Confidence Interval-Based Sample Size Determination Formulas and Some Mathematical Properties for Hierarchical Data

The use of hierarchical data (also called multilevel data or clustered data) is common in behavioural and psychological research when data of lower-level units (e.g., students, clients, repeated measures) are nested within clusters or higher units (e.g., classes, hospitals, individuals). Over the past twenty-five years we have seen great advances in methods for computing the sample sizes needed to obtain the desired statistical properties for such data in experimental evaluations. The present research provides closed-form and iterative formulas for sample size determination that can be used to ensure the desired width of confidence intervals for hierarchical data. Formulas are provided for a four-level hierarchical linear model that assumes slope variances and inclusion of covariates under both balanced and unbalanced designs. In addition, we address several mathematical properties relating to sample size determination for hierarchical data via the standard errors of experimental effect estimates. These include the relative impact of several indices (e.g., random intercept or slope variance at each level) on standard errors, asymptotic standard errors, minimum required values at the highest level, and generalized expressions of standard errors for designs with any-level randomization under any number of levels. Particularly, information for the minimum required values will help researchers to minimize the risk of conducting experiments that are statistically unlikely to show the presence of an experimental effect.

Keywords: sample size, confidence interval, hierarchical data, experimental design

1 Introduction

A hierarchical linear model (HLM[*1]) is a regression model for hierarchical data (also called multilevel data, clustered data, or grouped data). In behavioural and psychological research, we often encounter hierarchical data in which data of lower-level units (e.g., students, clients, and repeated measures) are nested within higher-level clusters (e.g., classes, hospitals, individuals), causing correlations among data within same higher-level clusters. The number of levels varies: sometimes three-level designs (e.g., students are nested within classes, and classes are nested within schools) or four-level designs (students are nested within classes, classes are nested within schools, and schools are nested within districts) are used to detect an experimental effect (e.g., Heo & Leon, 2008; Konstantopoulos, 2008a, 2008b; Muthén & Muthén, 1998-2010; Schochet, 2008; Skrondal & Rabe-Hesketh, 2004; Spybrook, Hedges, & Borenstein, 2014). When randomization is performed at the highest level, such experimental designs are typically called hierarchical designs (HD) or cluster randomized trials. In contrast, when randomization is not performed at the highest level, researchers typically use terminology such as randomized blocked designs (RBDs) or multisite (cluster) randomized trials (Spybrook et al., 2014). Randomization at higher levels is sometimes preferable for various procedural reasons (e.g., convenience of gathering data or ethical considerations), so there has been an increase in the literature on statistical power calculations for such cases (Heo & Leon,

[*1] HLMs are also called multilevel models (Goldstein, 2003; Hox, 2010; Rabe-Hesketh, Skrondal & Pickles, 2004; Singer & Willett, 2003; Skrondal & Rabe-Hesketh, 2004), mixed-effects models, or random-effects models (Laird & Ware, 1982).

2008; Usami, 2014; Spybrook et al. , 2014).

Over the past twenty-five years we have seen great advances in methods for estimating in advance the sample size that should be used to ensure that an experiment achieves the desired statistical power when trying to detect an experimental effect (e.g., Bloom, 2005; Dong & Maynard, 2013; Fazzari, Kim & Heo, 2014; Hedges & Borenstein, 2014; Hedges & Rhoads, 2010; Konstantopoulos, 2013; Liang & Pulver, 1996; Moerbeek & van Breukelen, 2000; Maas & Hox, 2005; Moerbeek, 2005; Raudenbush, 1997; Raudenbush & Liu, 2000; Roy, Bhaumik, Aryal, & Gibbons, 2007; Snijders & Bosker, 1993; Usami, 2014). Various methods for estimating required sample sizes are available, including software and mathematical formulas when using hierarchical data sets. Notably, Dong and Maynard (2013) have provided a powerful tool called *PowerUp!* that can be used for various experimental and quasi-experimental designs to compute minimum detectable effect sizes for existing studies and to estimate minimum required sample sizes for studies under design. Other software, such as Optimal Design Plus (*OD Plus*; Raudenbush et al., 2011) and *CRT Power* (Borenstein, Hedges, & Rothstein, 2012), are also of great help to applied researchers for computing required sample sizes, and can be effectively used regardless of the level of randomization. Spybrook et al. (2014) provide explicit connections between the languages, notation, and design parameters of OD Plus and CRT Power.

The present study aimed to contribute to this research area mainly in two ways. First, we provide closed-form and iterative sample size determination formulas that can be used to ensure the desired width of confidence intervals for hierarchical data. Over-reliance on null hypothesis

significance testing has been criticized, in that rejection of the null hypothesis itself does not provide useful information because, strictly speaking, the null hypothesis is rarely true in reality (e.g., Balluerka, Gomez, & Hidalgo, 2005; Cohen, 1994; Wasserstein & Lazar, 2016; Sedlmeier, 2009). The American Psychological Association (APA) has therefore recommended that researchers report confidence intervals (American Psychological Association, 2010) and the American Statistical Association (ASA) has recently released a "Statement on Statistical Significance and P-Values" giving six principles underlying the proper use and interpretation of *p*-values, saying that statistical significance does not measure the size of an effect or the importance of a result (American Statistical Association, 2016). Most previous research, however, has focused on sample size determination that only accounts for statistical power and total cost, failing to consider procedures for ensuring precision of experimental effect estimates (i.e., confidence intervals). Exceptionally, Usami (2014) derived confidence interval–based sample size determination formulas, focusing on three-level experimental designs. However, the derived formulas might provide biased estimates of required sample sizes, because the formulas rely on several assumptions that might not be realistic (e.g., assumptions that all residual variances are known in a random intercept model) and do not permit various experimental designs that are used in practice (e.g., four-level hierarchical designs and experimental designs that include covariates to reduce the magnitude of residual variances). For example, in the fields of policy evaluation and educational effectiveness research, outcome magnitudes (e.g., educational achievement) typically vary among higher-level units due to multiple reasons, including organizational environment (e.g., class climate, differences in teaching style of teachers) and geographical

conditions (e.g., city, suburban, or rural area). In addition, as we will explain below, this unit difference at higher levels is closely associated with standard errors of experimental effect estimates, so including covariates that can explain the difference (e.g., teacher's beliefs as measured on a psychological scale, size and male-to-female ratio in each class or school) is useful towards achieving desired precision of estimates. For these reasons, developing a new procedure that can be applied in such realistic situations and designs is undoubtedly desirable.

