	399	373	353
	.3		596	416	373	338	1022	563	471	402
10(9/3)	1		269	269	269	269	287	287	287	287
	.9		284	274	272	270	310	295	292	290
	.6		347	298	287	278	422	332	314	301
	.3		490	359	326	299	760	445	382	335
20(19/3)	1		236	236	236	236	242	242	242	242
	.9		245	239	238	237	254	246	244	243
	.6		288	254	247	241	317	267	257	249
	.3		383	295	273	255	504	329	294	268
∞	-		186	186	186	186	186	186	186	186

Note. $\alpha_1^{(g)}$ is the intercept mean in group g, d = standardized effect size, R = reliability, K = number of indicators, T = number of measurement occasions, D = duration of study, and f = frequency of observation.

6.4 Number of Occasions

Table 4 and Figure 3a show sample sizes for different numbers of equidistant occasions but with a fixed duration, *D=3*, and fixed standardized effect size, *d = 0.2*. The four-occasion models in Table 3 had f = 1 and x_t = 0, 1, 2, and 3. Here f varies e.g. f = 5/3 gives x_t = 0, 3/5, 6/5, 9/5, 12/5, 3 and f = 1/3 gives x_t = 0, 3. As expected, using only two measurements (f = 1/3) yields the highest sample sizes. Increasing frequency of observation results in lower required sample sizes but never below 186. Examining formulas (13), (14) and (15) we see that σ^2 decreases towards $\sigma^2_{\pi_2}$ as f increases with fixed D.

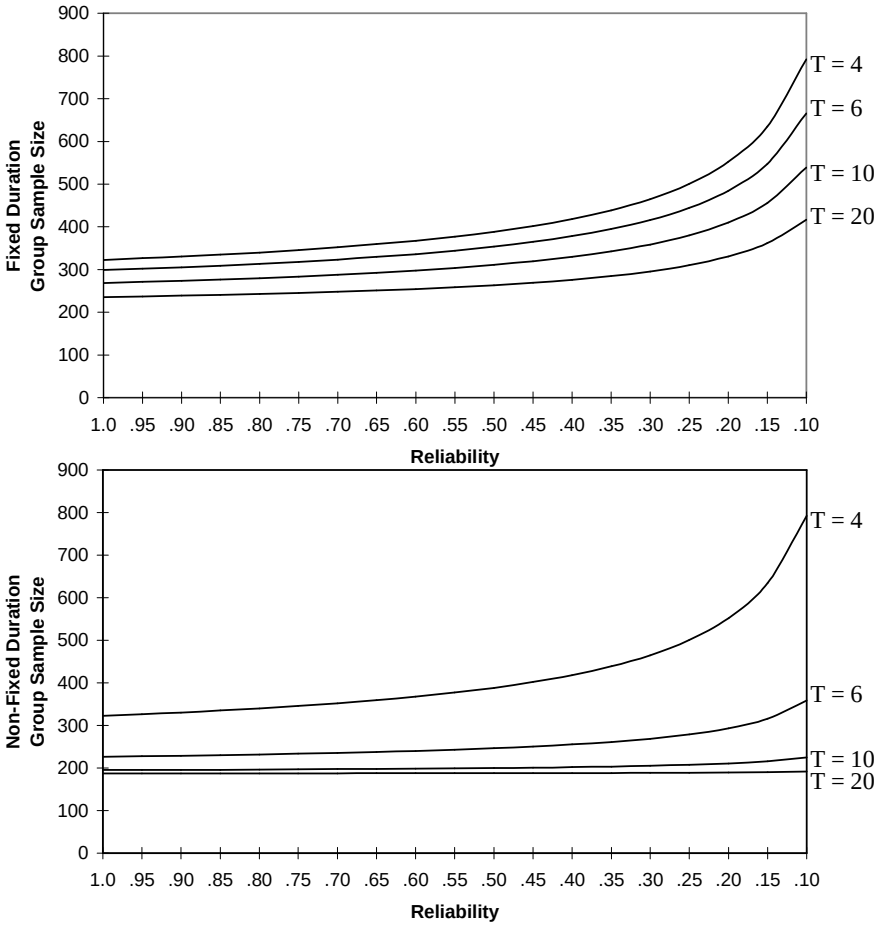

Figure 3a (top) and b (bottom). Group sample sizes as a function of reliability, fixed- or non-fixed duration, and number of measurement occasions for three-indicator models.

