Behavior Research Methods (2019) 51:1216–1243
https://doi.org/10.3758/s13428-018-1059-y

Sample size estimation for heterogeneous growth curve models with attrition

Guillermo Vallejo[1] · Manuel Ato[2] · M. Paula Fernández[1] · Pablo E. Livacic-Rojas[3]

Published online: 22 June 2018

Abstract
In this study, two approaches were employed to calculate how large the sample size needs to be in order to achieve a desired statistical power to detect a significant group-by-time interaction in longitudinal intervention studies—a power analysis method, based on derived formulas using ordinary least squares estimates, and an empirical method, based on restricted maximum likelihood estimates. The performance of both procedures was examined under four different scenarios: (a) complete data with homogeneous variances, (b) incomplete data with homogeneous variances, (c) complete data with heterogeneous variances, and (d) incomplete data with heterogeneous variances. Several interesting findings emerged from this research. First, in the presence of heterogeneity, larger sample sizes are required in order to attain a desired nominal power. The second interesting finding is that, when there is attrition, the sample size requirements can be quite large. However, when attrition is anticipated, derived formulas enable the power to be calculated on the basis of the final number of subjects that are expected to complete the study. The third major finding is that the direct mathematical formulas allow the user to rigorously determine the sample size required to achieve a specified power level. Therefore, when data can be assumed to be missing at random, the solution presented can be adopted, given that Monte Carlo studies have indicated that it is very satisfactory. We illustrate the proposed method using real data from two previously published datasets.

Keywords Multilevel model · Missing data · Heterogeneous variances · Sample size · Statistical power

Longitudinal studies are increasingly common in educational and psychological research settings. In some cases, subjects are measured repeatedly over time in order to examine their individual growth and the potential differences among them. In other cases, subjects assigned to different experimental conditions are treated for a specific period of time and when the study is finished they are compared with respect to their average growth rates. Whatever the purpose of the study, it is usual and reasonable to model the change in the response of interest assuming linear growth (Willett, 1988) and to express the effect of the intervention in terms of the difference in mean slopes or rates of change among groups over time.

A wide variety of methods based on classical linear models can be applied to the analysis of longitudinal data. However, the presence of imbalance, due to missing responses from some subjects or due to observations from the same subject being generally correlated, can lead to erroneous conclusions regarding hypotheses of interest. Among other reasons, this is why multilevel hierarchical linear models have become the method of choice for modeling the change in response over time and the factors influencing the change.

Modeling longitudinal data using a hierarchical system of regression equations requires sufficient experimental units in order to detect the effects of interest at the desired power level. Hence, it is advisable to determine the sample size when planning a longitudinal study. Numerous publications have explained how to calculate the sample size in this type of study (e.g., Heo, Xue, & Kim, 2013; Muthén & Curran, 1997; Raudenbush & Liu, 2001; Usami, 2014; Wänström, 2009). There are also many software packages (e.g., ACluster,

✉ Guillermo Vallejo
gvallejo@uniovi.es

[1] Department of Psychology, Universidad de Oviedo, Oviedo, Spain

[2] Department of Psychology, Universidad de Murcia, Murcia, Spain

[3] Department of Psychology, Universidad de Santiago de Chile, Santiago de Chile, Chile

nQuery, OptimalDesign, PASS, PinT, or RMASS2) that can be used to perform sample size/power calculations with multilevel data. However, very few publications have dealt with informing researchers on this topic about errors due to heterogeneous variances across treatment groups and/or when it is expected that some subjects will leave the study prematurely (Hedeker, Gibbons, & Waternaux, 1999; Heo, 2014; Roy, Bhaumik, Aryal & Gibbons, 2007; Vallejo, Ato, Fernández, Livacic-Rojas, & Tuero-Herrero, 2016).

Loss of subjects invariably occurs in longitudinal studies, potentially leading to inefficient analyses and invalid conclusions. The existence of heterogeneity has been found in several reviews of studies published in psychology journals (cf. Erceg-Hurn & Mirosevich, 2008). This phenomenon is not only likely to occur in nonrandomized intervention studies, but it can also occur in completely randomized experiments. Some common causes of heterogeneity in real data are problems related to measurement validity, research design, and analysis (e.g., unclear randomization, high dropout rates, small sample sizes, presence of floor or ceiling effects in treatment outcome measures, differential treatment effects across subjects, or bad data). Regardless of the potential sources of heterogeneity, neglecting heterogeneity when it is present can lead to inefficient and potentially misleading inferences about fixed effects. For more detailed information about why heterogeneity occurs in intervention studies, see Grissom and Kim (2012). Also, Keselman, Algina, Lix, Wilcox, and Deering (2008) discuss the impact that heterogeneous variances have on error probabilities.

For the derivation of the power function, it is generally assumed that all variance components included in the multilevel models are known. When suitable prior information is not available, specification of these random components is sometimes a difficult task. In these cases, a possible solution is to simplify the procedure of power analysis by assuming that some effects vary randomly between subjects or clusters, whereas others are constrained to be fixed effects (e.g., a model with nonrandomly varying slopes). These restrictions are sometimes specified in applied research (e.g., Heo & Leon, 2008, 2009). When a source of variation is completely ignored, however, this can lead to overly optimistic sample size and power calculations. For instance, if random-intercept models are used inappropriately, given that both random-intercept and -slope models need to be considered, there is a considerable risk of finding high apparent power, because the so-called random-intercept model generally has a poor control of the Type I error rate (Vallejo, Ato, & Valdés, 2008).

Usami (2014) has developed a procedure that can be applied in order to examine the statistical power to detect a significant group-by-time interaction in a two-level random-coefficient regression model, especially when no informative variance components are available. However, this author confined the development of the proposed method for investigating sample size requirement to detecting an intervention effect based on two groups for situations that assume a linear growth pattern of the outcomes over time, complete data for every subject, and homogeneous errors at both Levels 1 and 2. Subsequently, Vallejo et al. (2016) extended the procedure proposed by Usami (2014) to situations in which the presence of between-subjects heterogeneity can be reasonably predicted and the influence of attrition taken into consideration. However, the formulas derived by Vallejo et al. (2016) are restricted to models that assume a linear change in responses over time. Furthermore, the adequacy of the sample size determination formulas for heterogeneous and incomplete data has not been investigated.

The present study extended the work of Vallejo et al. (2016) so as to overcome the aforementioned limitations and, therefore, can be viewed as a generalization of the corresponding results of these authors. Specifically, our objective in this article is threefold: first, to extend the method originally proposed by Usami (2014) and later updated by Vallejo et al. (2016) to more complex growth models for power and sample size determinations; second, to carry out a Monte Carlo study to verify the statistical power achieved with the estimated sample sizes; and third, to check whether the theoretical statistical power based on estimates by ordinary least squares (OLS) differs from the empirical statistical power based on maximum likelihood (ML) estimates, by means of Monte Carlo simulations. In this study, we used restricted ML (REML) as the estimation method because, in multilevel modeling, REML estimates of variance components tend to be less biased than unrestricted ML estimates (Browne & Draper, 2000).

Formulation of a statistical model

Suppose we are interested in comparing the longitudinal trends of two groups, experimental versus control, in a numeric dependent variable. Considering that measures taken over time are nested in subjects, such data can be analyzed using a hierarchical regression model with two levels. At the first level, we represent the change we expect each subject of the population to experience during a specific period of time, whereas at the second level we describe the conjectured relationship between the parameters of individual growth and the explanatory variables that are assumed stable for the whole duration of the study.

Adopting an individual growth model in which change is a linear function of time, the Level 1 model can be formulated as follows:

$$Y_{it} = b_{0i} + b_{1i}X_{it} + e_{it}, \quad (1)$$

where Y_{it} denotes the response variable of the ith subject ($i = 1, \ldots, N$) at the tth measurement occasion ($t = 1, \ldots, T$), X_{it} defines the specific time (e.g., days) that this subject is observed, and random parameters b_{0i} (intercept), b_{1i} (slope or rate of change) and e_{it} (error term), respectively represent the true value of the subject's response at baseline, the rate of change during the period of data collection and the measurement error caused by the deviation from linearity. In the absence of missing data, we assume that $X_{it} = X_t$ for all i, and that measurements of the response from the baseline ($X_1 = 0$) to the last time point increase at time intervals whose length is equal to unity; so, $D = T - 1$. It is important to observe that starting the time coding with $T_1 = 1$ instead of $T_1 = 0$ would be equivalent, but more difficult to interpret, because the value zero is outside the range of observed measurement occasions.

At the second level, the parameters resulting from modeling the trajectories of individual change over time, are related to the explanatory variables that describe the differences between subjects in intercepts and slopes. If we have only one explanatory variable (e.g., a behavioral intervention to improve the language of autistic children), the Level 2 model becomes

$$b_{0i} = \beta_{00} + u_{0i}, \quad (2)$$

$$b_{1i} = \beta_{10} + \beta_{11}W_i + u_{1i}, \quad (3)$$

where the indicator variable of the intervention program is $W_i = 0$ if the ith Level 2 unit is assigned to the control group, and $W_i = 1$ if it is assigned to the experimental group. Because of the randomization of subjects to the two treatment groups, the Level 2 model for the intercept does not contain the value of group-level variable W_i and we assume a common mean response at time $t = 0$. In this model, β_{00} is the mean response in treatment and control group at baseline because no treatment main effect is assumed, β_{10} is the average rate of change of the control group and β_{11} is the difference between the average rates of change for the groups. As a result, the average rate of change of the experimental group corresponds to the sum of $\beta_{10} + \beta_{11}$ Random variables u_{0i} and u_{1i} are independent from e_{it} and it is assumed that they follow a bivariate normal distribution with mean zero, variances τ_{00} and τ_{11}, respectively, and covariance τ_{01}.

Note that Eq. 2 specifies no predictors for b_{0i}. Suppose, however, that this intercept depends on W_i. One might then formulate another form of the random-intercept model. Specifically, $b_{0i} = \beta_{00} + \beta_{01}W_i + u_{0i}$, where β_{01} is the main effect of the treatment W on b_{0i}. In this case, residual variance components τ_{00} and τ_{11}, represent the variability that remains in parameters b_{0i} and b_{1i} after controlling the effect due to the program.

By substituting Eqs. 2 and 3 into Eq. 1, the mixed or combined model can be expressed as follows:

$$Y_{it} = \beta_{00} + \beta_{10}X_{it} + \beta_{11}W_iX_{it} + (u_{1i}X_{it} + u_{0i} + e_{it}). \quad (4)$$

With no assumptions about group differences at baseline, Eq. 4 should also include W_i as a predictor. It is often assumed that errors e_{it}, conditional on u_{1i} and u_{0i}, are distributed normally and independently with mean zero and constant variance σ^2. In this study we also considered the presence of heterogeneous variance across treatment groups, although we hold that the distribution of errors is normal.

Under the combined model of Eq. 4, the expected value, variance, and covariance of the measurements Y_{it}, conditional on the explanatory variables, are given by

$$E(Y_{it}) = \beta_{00} + (\beta_{10} + \beta_{11}W_i)X_{it}, \quad (5)$$

$$Var(Y_{it}) = \tau_{00} + 2X_{it}\tau_{01} + X_{it}^2\tau_{11} + \sigma^2, \quad (6)$$

$$Cov(Y_{it}, Y_{it'}) = \tau_{00} + (X_{it} + X_{it'})\tau_{01} + X_{it}X_{it'}\tau_{11}. \quad (7)$$

If baseline values differ across groups, then Eq. 5 should also include the term $\beta_{01}W_i$ (For more details on these equations, see Appendix 1.)

If there are reasons to suspect that changes in the expected value of the outcome will deviate from linearity over the duration of the study, more complex models of growth can be considered. For example, if the average outcome increases monotonically with time until the improvement stabilizes, then we might consider the following curvilinear growth model:

$$Y_{it} = \beta_{00} + (\beta_{10} + \beta_{11}W_i)X_{it} + (\beta_{20} + \beta_{21}W_i)X_{it}^2 + (u_{0i} + u_{1i}X_{it} + u_{2i}X_{it}^2 + e_{it}). \quad (8)$$

Again, we can accept the groups as equivalent enough at the beginning of the study and omit a main effect of treatment from the model. To allow the intercepts (baselines) to differ by groups, we add the dummy variable treatment W to the model of Eq. 8.

In the model of Eq. 8, the expected value, variance, and covariance of the measurements Y_{it}, conditional on the explanatory variables, are now given by

$$E(Y_{it}) = \beta_{00} + (\beta_{10} + \beta_{11}W_i)X_{it} + (\beta_{20} + \beta_{21}W_i)X_{it}^2, \quad (9)$$

$$Var(Y_{it}) = \tau_{00} + 2X_{it}\tau_{01} + X_{it}^2\tau_{11} + 2X_{it}^2\tau_{02} + 2X_{it}^3\tau_{12} + X_{it}^4\tau_{22} + \sigma^2, \quad (10)$$

$$Cov(Y_{it}, Y_{it'}) = \tau_{00} + (X_{it} + X_{it'})\tau_{01} + X_{it}X_{it'}\tau_{11} + (X_{it}^2 + X_{it'}^2)\tau_{02} + (X_{it}X_{it'}^2 + X_{it}^2X_{it'})\tau_{12} + X_{it}^2X_{it'}^2\tau_{22}. \quad (11)$$

Equation 9 should also include the term $\beta_{01}W_i$ when the baseline mean responses are not assumed equal.

To simplify the calculations further, it is useful to re-express Eqs. 4 and 8 of the multilevel model in terms of matrices and vectors, as follows:

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_i, \quad (12)$$

where $\mathbf{y}_i$ is a $T \times 1$ vector of repeated observations for the ith subject, $\mathbf{X}_i(=\mathbf{Z}_i\mathbf{A}_i)$ is a $(T \times P)$ design matrix for the fixed effects, $\boldsymbol{\beta}$ is a vector $(P \times 1)$ of fixed effects, $\mathbf{Z}_i$ is a $(T \times Q)$ design matrix for the random effects, $\mathbf{u}_i$ is a $(Q \times 1)$ vector of random effects, and $\mathbf{e}_i$ is a $(T \times 1)$ vector of errors. Here, $\mathbf{Z}_i$is a within-subjects design cual's mean response changes over time, and $\mathbf{A}_i$ is a $(Q \times P)$ between-subjects design matrix that contains time-invariant explanatory variables.

With respect to errors and random effects, it is assumed that vectors $\mathbf{e}_i$ and $\mathbf{u}_i$ are normally distributed with mean 0 and variance and covariance matrices $\mathbf{R}_i$ and $\mathbf{T}$, respectively. Matrix $\mathbf{R}_i$ may take various forms, however, it is common to assume a model of conditional independence, that is, $\mathbf{R}_i = \sigma^2\mathbf{I}_T$, where $\mathbf{I}$ is a $T \times T$ identity matrix. These assumptions imply that, marginally, $\mathbf{y}_i \sim N(\mathbf{X}_i\beta, \mathbf{V}_i = \mathbf{Z}_i\mathbf{T}\mathbf{Z}_i'\mathbf{R}_i)$. When $\mathbf{V}_i$ is known, the generalized least squares estimator of vector $\boldsymbol{\beta}$ is given by $\hat{\beta} = \left(\sum_{i=1}^N \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^N \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i$ and its variance by $\left(\sum_{i=1}^N \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}$. In the usual case where $\mathbf{V}_i$ is unknown, then an approximation to the true covariance is given, replacing $\mathbf{V}_i$ with its estimator $\hat{\mathbf{V}}_i$.

Equations 5–7 and 9–11 are essential in order to plan a longitudinal study properly since, as we shall see later, they provide the machinery that allows us to carry out a correct power analysis. To estimate the sample size required to detect a statistically significant group-by-time interaction effect, it is necessary to specify the value of the parameters included in Eqs. 1–3 of the model. However, such a task is neither easy nor straightforward, given that in many cases it is impossible to surmise the value of the parameters without running the experiment. Hence, in practice, the use of existing methods for calculating the sample size is limited to situations in which researchers are able to anticipate a range of probable values of the parameters of interest from the results obtained in previous studies.