Second, we address several mathematical properties relating to sample size determination for hierarchical data via standard errors for experimental effect estimates. These include the relative impact of several indices (e.g., random intercept and slope variance at each level) on standard errors, asymptotic standard errors, minimum required values at the highest level, and generalized expressions of standard errors for designs with any-level randomization under any number of levels. These results are helpful for researchers searching for better research designs, because they can promote understanding about how required sample sizes change according to the research design. In addition, because the number of indices (or parameters) that must be specified for sample size determination increases in hierarchical experimental designs, such investigation can clarify the relative importance of accurate specification of each index (or parameter), contributing to making the whole procedure of sample size determination more efficient. We also derive the minimum requirements for sample sizes at the highest (i.e., four) level. This result will be especially useful for researchers, allowing them to minimize the risk of conducting experiments that have little chance to reveal the presence of an experimental effect statistically. For example, if large variance of outcomes (e.g., educational

achievement) among highest units (e.g., district) is expected in HD, researchers can evaluate the minimum required number of districts that should be sampled to achieve desired widths for confidence intervals, and how these numbers can be effectively reduced by re-examining research designs.

This article is organized into six sections. Section 2 discusses standard errors for experimental effect estimates in a four-level hierarchical model in a comprehensive manner. In Section 3, sample size determination formulas based on the desired width of confidence interval are provided. Examples of estimating the necessary sample size using the provided formulas are addressed in Section 4. Section 5 addresses several mathematical properties relating to sample size determination for hierarchical data via standard errors of the experimental effect estimates. From the results so far, we will address that the choice of RBD or HD yields different consequences in various aspects of sample size determination, and in many cases RBD is preferable to HD in terms of required sizes (i.e., smaller standard errors). The final section discusses prospects for the proposed method and future investigations. An appendix provides formulas for when the number of levels is less than four, and provides a derivation of standard errors. R code for calculating all required sizes is provided in the Online Supporting Materials.

2 Statistical model

In this section, we introduce standard errors for experimental effect estimates in four-level hierarchical models. For brevity, we first show standard errors based on a four-level random-intercept model, and expand these results to more general hierarchical models. That is, we will consider a

four-level hierarchical model that assumes both random intercept and slope with the inclusion of covariates. In addition, we confine our discussion to the case where numbers of lower-level units are equal for all higher clusters. Regarding this point, although cluster sizes are almost always different due to missing data (e.g., nonresponses) or other procedural limitations, in this case it is permissible to substitute the harmonic mean for the number of units per cluster, because it provides a good approximation for the calculation of required sizes (e.g., Dong & Maynard, 2013; Raudenbush, 1997).

Let Y_{ijkl} be the outcome (e.g., educational achievement) for a level-1 unit (index i, with $i \in \{1, 2, \ldots, n_1\}$; e.g., a student) nested within a level-2 unit (index j, with $j \in \{1, 2, \ldots, n_2\}$; e.g., a class) nested within a level-3 unit (index k, with $k \in \{1, 2, \ldots, n_3\}$; e.g., a school) nested within a level-4 unit (index l, with $l \in \{1, 2, \ldots, n_4\}$; e.g., a district). We assume the following four-level random intercept model for expressing Y_{ijkl}:

$$Y_{ijkl} = (\beta_0 + \delta X_{ijkl}) + (e_{ijkl} + e_{jkl} + e_{kl} + e_l). \tag{1}$$

Here, X_{ijkl} is an assignment indicator variable set to 1 to indicate assignment to an experimental group (e.g., new teaching method is used) and set to 0 to indicate assignment to a control group (e.g., existing teaching method is used). Let the proportion of units in the experimental group be P ($0 < P < 1$). Balanced design is satisfied only when $P = 0.5$. In RBD with two- and level-three randomizations and HD, we can simplify by letting $X_{ijkl} = X_{jkl}$, $X_{ijkl} = X_{kl}$, and $X_{ijkl} = X_l$, because clus-

ters are randomized at each level. So, the numbers of level-1 units assigned to an experimental group per higher-level cluster differ among designs. An example of X_{ijkl} where $n_1 = n_2 = n_3 = n_4 = 2$ is provided in Table 1.

In the equation above, β_0 is an overall control group mean, and e_{ijkl}, e_{jkl}, e_{kl}, and e_l are residuals, assumed to be independent of X_{ijkl} and each other. Additionally, these residuals are assumed to be distributed according to $e_{ijkl} \sim N(0, \sigma_1^2)$, $e_{jkl} \sim N(0, \sigma_2^2)$, $e_{kl} \sim N(0, \sigma_3^2)$, and $e_l \sim N(0, \sigma_4^2)$, respectively. Here, σ_1^2, σ_2^2, σ_3^2, and σ_4^2 are the variances of outcomes for the units of each level in their respective groups.