We can also increase the number of occasions by increasing the duration keeping the frequency constant. This is a situation where additional measurements correspond to additional time, and thus also to greater effect sizes, *d*, at the last time point. The effect of the number of occasions is illustrated in Table 5 and Figure 3b with standardized effect size $d = .2$ fixed at time point $t = 4$. The mean change in each interval, α_3, is constant at .0919 (which corresponds to $d = .2$ for the $T = 4$, $f = 1$ model). This model can thus be compared to Tables 1 and 3. Not surprisingly, increasing the duration by adding measurements results in substantially smaller sample sizes, although never below 186, whereas decreasing the duration results in substantially larger sample sizes.

Table 5. Required Group Sample Sizes for $\alpha_1^{(1)} = \alpha_1^{(2)}$ and $f = 1$

$T(D)$	d	R	K =	1	3	5	10
2(1)	0.0667	1		1580	1580	1580	1580
		.9		1833	1666	1632	1638
		.6		3005	2078	1883	1733
		.3		6243	3229	2595	2102
6(5)	0.3333	1		227	227	227	227
		.9		234	229	228	227
		.6		263	240	235	231
		.3		333	269	253	241
10(9)	0.6000	1		195	195	195	195
		.9		197	196	195	195
		.6		204	198	197	196
		.3		220	205	201	198
20(19)	1.2667	1		187	187	187	187
		.9		187	187	187	187
		.6		188	188	187	187
		.3		191	189	188	188
∞	-	-		186	186	186	186

Note. $\alpha_1^{(g)}$ is the intercept mean in group g, d = standardized effect size, R = reliability, K = number of indicators, T = number of measurement occasions, D = duration of study, and f = frequency of observation.

Formula (16) can be used to show that there is a limit in the required sample sizes, regardless of K and R, as T and/or f tends toward infinity. This corresponds to the case when we have succeded in identifying the latent variable π_2 exactly for the individuals. For example, with α_3 = .0919 (which corresponds do d = .2 for our basic model) n tends towards

$$\frac{(1.96 + .8416)^2 2 \times .1}{.0919^2} \approx 185.87,$$

which was seen in the tables. In other words, we cannot find a measurement plan with less than 186 participants per group to get a power of .8, regardless of the number of indicators, their reliabilities, or the number of measurements.

6.5 Non-Zero Intercept-Slope Covariance

All models so far have had zero covariance between the intercept and slope factors. A non-zero covariance may increase or decrease the required sample size. For example, covariances of .05, .1, and .2 (resulting in correlations (r) of .223, .447, and .895 respectively) for T = 4, D = 3, f = 1, R = .9, K = 3,

and $\alpha_3 = .0919$ result in n = 359, 378 and 382 respectively (compared to n = 330 for r = 0). Covariances $\sigma_{\pi_1\pi_2} = -.05$, -.1, and -.2 result in n = 291, 241, and 109 respectively. Equation (12) shows that the worst case occurs when $V_2\bar{x} = \sigma_{\pi_1\pi_2}$, i.e. when $\sigma_{\pi_1\pi_2} = .161$ in this case, and a minimum n is reached for r = -1, i.e. $\sigma_{\pi_1\pi_2} = -.22$. One has to keep in mind, however, that a positive covariance increases the variance at the last occasion. If one wants to use standardized effect size, d, one must also change α_3.

7. Choice of design

Looking at equation (16), we can examine the magnitudes of the effects of the various factors on sample size. For example, suppose we have a design with K = 2, D = 3, f = 1 (i.e. T = 4), R = .6, d = .2, significance level α = .05, and power = .8. The required sample size is 388 per group. Suppose that it is only possible to obtain 340 individuals per group. To keep the same power we need to use a design with K = 8, or f = 7/4, or prolong the study time to D = 5 keeping f fixed, or make more reliable measurements, R = .86, or use a lower level test α = .075, or accept a power of .75 (all else being constant). If costs permit, it is thus beneficial to increase the duration by taking another measurement at an additional time point. If the duration is fixed, however, increasing the frequency of observations has a larger effect on sample size than increasing the number of indicators. The costs of increasing the duration and/or the frequency is often higher, however, than the costs of more indicators.