In an attempt to optimize focus for a power analysis in studies in which linear growth is assumed, Usami (2014) suggests transforming the variance components associated with the model of Eq. 4 and the parameter related to the treatment (i. e., β_{11}) into statistical indices whose possible values could reasonably be specified in advance. These are reliability of measure at the baseline (ρ_1), standardized effect size at the last time point (d_L), level 2 residuals correlation (r_1) and ratio between the variance of outcomes at the end and at the beginning of study within groups (k_1). Formally,

$$\rho_1 = \frac{Var(u_{0i})}{Var(u_{0i} + e_{it})} = \frac{\tau_{00}}{\tau_{00} + \sigma^2}, \quad (13)$$

$$d_L = \frac{E(Y_{iT}|W_i = 1) - E(Y_{iT}|W_i = 0)}{\sqrt{Var(Y_{iT})}} = \frac{D\beta_{11}}{\sqrt{\tau_{00} + 2D\tau_{01} + D^2\tau_{11} + \sigma^2}}, \quad (14)$$

$$r_1 = \frac{Cov(u_{0i}, u_{1i})}{\sqrt{Var(u_{0i}u_{1i})}} = \frac{\tau_{01}}{\sqrt{\tau_{00}\tau_{11}}}, \quad (15)$$

and

$$k_1 = \frac{Var(Y_{iT})}{Var(Y_{i1})} = \frac{\tau_{00} + 2D\tau_{01} + D^2\tau_{11} + \sigma^2}{\tau_{00} + \sigma^2}. \quad (16)$$

It is important to note that the effect size parameter of Eq. 14 depends on the sum $\beta_{01} + D\beta_{11}$, rather than on the choice of β_{11} alone, when $\beta_{01} \neq 0$.

By solving Eqs. 15 and 16 simultaneously, the following components of variance and covariance are obtained (see Appendix 2):

$$\tau_{01} = \frac{-r_1^2\tau_{00} + r_1\sqrt{r_1^2\tau_{00}^2 + \tau_{00}(k_1 - 1)(\tau_{00} + \sigma^2)}}{D}, \quad (17)$$

$$\tau_{11} = \frac{2r_1^2\tau_{00} + (k_1 - 1)(\tau_{00} + \sigma^2) - 2r_1\sqrt{r_1^2\tau_{00}^2 + \tau_{00}(k_1 - 1)(\tau_{00} + \sigma^2)}}{D^2}. \quad (18)$$

At the same time, by replacing $Var(Y_{iT})$ in Eq. 14 with the value found for it in Eq. 16, the coefficient associated with the effect of linear treatment can be written as:

$$\beta_{11} = \frac{d_L\sqrt{k_1(\tau_{00} + \sigma^2)}}{D}. \quad (19)$$

Please note that if $\beta_{01} \neq 0$, then $\beta_{11} = \left(-\beta_{01} + d_L\sqrt{k_1(\tau_{00} + \sigma^2)}\right)/D$.

Without loss of generality, we can assume that the variance of the initial outcome is equal to 1 (i. e., $\tau_{00} + \sigma^2 = 1$). In this

case, Eqs. 13–19 reduce to that given by Usami (2014). The restriction above makes it possible to calculate the parameters of the model by specifying the values of ρ_1, d_L, r_1, and k_1. However, it should be noted that in this regard these indices can be detailed intuitively, which largely prevents the difficulty involved in exploratory studies in defining the values of the parameters before running the experiment. In addition, Usami found that the indices ρ_1, r_1, and k_1 have less influence on the sample size calculation than does d_L, in particular when $d_L > 0.4$.

So far, we have focused on a series of formulas derived in order to run a prospective analysis of power in models that assume linear growth. However, this approach can be extended to more complex curvilinear growth models, including polynomial and piecewise growth models. For instance, the outcome may follow a quadratic trend that would require the inclusion of the second-order treatment effect in the model (see Eq. 8).

The calculation of an appropriate sample size for detecting curvature in growth rates relies on transformation of the model parameters (i. e., τ_{02}, τ_{12}, τ_{22}, and β_{21}) into indices that can be specified from a literature review and conjecture. In addition to those specified in Eqs. 13–16, this new situation requires the inclusion of four additional indices. Using the results of Eqs. 9–11, these are defined as follows:

$$d_Q = \frac{E\left(Y_{iT}|W_i=1\right)-E\left(Y_{iT}W_i=0|\right)}{\sqrt{Var(Y_{iT})}} = \frac{D\beta_{11}+D^2\beta_{21}}{\sqrt{\tau_{00}+2D\tau_{01}+D^2\tau_{11}+2D^2\tau_{02}+2D^3\tau_{12}+D^4\tau_{22}+\sigma^2}}, \tag{20}$$

$$r_2 = \frac{Cov(u_{0i},u_{2i})}{\sqrt{Var(u_{0i}u_{2i})}} = \frac{\tau_{02}}{\sqrt{\tau_{00}\tau_{22}}}, \tag{21}$$

$$r_{12} = \frac{Cov(u_{1i},u_{2i})}{\sqrt{Var(u_{1i}u_{2i})}} = \frac{\tau_{12}}{\sqrt{\tau_{11}\tau_{22}}}, \tag{22}$$

and

$$k_2 = \frac{Var(Y_{iT})}{Var(Y_{i1})} = \frac{\tau_{00}+2D\tau_{01}+D^2\tau_{11}+2D^2\tau_{02}+2D^3\tau_{12}+D^4\tau_{22}+\sigma^2}{\tau_{00}+\sigma^2}. \tag{23}$$

Again, it is important to note that the effect size parameter of Eq. 20 depends on the sum $\beta_{01}+D\beta_{11}+D^2\beta_{21}$, rather than on the sum $D\beta_{11}+D^2\beta_{21}$, when $\beta_{01}\neq 0$.

By solving the Equation System 21–23, a series of equations of the form $ax^2+bx+c=0$ are obtained (see Appendix 2). The solutions or roots, which correspond to the components of variance we sought, can be obtained by solving each quadratic equation using the familiar formula of Bhaskara (cf. Puttaswamy, 2012):

$$\tau_{02} = \frac{-\beta_{02} \pm \sqrt{B_{02}^2-4A_{02}C_{02}}}{2A_{02}}, \tag{24}$$

where

$$A_{02} = D^4; B_{02} = 2D^2r_2^2\tau_{00}+2D^3r_{12}\sqrt{(\tau_{11}/\tau_{00})}r_2\tau_{00};$$
$$C_{02} = 2D\tau_{01}r_2^2\tau_{00}+D^2\tau_{11}r_2^2\tau_{00}-(k_2-1)\left(\tau_{00}+\sigma^2\right)r_2^2\tau_{00};$$

$$\tau_{12} = \frac{-B_{12} \pm \sqrt{B_{12}^2-4A_{12}C_{12}}}{2A_{12}}, \tag{25}$$

where

$$A_{12} = D^4; B_{12} = 2D^2r_2\sqrt{(\tau_{00}/\tau_{11})}r_{12}\tau_{11}+2D^3r_{12}^2\tau_{11}; C_{12} = 2D\tau_{01}r_{12}^2\tau_{11}+D^2\tau_{11}r_{12}^2\tau_{11}-(k_2-1)\left(\tau_{00}+\sigma^2\right)r_{12}^2\tau_{11}; \text{and}$$

$$\tau_{22} = \frac{-\beta_{22} \pm \sqrt{B_{22}^2-4A_{22}C_{22}}}{2A_{22}}, \tag{26}$$

where

$$A_{22} = D^8; B_{22} = -\left(2D^2r_2\sqrt{\tau_{00}}+D^3r_{12}\sqrt{\tau_{11}}\right)^2+2D^6\tau_{11}$$
$$+4D^5\tau_{01}-2D^4(k_2-1)\left(\tau_{00}+\sigma^2\right);$$
$$C_{22} = 4D^2\tau_{01}^2+D^4\tau_{11}^2+(k_2-1)^2\left(\tau_{00}+\sigma^2\right)^2+4D^3\tau_{01}\tau_{11}$$
$$-2(k_2-1)\left(\tau_{00}+\sigma^2\right)\left(2D\tau_{01}+D^2\tau_{11}\right).$$

Finally, by substituting in Eq. 20 the value found for $Var(Y_{iT})$ in Eq. 23, the coefficient for the quadratic treatment effect can be written as

$$\beta_{21} = \frac{d_Q\sqrt{k_2(\tau_{00}+\sigma^2)}-d_L\sqrt{k_1(\tau_{00}+\sigma^2)}}{D^2}. \tag{27}$$

In the presence of a main effect of the treatment *W*, the slope formula would have the same form as that provided in Eq. 27, because both d_L and d_Q contain information about β_{01}.

In Appendix 3 the machinery is provided that allows us to carry out a correct power analysis using piecewise models. Because the data from many longitudinal studies can be well-approximated using simple piecewise linear models with at most one or two knots that are located at judiciously chosen time points (Fitzmaurice, Laird, & Ware, 2011, p. 151), we only present a random two-slope piecewise model in which the entire growth period of the outcome under study is split into two parts: (1) linear growth from the baseline to the last time point in the study, and (2) linear growth from the breakpoint to the last time point. Obviously, when determining the sample size, it must be known ahead of time where the breakpoint is.

Estimation of the treatment effect and its variance

The goal of a longitudinal intervention study is to test whether there are differences between treatment conditions with respect to their average growth rates. If the change is conceptualized as a sustained linear process, then we must verify if iom

$\beta_{11} \neq 0$. With two groups (e.g., experimental *versus* control), the OLS estimator of β_{11} can be expressed as:

$$\hat{\beta}_{11} = \frac{\sum_{i=1}^{N_E}\sum_{t=1}^{T}\left(X_{it}-\overline{X_i}\right)Y_{it}}{\sum_{i=1}^{N_E}\sum_{t=1}^{T}\left(X_{it}-\overline{X_i}\right)^2} - \frac{\sum_{i=1}^{N_C}\sum_{t=1}^{T}\left(X_{it}-\overline{X_i}\right)Y_{it}}{\sum_{i=1}^{N_C}\sum_{t=1}^{T}\left(X_{it}-\overline{X_i}\right)^2}, \quad (28)$$

where N_E and N_C are the treatment and control group sample sizes, respectively. The generalization of Eq. 28 to more than one active treatment is not direct, but it is simple to derive (see Appendix 4).

To test the interaction effect between variables of Level 1, time, and Level 2, treatment, calculation of the variance of the β_{11} estimator is required. Using Eqs. 6 and 7, and considering that the variance of a difference reduces to the sum of variances of independent groups, ordinary algebra shows that (see Appendix 5):

$$Var\left(\hat{\beta}_{11}\right) = \frac{4}{N}\left(\frac{\sigma^2}{\sum_{i=1}^{T}\left(X_{it}-\overline{X}_i\right)^2} + \tau_{11}\right), \quad (29)$$

where N (= N_E + N_C) denotes the total number of units of second level included in study, with $N/2$ subjects in each group. The quantity $4/N$ on the right side of Eq. 30 should be replaced with $(1/Np_1p_2)$ to allow groups of unequal size, where $p_1 = N_C/N$ and $p_2 = N_E/N$.

If the T measures between $X_1 = 0$ and $X_T = D$ are equally spaced, Eq. 29 can be reformulated as follows (see Fitzmaurice et al., 2011):

$$Var\left(\hat{\beta}_{11}\right) = \frac{4}{N}\left(\frac{12\sigma^2(T-1)}{D^2T(T+1)} + \tau_{11}\right), \quad (30)$$

where $D = f^{-1}(T-1)$ and f is the frequency of observation per time unit, whereas V_1 and τ_{11} denote the variability in growth rates within and across subjects, respectively. The sum of $V_1 + \tau_{11}, \sigma_{b1}^2$ onward is a measure of variability in the estimation by OLS of the model slope (1).

When growth is assumed to be linear and $f = 1$(i. e., $X_t = 0, 1, 2\ldots, T-1$; $T = D+1$), the sample variance of the rate of change simplifies to $V_1 = 12\sigma^2/(T^3 - T)$. For more complex growth functions (e.g., quadratic function) and $f \neq 1$(e. g., $X_t = 0, 2, 4\ldots 2T-2$; $T = fD+1$), Raudenbush and Liu (2001) showed that the sample variance of the polynomial slope takes the following form

$$V_p = \frac{\sigma^2 f^{2p}(T-p-1)!}{l_p(T+p)!}, \quad (31)$$

where p denotes the polynomial order of the change of outcome and l_p is a constant whose values depend on the way of coding the time variable (sequential, centered or orthogonal). To model nonlinear relations across time it is beneficial to use orthogonal polynomials, since this reduces any form of collinearity that can result from using multiples of t as regressors.

Alternatively, the variance of any trend of interest (e.g., linear, quadratic, or cubic), regardless of the form assumed to characterize the covariance structure of measurement error, can be more easily obtained from the appropriate diagonal element of

$$Cov\left(\hat{\mathbf{b}}_i\right) = \left(\mathbf{Z}_i'\mathbf{V}_i^{-1}\mathbf{Z}_i\right)^{-1}, \quad (32)$$

where $\mathbf{Z}_i$ is a design matrix that specifies the change of outcome of any subject across the study (i.e., a constant, linear, quadratic, etc., function), $\mathbf{V}_i\left(= \mathbf{Z}_i\mathbf{T}_i\mathbf{Z}_i' + \mathbf{R}_i\right)$ is the covariance matrix of repeated measurements, $\mathbf{T}_i$ is the dispersion matrix of Level 2 random effects, and $\mathbf{R}_i$ is the covariance structure of Level 1 errors.

Additionally, a quick and easy way to test the effects that D and f will have on the power using the matrix formulation of

the model is to divide the linear trend component of matrix $\mathbf{Z}_i$ by f, that of the quadratic trend component by f^2 that of the cubic trend by f^2 and so on. Very often $f = 1$, but depending on the value of D, there are many possible alternative results (e.g., $f = 0.5$ or $f = 2$).

Statistical power analysis

The power to detect a specified treatment difference is defined as the probability of rejecting the null hypothesis of no treatment-by-linear-trend interaction $H_0 : \beta_{11} = 0$, given that it is in fact false ($\beta_{11} \neq 0$). Using Eqs. 28 and 30, this hypothesis can be tested with:

$$F_0 = \frac{\hat{\beta}_{11}^2}{Var\left(\hat{\beta}_{11}\right)}, \tag{33}$$

where $Var\left(\hat{\beta}_{11}\right) = \sigma_{b1}^2 / N p_1 p_2$, $p_1 = N_C / N$, $p_2 = N_E / N$, and $N = N_C + N_E$. The F_0 statistic follows the central F distribution when H_0 is true, but when H_0 is false it follows the noncentral F distribution with df_1 degrees of freedom in the numerator, df_2 degrees of freedom in the denominator, and noncentrality parameter λ which is defined as

$$\lambda = \frac{N p_1 p_2 \beta_{11}^2}{\sigma_{b1}^2}, \tag{34}$$

This strategy is both feasible and straightforward for studies in which there is good reason to assume that the groups have equal variances. However, as we previously indicated, it is possible that the assumption of Level 1 and/or Level 2 homogeneity of variances will be violated (see the example described in Vallejo, Fernández, Cuesta, & Livacic-Rojas, 2015, for details). Under the most general scenario, the noncentrality parameter λ is given by

$$\lambda^{\cdot} = \frac{N p_1 p_2 \beta_{11}^2}{\sigma_{b1^{(C)}}^2 + \sigma_{b1^{(E)}}^2}, \tag{35}$$

where $\sigma_{b1^{(C)}}^2 = p_2 \left[12\sigma_{(C)}^2 / \left(T^3 - T\right) + \tau_{11}^{(C)} \right]$ and $\sigma_{b1^{(E)}}^2 = p_1 \left[12\sigma_{(E)}^2 / \left(T^3 - T\right) + \tau_{11}^{(E)} \right]$.