From equation (1), it is evident that the mean of Y_{ijkl} for a given X_{ijkl} is

$$E(Y_{ijkl}|X_{ijkl}) = \beta_0 + \delta X_{ijkl}. \tag{2}$$

Additionally, the covariance of Y_{ijkl} and $Y_{i'j'k'l'}$ can be generally expressed as

$$cov(Y_{ijkl}, Y_{i'j'k'l'}|X_{ijkl}, X_{i'j'k'l'}) = 1(i = i' \& j = j' \& k = k' \& l = l')\sigma_1^2 + 1(j = j' \& k = k' \& l = l')\sigma_2^2$$
$$+1(k = k' \& l = l')\sigma_3^2 + 1(l = l')\sigma_4^2, \tag{3}$$

where $cov(\cdot)$ denotes covariance and $1(\cdot)$ is an indicator function that takes the value 1 if the conditions in the parentheses are satisfied and 0 otherwise. From equation (3), the variance of Y_{ijkl}

(namely, the covariance when $i = i'$, $j = j'$, $k = k'$, and $l = l'$) can be expressed as

$$var(Y_{ijkl}|X_{ijkl}) = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 = \sigma^2. \tag{4}$$

We define the standardized effect size Δ of an experimental effect δ following Cohen (1988), using the pooled standard deviation σ. [*2] Namely,

$$\Delta = \frac{\delta}{\sigma}. \tag{5}$$

The proportion of variance ρ_m for level m (m=1,2,3,4) can now be expressed as $\rho_m = \frac{\sigma_m^2}{\sigma_1^2+\sigma_2^2+\sigma_3^2+\sigma_4^2} = \frac{\sigma_m^2}{\sigma^2}$ [*3]. When outcome is standardized and then σ^2 is set to 1, from equation (5), $\Delta = \delta$ (i.e., a standardized effect size is equivalent to a raw experimental effect) and $\rho_m = \sigma_m^2$. The standard error of an experimental effect estimate $\hat{\delta}$ in each randomized trial can be expressed as

$$se(\hat{\delta}) = \begin{cases} se(\hat{\delta}_1) = \sigma\sqrt{\frac{\rho_1}{NP(1-P)}} & \text{(RBD with level-one randomization)} \\ se(\hat{\delta}_2) = \sigma\sqrt{\frac{n_1\rho_2+\rho_1}{NP(1-P)}} & \text{(RBD with level-two randomization)} \\ se(\hat{\delta}_3) = \sigma\sqrt{\frac{n_1n_2\rho_3+n_1\rho_2+\rho_1}{NP(1-P)}} & \text{(RBD with level-three randomization)} \\ se(\hat{\delta}_4) = \sigma\sqrt{\frac{n_1n_2n_3\rho_4+n_1n_2\rho_3+n_1\rho_2+\rho_1}{NP(1-P)}} & \text{(HD)} \end{cases} \tag{6}$$

[*2] As we will explain later, if random slopes are assumed in hierarchical model, $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2$ indicates variance of outcome in control group (i.e. $\sigma^2 = var(Y_{ijkl}|X_{ijkl} = 0)$).

[*3] Although we frame our discussion in terms of residual variances, in the hierarchical modelling literature intra-class correlations (ICCs) are popular indices that express correlations among data within the same higher-level clusters. ICCs for units in the same level-2, level-3, and level-4 clusters can be expressed using the residual variances as $ICC_2 = \sigma_2^2 + \sigma_3^2 + \sigma_4^2$, $ICC_3 = \sigma_3^2 + \sigma_4^2$, and $ICC_4 = \sigma_4^2$, respectively.

Here, $N = n_1n_2n_3n_4$, and $se(\hat{\delta}_m)$ $(m = 1, 2, 3, 4)$ is a standard error of $\hat{\delta}$ in level-m randomization. Because the derivations of these results in a four-level hierarchical model have not been comprehensively provided in the literature, we provide them in Appendix A. It is evident that $se(\hat{\delta}_3) = se(\hat{\delta}_4)$ when $\rho_4 = 0$, $se(\hat{\delta}_2) = se(\hat{\delta}_3)$ when $\rho_3 = 0$, and $se(\hat{\delta}_1) = se(\hat{\delta}_2)$ when $\rho_2 = 0$. Additionally, from equation (6) it is evident that a balanced design (where $P = 1/2$, namely, an equal number of units is allocated in each group) can achieve the smallest standard errors.

When heterogeneity of experimental effects (random slopes; e.g., a teaching effect is positive in some classes, schools, or districts, but negative in others) is assumed at each level (i.e., replacing δ with δ_{jkl}, δ_{kl}, or δ_l for level-one, level-two, level-three randomizations, respectively)[*4], the above standard errors become

$$se(\hat{\delta}) = \begin{cases} se(\hat{\delta}_1) = \sigma\sqrt{\frac{P(1-P)n_1n_2n_3\rho_4\omega_4+P(1-P)n_1n_2\rho_3\omega_3+P(1-P)n_1\rho_2\omega_2+\rho_1}{NP(1-P)}} \\ se(\hat{\delta}_2) = \sigma\sqrt{\frac{P(1-P)n_1n_2n_3\rho_4\omega_4+P(1-P)n_1n_2\rho_3\omega_3+n_1\rho_2+\rho_1}{NP(1-P)}} \\ se(\hat{\delta}_3) = \sigma\sqrt{\frac{P(1-P)n_1n_2n_3\rho_4\omega_4+n_1n_2\rho_3+n_1\rho_2+\rho_1}{NP(1-P)}} \\ se(\hat{\delta}_4) = \sigma\sqrt{\frac{n_1n_2n_3\rho_4+n_1n_2\rho_3+n_1\rho_2+\rho_1}{NP(1-P)}}. \end{cases} \tag{7}$$

[*4] If level-one RBD is used in a four-level hierarchical model that assumes random slopes, the analysis model can be expressed as

$$Y_{ijkl} = (\beta_0 + \delta X_{ijkl}) + [(r_{jkl} + r_{kl} + r_l)X_{ijkl} + (e_{ijkl} + e_{jkl} + e_{kl} + e_l)].$$