7.1 A hypothetical example

Say that we want to collect data to study group differences in spatial cognitive ability development for adults (e.g. differences between adults without dementia and adults with a mild case of dementia) and that we believe that there are small differences in the population. To determine sample sizes we use the parameters obtained by Finkel, Reynolds, McArdle, Gatz, and Pedersen (2003), who, among other things, studied[10] spatial ability of 590 adults. The measurements spanned over six years with three equidistant measurement occasions. Results for the Figure Logic (FL) and Card Rotations (CR) indicators were: $\alpha_{1FL} = 61.14$, $\alpha_{1CR} = 45.88$, $\alpha_{2FL} = -.46$, $\alpha_{2CR} = -.69$,

[10] They used the Swedish adoption/twin study of aging and fitted two slopes (one before age 65 and one after), however for illustration purposes we will only focus on the one-slope models.

$\sigma^2_{\pi_{1FL}} = 83.49$, $\sigma^2_{\pi_{1CR}} = 215.24$, $\sigma^2_{\pi_{2FL}} = .01$, $\sigma^2_{\pi_{2CR}} = .09$, $\sigma^2_{\varepsilon FL} = 82.56$, and $\sigma^2_{\varepsilon CR} = 81.02$ [11].

The estimates above are from separate first order analyses. However, knowing the reliabilities of the indicators, we can adjust the above estimates for use in a second order model. The means of the above variances, adjusted for .89 reliability (see Pedersen, Plomin, Nesselroade, and McClearn (1992) for reliabilities of these and other indicators assessing spatial ability), yield: $\sigma^2_{\pi_1} = 149.37$, $\sigma^2_{\pi_2} = .05$, $\sigma^2_{\eta} = 56.36$, and $\sigma^2_{\varepsilon} = 25.43$. For $T = 3$, $f = 1/3$, $D = 6$, $K = 1$, $\alpha_i^{(1)} \neq \alpha_i^{(2)}$, and $d = .2$ we get the following sample sizes:

$$n = 2\left(\frac{1.96 + .8416}{0.48}\right)^2\left(\frac{56.36 + 25.43}{3^2\left(3^3 - 3\right)/12} + .05\right) = 312.76$$

We will thus need approximately 313 participants in each group to have .8 power to detect small differences in slopes using one indicator per measurement occasion, at three occasions (x_t = 0, 3, and 6) with no assumptions on equal intercept means. If we increase the number of indicators to two, we will need 265 per group, and we can decrease n to 236 by using five indicators per occasion. We can also increase the frequency of observation, f. Using one indicator and x_t = 0, 2, 4, 6 (f = 1/2) or x_t = 0, 1, 2, 3, 4, 5, 6 (f = 1) results in 282 and 202 per group respectively. However, if we add four indicators to each occasion as well, the results are 213 and 153 respectively. We can also increase the duration. Measuring one indicator at x_t = 0, 3, 6, and 9 (f = 1/3) results in sample sizes of 127 (96 for five indicators). The best thing we can do is to extend the study period, if it is feasible.

8. Discussion

8.1 Results in relation to previous work

We presented a sample size formula that can be used to calculate required sample sizes to study group differences in latent constructs. Several authors have presented algebraic expressions for related models (e.g. Liu and Liang, 1997; Liu et al. (2000); Liu et al., 2002; Rochon, 1991), however most expressions have been in terms of the correlation between repeated measures. When models such as latent growth curve- or multilevel models are used, variance expressions in terms of the random components of these models might then be used to study design issues. We extended a variance formula

[11] Correlations between the intercept and slope factors were also found, however these are ignored here, and the example should only be thought of as an illustration.

by Raudenbush and Liu (2001) to include growth in latent constructs, which enabled us to study effects of the number of indicators and their reliabilities on sample size.

In settings where randomization to groups have occurred, the group intercepts may be assumed equal. The difference in mean slopes across groups can then be estimated with more precision. Our formulas also incorporated the possibility of this extra precision. We found that this extra precision is often substantial but largest when there are only a few measurement occasions.

The results were illustrated in different situations where many different design factors were varied. Sample sizes were found to decrease with increases in effect size, the number and reliability of the indicators, duration of time, and frequency of observation. This is not astonishing and in this respect the results agreed with previous studies (e.g. Fan, 2003; Hertzog et al., 2006; Muthén & Curran, 1997; Raudenbush & Liu, 2001). We illustrated the relative importance of the different factors and noted some facts that have not been reported elsewhere for LGC models.