Regardless of the values of f and D and of the number of groups to be compared, as well as in the possible presence of heterogeneity, λ can also be computed using a method similar to the one that Shieh (2003) suggested under the multivariate general linear model. Specifically,

$$\lambda = tr\left[\left(\mathbf{AVA}'\right)^{-1} \left(\mathbf{CBA}'\right)' \left(\mathbf{CM}^{-1}\mathbf{C}'\right)^{-1} \left(\mathbf{CBA}'\right) \right], \tag{36}$$

where tr denotes the trace of matrix $[\cdot]$, $\mathbf{A} = (\mathbf{1}_{NG} | -\mathbf{1}_{NG})$ and $\mathbf{C} = (\mathbf{1}_{NG-1} | -\mathbf{1}_{NG-1})$ are contrast matrices between subjects with a complete row range, $\mathbf{1}_{NG}$ is a column vector of ones, $\mathbf{1}_{NG}$ is an identity matrix, and the symbol | represents the augmented matrix resulting from appending the columns of matrices $\mathbf{A}$ and $\mathbf{C}$. The expected values matrix across T measurements, $\mathbf{B} = [\mu_{(C)0} \ldots \mu_{(C)T-1}; \mu_{(E)0} \ldots \mu_{(E)T-1}]$, can be easily obtained from Eq. 5 by fixing $\beta_{00} = \beta_{10} = 0$, $\mathbf{M}$ is a diagonal matrix whose elements are the number of subjects in each group [in our case, $\mathbf{M} = \text{diag}(N_C, N_E)$,], and the $\mathbf{V}$ matrix is constructed using Eqs. 6 and 7. If the group variance components are heterogeneous, then $\mathbf{V} = p_2\mathbf{V}_{(C)} + p_1\mathbf{V}_{\_(E)}$. The described method to compute λ is limited to Model 1; however, nothing prevents this from being extended to other contexts. For example, under Model 8, one would proceed in a similar way, but using Eqs. 9–11.

That said, the procedure used here to calculate the power of the statistical test F_0 to compare groups in terms of linear rates of change involves the following steps:

1. Define the significance level α and sample sizes of the control and experimental groups—that is, N_C and N_E. Without loss of generality, we can establish that $\beta_{00} = \beta_{10} = 0$ (or, alternatively, $\beta_{00} = \beta_{10} = \beta_{20} = 0$, in the case of the quadratic growth model).
2. Set the values of indices ρ_1, d_L, r_1 and k_1 (or, alternatively, ρ_1, d_L, d_Q, r_1, r_2, r_{12}, k_1, and k_2, in the case of the quadratic growth model), determining the values of parameters σ^2, τ_{00}, τ_{01}, τ_{11} and β_{11} (or σ^2, τ_{00}, τ_{01}, $\tau_{11}\tau_{02}$, τ_{12}, τ_{22}, β_{11}, and β_{21}, in the case of the quadratic growth model), and calculate the λ parameter defined in Eqs. 34–36.
3. Specify the critical value of the inverse of the F central distribution function, namely:

$$F_c = \text{FINV}(1-\alpha, df_1, df_2).$$

4. Calculate the probability that the F_0 ratio exceeds the critical value F_C when H_0 is false. Under the alternative hypothesis (H_1), the power function associated with the F_0 test is given by $1 - \beta = P[F'(df_1, df_2, \lambda) > F_C]$, where $F'(df_1, df_2, \lambda)$ denotes a noncentral F random variable with degrees of freedom (df_1, df_2) and noncentrality parameter λ, and β denotes the probability of a Type II error.

Determination of sample size

There are several approaches to determining the sample size, including Bayesian and frequentist methods that focus on estimation instead of hypothesis testing. However, the most popular approach involves calculating the power of a statistical test, that is, the probability of rejecting H_0 when H_1 is true.

Required sample size for two groups

Let us assume that we want to determine the sample size to detect differences between two groups. Hypothesis $H_0 : \beta_{11} = 0$ is rejected if the estimator of β_{11} exceeds the critical value $\left(\hat{\beta}_{11} > c\right)$. In accordance with Amatya, Bhaumik, and Gibbons (2013), this value defines the limit between the acceptance and rejection regions and is set under the following two conditions:

$$P\left(\hat{\beta_{11}} > c = 0 + Z_{1-(\alpha/2)}\sqrt{(Np_1p_2)^{-1}\sigma^2_{b1}}|\mathrm{H}_0 true\right) = \alpha, \quad (37)$$

$$P\left(\hat{\beta_{11}} > c = \beta_{11} - Z_{1-\beta}\sqrt{(Np_1p_2)^{-1}\sigma^2_{b1}}|\mathrm{H}_1 true\right) = 1-\beta. \quad (38)$$

Equating Eqs. 37 and 38, since the critical value c is assumed identical under both statistical hypotheses, and solving for N, we obtain the formula that informs us of the sample size required in order to achieve the desired power (see Appendix 6). Specifically,

$$N = \frac{\left(Z_{1-(\alpha/2)} + Z_{1-\beta}\right)^2 \sigma^2_{b1}}{\beta^2_{11} p_1 p_2} \quad (39)$$

where $Z_{1-(\alpha/2)}$ and $Z_{1-\beta}$ are 100 $(1-\alpha/2)$ and 100 $(1-\beta)$ percentiles of the standard normal distribution for a bilateral test.

Required sample size for multiple groups

Determining the sample size needed to compare the trends of an arbitrary number of groups is a relatively simple procedure, but one that is seldom documented in longitudinal studies. For this purpose, Eq. 39 can be rewritten as

$$N = \frac{\left(Z_{1-(a/2)} + Z_{1-\beta}\right)^2}{1/\mathrm{Tr}\left[\left(\mathbf{AVA}'\right)^{-1}\left(\mathbf{CBA}'\right)'\left(\mathbf{CP}^{-1}\mathbf{C}'\right)^{-1}\left(\mathbf{CBA}'\right)\right]}, \quad (40)$$

where $\mathbf{P} = \mathrm{diag}(p_1, p_2, \ldots, p_J)$. The remaining terms have been defined previously.

Required sample size for two or more groups with unequal variances

The sample size calculation specified in Eq. 39 assumes homogeneous errors at both Levels 1 and 2. When it is suspected that the variance components may differ depending on the participation of subjects in the training program, the required sample becomes:

$$N^* = \left(\frac{\left[Z_{1-(\alpha/2)} + Z_{1-\beta}\right]^2}{\beta^2_{11} p_1 p_2}\right)\left(p_2\sigma^{2(C)}_{b1} + p_1\sigma^{2(E)}_{b1}\right). \quad (41)$$

As was the case in the homogeneous model, the determination of the sample size in models with heterogeneous variances with an arbitrary number of groups also requires the modification of Eq. 41. For example, the value of N to detect differences among the trends of three groups can be obtained as

$$N^* = \frac{\left(Z_{1-(\alpha/2)} + Z_{1-\beta}\right)^2}{1/\mathrm{Tr}\left[\mathbf{AV}^*\mathbf{A}'\right]^{-1}\left(\mathbf{CBA}'\right)'\left(\mathbf{CP}^{-1}\mathbf{C}'\right)^{-1}\left(\mathbf{CBA}'\right)}, \quad (42)$$

where $\mathbf{V}^{\wedge *.} = p_1\mathbf{V}_1 + p_2\mathbf{V}_2 + p_3\mathbf{V}_3$. If it is suspected that treatment groups are unbalanced, then $\mathbf{V}^{\wedge *.} = [(p_2p_3)/p^{\wedge *.}]\mathbf{V}_1 + [(p_1p_3)/p^{\wedge *.}]\mathbf{V}_2 + [(p_1p_2)/p^{\wedge *.}]\mathbf{V}_3$, with $p^{\wedge *.} = p_1p_2 + p_1p_3 + p_2p_3$.

Required sample size for missing data

So far we have focused on how to determine the sample size assuming complete cases. However, dropout (also called *attrition*) is an inevitable problem in most longitudinal studies. The occurrence of missing values can produce biased estimates and can reduce statistical power, leading to inefficient analyses and invalid conclusions. When the rate of attrition is anticipated, a required sample size may be calculated on the basis of the final number of subjects that are expected to complete the study.

In the case of missing data, the formula described above to calculate the variance in the slopes of the subjects, $\sigma^2_b = Var\left(\hat{b}_i\right)$, may no longer be applicable or may not be realistic (Fitzmaurice et al., 2011). For this reason, we need a solution that mitigates the negative impact exerted by the attrition of the sample on the validity of the inferences and of the conclusions reached.

A method for modeling early leaving of a study reasonably is to divide, element by element, the $\mathbf{V}_i$ matrix of Eq. 32 by the

matrix that identifies the missing data pattern L. In this regard, O'Kelly and Ratitch (2014) clarified that in studies related to the health area it is more common for subjects to drop out of the study prematurely than temporarily. In this situation—that is, of attrition or dropping out definitively—the variance of the estimator rate of change can be obtained from the appropriate diagonal element of

$$Cov\left(b_i^*\right) = \left[\mathbf{Z}_i'\left(\mathbf{V}_i^{-1} \varnothing \mathbf{L}\right)\mathbf{Z}_i\right]^{-1} \tag{43}$$

where $\varnothing$ denotes the operator of the Hadamard division.

The choice of $\mathbf{L}$ matrix will depend on the loss model that we wish to emphasize. However, if we are interested in modeling the pattern of missingness found most frequently in applied research—that is, the monotone—a reasonable choice of $\mathbf{L}$ matrix would be one in which each element of the main diagonal informs us of the proportion of subjects who remain in the study over time (i.e., $1, r, r^2, \ldots, r^{t-1}$), and the remaining elements of the assumed survival rate (i.e., r). For the homogeneous model, the suggested procedure provides results similar to those obtained using the method described in Hedeker, Gibbons, and Waternaux (1999).

Method

Theoretical and Monte Carlo studies were conducted in order to determine the optimal sample size (N) for a study that ensures adequate statistical power for rejecting the null hypothesis of $\beta_{11} = 0$, as well as the accuracy of the estimates, assuming homogeneous ($\mathbf{V}_2 = \mathbf{V}_1$) or heterogeneous ($\mathbf{V}_2 = 2\mathbf{V}_1$) group variances at each of the levels of the model and missing data due to subject dropouts before the completion of the study after baseline. For this purpose we proceed as follows. Initially, using the formulas derived in Eqs. 38 and 40 we carried out a theoretical study to examine the effect of heterogeneity and attrition on determining the appropriate N when the significance level $\alpha = 0.05$ and the nominal statistical power $1 - \beta = 0.80$. Five factors were manipulated and completely crossed in the study for a total of 108 investigated conditions: reliability of measurement at the first time point ($\rho_1 = 0.1, 0.5$), Level 2 residual correlation ($r_1 = -0.5, 0, 0.5$), number of repeated measurements ($T = 4, 8$), proportion of imbalance between the group sizes ($\Delta = 0.5, 0.35, 0.2$), and standardized effect size at the last time point ($d_L = 0.4, 0.5, 0.6$). According to Cohen (1988), standardized mean differences of 0.2, 0.5, and 0.8 correspond to small, medium, and large magnitudes of an effect, respectively. The ratio between the variances of the outcomes at the end and at the beginning of the study remained constant ($k_1 = 25$) under each of the conditions. Later, a Monte Carlo study was carried out to verify the statistical power achieved with the estimated sample sizes.

Data generation

Datasets were simulated on the basis of the two-level model shown in Eqs. 1–3. At the first level, a continuous outcome was generated as a linear function of time. The intercept and one Level 1 variable were simulated to vary randomly as a function of treatment at the second level. Each explanatory variable X and W was generated to be standard normal. Later, we dichotomized the W variable by an arbitrary threshold (i.e., the mean of all observed data). The error terms were generated as independent normal random variables with means zero and the variances obtained from the values specified above for the manipulated factors. We used SAS version 9.4 (SAS, 2016) for the simulations.

For each of the 108 investigated conditions, 1,000 sets of raw data were generated and analyzed during the simulation process. In our simulation study, two different situations were considered: with no missing data at each of the time points and time-related dropout with cumulative missing data rates of 27% at the fourth occasion and 52% at the eighth occasion. Both with complete and with missing data, the analyses were carried out twice by REML methods using SAS PROC MIXED, once assuming homogeneity and once modeling the variances, in order to investigate the results of incorporating heterogeneity into the models.

Here we will focus on sample size determination in the presence of a monotone missing data pattern that spans the missing-at-random (MAR) model. For our dropout MAR mechanism, the data point for subject i was missing at time t and the subsequent times if $U_{it} < \Phi[\lambda_t + Y_{i(t-1)}]$, where U_{it} is a uniform random variable and Φ is the cumulative normal distribution function. The values of λ_t in the above mechanisms were chosen to yield time-related dropout rates of 0%, 10%, 19%, and 27% for the four respective occasions, and time-related dropout rates of 0%, 10%, 19%, 27%, 34%, 41%, 47%, and 52% for the eight respective occasions.

Evaluation criteria

To determine the accuracy and precision of the strategies being compared (i.e., sample size calculations using derived formula based on OLS estimates and simulations based on REML estimates), we examined their performance in terms of the following quantities:

1. *Relative bias* To find out whether a parameter tends to be over- or underestimated, the relative bias index was used

Table 1 Sample sizes to obtain theoretical power of at least 80% and the empirical power, with complete data and homogeneous Level 1 and 2 variances

				Δ= .50			Δ= .35			Δ= .20		
r_1	d_L	T	ρ_1	N	1-β	$1-\hat{\beta}$	N	1-β	$1-\hat{\beta}$	N	1-β	$1-\hat{\beta}$
– .5	.4	4	.1	214	.801	.815	235	.800	.803	334	.800	.789
– .5	.4	4	.5	222	.801	.796	244	.800	.802	346	.801	.802
– .5	.4	8	.1	210	.802	.809	230	.800	.783	327	.800	.819
– .5	.4	8	.5	219	.802	.810	241	.801	.779	342	.801	.786
– .5	.5	4	.1	138	.801	.807	151	.802	.786	214	.801	.792
– .5	.5	4	.5	142	.801	.804	156	.803	.809	222	.801	.795
– .5	.5	8	.1	136	.801	.798	148	.803	.808	210	.802	.812
– .5	.5	8	.5	142	.802	.796	155	.802	.793	219	.802	.778
– .5	.6	4	.1	96	.800	.802	105	.803	.806	149	.802	.796
– .5	.6	4	.5	98	.801	.802	109	.801	.778	154	.803	.798
– .5	.6	8	.1	94	.800	.799	103	.803	.811	146	.802	.797
– .5	.6	8	.5	98	.801	.796	108	.802	.798	153	.802	.788
.0	.4	4	.1	203	.802	.802	222	.800	.778	316	.801	.782
.0	.4	4	.5	197	.801	.806	216	.801	.789	307	.801	.802
.0	.4	8	.1	198	.801	.806	218	.802	.808	309	.801	.813
.0	.4	8	.5	194	.800	.786	214	.802	.817	303	.801	.822
.0	.5	4	.1	130	.801	.793	143	.802	.815	203	.802	.800
.0	.5	4	.5	127	.803	.804	139	.802	.801	197	.801	.799
.0	.5	8	.1	127	.801	.785	140	.802	.812	198	.801	.807
.0	.5	8	.5	125	.802	.812	137	.801	.809	194	.800	.803
.0	.6	4	.1	91	.803	.804	100	.803	.814	141	.801	.795
.0	.6	4	.5	88	.801	.802	97	.802	.805	137	.801	.781
.0	.6	8	.1	89	.803	.814	98	.804	.811	138	.801	.805
.0	.6	8	.5	87	.801	.793	96	.803	.805	135	.800	.789
.5	.4	4	.1	191	.802	.806	210	.802	.809	297	.800	.802
.5	.4	4	.5	172	.802	.789	188	.800	.796	267	.800	.804
.5	.4	8	.1	186	.800	.802	205	.802	.819	290	.800	.806
.5	.4	8	.5	169	.801	.802	186	.802	.791	263	.800	.818
.5	.5	4	.1	123	.803	.804	135	.803	.813	191	.802	.793
.5	.5	4	.5	110	.801	.798	121	.801	.793	172	.802	.808
.5	.5	8	.1	120	.802	.805	131	.800	.807	186	.800	.779
.5	.5	8	.5	109	.803	.805	119	.800	.814	169	.801	.809
.5	.6	4	.1	86	.804	.793	94	.802	.811	133	.802	.819
.5	.6	4	.5	77	.802	.805	85	.804	.803	120	.803	.793
.5	.6	8	.1	84	.804	.807	92	.803	.801	130	.802	.812
.5	.6	8	.5	76	.803	.806	83	.801	.803	118	.802	.814

Δ = proportion of imbalance between group sizes; r_1 = Level 2 residual correlation; d_L = standardized effect size at the last time point; T = number of repeated measurements; ρ_1 = reliability of measurement at the first time point; $1-\beta$ = theoretical power; $\hat{\beta}$ = empirical power.

in this study. If the parameter of interest was φ=(1-β), the percentage relative bias was $100 \times \left[\left(E\left(\hat{\phi}\right)-\phi\right)/\phi\right]$, where $E\left(\hat{\phi}\right)$ was computed as the average parameter estimate across valid replications. We have not been able to find any formal criteria in the literature for when a relative bias is too big, so in this article, a relative bias less than 10% was considered acceptable.