Here, r_{jkl}, r_{kl} and r_l are residuals indicating random slopes at the second, third, and fourth levels, respectively, and are assumed to be independent of X_{ijkl}. Additionally, they are assumed to be distributed according to $r_{jkl} \sim N(0, \sigma^2_{s2})$, $r_{kl} \sim N(0, \sigma^2_{s3})$, and $r_l \sim N(0, \sigma^2_{s4})$. Here, σ^2_{s2}, σ^2_{s3}, and σ^2_{s4} are the slope variances at each level. Note that in this model residual variances $\sigma^2_m (m = 1, 2, 3, 4)$ can be interpreted as random intercept variances, and $\sigma^2 = \sigma^2_1+\sigma^2_2+\sigma^2_3+\sigma^2_4$ is a variance of Y in a control group (i.e., $\sigma^2 = var(Y_{ijkl}|X_{ijkl} = 0)$). If a level-two RBD is used, this implies $r_{jkl} = 0$, indicating $\sigma^2_{s2} = 0$. Likewise, when a level-three RBD is used, it means that $r_{jkl} = r_{kl} = 0$ and $\sigma^2_{s2} = \sigma^2_{s3} = 0$. Finally, when HD is used, we have $r_{jkl} = r_{kl} = r_l = 0$ and $\sigma^2_{s2} = \sigma^2_{s3} = \sigma^2_{s4} = 0$, so this setting is essentially equivalent to the random intercept model of equation (1).

Note that the amount of $se(\hat{\delta}_4)$ (i.e., HD) is unchanged regardless of this assumption. Here, $\omega_m = \sigma^2_{sm}/\sigma^2_m$ $(m = 2, 3, 4)$ denotes the ratio of the variance of the experimental effect (i.e., random slope variances σ^2_{sm}) to the residual variance (random intercept variances σ^2_m) in level-m units. $\sigma^2 = \sigma^2_1 + \sigma^2_2 + \sigma^2_3 + \sigma^2_4$ is a variance of Y in a control group in this model (i.e., $\sigma^2 = var(Y_{ijkl}|X_{ijkl} = 0)$). See the footnote 4 on this point. Equation (7) indicates that random slope variances ($\sigma^2_{sm} = \sigma^2_m\omega_m$) at levels higher than the level of randomization (e.g., variances of the teaching effect among districts (σ^2_{s4}) and schools (σ^2_{s3}) when classes (level-two) are randomized) inflate standard errors of experimental effects estimates. In particular, in equation (7), ρ_4 (or σ^2_4, random intercept variances or variances of educational achievement among districts in a control group) and ω_4 (variances of teaching effects among districts) are associated with the sizes of each level-4 unit ($n_1n_2n_3$; e.g., the number of students in each district) and thus they can be considered as factors that are influential for standard errors in each randomization. Note that equation (7) is equivalent to (6) when $\omega_2 = \omega_3 = \omega_4 = 0$ (i.e., there are no slope variances at each level). We discuss the derivations of these results in Appendix A.

When covariates are included to reduce the magnitudes of random intercept or slope variances at each level[*5], the results of Dong and Maynard (2013) show that the standard errors of the estimate

[*5] If level-one RBD is used in a four-level hierarchical model that assumes random slopes, an analysis model that includes covariates can be expressed as

$$Y_{ijkl} = (\beta_0 + \delta X_{ijkl} + \sum_{p=1}^{g_1}\beta^{(I)}_{1p}Z^p_{ijkl} + \sum_{p=1}^{g_2}\beta^{(I)}_{2p}Z^p_{jkl} + \sum_{p=1}^{g_3}\beta^{(I)}_{3p}Z^p_{kl} + \sum_{p=1}^{g_4}\beta^{(I)}_{4p}Z^p_{l}$$
$$+ \sum_{p=1}^{g_2}\beta^{(S)}_{2p}X_{ijkl}Z^p_{jkl} + \sum_{p=1}^{g_3}\beta^{(S)}_{3p}X_{ijkl}Z^p_{kl} + \sum_{p=1}^{g_4}\beta^{(S)}_{4p}X_{ijkl}Z^p_{l}) + [(r_{jkl} + r_{kl} + r_l)X_{ijkl} + (e_{ijkl} + e_{jkl} + e_{kl} + e_l)].$$

Here, Z^p_{ijkl}, Z^p_{jkl}, Z^p_{kl} and Z^p_l are p-th covariates at the first, second, third, and fourth levels, respectively, and are assumed to be independent of residuals. $g_m(m = 1, 2, 3, 4)$ denotes the number of covariates at level-m units. $\beta^{(I)}_{mp}$ $(m = 1, 2, 3, 4)$

$\hat{\delta}$ in a four-level hierarchical model that assumes random slopes can be expressed as

$$se(\hat{\delta}) = \begin{cases} se(\hat{\delta}_1) = \sigma\sqrt{\frac{P(1-P)n_1n_2n_3\rho_4\omega_4(1-R^2_{s4})+P(1-P)n_1n_2\rho_3\omega_3(1-R^2_{s3})+P(1-P)n_1\rho_2\omega_2(1-R^2_{s2})+\rho_1(1-R^2_1)}{NP(1-P)}} \\ se(\hat{\delta}_2) = \sigma\sqrt{\frac{P(1-P)n_1n_2n_3\rho_4\omega_4(1-R^2_{s4})+P(1-P)n_1n_2\rho_3\omega_3(1-R^2_{s3})+n_1\rho_2(1-R^2_2)+\rho_1(1-R^2_1)}{NP(1-P)}} \\ se(\hat{\delta}_3) = \sigma\sqrt{\frac{P(1-P)n_1n_2n_3\rho_4\omega_4(1-R^2_{s4})+n_1n_2\rho_3(1-R^2_3)+n_1\rho_2(1-R^2_2)+\rho_1(1-R^2_1)}{NP(1-P)}} \\ se(\hat{\delta}_4) = \sigma\sqrt{\frac{n_1n_2n_3\rho_4(1-R^2_4)+n_1n_2\rho_3(1-R^2_3)+n_1\rho_2(1-R^2_2)+\rho_1(1-R^2_1)}{NP(1-P)}} \end{cases} \tag{8}$$