Hertzog et al. (2006) studied first order LGC models. They looked at the power to detect covariances between slopes of parallel processes and found that growth curve reliability, i.e. the proportion of variance explained by the structural model at the first occasion, had positive effects on power. They suspected that multiple indicator models would increase the power in detecting inter-individual differences. This occurred in the present study, although sample sizes for differences in slopes between groups was examined and not sample sizes for slope covariances. Duncan et al. (2002) suggested that using multiple indicators should increase power of the model because it accounts for measurement error and thus “refines” the developmental process. The present study also found that if the observed variables perfectly measure the latent construct of interest, one measure per construct is enough. If the reliability is not perfect, however, more measures can make up for this unreliability.

Some previous studies on power in LGC models found only small effects of the frequency of observation with fixed duration (e.g. Fan & Fan, 2005). This can happen when the residual variance at each occasion is small compared to the slope variance (see e.g. Raudenbush & Liu, 2001). Increasing the number of measurement occasions within the interval will, in that case, not make a big difference because there is not much to be gained.

The sample sizes found in the illustrations may be considered fairly large and thus discouraging. However, sample sizes for larger effects or other parameter values can be derived by the same formula. Although this study

did not illustrate the effects of varying the variance among individual slopes or occasion factors, the formulas still hold. They show that larger variances result in larger sample sizes, as Raduenbush and Liu (2001) also have noted. In addition, sample size formulas from Diggle et al. (2002) and Liu et al. (2000) have shown that sample sizes depend on the correlations between observations over time.

8.2 Practical considerations

LGC models or multilevel models are increasingly used in the study of growth. These models can be used when we follow one or several groups of individuals over time. They can be used in intervention research such as when we are evaluating a new program. Because these types of studies often come with considerable costs (we observe participants several times, interventions might be expensive etc.), design considerations such as decisions on sample size and number of measurement occasions can therefore be valuable. Although we can use Satorra-Saris (1985) approximations or simulations prior to data collection to determine the needed sample size for these models, using formulas presented in this paper can be quicker and easier. These formulas also allow researchers to see how different factors interact in their impacts on sample size. We can easily change design aspects, such as the number of occasions and indicators, in the formula and see what happens to the sample size.

We compared some of our results to Satorra and Saris (1985) approximations and simulations. They were found to be similar for all checked sample sizes. The formula is based on a known covariance structure, however, and a reservation for small sample sizes is therefore necessary. If sample size calculations result in small n's (e.g. $n < 20$), results may be improved by using t percentiles. On the other hand, if the covariance matrix is not known when designing the study one can never be sure of finding the correct design in advance. In addition, the results from our formulas are approximately true for non-normal models as long as the linear structure is the same and the sample sizes so large that normal approximation is allowed.

When we are interested in differences in growth between predetermined groups, such as males and females, or children of married and divorced parents, we can still use formulas presented in this paper. The variance formula becomes less complex when we cannot incorporate the extra precision of knowing that the groups were equal at one point in time. However, when the means of the intercepts differ between groups, that is when the groups differ at the start of the study, the standardized effect size, d, as defined by (18) may not be appropriate. Other standardized effect sizes, such as those used by Raudenbush and Liu (2001) might then be used instead.

8.3 Future work

We gave a farly general approach based on matrices and linear models. For the special case of second order linear models a simplified algebraic formula was given. Although we examined several factors that affect sample size, there were several factors we did not examine in detail. We did not examine models with unequal group variances. Muthén and Curran (1997) looked at groups with different variances and showed that power was greatest for studies that were nearly balanced but with a slight overrepresentation in the treatment group. This was because the treatment group had a larger variance than the control group. The results were the opposite when they changed the design so that the control group had the larger variance. A treatment that increases the variance in the treatment group will increase the sample size needed for sufficient power. For example, in terms of Curran and Muthén (1999) and Muthén and Curran´s (1997) added curve factor in the treatment group, we may have a positive added slope variance, $\sigma^2_{\pi_3}$. If $\sigma^2_{\eta_3} = .1$ in the $T = 4$, $f = 1$, $K = 3$, $R = .9$, model, the required sample size per group increases from 330 to 422. This is not surprising, however, because the total variance then increases.