2. *Approximate 95% coverage rates* This refers to the number of times that the absolute difference between the

theoretical and empirical power across the examined conditions falls outside of approximately two standard errors (*SE*). The *SE*s reported for the empirical estimates of power were estimated by $\sqrt{pq/m}$, where p is the theoretical probability of a Type II error, q equals 1–p, and m is the number of simulations carried out in the numerical experiment.

Results

Tables 1, 2, 3, and 4 show the sample sizes obtained by the proposed method to achieve theoretical power of at least 80% and the simulation-based empirical power estimates. Table 1 gives the results for complete data with homogeneous variances, Table 2 gives the results for complete data with heterogeneous variances, Table 3 gives the results for incomplete data with homogeneous variances, and Table 4 gives the results for incomplete data with heterogeneous variances. Hereafter, these are known as Scenarios A, B, C, and D, respectively.

As can be seen from Table 1, the sample size needed to achieve 80% power with a two-sided Type I error rate of 5% decreases substantially with small increases in the effect size at the last time point (d_L), whereas the influences of the number of repeated measurements (T), the Level 2 residual correlation (r_1), and the reliability of measurement at the first time point (ρ_1) are not so obvious. Although the effects of T, r_1, and ρ_1 are relatively small on statistical power, larger values of these factors show a positive effect on statistical power. It is also shown in Table 1 that the sample size increases with an increasing degree of imbalance between the group sizes. In fact, high levels of imbalance (i.e., $\Delta = .2$) cause a notable increase in the sample size needed to maintain a specific statistical power of 80%. A similar tendency is observed for the same conditions under the remaining scenarios (i.e., B, C, and D).

Table 2 presents the results for complete data in the presence of heterogeneity of variances (Scenario B). When the sample size estimates of Table 1 are compared to those of Table 2, we find that the mere presence of a small degree of heterogeneity in the Level 1 and 2 random effects ($\mathbf{V}_2 = 2\mathbf{V}_1$) leads to a noticeable increase in the sample size necessary to achieve at least 80% power, even when the group sizes are equal. Table 3 lists the necessary sample sizes to reach the preset value of power when the assumption ofthe homogeneity of Level 1 and 2 variances is satisfied but attrition is present (Scenario C). As we stated previously, in this study we have assumed that the dropout rate of subjects from baseline to the last time point of interest is 10% in each group. As compared to the case of equal variances and complete data (Scenario A), it may be observed that dropout rates of 10% over time require that the sample size increase by 20%–25% in order to reach a similar power. Finally, the sample sizes required to accommodate the dropout rate in the presence of heterogeneity of variances (Scenario D) are given in Table 4. All results displayed in this table agree with the previous findings from a qualitative point of view; however, as one would expect, a larger sample is required under this scenario to reach the same level of power.

Table 5 shows the percentages of relative bias by ρ_1, d_L, and T, collapsed across Level 2 residual correlations (r_1). The results yielded negligible levels of bias (less than ±0.05% to 1.5% of the true population parameter, on average) in the vast majority of the 108 conditions examined. The levels of bias of predicted theoretical power were always less than 1%, regardless of the investigated conditions, whereas the mean relative bias for the empirical estimates of power remained under 3.6% in all cells, and it exceeded 3% in only five cases. In fact, there were no statistically significant differences in bias for the power estimates in any of the simulated conditions.

The empirical estimates of power can also be compared to the theoretical values stated in Tables 1, 2, 3 and 4. The highest absolute difference was .024 among the 108 conditions displayed in Table 1, .026 among the 108 conditions displayed in Table 2, .039 among the 108 conditions displayed in Table 3, and .038 among the 108 conditions displayed in Table 4. Under Scenarios A and B, the discrepancies between theoretical prediction and empirical results are negligible, since 99% of the power estimates fall within two standard deviation limits (i.e., between .775 and .825). On the other hand, our results also indicate that, for Scenarios C and D, about 85% of power estimates fall within the confidence intervals when $T = 4$, while only 5% of absolute differences were beyond two standard deviations when $T = 8$. Therefore, the derived formulas allow the user to rigorously determine the sample size required to yield a certain power for both complete and incomplete data, both assuming homogeneity and when incorporating heterogeneity into the multilevel model.

Empirical illustration using two real longitudinal data examples

To illustrate how the derived formulas for sample size calculations that can be used for a study ensure adequate power to detect statistical significance under different models and conditions (e.g., linear and quadratic, homogeneous and heterogeneous, or complete and missing data), we rely on the data of two longitudinal studies carried out by Núñez, Rosário,

Table 2 Sample sizes to obtain theoretical power of at least 80% and the empirical power, with complete data and heterogeneous Level 1 and 2 variances

				Δ = .50			Δ = .35			Δ = .20		
r_1	d_L	T	ρ_1	N	1-β	$1-\hat{\beta}$	N	1-β	$1-\hat{\beta}$	N	1-β	$1-\hat{\beta}$
– .5	.4	4	.1	321	.801	.803	317	.800	.800	400	.800	.792
– .5	.4	4	.5	332	.800	.787	328	.801	.786	414	.801	.779
– .5	.4	8	.1	314	.800	.801	311	.801	.807	392	.801	.805
– .5	.4	8	.5	328	.801	.793	325	.801	.803	410	.800	.790
– .5	.5	4	.1	206	.802	.789	204	.800	.805	257	.801	.794
– .5	.5	4	.5	213	.801	.810	211	.801	.801	266	.801	.785
– .5	.5	8	.1	202	.800	.802	199	.801	.796	252	.802	.811
– .5	.5	8	.5	211	.801	.795	208	.800	.788	263	.801	.818
– .5	.6	4	.1	143	.802	.804	142	.801	.802	179	.802	.797
– .5	.6	4	.5	148	.801	.798	147	.802	.787	185	.802	.803
– .5	.6	8	.1	140	.802	.801	139	.800	.802	175	.802	.789
– .5	.6	8	.5	147	.802	.797	145	.802	.789	183	.802	.776
.0	.4	4	.1	303	.801	.809	300	.801	.812	379	.801	.799
.0	.4	4	.5	295	.801	.818	291	.800	.796	368	.801	.797
.0	.4	8	.1	296	.800	.800	293	.801	.818	370	.800	.793
.0	.4	8	.5	291	.801	.801	288	.801	.807	363	.800	.800
.0	.5	4	.1	195	.802	.819	192	.800	.793	243	.801	.804
.0	.5	4	.5	189	.801	.807	187	.801	.803	236	.801	.819
.0	.5	8	.1	190	.801	.801	188	.801	.802	237	.800	.797
.0	.5	8	.5	187	.802	.805	185	.802	.804	233	.801	.814
.0	.6	4	.1	136	.803	.781	134	.801	.814	169	.801	.803
.0	.6	4	.5	132	.802	.783	130	.801	.787	164	.800	.821
.0	.6	8	.1	133	.803	.805	131	.801	.813	165	.801	.793
.0	.6	8	.5	130	.801	.803	129	.803	.808	162	.801	.798
.5	.4	4	.1	285	.800	.783	282	.800	.798	356	.800	.794
.5	.4	4	.5	257	.801	.792	254	.801	.806	321	.801	.789
.5	.4	8	.1	279	.801	.796	276	.801	.794	348	.801	.784
.5	.4	8	.5	253	.801	.801	250	.801	.819	316	.801	.799
.5	.5	4	.1	183	.801	.802	181	.801	.799	229	.802	.801
.5	.5	4	.5	165	.802	.825	163	.801	.810	206	.802	.811
.5	.5	8	.1	179	.801	.800	177	.801	.811	223	.800	.807
.5	.5	8	.5	162	.800	.802	161	.802	.810	203	.802	.814
.5	.6	4	.1	128	.803	.806	126	.801	.801	159	.801	.786
.5	.6	4	.5	115	.802	.807	114	.803	.798	143	.801	.802
.5	.6	8	.1	125	.803	.807	123	.801	.789	156	.803	.801
.5	.6	8	.5	113	.801	.802	112	.802	.795	141	.801	.789

See the note to Table 1.

Vallejo, and González-Pienda (2013) and Rosário et al. (2017). In the first study, a linear change model was a reasonable assumption, whereas in the second study a quadratic model provides a more suitable choice to represent the shape of change. Consistent with common practice in empirical applications of growth curve models, the Level 1 predictors (i.e., Time and/or Time^2) are assumed to be free of measurement error; if errors do exist, they would generally attenuate the estimate of the regression coefficients relative to their population values.

The first example (Núñez et al., 2013) examined the effectiveness of a school-based mentoring program on student self-regulated learning strategies. In this study program effects were tested in 94 sixth grade students assigned randomly to two experimental conditions, evaluated at the beginning of the study and after 3, 6, and 9 months. Thus, if we measure the

passage of time quarterly, this design involves $f = 1$ (the frequency of observation per unit of time is equal to one), $D = 3$ (the study lasts three quarters), and $T = fD + 1$ (the number of measurement occasions is four).

After reanalyzing the data of Núñez et al. (2013), without assuming that the groups' average responses are equal at baseline and using SAS PROC MIXED, the following estimates were obtained: $\hat{\tau}_{00} = .0708, \hat{\tau}_{01} = .0048, \hat{\tau}_{11} = .0050, \hat{\sigma}^2 = 0.865, \hat{\beta}_{01} = .1169$ and $\hat{\beta_{11}} = .0804$. Here, time was treated as a continuous variable centered on its overall mean, rather than as a classification variable, as in the original study. Substituting these estimates in Eqs. 13–16, yields estimates of the reliability of measurement at the first time point $(\hat{\rho}_1 = .45)$ standardized effect size at the last time point $\left(\hat{d}_L = .75\right)$, proportion of variance of outcomes between the first and the last time points $\left(\hat{k}_1 = 1.47\right)$ and slope-intercept correlation $(\hat{r}_1 = .25)$. In turn, using Eqs. 30 and 34 the variance of the slope $\left(\hat{\sigma}^2_{b1} = .0223\right)$. and the non-centrality parameter $\left(\hat{\lambda} = 6.81\right)$ are estimated. Inspection of a table for the noncentral F distribution (see, e.g., Ato & Vallejo, 2015) at the .05 significance level with $\left(\hat{\lambda} = 6.81\right)$ and with (1,280) degrees of freedom yields a power of $\hat{\varphi}\cong.74$. Also, standard software (e.g., SAS PROC IML) can be used to estimate this value. Next, we removed 28 data points to yield approximate dropout rates of 0%, 5%, 9%, and 13% for the four time points. In this particular application, the variance of the slope, $\hat{\sigma}^2_{b1}$ was .0246 and the non-centrality parameter, $\hat{\lambda}$, 6.18. Using these results and tables of noncentral F distribution, the power is found to be approximately .70. The corresponding estimates of σ^2_{b1}, λ and φ with heterogeneous errors (ratio 1:3) were .0446, 3.4, and .45, respectively.

Given that, in all three cases described, a power below the often-mentioned benchmark of .80 (Cohen, 1988) was obtained, it was necessary to determine the new sample size that would have allowed us to replicate the differences between treatment conditions, with respect to their average linear growth rates, under each of the situations described. From Eq. 39, with $Z_{1-(\alpha/2)} = 1.96$ and $Z_{1-\beta} = .84$, we see that the total sample sizes needed to achieve 80% power with a 5% significance level were 109, 120, and 217, respectively. So far, we have only considered power results for comparing groups on linear rates of change. Yet the rate of change can also be nonlinear.

Next we considered data from the longitudinal randomized design, conducted by Rosário et al. (2017) with 182 fourth grade students, to examine whether the students' writing quality differed when they wrote journals on a weekly basis, as compared with a control group. In the study, the subjects were measured at baseline and weekly for up to 12 weeks. With regard to the quality of writing compositions, Rosário et al. found that providing extra writing opportunities (i.e., writing journals) had a statistically significant impact on instantaneous rate of change at one specific moment and curvature. We suppose that our interest would lie in replicating the difference in the average acceleration rates between the two groups. Thus, we will first check whether there is sufficient statistical power to detect the described effects.

As in the previous example, we briefly considered three cases: a complete set of data with homogeneous errors; an incomplete set of data with homogeneous errors; and a complete set of data with heterogeneous errors. After analyzing the data using PROC MIXED, the following estimates were obtained: $\hat{\tau}_{00} = 45.0677, \hat{\tau}_{01} = 1.0519, \hat{\tau}_{11} = .3254, \hat{\tau}_{02} = .2867, \hat{\tau}_{21} = .0081, \hat{\tau}_{22} = .0081, \hat{\sigma}^2 = 21.1842, \beta_{11} = .2238,$ *and* $\beta_{11} = .2238,$ and $\hat{\beta}_{21} = -.0446$. Substituting these estimates into Eqs. 13, 20–23, 32, and 36, the indices and parameter estimates can be calculated as $\hat{\rho}_1 = .6802, \hat{d}_Q = -.3106, \hat{k}_1 = 1.3262, \hat{k}_2 = 2.1824, \hat{r}_1 = -.2747, \hat{r}_2 = -.4756, \hat{r}_{12} = -.1574, \hat{\sigma}^2_{b2} = .0186,$ and $\hat{\lambda} = 4.8533$. Inspection of noncentral F tables at the .05 significance level with $\hat{\lambda} = 4.8533$ and with (1, 2178) degrees of freedom yields a power of $\hat{\varphi}\cong.60$. Removing 594 data points from the original study according to a monotone dropout pattern, which represents a 5% dropout, we obtained $\hat{\sigma}^2_{b2} = 0.257$, $\hat{\lambda} = 3.5096$, and $\hat{\varphi}\cong0.47$. In the presence of heterogeneity of variances (ratio 1:3), however, we obtained $\hat{\sigma}^2_{b2} = 0.341$, $\hat{\lambda} = 2.4267$, and $\hat{\varphi} = .34$. According to the convention suggested by Cohen (1988), in all three cases an unsatisfactory level of statistical power was obtained. Thus, it was necessary to calculate the sample size that would have allowed us to replicate the differences between treatment conditions, with respect to their average acceleration rates, under each of the situations described. From Eq. 39, with $Z_{1-(\alpha/2)} = 1.96$ and $Z_{1-\beta} = .84$, we established that the total sample sizes needed to ensure adequate power were 295, 408, and 589, respectively.

Although we have omitted the original data due to limitations of space, the databases for the two examples are available from the first author upon request, and Appendix 7 provides the SAS codes used to perform the sample size and power calculations for Examples 1 and 2.