Here, R^2_m $(m = 1, 2, 3, 4)$ indicates the proportion of random intercept variances at level-m units explained by level-m covariates (i.e., coefficient of determination for random intercepts). In other words, $1 - R^2_m$ reflects the magnitude of conditional random intercept variance after accounting for covariates at level-m units. R^2_{sm} $(m = 2, 3, 4)$ indicates the proportion of the variance between level-m units of the experimental effect (i.e., random slope variances) explained by level-m covariates (i.e., coefficient of determination for random slopes). In other words, $1 - R^2_{sm}$ reflects the magnitude of conditional random slope variance after accounting for covariates at level-m units. Therefore, in equation (8), $\sigma^2_{m|z} = \sigma^2_m(1 - R^2_m)$ $(m = 1, 2, 3, 4)$ indicates the (conditional) random intercept variance between level-m units that cannot be explained by level-m covariates. Likewise,

and $\beta^{(S)}_{mp}(m = 2, 3, 4)$ denote p-th (fixed) regression coefficient to account for random intercept and slope variances at level-m units, respectively. r_{jkl}, r_{kl} and r_l are residuals indicating random slopes after including covariates at the second, third, and fourth levels, respectively, and are assumed to be independent of X_{ijkl} and covariates. Additionally, they are assumed to be distributed according to $r_{jkl} \sim N(0, \sigma^2_{s2|z})$, $r_{kl} \sim N(0, \sigma^2_{s3|z})$, and $r_l \sim N(0, \sigma^2_{s4|z})$. Here, $\sigma^2_{sm|z}$ $(m = 2, 3, 4)$ are the (conditional) random slope variance between level-m units that cannot be explained by level-m covariates. Likewise, e_{ijkl}, e_{jkl}, e_{kl} and e_l are residuals indicating random intercepts after including covariates at respective levels. Additionally, they are assumed to be distributed according to $e_{ijkl} \sim N(0, \sigma^2_{1|z})$, $e_{jkl} \sim N(0, \sigma^2_{2|z})$, $e_{kl} \sim N(0, \sigma^2_{3|z})$, and $e_l \sim N(0, \sigma^2_{4|z})$. Here, $\sigma^2_{m|z}$ $(m = 1, 2, 3, 4)$ are the (conditional) random intercept variance between level-m units that cannot be explained by level-m covariates. If a level-two RBD is used, this implies $r_{jkl} = 0$, indicating $\sigma^2_{s2|z} = 0$. Likewise, when a level-three RBD is used, it means that $r_{jkl} = r_{kl} = 0$ and $\sigma^2_{s2|z} = \sigma^2_{s3|z} = 0$. Finally, when HD is used, we have $r_{jkl} = r_{kl} = r_l = 0$ and $\sigma^2_{s2|z} = \sigma^2_{s3|z} = \sigma^2_{s4|z} = 0$.

$\sigma^2_{sm|z} = \sigma^2_m \omega_m (1-R^2_{sm}) = \sigma^2_{sm}(1-R^2_{sm})$ ($m = 2, 3, 4$) indicates the (conditional) random slope variance between level-m units that cannot be explained by level-m covariates. Thus, the essential difference between equations (7) and (8) is that the former uses unconditional variances (i.e., σ^2_m and σ^2_{sm}) while the latter uses conditional variances after including covariates (i.e., $\sigma^2_{m|z}$ and $\sigma^2_{sm|z}$). Note that in this model $\sigma^2 = \sigma^2_1 + \sigma^2_2 + \sigma^2_3 + \sigma^2_4$ is variance of Y in control group (i.e., $\sigma^2 = var(Y_{ijkl}|X_{ijkl} = 0)$).

Obviously, larger R^2_m and R^2_{sm} can decrease the standard error of experimental effects estimates. In particular, R^2_4 (in HD) and R^2_{s4} (in RBD) are associated with the sizes of each level-4 unit ($n_1 n_2 n_3$; e.g., the number of students in each district), so they can be considered as factors that have a large influence on standard errors. For the same reason, as in equation (7), ρ_4 (or σ^2_4: random intercept variances, such as variances of educational achievement among districts in a control group), ω_4 (variances of teaching effects among districts), and P (the proportion of units in the experimental group) are also influential for standard errors in RBD. In HD, ρ_4 and P are influential on standard errors (i.e., $se(\hat{\delta}_4)$). Although the factors that are influential on standard errors are different according to research designs, it is common that precise specifications of these indices are especially important in estimating required sizes. Readers can see Dong and Maynard (2013) for a more detailed explanation regarding a four-level hierarchical model that assumes random slopes with covariates. We also discuss the derivations of these results in Appendix A.

3 Generalized formulas for desired width of confidence interval

A $100(1-\alpha)\%$ confidence interval for δ is expressed as

$$\hat{\delta} - t_{1-\alpha/2,df}\, se(\hat{\delta}) \leqq \delta \leqq \hat{\delta} + t_{1-\alpha/2,df}\, se(\hat{\delta}). \tag{9}$$

Here, $t_{\alpha,df}$ denotes the $100\alpha\%$ point of a t-distribution with df degrees of freedom. The value for df differs according to the difference in levels of randomization: $df = n_4 - g_4 - 1$ for RBD, and $df = n_4 - g_4 - 2$ for HD, respectively, where g_4 denotes the number of covariates in the fourth level. If only the population of each district is included as a covariate for district level, $g_4 = 1$. From equation (9), the width of a confidence interval can be evaluated as

$$2t_{1-\alpha/2,df}\, se(\hat{\delta}). \tag{10}$$

Note that this is also the width of the confidence interval for $\hat{\Delta}$ when σ^2 (indicating variance of outcome if random slopes are not assumed, or variance of outcome in control group if assumed) is set to 1 (i.e., standardized). When the desired width of a confidence interval is specified as L, using equations (8) and (10), the relation $2t_{1-\alpha/2,df}\, se(\hat{\delta}) \leq L$ can be re-expressed for each level unit in