This study did not cover other more complex models such as models with correlated errors or models with interactions among latent variables[12], non-linear models, models with more than two groups and so forth. Further study is needed to examine these models. For example, Fan (2003) predicts that the differences in power between LGC models and repeated measures ANOVA will increase for non-linear models because LGC models can estimate time scores (factor loadings). These models can also be extended in various other ways. For example, they can be constructed to include regressions among latent factors, additional growth processes, piecewise growth factors, covariates etc. The formulas in this paper can be extended to incorporate covariates or other factors. The generalisation of the matrix expressions is fairly straightforward but the algebraic formulas will get more complex. Raudenbush and Liu's (2001) formulas for higher order trends can also be extended to include growth in latent factors as well as intervention settings with equal intercepts.

[12] See e.g. Muthén and Curran (1997) and Curran and Muthén (1999) for interaction interpretations and power computations for first order models.

9. References

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Curran, P. J., & Bollen, K. A. (2001). The best of both worlds. Combining autoregressive and latent curve models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp 105-135). Washington DC: American Psychological Association.

Curran, P. J., & Muthén, B. O. (1999). The application of latent curve analysis to testing developmental theories in intervention research. *American Journal of Community Psychology, 27*, 567-595.

Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi- square distribution in misspecified structural equation models: finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research, 37*, 1-36.

Desu, M. M., & Raghavarao, D. (1990). *Sample size methodology*. Boston: Academic Press.

Diggle, P. J., Heagerty, P. J., Liang, K., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford Statistical Science Series. NY: Oxford University Press Inc.

Duncan, S. C., & Duncan, T. E. (1996). A multivariate growth curve analysis of adolescent substance use. *Structural Equation Modeling, 3*, 323-347.

Duncan, T. E., Duncan, S. C., Strycker, L. A., & Li, F. (2002). A latent variable framework for power estimation within intervention contexts. *Journal of Psychopathology and Behavioral Assessment, 24*, 1-12.

Duncan, T. E., Duncan, S. C., Strycker, L. A., Li, F., & Alpert, A. (1999). *An introduction to latent variable growth curve modeling: concepts, issues, and applications*. In Marcoulides, G. A. (Ed.). Quantitative Methodology Series. Methodology for Business and Management. London: Lawrence Erlbaum.

Fan, X. (2003). Power of latent growth modeling for detecting group differences in linear growth trajectory parameters. *Structural Equation Modeling: an Interdisciplinary Journal, 10*, 380-400.

Fan, X. & Fan, X. (2005). Power of latent growth modeling for detecting linear growth: number of measurements and comparison with other analytic approaches. *Journal of Experimental Education, 73*, 121-139.

Ferrer, E., & McArdle, J. J. (2003). Alternative structural models for multivariate longitudinal data analysis. *Structural Equation Modeling, 10*, 493-524.

Finkel, D., Reynolds, C. A., McArdle, J. J., Gatz, M., & Pedersen, N. L. (2003). Latent growth curve analysis of accelerating decline in cognitive abilities in late adulthood. *Developmental Psychology, 39*, 535-550.

Friedman, L. M., Furberg, C. D., & DeMets, D. L. (1985). *Fundamentals of clinical trials* (2nd ed.). Littleton, MA: PSG Publishing.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). New York: Halsted.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66*, 373-388.

Hertzog, C., Ghisletta, P., Lindenberger, U., & Oertzen, T. (2006). On the power of multivariate latent growth curve models to detect correlated change. *Psychological Methods, 11*, 244-252.

Hong, S, & Ho, H. (2005). Direct and indirect effects of parental involvement on student achievement: second-order latent growth modeling across ethnic groups. *Journal of Educational Psychology, 97*, 32-42.

Hox, J. J. (2000). Multilevel analyses of grouped and longitudinal data. In T. D. Little, K. U. Schnabel, and J. Baumert (Eds.). *Modeling longitudinal and multilevel data. Practical issues, applied approaches and specific examples* (pp.15-32). Mahwah, NJ: Lawrence Erlbaum Associates.

Hox, J., & Stoel, R. D. (2005). Multilevel and SEM approaches to growth curve modeling. In B. S. Everett, & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioural science* (pp 1296-1305). Chisester: John Wiley and Sons, Ltd.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Associates.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lance, C., E., Vandenberg, R., J., & Self, R., M. (2000). Latent growth models of individual change: the case of the newcomer adjustment. *Organizational Behavior and Human Decision Processes, 83*, 107-140.