Discussion and conclusion

Sample size calculations to provide specified power levels were performed in four different scenarios, each involving 108 treatment combinations, through the use of mathematical formulas and numerical simulations. Our results indicate that both the analytic and empirical method provide virtually identical estimates of power across all examined conditions. The

Table 3 Sample sizes to obtain theoretical power of at least 80% and the empirical power, with incomplete data and homogeneous Level 1 and 2 variances

				Δ = .50			Δ = .35			Δ = .20		
r_1	d_L	T	ρ_1	N	1-β	$1-\hat{\beta}$	N	1-β	$1-\hat{\beta}$	N	1-β	$1-\hat{\beta}$
– .5	.4	4	.1	260	.801	.824	285	.801	.832	404	.800	.819
– .5	.4	4	.5	264	.801	.819	288	.801	.839	410	.800	.826
– .5	.4	8	.1	266	.800	.796	291	.800	.812	414	.800	.798
– .5	.4	8	.5	270	.801	.823	296	.800	.807	419	.801	.794
– .5	.5	4	.1	168	.801	.791	184	.800	.808	260	.801	.824
– .5	.5	4	.5	169	.802	.799	186	.802	.817	263	.801	.816
– .5	.5	8	.1	172	.801	.797	188	.802	.782	266	.800	.821
– .5	.5	8	.5	174	.802	.809	189	.801	.805	269	.801	.797
– .5	.6	4	.1	118	.800	.818	128	.802	.784	180	.801	.796
– .5	.6	4	.5	118	.803	.822	130	.800	.796	183	.801	.827
– .5	.6	8	.1	120	.801	.808	130	.802	.802	185	.800	.780
– .5	.6	8	.5	122	.800	.804	132	.801	.804	187	.800	.775
.0	.4	4	.1	246	.800	.794	270	.800	.836	384	.801	.802
.0	.4	4	.5	236	.802	.798	259	.801	.802	368	.800	.819
.0	.4	8	.1	254	.801	.791	278	.801	.793	394	.800	.800
.0	.4	8	.5	244	.802	.798	266	.800	.821	378	.801	.804
.0	.5	4	.1	160	.801	.797	174	.802	.841	246	.800	.824
.0	.5	4	.5	152	.801	.828	166	.801	.837	236	.802	.832
.0	.5	8	.1	164	.800	.801	178	.801	.800	254	.801	.801
.0	.5	8	.5	156	.802	.813	172	.801	.798	244	.802	.819
.0	.6	4	.1	112	.801	.820	122	.802	.816	172	.800	.826
.0	.6	4	.5	106	.804	.812	116	.803	.828	164	.801	.824
.0	.6	8	.1	114	.801	.801	124	.801	.782	176	.801	.810
.0	.6	8	.5	110	.803	.794	120	.801	.819	170	.802	.791
.5	.4	4	.1	234	.800	.794	256	.800	.816	364	.801	.812
.5	.4	4	.5	208	.802	.839	228	.801	.812	324	.801	.825
.5	.4	8	.1	242	.800	.810	264	.801	.818	376	.801	.802
.5	.4	8	.5	216	.802	.833	237	.801	.825	336	.801	.833
.5	.5	4	.1	152	.802	.831	165	.802	.835	234	.800	.834
.5	.5	4	.5	134	.800	.808	147	.803	.820	208	.802	.819
.5	.5	8	.1	155	.800	.809	169	.800	.839	240	.800	.815
.5	.5	8	.5	139	.803	.821	152	.801	.824	216	.802	.811
.5	.6	4	.1	106	.804	.822	116	.803	.833	163	.803	.837
.5	.6	4	.5	94	.802	.835	104	.801	.808	145	.802	.812
.5	.6	8	.1	108	.803	.791	118	.801	.815	168	.800	.814
.5	.6	8	.5	98	.804	.835	107	.802	.824	150	.801	.829

See the note to Table 1.

empirical estimates were below the theoretical estimates in 124 of the 432 cells of the design (28.7%), but the differences were practically insignificant. As we mentioned above, the mean relative bias for the empirical estimates of power remained under 3.6% in all cells and, with few exceptions, the estimates of power fall inside the boundaries of a 95% confidence interval for the theoretical values, suggesting that the trend described above is due to chance. Consistent with the results of Heo et al. (2013), the data indicate that the derived formulas of power are well-validated by simulation studies, which show that the values of theoretical power are very close to those of the empirical power.

In Scenario A, in which complete data across time and homogeneous variances were available, our results revealed that the effect size and a large degree of imbalance between group sizes had decisive impacts on the sample size

Table 4 Sample sizes to obtain theoretical power of at least 80% and the empirical power, with incomplete data and heterogeneous Level 1 and 2 variances

				Δ = .50			Δ = .35			Δ = .20		
r_1	d_L	T	ρ_1	N	1-β	$1-\hat{\beta}$	N	1-β	$1-\hat{\beta}$	N	1-β	$1-\hat{\beta}$
−.5	.4	4	.1	388	.800	.810	384	.801	.815	484	.800	.820
−.5	.4	4	.5	393	.801	.812	390	.800	.817	492	.800	.792
−.5	.4	8	.1	397	.801	.796	394	.800	.791	496	.801	.782
−.5	.4	8	.5	402	.801	.798	398	.800	.804	504	.801	.777
−.5	.5	4	.1	249	.801	.813	246	.801	.829	312	.801	.818
−.5	.5	4	.5	252	.801	.808	250	.801	.813	316	.801	.812
−.5	.5	8	.1	255	.801	.818	252	.801	.824	318	.800	.785
−.5	.5	8	.5	258	.801	.787	256	.801	.793	322	.801	.801
−.5	.6	4	.1	174	.802	.822	172	.801	.825	216	.801	.799
−.5	.6	4	.5	176	.801	.798	174	.801	.818	220	.801	.824
−.5	.6	8	.1	178	.801	.797	176	.801	.806	222	.801	.786
−.5	.6	8	.5	180	.801	.803	178	.801	.795	224	.801	.778
.0	.4	4	.1	369	.801	.819	366	.801	.817	460	.800	.825
.0	.4	4	.5	353	.801	.800	350	.801	.793	440	.800	.829
.0	.4	8	.1	379	.801	.809	374	.800	.798	474	.801	.807
.0	.4	8	.5	363	.801	.818	360	.801	.811	454	.800	.808
.0	.5	4	.1	236	.800	.821	234	.801	.808	296	.800	.824
.0	.5	4	.5	226	.801	.827	224	.802	.809	282	.800	.803
.0	.5	8	.1	243	.801	.794	240	.801	.792	304	.801	.805
.0	.5	8	.5	233	.801	.782	230	.800	.807	292	.801	.812
.0	.6	4	.1	165	.802	.825	164	.802	.831	206	.800	.831
.0	.6	4	.5	158	.803	.828	156	.802	.826	198	.802	.831
.0	.6	8	.1	169	.801	.806	168	.801	.788	212	.801	.797
.0	.6	8	.5	162	.801	.821	160	.800	.805	202	.800	.811
.5	.4	4	.1	349	.800	.816	348	.800	.824	436	.800	.806
.5	.4	4	.5	311	.801	.814	308	.800	.822	388	.800	.826
.5	.4	8	.1	360	.801	.807	356	.801	.775	450	.800	.809
.5	.4	8	.5	323	.801	.805	320	.801	.831	404	.801	.820
.5	.5	4	.1	224	.801	.807	222	.802	.806	280	.801	.811
.5	.5	4	.5	200	.800	.811	198	.801	.827	250	.801	.839
.5	.5	8	.1	231	.801	.796	228	.800	.822	288	.801	.798
.5	.5	8	.5	207	.801	.815	206	.802	.830	258	.800	.818
.5	.6	4	.1	156	.801	.819	154	.800	.813	196	.802	.825
.5	.6	4	.5	140	.802	.820	138	.800	.825	174	.800	.827
.5	.6	8	.1	161	.802	.817	160	.801	.811	202	.802	.810
.5	.6	8	.5	144	.801	.805	144	.802	.829	180	.801	.829

See the note to Table 1.

determination. For instance, when the groups had markedly different sizes (i.e., one group was four times the size of the other), the sample size was required to increase by approximately 50% in order to achieve the same power as in the balanced case; whereas, for an effect size of .40, the sample size that was required to achieve a power comparable to an effect size of .60 was close to a 100% increase. Therefore, careful attention should be paid with regard to the choice among possible population effect sizes and unequal randomization when planning a study. A conservative approach would be to consider the most plausible and choose the smallest effect size among them. On the other hand, the effect of the correlation of the Level 2 residuals and the reliability of measurement at the first time point was not trivial, but the consequences were much less severe. As compared with other similar studies, these results match, to a large degree, the numerical results reported by Usami (2014) using a method proposed by Satorra and Saris (1985) in the context of structural equation modeling.

In the remaining scenarios, our two main findings can be summarized as follows. Firstly, in the presence of heterogeneity in the Level 1 and 2 random effects, larger sample sizes are required in order to obtain the desired nominal power, even for complete and balanced data. One important caveat is that the results were only obtained by the proposed method under positively paired conditions. A positive pairing implies that the treatment condition that has the smallest number of subjects is associated with the smallest variance, whereas the opposite occurs for a negative pairing. Unfortunately, with an unbalanced design similar to that employed in our work (Livacic-Rojas, Vallejo, Fernández, & Tuero, 2017; Vallejo et al., 2008), the tendency to be conservative is worse under negatively paired conditions. The second finding is that, when there is attrition, sample size requirements can be quite large.

Table 5 Percentages of relative bias for predicted theoretical and empirical powers

			Scenario A			Scenario B			Scenario C			Scenario D		
d_L	T	ρ_1	Δ=.5	Δ=.35	Δ=.2	Δ=.5	Δ=.35	Δ=.2	Δ=.5	Δ=.35	Δ=.2	Δ=.5	Δ=.35	Δ=.2
Theoretical power														
.4	4	.1	.2	.1	.0	.1	.0	.0	.0	.0	.1	.0	.1	.0
.4	4	.5	.2	.0	.1	.1	.1	.1	.2	.1	.0	.1	.0	.0
.4	8	.1	.1	.2	.0	.0	.1	.1	.0	.1	.0	.1	.0	.1
.4	8	.5	.1	.2	.1	.1	.1	.0	.2	.0	.1	.1	.1	.1
.5	4	.1	.2	.3	.2	.2	.0	.2	.2	.2	.0	.1	.2	.1
.5	4	.5	.2	.3	.2	.2	.1	.2	.1	.3	.2	.1	.2	.1
.5	8	.1	.2	.2	.1	.1	.1	.1	.0	.1	.0	.1	.1	.1
.5	8	.5	.3	.1	.1	.1	.2	.2	.3	.1	.2	.1	.1	.1
.6	4	.1	.3	.3	.2	.3	.1	.2	.2	.3	.2	.2	.1	.1
.6	4	.5	.2	.3	.3	.2	.3	.1	.4	.2	.2	.3	.1	.1
.6	8	.1	.3	.4	.2	.3	.1	.3	.2	.2	.0	.2	.1	.2
.6	8	.5	.2	.3	.2	.2	.3	.2	.3	.2	.1	.1	.1	.1
Empirical power														
.4	4	.1	1.0	– 0.4	– 1.1	– 0.2	0 .4	– 0.6	0.5	3.5	1.4	1.9	2.3	2.1
.4	4	.5	– 0.4	– 0.5	0.3	– 0.1	– 0.5	– 1.5	2.3	2.2	2.9	1.1	1.3	2.0
.4	8	.1	0.7	0.4	1.6	– 0.1	0.8	– 0.8	– 0.1	1.0	0.1	0.5	– 1.5	– 0.1
.4	8	.5	– 0.1	– 0.5	1.1	– 0.2	1.2	– 0.5	2.3	2.2	1.3	0.9	1.9	0.2
.5	4	.1	0.2	0.6	– 0.6	0.4	– 0.1	0.0	0.8	3.4	3.4	1.7	1.8	2.2
.5	4	.5	0.3	0.1	0.1	1.8	0.6	0.6	1.5	3.1	2.8	1.9	2.0	2.3
.5	8	.1	– 0.5	1.1	– 0.1	0.1	0.4	0.6	0.3	0.9	1.5	0.3	1.6	– 0.5
.5	8	.5	0.5	0.7	– 0.4	0.1	0.1	1.9	1.8	1.1	1.1	– 0.7	1.3	1.3
.6	4	.1	0.0	1.3	0.4	– 0.4	0.7	– 0.6	2.5	0.14	2.5	2.8	2.9	2.3
.6	4	.5	0.4	– 0.6	– 1.2	– 0.5	– 1.2	1.1	2.9	1.3	2.6	1.9	2.9	3.4
.6	8	.1	0.8	1.0	0.6	0.5	0.2	– 0.7	0.0	0.1	0.2	0.8	0.2	– 0.3
.6	8	.5	– 0.2	0.3	– 0.4	0.1	– 0.3	– 1.5	1.4	2.0	– 0.2	1.2	1.2	0.8

See the note to Table 1.

As one can easily imagine, however, it is not clear what is sufficiently large with regard to sample size in order to make valid inferences about the parameter of interest. In many cases an increase of 5% or 10% may be sufficient, but depending on the expected rate of attrition, the appropriate percentage could vary. In the present study we observed that with dropout rates of 10% at every time point (e.g., a condition with eight time points would retain approximately 50% of the original sample at the last time point), the sample size would be required to increase by 20%–25% in order to reach a power that was equivalent to the case of complete data. In any case, when attrition is anticipated, the formulas we derived allow the power to be calculated on the basis of the final number of subjects that are expected to complete the study.

Although the numerical results may change slightly depending on the statistical package and the number of iterations or the algorithm used to estimate the parameters, the simulations presented in this article strongly suggest that on the whole the empirical power based on REML estimates is in fairly good agreement with the theoretical power based on OLS estimates. However, it has also become clear from the present study that, with complex statistical models, sample size estimation using simulations may be needed. One reason why the Monte Carlo power method may be preferred over a theoretical method in some cases is because of its great flexibility to be applied to almost any kind of data, regardless of whether all the model assumptions are satisfied, the type of covariates present, and the attrition rate expected. In fact, the sample size calculation through simulation can easily be extended to more complex linear mixed models or generalized linear mixed models, both univariate and multivariate.

Recommendations

As we noted earlier, when performing a prospective power analysis and no information is available regarding the growth model parameters, researchers may explicitly specify

parameters by indirectly setting four types of indices (ρ_1, k_1, d_L, r_1) for a linear trend. In some cases, this is a reasonable approximation, but in other cases it may become a tricky task. Hence, a range of values often need to be considered.

1. Reliability (ρ_1) depends on what measure is being used. The reader should note, however, that questionnaire measures, which represent one of the most important tools available for data collection in the educational and social sciences, appear to have relatively low reliability. Hence, reliabilities in the .4–.7 range would provide a reasonable starting point when planning research.
2. Empirical studies have indicated that, under most situations likely to be encountered by behavioral science researchers, the ratio between the variances of the outcomes at the end and at the beginning of a study (k_1) could be more than five times smaller than the value we have examined (cf. Hertzog, Lindenberger, Ghisletta, & von Oertzen, 2008). Thus, the sample size requirements will be less demanding than those shown in the tables.
3. The average effect size (d_L) found in published meta-analyses in psychology is around $d_L = 0.50$ (see Bakker, van Dijk, & Wicherts, 2012). An effect size in the range of 0.4–0.6 is regarded as typical. We have not been able to find any guidelines on how to select these effect sizes for a quadratic growth model. Although this issue is an open question and should be investigated, provisionally we have assumed an effect size of one-half of a standard deviation unit for the rate of acceleration (i.e., $d_Q = 0.50$).
4. Although the correlation between the starting point and the rate of change over time (r_1) is not known, precisely different authors (Hertzog et al., 2008; Hox, 2010) have suggested that it is unlikely that this correlation would reach values close to zero in a given population. Hence, correlations in the .25–.50 range would be values that are reasonable to choose when planning a longitudinal study.