level-one RBD as

$$n_1 > \frac{4\sigma^2(1-R_1^2)\rho_1 t_{1-\alpha/2,df}^2}{P(1-P)[L^2 n_2 n_3 n_4 - 4(1-R_{s4}^2)n_2 n_3 \rho_4 \omega_4 t_{1-\alpha/2,df}^2 - 4(1-R_{s3}^2)n_2 \rho_3 \omega_3 t_{1-\alpha/2,df}^2 - 4(1-R_{s2}^2)\rho_2 \omega_2 t_{1-\alpha/2,df}^2]}, \quad (11)$$

$$n_2 > \frac{4\sigma^2[P(1-P)(1-R_{s2}^2)n_1\rho_2\omega_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{P(1-P)n_1[L^2 n_3 n_4 - 4(1-R_{s4}^2)n_3\rho_4\omega_4 t_{1-\alpha/2,df}^2 - 4(1-R_{s3}^2)\rho_3\omega_3 t_{1-\alpha/2,df}^2]}, \quad (12)$$

$$n_3 > \frac{4\sigma^2[P(1-P)(1-R_{s3}^2)n_1 n_2\rho_3\omega_3 + P(1-P)(1-R_{s2}^2)n_1\rho_2\omega_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{P(1-P)n_1 n_2[L^2 n_4 - 4(1-R_{s4}^2)\rho_4\omega_4 t_{1-\alpha/2,df}^2]}, \quad (13)$$

$$n_4 > \frac{4\sigma^2[P(1-P)(1-R_{s4}^2)n_1 n_2 n_3\rho_4\omega_4 + P(1-P)(1-R_{s3}^2)n_1 n_2\rho_3\omega_3 + P(1-P)(1-R_{s2}^2)n_1\rho_2\omega_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{L^2 P(1-P)n_1 n_2 n_3}, \quad (14)$$

The corresponding results are

$$n_1 > \frac{4\sigma^2(1-R_1^2)\rho_1 t_{1-\alpha/2,df}^2}{L^2 P(1-P)n_2 n_3 n_4 - 4P(1-P)(1-R_{s4}^2)n_2 n_3\rho_4\omega_4 t_{1-\alpha/2,df}^2 - 4P(1-P)(1-R_{s3}^2)n_2\rho_3\omega_3 t_{1-\alpha/2,df}^2 - 4(1-R_2^2)\rho_2 t_{1-\alpha/2,df}^2}, \quad (15)$$

$$n_2 > \frac{4\sigma^2[(1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{P(1-P)n_1[L^2 n_3 n_4 - 4(1-R_{s4}^2)n_3\rho_4\omega_4 t_{1-\alpha/2,df}^2 - 4(1-R_{s3}^2)\rho_3\omega_3 t_{1-\alpha/2,df}^2]}, \quad (16)$$

$$n_3 > \frac{4\sigma^2[P(1-P)(1-R_{s3}^2)n_1 n_2\rho_3\omega_3 + (1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{P(1-P)n_1 n_2[L^2 n_4 - 4(1-R_{s4}^2)\rho_4\omega_4 t_{1-\alpha/2,df}^2]}, \quad (17)$$

$$n_4 > \frac{4\sigma^2[P(1-P)(1-R_{s4}^2)n_1 n_2 n_3\rho_4\omega_4 + P(1-P)(1-R_{s3}^2)n_1 n_2\rho_3\omega_3 + (1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{L^2 P(1-P)n_1 n_2 n_3} \quad (18)$$

for level-two RBD,

$$n_1 > \frac{4\sigma^2(1-R_1^2)\rho_1 t_{1-\alpha/2,df}^2}{L^2 P(1-P)n_2 n_3 n_4 - 4P(1-P)(1-R_{s4}^2)n_2 n_3\rho_4\omega_4 t_{1-\alpha/2,df}^2 - 4(1-R_3^2)n_2\rho_3 t_{1-\alpha/2,df}^2 - 4(1-R_2^2)\rho_2 t_{1-\alpha/2,df}^2}, \quad (19)$$

$$n_2 > \frac{4\sigma^2[(1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{n_1[L^2 P(1-P)n_3 n_4 - 4P(1-P)(1-R_{s4}^2)n_3\rho_4\omega_4 t_{1-\alpha/2,df}^2 - 4(1-R_3^2)\rho_3 t_{1-\alpha/2,df}^2]}, \quad (20)$$

$$n_3 > \frac{4\sigma^2[(1-R_3^2)n_1 n_2\rho_3 + (1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{P(1-P)n_1 n_2[L^2 n_4 - 4(1-R_{s4}^2)\rho_4\omega_4 t_{1-\alpha/2,df}^2]}, \quad (21)$$

$$n_4 > \frac{4\sigma^2[P(1-P)(1-R_{s4}^2)n_1 n_2 n_3\rho_4\omega_4 + (1-R_3^2)n_1 n_2\rho_3 + (1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{L^2 P(1-P)n_1 n_2 n_3} \quad (22)$$

for level-three RBD, and

$$n_1 > \frac{4\sigma^2(1-R_1^2)\rho_1 t_{1-\alpha/2,df}^2}{L^2P(1-P)n_2n_3n_4 - 4(1-R_4^2)n_2n_3\rho_4 t_{1-\alpha/2,df}^2 - 4(1-R_3^2)n_2\rho_3 t_{1-\alpha/2,df}^2 - 4(1-R_2^2)\rho_2 t_{1-\alpha/2,df}^2}, \quad (23)$$

$$n_2 > \frac{4\sigma^2[(1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{n_1[L^2P(1-P)n_3n_4 - 4(1-R_4^2)n_3\rho_4 t_{1-\alpha/2,df}^2 - 4(1-R_3^2)\rho_3 t_{1-\alpha/2,df}^2]}, \quad (24)$$

$$n_3 > \frac{4\sigma^2[(1-R_3^2)n_1n_2\rho_3 + (1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{n_1n_2[L^2P(1-P)n_4 - 4(1-R_4^2)\rho_4 t_{1-\alpha/2,df}^2]}, \quad (25)$$

$$n_4 > \frac{4\sigma^2[(1-R_4^2)n_1n_2n_3\rho_4 + (1-R_3^2)n_1n_2\rho_3 + (1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{L^2P(1-P)n_1n_2n_3} \quad (26)$$

for HD. Because df is a function of n_4, the above formulas relating to n_4 (i.e., equations 14, 18, 22, and 26) cannot be used directly. However, as we will show in the example below, the minimum required n_4 can be iteratively calculated from these formulas using the R code provided in the Online Supporting Materials. Appendix B provides similar formulas as well as standard errors in cases where the number of levels is either two or three.