Liu, A., Boyett, J. M., & Xiong, X. (2000). Sample Size calculation for planning group sequential longitudinal trials. *Statistics in Medicine, 19*, 205-220.

Liu, G., & Liang, K. (1997). Sample size calculations for studies with correlated observations. *Biometrics, 53*, 937-947.

Liu, A., Shih, W. J., & Gehan, E. (2002). Sample size and power determination for clustered repeated measurements. *Statistics in Medicine, 21*, 1787-1801.

Loehlin, J. C. (1992). *Latent variable models: An introduction to factor, path, and structural analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Longford, N. T. (1993). *Random coefficient models*. Oxford, UK: Clarendon.

MacCallum, R. C. (2000). Modeling multivariate change. In T. D. Little, K. U. Schnabel, and J. Baumert (Eds.). *Modeling longitudinal and multilevel data. Practical issues, applied approaches and specific examples* (pp.51-68). Mahwah, NJ: Lawrence Erlbaum Associates.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: power analysis and null hypotheses. *Psychological Methods, 11*, 19-35.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

MacCallum, R. C., & Hong, S. (1997). Power analysis in covariance structure modeling using GFI and AGFI. *Multivariate Behavioral Research, 32*, 193-210.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In R. B. Cattell & J. Nesselroade (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 564-614). New York: Plenum Press.

McArdle, J. J., & Bell, R. Q. (2000). An introduction to latent growth models for developmental data analysis. In T. D. Little, K. U. Schnabel, and J. Baumert (Eds.). *Modeling longitudinal and multilevel data. Practical issues, applied ap-*

proaches and specific examples (pp. 69-107). Mahwah, NJ: Lawrence Erlbaum Associates.

McDonald, R. P. (1999). *Test theory: a unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107-122.

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: a latent variable framework for analysis and power estimation. *Psychological Methods, 2*, 371-402.

Muthén, L. K., & Muthén, B. O. (1998-2006). *Mplus User's Guide*. Fourth Edition. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4*, 599-620.

Pedersen, N. L., Plomin, R., Nesselroade, J. R., & McClearn, G. E. (1992). Quantitative genetic analysis of cognitive abilities during the second half of the lifespan. *Psychological Science, 3*, 346-353.

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics, 14*, 1-17.

Raudenbush, S. W. (2002). Toward a coherent framework for comparing trajectories of individual change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp.35-64). Washington, DC: American Psychological Association.

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

Raudenbush, S. W., & Liu, X. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6*, 387-401.

Raykov, T. (2000). Modeling simultaneously individual and group patterns of ability growth or decline. In: T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling Longitudinal and Multilevel Data* (pp. 127-146). Mahwaj, NJ: Lawrence Erlbaum.

Rochon, J. (1991). Sample size calculations for two-group repeated-measures experiments. *Biometrics, 47, 1383-1398*.

Rochon, J. (1998). Application of GEE procedures for sample size calculations in repeated measures experiments. *Statistics in Medicine, 17*, 1643-1658.

Satorra, A., & Saris, W. E. (1985). The power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50*, 83-90.

Sayer, A. G., & Cumsille, P. E. (2002). Second-order latent growth models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp 177-200). Washington DC: American Psychological Association.

Shieh, S. (2003). A comparative study of power and sample size calculations for multivariate general linear models. *Multivariate Behavioral Research, 38*, 285-307.

Stoel, R. D., Peetsma, T. T., & Roeleveld, J. (2003). Relations between the development of school investment, self-confidence, and language achievement in elementary education: a multivariate latent growth curve approach. *Learning and Individual differences, 13*, 313-333.

Stoel, R. D., van Der Wittenboer, G, & Hox, J. J. (2003). Analyzing longitudinal data using multilevel regression and latent growth curve analysis. *Metodologia de las Ciencas del Comportamiento, 5*, 21.42.

Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika, 23*, 19-23.

Willet, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116*, 363- 381.