Finally, for completeness, three caveats are included. First, it should be clear that the sample size requirements to detect an intervention effect are study-specific. Second, although longitudinal studies often involve small samples, it is very important to emphasize that large samples sizes make small effect sizes detectable. Therefore, researchers interested in carrying out studies that have sufficient power to reject the null hypothesis should avoid using small sample sizes whenever possible. This is especially the case when they are unable to specify a minimum effect size that would have either practical or theoretical significance. Third, it should be noted that the reliabilities studied (i.e., .1 and .5) are on the low side. Since unreliability affects statistical power, it becomes obvious that more positive results should be obtained with higher initial reliabilities. If reliability is improved to .80, for example, the potential reduction in sample size that could be achieved would be approximately 20%. Hence, researchers should make an effort to reduce the effects of measurement error.

Limitations of this study

In our simulation study we saw that the theoretical power values based on the sample size formulas derived using the OLS estimates were nearly identical to the empirical power based on the ML estimates, even with a combination of heterogeneous variances and missing data. However, readers should note that the generalization of our results is limited to situations in which the mechanism for missing data is MAR. When missing data due to attrition are driven by an MAR mechanism, the standard likelihood-based method provides valid inferences about differences in growth rates between groups. In contrast, when the missing data are not MAR (NMAR), the likelihood-based method yields erroneous inferences (failure to control the Type I error rate and to provide altered power). Thus, caution should be exercised if the missingness is thought to be NMAR. To improve the validity of estimates, it is recommended that researchers determine why data are missing and build models that include covariates that may be predictive of dropping out.

An additional limitation of our study is that the results and recommendations are based on assuming normality for the continuous outcome variable. The effect of nonnormality on the power would not be of much consequence in the case of near-normal populations. However, the presence of a fair degree of skewness and/or kurtosis, as is not uncommon in educational and psychological studies (see, e.g., Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013; Cain,Zhang, & Yuan, 2017; Micceri, 1989), would lead to a more conservative alpha level and, thus, to more demanding sample size requirements.

Finally, for computational simplicity, we assumed that the model used only included one categorical variable (e.g., the program studied). However, it is possible to increase precision in the estimation of treatment effects if effective covariates are used in the design. In fact, continuous variables are sometimes included in longitudinal studies as predictors or baseline covariates. In general, as long as the covariates are independent of the group assignments and do not modify the group effects, making an adjustment for baseline response will increase statistical power, because it can be expected that the adjustment will reduce the between- and within-subjects variability.

Author note We are grateful to the Editor, Associate Editor Wei Wu, and reviewers for their constructive suggestions on an draft of this manuscript.

This work has been funded by the Spanish Ministry of Science and Innovation (Ref: PSI-2015-67630-P) and the

Chilean National Fund for Scientific and Technological Development (FONDECYT. Ref.: 1170642).

Appendix 1

$$Var\left(Y_{ij}|X_{1ij},\ldots,X_{pij}\right)=E\left[Y_{ij}-E\left(Y_{ij}\right)\right]^2=E\left(Y_{ij}\right)^2-E\left[\left(Y_{ij}\right)\right]^2$$

$$=E\left(u_{0j}+\sum_{p=1}^{P}u_{pj}X_{pij}+e_{ij}\right)^2$$

$$=E\left(u_{0j}^2+\sum_{p=1}^{P}u_{pj}^2X_{pij}^2+2\sum_{p=1}^{P}u_{0j}u_{pj}X_{pij}+e_{ij}^2\right)$$

$$=E\left(u_{0j}^2+\sum_{p=1}^{P}X_{pij}^2E\left(u_{pj}^2\right)+2\sum_{p=1}^{P}X_{pij}E\left(u_{0j}u_{pj}\right)+E\left(e_{ij}^2\right)\right)$$

$$=\tau_{00}+2\sum_{p=1}^{P}\tau_{0p}X_{pij}+\sum_{p=1}^{P}\tau_{pp}X_{pij}^2+\sigma^2$$

$$Cov\left(Y_{ij},Y_{i,j}\middle|X_{1ij},\ldots,X_{pij}\right)=E\left[\left(Y_{ij}-E\left(Y_{ij}\right)\right)\left(Y_{i,j}-E\left(Y_{i,j}\right)\right)\right]=E\left(Y_{ij}Y_{i,j}\right)-E\left(Y_{ij}\right)E(Y_i,j)$$

$$=E\left[\left(u_{0j}+\sum_{p=1}^{P}u_{pj}X_{pij}+e_{ij}\right)\left(u_{0j}+\sum_{p=1}^{P}u_{pj}X_{pi,j}+e_{i,j}\right)\right.$$

$$\left.=E\left(u_{0j}^2\right)+\sum_{p=1}^{P}X_{pij}E\left(u_{0j},u_{pj}\right)+\sum_{p=1}^{P}X_{pi,j}E\left(u_{0j},u_{pj}\right)+\sum_{p=1}^{P}\left(X_{pij}X_{pi,j}\right)E\left(u_{pj}^2\right)\right]$$

$$=\tau_{00}+\sum_{p=1}^{P}\tau_{0p}\left(X_{pij}+X_{pi,j}\right)+\sum_{p=1}^{P}\tau_{pp}\left(X_{pij}X_{pi,j}\right)$$

Appendix 2

Calculation of variance component τ_{01}

$$r_1=\frac{\tau_{01}}{\sqrt{\tau_{00}\tau_{11}}};k_1=\frac{\tau_{00}+2D\tau_{01}+D^2\tau_{11}+\sigma^2}{\tau_{00}+\sigma^2}$$

$$\Rightarrow\begin{vmatrix}k_1\left(\tau_{00}+\sigma^2\right)=\tau_{00}+2D\tau_{01}+D^2\tau_{11}+\sigma^2\\ \tau_{11}=\frac{\tau_{01}^2}{r_1^2\tau_{00}}\end{vmatrix}$$

$$\Rightarrow(k_1-1)(\tau_{00}+\sigma^2)=2D\tau_{01}+\tfrac{D^2\tau_{01}^2}{r_1^2\tau_{00}}$$

$$\Rightarrow(k_1-1)(\tau_{00}+\sigma^2)r_1^2\tau_{00}=2D\tau_{01}r_1^2\tau_{00}+D^2\tau_{01}^2$$

$$\Rightarrow D^2\tau_{01}^2+2D\tau_{01}r_1^2\tau_{00}-\left(k_1-1\right)(\tau_{00}+\sigma^2)r_1^2\tau_{00}=0$$

τ_{01} parameter is one of the roots of the quadratic function

$$\tau_{01}=\left(-B_{01}\pm\sqrt{B_{01}^2-4A_{01}C_{01}}/2A_{01}\right)$$

where

$$A_{01}=D^2,B_{01}=2Dr_1^2\tau_{00}\ and\ C_{01}$$

$$=-(k_1-1)\left(\tau_{00}+\sigma^2\right)r_1^2\tau_{00}$$

Calculation of variance component τ_{11}

$$r_1=\frac{\tau_{01}}{\sqrt{\tau_{00}\tau_{11}}};k_1=\frac{\tau_{00}+2D\tau_{01}+D^2\tau_{11}+\sigma^2}{\tau_{00}+\sigma^2}$$

$$\Rightarrow\begin{vmatrix}k_1\left(\tau_{00}+\sigma^2\right)=\tau_{00}+2D\tau_{01}+D^2\tau_{11}+\sigma^2\\ \tau_{01}=r_1\sqrt{\tau_{00}\tau_{11}}\end{vmatrix}$$

$$\Rightarrow(k_1-1)\left(\tau_{00}+\sigma^2\right)=2Dr_1\sqrt{\tau_{00}\tau_{11}}+D^2\tau_{11}$$

$$\Rightarrow(2Dr_1\sqrt{\tau_{00}\tau_{11}})^2=\left[(k_1-1)\left(\tau_{00}+\sigma^2\right)-D^2\tau_{01}^2\right]^2$$

$$\Rightarrow D^4\tau_{11}^2-4D^2r_1^2\tau_{00}\tau_{11}-2(k_1-1)\left(\tau_{00}+\sigma^2\right)$$

$$D^2\tau_{11}+(k_1-1)^2\left(\tau_{00}+\sigma^2\right)^2=0$$

τ_{11} parameter is one of the roots of the quadratic function

$$\tau_{11} = \left(-B_{11} \pm \sqrt{B_{11}^2 - 4A_{11}C_{11}}\right)/2A_{11}$$

where

$$A_{11} = D^4, B_{01} = -4D^2 r_1^2 \tau_{00} - 2(k_2 - 1)\left(\tau_{00} + \sigma^2\right)D^2 \text{ and } C_{11} = (k_1 - 1)^2\left(\tau_{00} + \sigma^2\right)^2.$$

Calculation of variance component τ_{02}

$$r_2 = \frac{\tau_{02}}{\sqrt{\tau_{00}\tau_{22}}}; r_{12} = \frac{\tau_{12}}{\sqrt{\tau_{11}\tau_{22}}}; k_2 = \frac{\tau_{00} + 2D\tau_{01} + D^2\tau_{11} + 2D^2\tau_{02} + 2D^3\tau_{12} + D^4\tau_{22} + \sigma^2}{\tau_{00} + \sigma^2}$$

$$\Rightarrow \left| \begin{array}{l} k_2\left(\tau_{00} + \sigma^2\right) = \tau_{00} + 2D\tau_{01} + D^2\tau_{11} + 2D^2\tau_{02} + 2D^3\tau_{12} + D^4\tau_{22} + \sigma^2 \\ \tau_{12} = r_{12}\sqrt{\tau_{11}\tau_{22}} \text{ y } \tau_{22} = \dfrac{\tau_{02}^2}{r_2^2\tau_{00}} \end{array} \right.$$

$$\Rightarrow (k_2\text{-}1)\left(\tau_{00} + \sigma^2\right) = 2D\tau_{01} + D^2\tau_{11} + 2D^2\tau_{02} + 2D^3 r_{12}\sqrt{\frac{\tau_{11}\tau_{02}^2}{r_2^2\tau_{00}}} + \frac{D^4\tau_{02}^2}{r_2^2\tau_{00}}$$

$$\Rightarrow (k_2\text{-}1)\left(\tau_{00} + \sigma^2\right)r_2^2\tau_{00} = 2D\tau_{01}r_2^2\tau_{00} + D^2\tau_{11}r_2^2\tau_{00} + 2D^2\tau_{02}r_2^2\tau_{00} + 2D^3 r_{12}\sqrt{\frac{\tau_{11}\tau_{02}^2}{r_2^2\tau_{00}}}r_2^2\tau_{00} + D^4\tau_{02}^2$$

$$\Rightarrow D^4\tau_{02}^2 + 2D^2\tau_{02}r_2^2\tau_{00} + 2D^3 r_{12}\sqrt{(\tau_{11}/\tau_{00})}r_2\tau_{00}\tau_{02} + 2D\tau_{01}r_2^2\tau_{00} + D^2\tau_{11}r_2^2\tau_{00}\text{-}(k_2\text{-}1)\left(\tau_{00} + \sigma^2\right)r_2^2\tau_{00} = 0$$

τ_{02} parameter is one of the roots of quadratic function

$$\tau_{02} = \left(-B_{02} \pm \sqrt{B_{02}^2 - 4A_{02}C_{02}}\right)/2A_{02}$$

where

$$A_{02} = D^4, B_{02} = 2D^2 r_2^2\tau_{00} + 2D^3 r_{12}\sqrt{(\tau_{11}/\tau_{00})}r_2\tau_{00} \text{ and}$$

$$C_{02} = 2D\tau_{01}r_2^2\tau_{00} + D^2\tau_{11}r_2^2\tau_{00}\text{-}(k_2\text{-}1)\left(\tau_{00} + \sigma^2\right)r_2^2\tau_{00}.$$

Calculation of variance component τ_{12}

$$\Rightarrow \left| \begin{array}{l} k_2\left(\tau_{00} + \sigma^2\right) = \tau_{00} + 2D\tau_{01} + D^2\tau_{11} + 2D^2\tau_{02} + 2D^3\tau_{12} + D^4\tau_{22} + \sigma^2 \\ \tau_{02} = r_2\sqrt{\tau_{00}\tau_{22}} \text{ y } \tau_{22} = \dfrac{\tau_{12}^2}{r_{12}^2\tau_{11}} \end{array} \right.$$

$$\Rightarrow (k_2\text{-}1)\left(\tau_{00} + \sigma^2\right) = 2D\tau_{01} + D^2\tau_{11} + 2D^2 r_2\sqrt{\tau_{00}\tau_{22}} + 2D^3\tau_{12} + \frac{D^4\tau_{12}^2}{r_{12}^2\tau_{11}}$$

$$\Rightarrow (k_2\text{-}1)\left(\tau_{00} + \sigma^2\right)r_{12}^2 r_{11} = 2D\tau_{01}r_{12}^2\tau_{11} + D^2\tau_{11}r_{12}^2\tau_{11} + 2D^2 r_2\sqrt{\frac{\tau_{00}\tau_{12}^2}{r_{12}^2\tau_{11}}}r_{12}^2\tau_{11} + 2D^3\tau_{12}r_{12}^2\tau_{11} + D^4\tau_{12}^2$$

$$\Rightarrow D^4\tau_{12}^2 + 2D^2 r_2\sqrt{(\tau_{00}/\tau_{11})}r_{12}\tau_{11}\tau_{12} + 2D^3\tau_{12}r_{12}^2\tau_{11} + 2D\tau_{01}r_{12}^2\tau_{11} + D^2\tau_{11}r_{12}^2\tau_{11}\text{-}(k_2\text{-}1)\left(\tau_{00} + \sigma^2\right)r_{12}^2\tau_{11} = 0$$

τ_{12} parameter is one of the roots of quadratic function

$$\tau_{12} = \left(-B_{12} \pm \sqrt{B_{12}^2 - 4A_{12}C_{12}}\right)/2A_{12}$$

where

$$A_{12} = D^4, B_{12} = 2D^2 r_2\sqrt{(\tau_{00}/\tau_{11})}r_{12}\tau_{11} + 2D^3 r_{12}^2\tau_{11} \text{ and}$$

$$C_{12} = 2D\tau_{01}r_{12}^2\tau_{11} + D^2\tau_{11}r_{12}^2\tau_{11} - (k_2 - 1)\left(\tau_{00} + \sigma^2\right)r_{12}^2\tau_{11}.$$

Calculation of variance component τ_{22}

$$r_2 = \frac{\tau_{02}}{\sqrt{\tau_{00}\tau_{22}}}; r_{12} = \frac{\tau_{12}}{\tau_{11}\tau_{22}}; k_2 = \frac{\tau_{00} + 2D\tau_{01} + D^2\tau_{11} + 2D^2\tau_{02} + 2D^3\tau_{12} + D^4\tau_{22} + \sigma^2}{\tau_{00} + \sigma^2}$$

$$\Rightarrow \left| \begin{array}{l} k_2\left(\tau_{00} + \sigma^2\right) = \tau_{00} = 2D\tau_{01} + D^2\tau_{11} + 2D^2\tau_{02} + 2D^3\tau_{12} + D^4\tau_{22} + \sigma^2 \\ \tau_{02} = r_2\sqrt{\tau_{00}\tau_{22}}\, y\tau_{12} = r_{12}\sqrt{\tau_{11}\tau_{22}} \end{array} \right.$$

$$\Rightarrow (k_2{-}1)\left(\tau_{00} + \sigma^2\right) = 2D\tau_{01} + D^2\tau_{11} + 2D^2 r_2\sqrt{\tau_{00}\tau_{22}} + 2D^3 r_{12}\sqrt{\tau_{11}\tau_{22}} + D^4\tau_{22}$$

$$\Rightarrow (k_2{-}1)\left(\tau_{00} + \sigma^2\right) = 2D\tau_{01} + D^2\tau_{11} + D^4\tau_{22} = 2D^2 r_2\sqrt{\tau_{00}\tau_{22}} + 2D^3 r_{12}\sqrt{\tau_{11}\tau_{22}}$$

$$\Rightarrow \left(2D^2 r_2\right)\sqrt{\tau_{00}\tau_{22}} = 2D^3 r_{12}\sqrt{\tau_{11}\tau_{22}}\Big)^2 = \left[(k_2{-}1)\left(\tau_{00} + \sigma^2\right){-}2D\tau_{01}{-}D^2\tau_{11}{-}D^4\tau_{22}\right]^2$$

$$\Rightarrow D^8\tau_{22}^2{-}\tau_{22}\left(2D^2 r_2\sqrt{\tau_{00}} + 2D^3 r_{12}\sqrt{\tau_{11}}\right)^2 + 2D^6\tau_{11}\tau_{22} + 4D^5\tau_{01}\tau_{22}{-}2(k_2{-}1)\left(\tau_{00} + \sigma^2\right)D^4\tau_{22} + 4D^2\tau_{01}^2 + D^4\tau_{11}^2$$

$$+(k_2{-}1)^2\left(\tau_{00} + \sigma^2\right)^2{-}2(k_2{-}1)\left(\tau_{00} + \sigma^2\right) \times \left(2D\tau_{01} + D^2\tau_{11}\right) + 4D^3\tau_{01}\tau_{11} = 0$$