Note that researchers can evaluate the required sizes from formulas to achieve a desired width of the confidence interval for both raw experimental effects (δ) and standardized experimental effects (i.e., standardized effect size Δ), by adjusting the values of σ^2 and L. Specifically, the latter can be obtained by setting σ^2 (indicating variance of outcome if random slopes are not assumed, or variance of outcome in control group if assumed) to 1, as we will illustrate in the examples below.

4 Examples

Here, we consider a hypothetical research study that investigates whether daily life guidance from teachers discourages late bedtimes among adolescent students (level-1) in classes (level-2) in

schools (level-3) in districts (level-4). Sleep research has shown that a late bedtime may have a significant effect on the mental health of adolescents (e.g., Gangwisch, Babiss, Malaspina et al., 2010; Merikanto, Lahti, Puusniekka et al., 2013; Tochigi, Usami, Matamura et al., 2015).

We assume that the mental health of students is evaluated using the General Health Questionnaire 12 (GHQ12; Goldberg, Rickels, Downing et al., 1976), which is one of the most widely used self-reporting tools used to screen for non-psychotic psychiatric symptoms, particularly for symptoms of anxiety and depression (e.g., Tochigi et al., 2015). We also assume that classes from different schools and districts are assigned to either experimental or control groups to evaluate the effect of receiving life guidance from a teacher (i.e., we use RBD with level-two randomization). Previous studies (e.g., Tochigi et al., 2015) have found relatively small effect sizes for relations between late bedtime and mental health (from 0.06 to 0.13 for standardized cross-lagged coefficients for samples of adolescents in grades 7–12), so in this example the size of the standardized experimental effect Δ is assumed to be $\Delta = 0.20$. It is widely recognized that there are large individual differences in mental health in adolescents (e.g., Matamura et al., 2014; Tochigi et al., 2015), so the variance of means of GHQ scores (in a control group) is assumed to be large among students within classes, while variances are assumed to be relatively small among classes, schools and districts. To characterize this, residual variances (random intercept variances) are specified as $\sigma_1^2 = 4$ (i.e., student differences), $\sigma_2^2 = 0.20$ (i.e., class differences), $\sigma_3^2 = 0.05$ (i.e., school differences), and $\sigma_4^2 = 0.05$ (i.e., district differences). Consequently, $\sigma^2 = 4 + 0.20 + 0.05 + 0.05 = 4.30$ ($\sigma = 2.074$), and the proportions of these variances are $\rho_1 = 4/(4+0.20+0.05+0.05) = 0.930$, $\rho_2 = 0.20/(4+0.20+0.05+0.05) = 0.046$,

and $\rho_3 = \rho_4 = 0.05/(4 + 0.20 + 0.05 + 0.05) = 0.012$. These settings imply that the raw experimental effect is $\delta = \Delta \times \sigma = 0.20 \times 2.074 = 0.415$.

Suppose that the desired width of a confidence interval for the standardized experimental effect is set as $L = 0.20$ (i.e., aiming to achieve a confidence interval like $0.20 - 0.10 = 0.10 \leq \Delta \leq 0.30 = 0.20 + 0.10$, this setting is equivalent to the confidence interval of raw experimental effect: $0.415 - 0.10 \times 2.074 = 0.208 \leq \delta \leq 0.622 = 0.415 + 0.10 \times 2.074$, indicating $L = 2 \times 0.10 \times 2.074 = 0.415$), and that the two-sided significance level is set as $\alpha = .05$. Under this RBD with level-two randomization with $n_1 = 30$ (expected harmonic mean of the number of sampled students in each class), $n_2 = 6$ (expected harmonic mean of the number of sampled classes in each school), $n_3 = 5$ (expected harmonic mean of the number of sampled schools in each district), $P = 0.5$ (i.e., balanced design), $R_1^2 = R_2^2 = R_{s3}^2 = R_{s4}^2 = 0.25$ (i.e., the proportions of variances for random intercepts or slopes in level-*m* units explained by level-*m* covariates are relatively large), g_4=3 (i.e., the number of district level covariates is three), $\omega_3 = \omega_4 = 0.10$ (i.e., random slope variances are present but are much smaller than the random intercept variances at each level), taken as fixed, the R function *L*4*random*2*n*4 based on equation (18) provided in the Online Supporting Materials gives

```
L4random2n4(L=0.20,alpha=0.05,g4=3,n1=30,n2=6,n3=5,rho1=0.930,rho2=0.046,rho3=0.012,rho4=0.012,R1sq=0.25,R2sq=0.25,Rs3sq=0.25,Rs4sq=0.25,w3=0.10,w4=0.10,P=0.5,sigma=1)

[1]  8
```

indicating that the minimum number of required districts (i.e., n_4) is 8 for a confidence interval of Δ (i.e., $\sigma^2 = 1$).