Appendix

In a general analysis of several groups with some common parameters, let α_c be the common parameters and $\alpha_g^{(g)}$ be the group-specific parameters (including τ). We then have in each group $(\alpha_C^*, \alpha^{(g)*}_g)'$ is normal with mean $(\alpha_C, \alpha^{(g)}_g)$ and variance $(\Lambda_0^{(g)\prime} V^{(g)-1}\Lambda_0^{(g)})^{-1}/n^{(g)}$. Stacking these expressions above each other we have that $(\alpha_C^*, \alpha^{(1)*}_g, \alpha_C^*, \alpha^{(2)*}_g, \ldots, \alpha_C^*, \alpha^{(G)*}_g)'$ is normal with mean $(\alpha_C, \alpha^{(1)}_g, \alpha_C, \alpha^{(2)}_g, \ldots, \alpha_C, \alpha^{(G)}_g)'$ and covariance matrix block diagonal with blocks $(\Lambda_0^{(g)\prime} V^{(g)-1} \Lambda_0^{(g)})^{-1}/n^{(g)}$. These can be combined to form a least squares estimate of the parameters and to find their variance.

Let us consider an example with two groups. Let Y_{1i} be an observation on individual i in group 1, $i = 1, \ldots, n^{(1)}$, and Y_{2i} be an observation on individual i in group 2, $i = 1, \ldots, n^{(2)}$.

$$Y_{1i} \in N\left(\begin{bmatrix}\alpha_c \\ \alpha^{(1)}\end{bmatrix} X^{(1)}, \Sigma^{(1)}\right), \; i = 1, \ldots, n^{(1)} \tag{a}$$

$$Y_{2i} \in N\left(\begin{bmatrix}\alpha_c \\ \alpha^{(2)}\end{bmatrix} X^{(2)}, \Sigma^{(2)}\right), \; i = 1, \ldots, n^{(2)} \tag{b}$$

From (a) we can write

$$\begin{bmatrix}\alpha_c^* \\ \alpha^{(1)*}\end{bmatrix} \in N\left(\begin{bmatrix}\alpha_c \\ \alpha^{(1)}\end{bmatrix}, \begin{bmatrix}\Gamma^{(1)} & \kappa^{(1)} \\ \kappa^{(1)} & \sigma^{(1)}\end{bmatrix} \frac{1}{n^{(1)}}\right) \tag{c}$$

and from (b) we can write

$$\begin{bmatrix}\alpha_c^* \\ \alpha^{(2)*}\end{bmatrix} \in N\left(\begin{bmatrix}\alpha_c \\ \alpha^{(2)}\end{bmatrix}, \begin{bmatrix}\Gamma^{(2)} & \kappa^{(2)} \\ \kappa^{(2)} & \sigma^{(2)}\end{bmatrix} \frac{1}{n^{(2)}}\right). \tag{d}$$

We can stack (c) and (d) on top of each other

$$\begin{bmatrix} \alpha_c^* \\ \alpha^{(1)*} \\ \alpha_c^* \\ \alpha^{(2)*} \end{bmatrix} \in N\left(\begin{bmatrix} \alpha_c \\ \alpha^{(1)} \\ \alpha_c \\ \alpha^{(2)} \end{bmatrix}, \begin{bmatrix} \Gamma^{(1)}/n^{(1)} & \kappa^{(1)}/n^{(1)} & 0 & 0 \\ \kappa^{(1)}/n^{(1)} & \sigma^{(1)}/n^{(1)} & 0 & 0 \\ 0 & 0 & \Gamma^{(2)}/n^{(2)} & \kappa^{(2)}/n^{(2)} \\ 0 & 0 & \kappa^{(2)}/n^{(2)} & \sigma^{(2)}/n^{(2)} \end{bmatrix} \right). \qquad \text{(e)}$$

We will call the covariance matrix in (e) *R*. The variance of the generalized least squares estimate of $\alpha^{(1)} - \alpha^{(2)}$ is $\left[Z' R^{-1} Z\right]^{-1}_{3,3}$ where

$$\begin{bmatrix} \alpha_c \\ \alpha^{(1)} \\ \alpha_c \\ \alpha^{(2)} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \\ 0 & 1/2 & -1/2 \end{bmatrix} \begin{bmatrix} \alpha_c \\ \alpha^{(1)} + \alpha^{(2)} \\ \alpha^{(1)} - \alpha^{(2)} \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \\ 0 & 1/2 & -1/2 \end{bmatrix} = Z.$$