τ_{22} parameter is one of the roots of quadratic function $\tau_{22} = \frac{-B_{22} \pm \sqrt{B_{22}^2 - 4A_{11}C_{22}}}{2A_{22}}$

where

$$A_{22} = D^8, B_{22} = -\left(2D^2 r_2\sqrt{\tau_{00}} + 2D^3 r_{12}\sqrt{\tau_{11}}\right)^2 + 2D^6\tau_{11}$$

$$+4D^5\tau_{01}{-}2D^4(k_2{-}1)\left(\tau_{00} + \sigma^2\right) \text{and } C_{22} = 4D^2\tau_{01}^2$$

$$+D^4\tau_{11}^2 + (k_2{-}1)^2\left(\tau_{00} + \sigma^2\right)^2{-}2(k_2{-}1)\left(\tau_{00} + \sigma^2\right)$$

$$\times\left(2D\tau_{01} + D^2\tau_{11}\right) + 4D^3\tau_{01}\tau_{11}.$$

Appendix 3

To extend the proposed method to addressing sample size calculations for future data, the following piecewise growth model for two separate linear pieces will be used:

Level 1:

$$Y_{it} = b_{0i} + b_{1i}X_{1it} + b_{2i}X_{2it} + e_{it}$$

Level 2:

$$b_{0i} = \beta_{00} + u_{0i},$$
$$b_{1i} = \beta_{10} + \beta_{11}W_i + u_{1i},$$
$$b_{2i} = \beta_{20} + \beta_{21}W_i + u_{2i},$$

Note that because of the randomization of subjects to the two treatment groups, Level 2 for the intercept does not contain the value of the group-level variable W_i, and we assume a common mean response at baseline. Substituting the corresponding Level 2 equations into the Level 1 equation, we get the combined model:

$$Y_{it} = \beta_{00} + (\beta_{10} + \beta_{11}W_i)X_{1it} + (\beta_{20} + \beta_{21}W_i)X_{2it} + (u_{0i} + u_{1i}X_{1ti} + u_{2i}X_{2ti} + e_{it}),$$

where X_{1it} and X_{2it} are coded variables to represent the piecewise regression. In this case, X_{1it} denotes the time of the tth measurement on the ith subject, while the variable X_{2it} would be coded as $X_{2it} = X_{it}$ if $(X_{it} - Bp) > 0$ and $X_{2it} = 0$ *if* $(X_{it} - Bp) \le 0$.

From this single equation model, the expected value and variance–covariance structures of Y_{it} given W_i can be expressed as

$$E(Y_{it}|W_i) = \beta_{00} + (\beta_{10} + \beta_{11}W_i)X_{1it} + (\beta_{20} + \beta_{21}W_i)X_{2it},$$

$$Var(Y_{it}|W_i) = \tau_{00} + 2X_{1it}\tau_{01} + X_{1it}^2\tau_{11} + 2X_{2it}\tau_{02} + 2X_{1it}X_{2it}\tau_{12} + X_{2it}^2\tau_{22} + \sigma^2,$$

$$Cov(Y_{it}, Y_{it}, |W_i) = \tau_{00} + \left(X_{1it} + X_{1it'}\right)\tau_{01} + X_{1it}X_{1it'}\tau_{11} + \left(X_{2it} + X_{2it'}\right)\tau_{02} + \left(X_{1it}X_{2it'} + X_{2it}X_{1it'}\right)\tau_{12} + \left(X_{2it}X_{2it'}\right)\tau_{22}.$$

To generalize the proposed procedure to more complicated piecewise models, it is fundamental to use transformed indices that are easy to specify. In addition to those specified in Eqs. 13–16 and 19 (i. e., ρ_1, d_L, r_1,

k_1 and β_{11}), this new situation requires the indices to be defined as follows:

$$d_{PW} = \frac{D_1\beta_{11} + D_2^2\beta_{21}}{\sqrt{\tau_{00} + 2D_1\tau_{01} + D_1^2\tau_{11} + 2D_2\tau_{02} + 2D_1D_2\tau_{12} + D_2^2\tau_{22} + \sigma^2}},$$

$$r_2 = \frac{Cov\quad(u_{0i}, u_{2i})}{\sqrt{Var\quad(u_{0i}\quad u_{2i})}} = \frac{\tau_{02}}{\sqrt{\tau_{00}\quad\tau_{22}}},$$

$$r_{12} = \frac{Cov\quad(u_{0i}, u_{2i})}{\sqrt{Var\quad(u_{0i}\quad u_{2i})}} = \frac{\tau_{12}}{\sqrt{\tau_{11}\quad\tau_{22}}},$$

and

$$k_2 = \frac{Var\quad(Y_{iT})}{Var\quad(Y_{i1})} = \frac{\tau_{00} + 2D_1\tau_{01} + D_1^2\tau_{11} + 2D_2\tau_{02} + 2D_1D_2\tau_{12} + D_2^2\tau_{22} + \sigma^2}{\tau_{00} + \sigma^2}.$$

By solving simultaneous equations following a procedure similar to that described in Appendix 2, we obtain the components of variance τ_{02}, τ_{12} *and* τ_{22} associated with the second of the two piecewise slopes of the linear growth model. Specifically,

$$\tau_{02} = \frac{-B_{02} \pm \sqrt{B_{02}^2 - 4A_{02}C_{02}}}{2A_{02}},$$

where $A_{02} = D_2^2; B_{02} = 2D_2r_2^2\tau_{00} + 2D_1D_2r_{12}\sqrt{(\tau_{11}/\tau_{00})}r_2\tau_{00}; C_{02} = 2D_1\tau_{01}r_2^2\tau_{00} + D_1^2\tau_{11}r_2^2\tau_{00} - (k_2-1)(\tau_{00} + \sigma^2)\, r_2{}^2\tau_{00};$

$$\tau_{12} = \frac{-B_{12} \pm \sqrt{B_{12}^2 - 4A_{12}C_{12}}}{2A_{12}},$$

where $A_{12} = D_2^2; B_{12} = 2D_2r_2\sqrt{(\tau_{00}/\tau_{11})}r_{12}\tau_{11} + 2D_1D_2 r_{12}^2\tau_{11}; C_{12} = 2D\tau_{01}r_{12}^2\tau_{11} + D^2\tau_{11}\, r_{12}^2\tau_{11} - (k_2-1)(\tau_{00} + \sigma^2)\, r_{12}^2\tau_{11}$; and

$$\tau_{22} = \frac{-B_{22} \pm \sqrt{B_{22}^2 - 4A_{22}C_{22}}}{2A_{22}},$$

where $A_{22} = D_2^4; B_{22} = -\left(2D_2r_2\sqrt{\tau_{00}} + 2D_1\quad D_2r_{12}\sqrt{\tau_{11}}\right)^2 + 2D_1^2\, D_2^2\tau_{11} + 4D_1D_2^2\tau_{01} - 2D_2^2\ (k_2-1)\left(\tau_{00} + \sigma^2\right)$; $C_{22} = 4\, D_1^2\tau_{01}^2 + D_1^4\tau_{11}^2 + (k_2-1)\quad{}^2(\tau_{00} + \sigma^2)\quad{}^2 + 4D_1^3\tau_{01}\tau_{11} - 2\,(k_2-1)\,(\tau_{00} + \sigma^2)\,\left(2D_1\tau_{01} + D_1^2\tau_{11}\right)$.

Finally, by substituting in the effect size formula (*i. e.*, d_{PW}) the value found for *Var* (Y_{iT}) in the formula used to define the ratio between the variances of outcomes at the beginning and end of the study (*i. e.*, k_2), the coefficient associated with differences between treatment conditions from the breakpoint to the end of the study can be written as

$$\beta_{21} = \frac{d_{PW}\sqrt{k_2(\tau_{00} + \sigma^2)} - d_L\sqrt{k_1(\tau_{00} + \sigma^2)}}{D_2}.$$

The components of variance associated with the first of the two piecewise slopes (i. e., τ_{01} *and* τ_{11}) coincide with those obtained for the linear growth model with D replaced by D_1.

Appendix 4

OLS estimation of the effect of the interaction of treatment by time with three groups:

$$\hat{\beta}_{11} = \left\{ \left[\frac{\sum_{i=1}^{N_2}\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)Y_{it}}{\sum_{i=1}^{N_2}\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)^2} - \frac{\sum_{i=1}^{N_1}\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)Y_{it}}{\sum_{i=1}^{N_1}\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)^2} \right] + \left[\frac{\sum_{i=1}^{N_3}\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)Y_{it}}{\sum_{i=1}^{N_3}\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)^2} - \frac{\sum_{i=1}^{N_1}\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)Y_{it}}{\sum_{i=1}^{N_1}\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)^2} \right] + \left[\frac{\sum_{i=1}^{N_3}\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)Y_{it}}{\sum_{i=1}^{N_3}\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)^2} - \frac{\sum_{i=1}^{N_2}\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)Y_{it}}{\sum_{i=1}^{N_2}\sum_{t=1}^{T}\left(X_{it} - \overline{X}_i\right)^2} \right] \right\} / 4$$

With four groups we would proceed similarly, but dividing instead by 10. With five groups, we should divide by 20.

Appendix 5

If we assume that the subject outcomes are independent, the variance of the numerator of Eq. 28 can be decomposed as follows:

$$\mathrm{Var}\left(\sum_{i=1}^{N_j}\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)Y_{it}\right)=\mathrm{A}+\mathrm{B}.$$

Developing the terms of the second member of the above expression, we obtain

$$\begin{aligned}
\mathrm{A}&=\sum_{i=1}^{N_j}\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^2 Var\left(Y_{it}\right)=\sum_{i=1}^{N_j}\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^2\left(\tau_{00}+2_{it}\tau_{01}+X_{it}^2\tau_{11}+\sigma^2\right)\\
&=\sum_{i=1}^{N_j}\sum_{t=1}^{T}\left[\left(X_{it}-\overline{X}_i\right)^2\tau_{00}+\left(X_{it}-\overline{X}_i\right)^2 2X_{it}\tau_{01}+\left(X_{it}-\overline{X}_i\right)^2 X_{it}^2\tau_{11}+\left(X_{it}-\overline{X}_i\right)^2\sigma^2\right]\\
&=\left(\tau_{00}+\sigma^2\right)N_j\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^2+2\tau_{01}N_j\sum_{t=1}^{T}X_{it}\left(X_{it}-\overline{X}_i\right)^2+\tau_{11}N_j\sum_{t=1}^{T}X_{it}^2\left(X_{it}-\overline{X}_{\mathrm{i}}\right)^2\\
\mathrm{B}&=\sum_{i=1}^{N_j}\sum_{t=1}^{T}\sum_{t'\neq t}^{T}\left(X_{it}-\overline{X}_i\right)\left(X_{it'}-\overline{X}_i\right)Cov(Y_{it},Y_{it},)=\sum_{i=1}^{N_j}\sum_{t=1}^{T}\sum_{t'\neq t}^{T}Z_tZ_t,\left[\tau_{00}+(\mathrm{X}_{it}+\mathrm{X}_{it},)\tau_{01}+\mathrm{X}_{it}\mathrm{X}_{it},\tau_{11}\right],\\
&=\tau_{11}N_j\sum_{t=1}^{T}\sum_{t'\neq t}^{T}\left(X_{it}-\overline{X}_i\right)\left(X_{it},-\overline{X}_i\right)X_{it}X_{it},-\tau_{00}N_j\sum_{t=1}^{T}\left(X_{it},-\overline{X}_i\right)^2-2\tau_{01}N_j\sum_{t=1}^{T}X_{it}\left(X_{it}-\overline{X}_i\right)^2
\end{aligned}$$

where it follows that:

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1t}^{N_j}\sum_{t=1}^{T}\left(X_{it}-\overline{X_i}\right)Y_{it}\right)&=\left(\tau_{00}+\sigma^2=\tau_{00}\right)N_j\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^2+2\tau_{01}N_j\sum_{t=1}^{T}X_{it}\left(X_{it}-\overline{X_i}\right)^2-\\
&\quad 2\tau_{01}N_j\sum_{t=1}^{T}X_{it}\left(X_{it}-\overline{X}_i\right)^2+\tau_{11}N_j\left[\sum_{t=1}^{T}X_{it}^2\left(X_{it}-\overline{X}_i\right)^2+\sum_{t=1}^{T}\sum_{t'\neq t}^{T}\left(X_{it}-\overline{X}_i\right)\left(X_{it},-\overline{X}_i\right)X_{it}X_{it},\right]\\
&=\sigma^2N_j\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^2+\tau_{11}N_j\left[\sum_{t=1}^{T}X_{it}^2\left(X_{it}-\overline{X}_i\right)^2+\sum_{t=1}^{T}\sum_{t'\neq t}^{T}\left(X_{it}-\overline{X}_i\right)\left(X_{it},-\overline{X}_i\right)X_{it}X_{it},\right]\\
&=\sigma^2N_j\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^2+\tau_{11}N_j\sum_{t}^{T}\sum_{t'}^{T}\left(X_{it}-\overline{X}_i\right)\left(X_{it},-\overline{X}_i\right)X_{it}X_{it},\\
&=\sigma^2N_j\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^2+\tau_{11}N_j\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^4
\end{aligned}$$

Therefore, the variance of the estimator of the linear slope for *j*th group, $\hat{\delta}_j (j = 0, 1)$, is

$$\mathrm{Var}\left(\hat{\delta}_j\right) = \frac{\mathrm{Var}\left(\sum_{i=1}^{N_j}\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)Y_{it}\right)}{\mathrm{Var}\left(\sum_{i=1}^{N_j}\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^2\right)} = \frac{\sigma^2 N_j \sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^2}{N_j^2\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^4} + \frac{\tau_{11} N_j \sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^4}{N_j^2\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^4}$$

$$= \frac{\sigma^2}{N_j\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^2} + \frac{\tau_{11}}{N_j} = \frac{1}{N_j}\left(\frac{\sigma^2}{\sum_{t=1}^{T}\left(X_{it}-\overline{X}_i\right)^2} + \tau_{11}\right)$$

Finally, using the properties of variance when variables are independent, we arrive at

$$\mathrm{Var}\left(\hat{\beta}_{11}\right) = \mathrm{Var}\left(\hat{\delta}_1\right) + Var\left(\hat{\delta}_0\right) = \frac{4}{N}\left(\frac{\sigma^2}{\sum_{t=1}^{T}\left(X_{\mathrm{it}}-\overline{X}_{\mathrm{i}}\right)^2} + \tau_{11}\right)$$

Remember that N denotes here the total number of Level 2 units included in study, with $N/2$ subjects in each group. The quantity $4/N$ on the right side of variance formula should be replaced by $(1/Np_1p_2)$ in order to allow for groups of unequal size, where $p_1 = N_C/N$ and $p_2 = N_E/N''$.

Appendix 6

Equating Eqs. 36 and 37 and solving for N, we get

$$Z_{1-\alpha/2}\sqrt{\sigma_b^2/Np_1p_2} = \beta_{11} - Z_{1-\beta}\sqrt{\sigma_b^2/Np_1p_2}$$
$$\beta_{11} = Z_{1-\alpha/2}\sqrt{\sigma_b^2/Np_1p_2} + Z_{1-\beta}\sqrt{\sigma_b^2/Np_1p_2}$$
$$\beta_{11} = \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)\sqrt{\sigma_b^2/Np_1p_{22}}$$
$$\beta_{11}^2 Np_1p_2 = \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2\sigma_b^2$$
$$N = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2\sigma_b^2}{\beta_{11}^2 p_1 p_2}$$

Appendix 7

```
*SAS program that reproduced the same results as the first example
(Núñez et al., 2013) which have been described previously (linear
model);
proc iml;
%let za=probit(1-alpha/2);
%let zb=probit(1-beta);
%let t01=(r1*(sqrt(r1##2*t00##2+t00*(k1-1)*(t00+sig)))-r1##2*t00)/d;
%let t11=((sqrt(r1##2*t00+(k1-1)*(t00+sig))-r1*sqrt(t00))##2)/d##2;
%let
lambda1=trace(inv(a*(v1)*a`)*((c*b*a`)`*inv(c*inv(x`*x)*c`)*(c*b*a`)))
;
%let lambda2=(n*p1*p2*b11##2)/(p2*cov+p1*cov);
%let lambda_a=(n_a*p1*p2*b11##2)/(p2*cov+p1*cov);
%let power=1-probf(finv(1-alpha,ndf,ddf),ndf,ddf,lambda1);
%let power_a=1-probf(finv(1-alpha,ndf,ddf_a),ndf,ddf_a,lambda_a);
%let icc=t00/(t00+sig);
%let ddf=n*t-2;
%let ddf_a=n_a*t-2;

alpha=0.05;beta=0.2;
t00=0.07076;sig=.08649;icc=&icc;d_l=0.74543;k1=1.46834;r1=0.25231;
b01=0.1169;n1=47;n2=47;n=n1+n2;t=4;d=t-1;p1=n1/n;p2=1-p1;f=(t-1)/d;
gc=0;ge=1;ngr=2;ndf=1;ddf=&ddf;l_lin=1;

do case=1 to 3 by 1;
if case=1 then do;heter=1;prop_missing=0;rho=1-prop_missing;end;
if case=2 then do;heter=1;prop_missing=.05;rho=1-prop_missing;end;
if case=3 then do;heter=3;prop_missing=0;rho=1-prop_missing;end;

t01=&t01;
t11=&t11;
tt=(t00||t01)`||(t01||t11)`;e=(i(t)*sig);
z0=(j(1,t,1))`;z1=((do(0,d,1))-sum((do(0,d,1))/t))`;z=(z0||z1)`;
v=z`*tt*z+e;
b11=(-b01+d_l*(sqrt(k1*(t00+sig))))/d;

*Matrix of missingness;
start ar1(t,v,rho);
      u = cuprod(j(1,t-1,rho));
      return(v##2#toeplitz(1||u));
finish;
ar1 = ar1(t,1,rho);
perr=ar1[1,];
matriz_p=j(t,t,1);
      do i=1 to t;
      do j=1 to i;
matriz_p[i,j]=sqrt(perr[i])*sqrt(perr[j]);
```

```
matriz_p[j,i]=matriz_p[i,j];
end; end;
m_m=j(t,t,rho);
start setdiag(m_m, v);
   diagidx = do(1,nrow(m_m)*ncol(m_m), ncol(m_m)+1);
   m_m[diagidx] = v;
finish;
run setdiag(m_m, 0);
b_d=diag(matriz_p);
m_p=m_m+b_d;

*Calculating noncentrality parameter and power values;
start ini (a,c,b,x) global(ngr,t,z1,b11,gc,ge,n,n1,n2);
      nx=(n1-0)||(n2-0);
      do i=1 to ncol(nx);
      nx1=j(nx[i],1,i);
      if i=1 then x=nx1;
            else x=x//nx1;
      end;
      x=design(x);
      a=(j(t-1,1,1)||(-i(t-1)));
      b=((b11*gc)*z1)`//((b11*ge)*z1)`;
      c = j(ngr-1,1,1)||(-i(ngr-1));
finish;
call ini (a,c,b,x);

v1=(p2*v+p1*v*heter)/m_pcov=(inv(z*inv(v1)*z`))[2,2];
v_l1=(sig*f##(2*l_lin)*(fact(t-l_lin-
1)))/((1/12)*(fact(t+l_lin)))+t11;
v_l2=v_l1*heter;
cov_l=(p2*v_l1+p1*v_l2)/(rho##2);
lambda1=&lambda1;
power=&power;
lambda2=&lambda2;

*Total sample size calculation;
n_n=(((&za+&zb)##2)*cov)/(b11##2*p1*p2);
nt=(ceil(n_n));          n_t=ceil(nt);
nc=((n_t)*p1);           n_c=ceil(nc);
ne=(n_t-n_c);            n_e=ceil(ne);
n_a=n_c+n_e;

*Adequate power to detect statistical significance;
ddf_a=&ddf_a;
lambda_a=&lambda_a;
power_a=&power_a;

print lambda1 lambda2 cov power n_a lambda_a power_a;
end;
proc print;

*SAS program that reproduced the same results as the second example
(Rosário et al., 2017) which have been described previously (quadratic
model);
proc iml;
%let za=probit(1-alpha/2);
%let zb=probit(1-beta);
%let t01=(r1*(sqrt(r1##2*t00##2+t00*(k1-1)*(t00+sig)))-r1##2*t00)/d;
%let t11=((sqrt(r1##2*t00+(k1-1)*(t00+sig))-r1*sqrt(t00))##2)/d##2;
```

```
%let
lambda1=trace(inv(a*(v1)*a`)*((c*b*a`)`*inv(c*inv(x`*x)*c`)*(c*b*a`)))
;
%let lambda2=(n*p1*p2*b21##2)/(p2*cov+p1*cov);
%let lambda_a=(n_a*p1*p2*b21##2)/(p2*cov+p1*cov);
%let power=1-probf(finv(1-alpha,ndf,ddf),ndf,ddf,lambda1);
%let power_a=1-probf(finv(1-alpha,ndf,ddf_a),ndf,ddf_a,lambda_a);
%let icc=t00/(t00+sig);
%let ddf=n*t-2;
%let ddf_a=n_a*t-2;

alpha=0.05;beta=0.2;
t00=45.0677;sig=21.1842;icc=&icc;d_l=.2866;d_q=-.3106;
k1=1.3262;k2=2.1824;r1=-0.2747;r2=-0.4756;r12=-0.1574;
n1=91;n2=91;n=n1+n2;p1=n1/n;p2=1-p1;t=13;d=t-1;f=(t-1)/d;
gc=0;ge=1;ngr=2;ndf=1;ddf=&ddf;l_lin=1;c_cua=2;

do case=1 to 3 by 1;
if case=1 then do;heter=1;prop_missing=0;rho=1-prop_missing;end;
if case=2 then do;heter=1;prop_missing=.05;rho=1-prop_missing;end;
if case=3 then do;heter=3;prop_missing=0;rho=1-prop_missing;end;

*Calculation of an adequate t01 t11 t02 t12 t22;
t01=&t01;
t11=&t11;

a=d##4;
b=2*d##2*r2##2*t00+2*d##3*r12*sqrt(t11/t00)*r2*t00;
c=2*d*t01*r2##2*t00+d##2*t11*r2##2*t00-(k2-1)*(t00+sig)*r2##2*t00;
t02=(-b-sqrt(b##2-4*a*c))/(2*a);

a=d##4;
b=2*d##3*r12##2*t11+2*d##2*r2*sqrt(t00/t11)*r12*t11;
c=2*d*t01*r12##2*t11+d##2*t11*r12##2*t11-(k2-1)*(t00+sig)*r12##2*t11;
t12=(-b-sqrt(b##2-4*a*c))/(2*a);

a=d##8;
b=-
(2*d##2*r2*sqrt(t00)+2*d##3*r12*sqrt(t11))##2+2*d##6*t11+4*d##5*t01-
2*d##4*(k2-1)*(t00+sig);
c=4*d##2*t01##2+d##4*t11##2+((k2-1)*(t00+sig))##2+4*d##3*t01*t11-
2*(k2-1)*(t00+sig)*(2*d*t01+d##2*t11);
t22=(-b+sqrt(b##2-4*a*c))/(2*a);

r=sig*(i(t));
tt=(t00||t01||t02)//(t01||t11||t12)//(t02||t12||t22);
z0=(j(1,t,1))`;z1=(((do(0,d,1))-
sum((do(0,d,1))/t))/f)`;z2=((z1##2)/f##2);z=(z0||z1||z2)`;
v=z`*tt*z+r;
tt_c=(t00||t02)//(t02||t22);
z_c=(z0||z2)`;
v_c=z_c`*tt_c*z_c+r;
b11=(d_l*sqrt(k1*(t00+sig)))/d;
b21=((d_q*sqrt(k2*(t00+sig)))-(d_l*sqrt(k1*(t00+sig))))/d##2;

*Matrix of missingness;
start ar1(t,v,rho);
      u = cuprod(j(1,t-1,rho));
      return(v##2#toeplitz(1||u));
finish;
ar1 = ar1(t,1,rho);
```

```
perr=ar1[1,];
matriz_p=j(t,t,1);
      do i=1 to t;
      do j=1 to i;
matriz_p[i,j]=sqrt(perr[i])*sqrt(perr[j]);
matriz_p[j,i]=matriz_p[i,j];
end; end;
m_m=j(t,t,rho);
start setdiag(m_m, v);
   diagidx = do(1,nrow(m_m)*ncol(m_m), ncol(m_m)+1);
   m_m[diagidx] = v;
finish;
run setdiag(m_m, 0);
b_d=diag(matriz_p);
m_p=m_m+b_d;

*Calculating noncentrality parameter and power values;
start ini (a,c,b,x) global(ngr,t,z2,b21,gc,ge,n,n1,n2);
      nx=(n1-0)||(n2-0);
      do i=1 to ncol(nx);
      nx1=j(nx[i],1,i);
      if i=1 then x=nx1;
            else x=x//nx1;
      end;
      x=design(x);
      a=(j(t-1,1,1)||(-i(t-1)));
      b=((b21*gc)*z2)`//((b21*ge)*z2)`;
      c = j(ngr-1,1,1)||(-i(ngr-1));
finish;
call ini (a,c,b,x);

v1=(p2*v_c+p1*v_c*heter)/m_p;
cov=(inv(z_c*inv(v1)*z_c`))[2,2];
v_l1=(sig*f##(2*l_lin)*(fact(t-l_lin-
1)))/((1/12)*(fact(t+l_lin)))+t11;v_l2=v_l1*heter;
cov_l=(p2*v_l1+p1*v_l2);
v_c1=(sig*f##(2*c_cua)*(fact(t-c_cua-
1)))/((1/180)*(fact(t+c_cua)))+t22;v_c2=v_c1*heter;
cov_c=(p2*v_c1+p1*v_c2);
lambda1=&lambda1;
power=&power;
lambda2=&lambda2;

*Total sample size calculation;
n_n=(((&za+&zb)##2)*cov)/(b21##2*p1*p2);
nt=(ceil(n_n));          n_t=ceil(nt);
nc=((n_t)*p1);           n_c=ceil(nc);
ne=(n_t-n_c);            n_e=ceil(ne);
n_a=n_c+n_e;

*Adequate power to detect statistical significance;
ddf_a=&ddf_a;
lambda_a=&lambda_a;
power_a=&power_a;

print lambda1 lambda2 cov power n_a lambda_a power_a;
end;
proc print;
```

References

Amatya, A., Bhaumik, D., & Gibbons, R. D. (2013). Sample size determination for clustered count data. *Statistics in Medicine*, *32*, 4162–4179.

Ato, M., & Vallejo, G. (2015). *Diseños de investigación en psicología* [Research designs in psychology]. Madrid, Spain: Pirámide.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. https://doi.org/10.1177/1745691612459060

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, *9*, 78–84.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*, 391–420.

Cain, M. K., Zhang, Z., & Yuan, K. H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods 49*, 1716–1735.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*, 591–601.

Fitzmaurice, G., Laird, N., & Ware, J. (2011). *Applied longitudinal analysis* (2nd ed.). Hoboken, NJ: Wiley.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research*: *Univariate and multivariate applications* (2nd ed.). New York, NY: Routledge

Hedeker, D., Gibbons, R. D., & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition. *Journal of Educational and Behavioral Statistics*, *24*, 70–93.

Heo, M. (2014). Impact of subject attrition on sample size determinations for longitudinal cluster randomized clinical trials. *Journal of Biopharmaceutical Statistics*, *24*, 507–522.

Heo, M., & Leon, A. C. (2008). Statistical power and sample size requirements for three level hierarchical cluster randomized trials. *Biometrics*, *64*, 1256–1262.

Heo, M., & Leon, A. C. (2009). Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials. *Statistics in Medicine*, *28*, 1017–1027.

Heo, M., Xue, X., & Kim, M. Y. (2013). Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials with random slopes. *Computational Statistics and Data Analysis*, *60*, 169–178.

Hertzog, C., von Oertzen, T., Ghisletta, P., & Lindenberger, U. (2008). Evaluating the power of latent growth curve models to detect individual differences in change. *Structural Equation Modeling*, *15*, 541–563. https://doi.org/10.1080/10705510802338983

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications* (2nd ed.). New York, NY: Routledge.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, *13*, 110–129.

Livacic-Rojas, P. E., Vallejo, G., Fernández, M. P., & Tuero, E. (2017). Power of modified Brown–Forsythe and mixed-model approaches in split-plot designs. *Methodology*, *13*, 9–22.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *92*, 778–785.

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. https://doi.org/10.1037/1082-989X.2.4.371

Núñez, J. C., Rosário, P., Vallejo, G., & González-Pienda, J. A. (2013). A longitudinal assessment of the effectiveness of a school-based mentoring program in middle school. *Contemporary Educational Psychology*, *38*, 11–21.

O'Kelly, M., & Ratitch, B. (2014). Analysis under missing-not-at-random assumptions. In M. O'Kelly & B. Ratitch (Eds.), *Clinical trials with missing data: A guide for practitioners* (pp. 257–368). New York, NY: Wiley.

Puttaswamy, T. K. (2012). *Mathematical achievements of pre-modern Indian mathematicians*. New York, NY: Elsevier.

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.

Rosário, P., Högemann, J., Nuñez, J. C., Vallejo, G., Cunha, J, Oliveira, V.,… Rodríguez, C. (2017). Writing week-journals to improve the writing quality of fourth-graders' compositions. *Reading and Writing*, *30*, 1001–1032.

Roy, A., Bhaumik, D. K., Aryal, S., & Gibbons, R. D. (2007). Sample size determination for hierarchical longitudinal designs with differential attrition rates. *Biometrics*, *63*, 699–707.

SAS Institute, Inc. (2016). *SAS/STAT® 14.2 user's guide*. Cary, NC: SAS Institute, Inc.

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

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

Usami, S. (2014). A convenient method and numerical tables for sample size determination in longitudinal-experimental research using multilevel models. *Behavior Research Methods*, *46*, 1207–1219. https://doi.org/10.3758/s13428-013-0432-0

Vallejo, G., Ato, M., Fernández, M. P., Livacic-Rojas, P. E., & Tuero-Herrero, E. (2016). Power analysis to detect the treatment effect in longitudinal studies with heterogeneous errors and incomplete data. *Psicothema*, *28*, 330–339.

Vallejo, G., Ato, M., & Valdés, T. (2008). Consequences of misspecifiying the error covariance structure in linear mixed models for longitudinal data, *Methodology*, *4*, 10–21.

Vallejo, G., Fernández, M. P., Cuesta, M., & Livacic-Rojas, P. E. (2015). Effects of modeling the heterogeneity on inferences drawn from multilevel designs. *Multivariate Behavioral Research*, *50*, 75–90.

Wänström, L. (2009). Sample sizes for two-group second-order latent growth curve models. *Multivariate Behavioral Research*, *44*, 588–619.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, pp. 345–442), Washington, DC: American Educational Research Association.