As noted above, in RBD R^2_{s4}, ρ_4, ω_4, and P are influential on standard errors in estimating n_4, so precise specifications of these indices are especially important. If we evaluate the required size besed on the confidence interval of raw experimental effect δ, by setting $L = 0.415$ and $\sigma = 2.074$ the same results can be obtained as

```
L4random2n4(L=0.415,alpha=0.05,g4=3,n1=30,n2=6,n3=5,rho1=0.930,rho2=0.046,rho3=0.012,rho4=0.012,R1sq=0.25,R2sq=0.25,Rs3sq=0.25,Rs4sq=0.25,w3=0.10,w4=0.10,P=0.5,sigma=2.074)

[1]  8
```

Because useful information about these indices are not readily available from previous research results, in this example the minimum required n_4 is again calculated using the same function $L4random2n4$ under the various conditions of $R^2_{s4} = 0.10, 0.20, 0.30, 0.40, 0.50$ and $\omega_4 = 0.10, 0.20, 0.30, 0.40, 0.50$, with other conditions remaining unchanged. As we have observed in equation (8), these indices are associated with the sizes of each level-4 unit ($n_1 n_2 n_3$; e.g., the number of students in each district), so they can be considered as factors that have a large influence on the calculation results of required size. The results indicate that the minimum required number ranges from 7 to 9. Thus, if researchers choose to take a conservative approach and assume lower R^2_{s4} or higher ω_4, sampling $n_4 = 9$ districts might be preferable. This result also indicates that the minimum required n_4 does not change much (i.e., $n_4 = 8$ or $n_4 = 9$) even when researchers cannot specify precise values for these indices. The issue regarding precise specification of indices will be discussed in more detail later.

If the number of levels is three with level-two randomization (e.g., students are nested within classes, and classes are nested in schools), another R function *L3random2n3* provided in the Online Supporting Materials can be used to calculate required sizes in level-3 units. Under the same conditions as in the previous example (i.e., $L = 0.20$ for the standardized experimental effect, $\alpha = .05$, $g_3 = 3$, $n_1 = 30$ (for the number of students), $n_2 = 6$ (for the number of classes), $\rho_1 = 0.070/(1 - 0.012) = 0.941$, $\rho_2 = 0.046/(1 - 0.012) = 0.047$, $\rho_3 = 1 - \rho_1 - \rho_2 = 0.012/(1 - 0.012) = 0.012$, $R_1^2 = R_2^2 = R_{s3}^2 = 0.25$, $\omega_3 = 0.10$, and $P = 0.5$), then *L3random2n3* gives

```
L3random2n3(L=0.20,alpha=0.05,g3=3,n1=30,n2=6,rho1=0.941,rho2=0.047,rho3=0.012,R1sq=0.25,R2sq=0.25,Rs3sq=0.25,w3=0.10,P=0.5,sigma=1)

[1]  19
```

indicating that the minimum number of required schools (i.e., n_3) is 19. Now consider the situation where the proportion of teachers specializing in health and physical education is limited in each school by setting $P = 0.1$. In this case, the total number of required schools (n_3) is calculated to be 45, which is more than double the number calculated for balanced designs, indicating that a difference in the value of P has a large effect on the estimation results.

5 Some mathematical results regarding standard errors and required sizes

5.1 Comparing standard errors in each design

From equation (8),

$$se^2(\hat{\delta}_2) - se^2(\hat{\delta}_1) = \frac{\sigma^2\rho_2[(1-R_2^2) - P(1-P)(1-R_{s2}^2)\omega_2]}{n_2n_3n_4P(1-P)}, \tag{27}$$

$$se^2(\hat{\delta}_3) - se^2(\hat{\delta}_2) = \frac{\sigma^2\rho_3[(1-R_3^2) - P(1-P)(1-R_{s3}^2)\omega_3]}{n_3n_4P(1-P)}, \tag{28}$$

$$se^2(\hat{\delta}_4) - se^2(\hat{\delta}_3) = \frac{\sigma^2\rho_4[(1-R_4^2) - P(1-P)(1-R_{s4}^2)\omega_4]}{n_4P(1-P)}. \tag{29}$$

It is known that the relations $se(\hat{\delta}_4) \geq se(\hat{\delta}_3) \geq se(\hat{\delta}_2) \geq se(\hat{\delta}_1)$ are always fulfilled when a random intercept model is used (e.g., $\omega_2 = \omega_3 = \omega_4 = 0$; see Usami, 2014, for a three-level hierarchical model). However, when slope variances are present, such a simple relation is satisfied only when ω_m satisfies

$$\omega_m \leq \frac{1-R_m^2}{P(1-P)(1-R_{sm}^2)} \tag{30}$$

for $m = 2, 3, 4$.

5.2 Relative effects of changing indices on standard errors

From equation (8), it can be shown that

$$\frac{\partial se^2(\hat{\delta}_1)}{\partial \rho_1} = \frac{\partial se^2(\hat{\delta}_2)}{\partial \rho_1} = \frac{\partial se^2(\hat{\delta}_3)}{\partial \rho_1} = \frac{\partial se^2(\hat{\delta}_4)}{\partial \rho_1} \geq 0, \tag{31}$$

$$\frac{\partial se^2(\hat{\delta}_1)}{\partial \rho_2} \geq \frac{\partial se^2(\hat{\delta}_2)}{\partial \rho_2} = \frac{\partial se^2(\hat{\delta}_3)}{\partial \rho_2} = \frac{\partial se^2(\hat{\delta}_4)}{\partial \rho_2} \geq 0, \tag{32}$$

$$\frac{\partial se^2(\hat{\delta}_1)}{\partial \rho_3} = \frac{\partial se^2(\hat{\delta}_2)}{\partial \rho_3} \geq \frac{\partial se^2(\hat{\delta}_3)}{\partial \rho_3} = \frac{\partial se^2(\hat{\delta}_4)}{\partial \rho_3} \geq 0, \tag{33}$$

$$\frac{\partial se^2(\hat{\delta}_1)}{\partial \rho_4} = \frac{\partial se^2(\hat{\delta}_2)}{\partial \rho_4} = \frac{\partial se^2(\hat{\delta}_3)}{\partial \rho_4} \geq \frac{\partial se^2(\hat{\delta}_4)}{\partial \rho_4} \geq 0. \tag{34}$$

In a four-level hierarchical design, increasing residual variances (random intercept variances) always increases the standard errors of experimental effect estimates, and the influences are equal or larger in lower-level randomization. Note that such a magnitude relationship cannot be observed when comparing the influences of residual variances from different levels (i.e., comparing $\frac{\partial se^2(\hat{\delta}_p)}{\partial \rho_m}$ and $\frac{\partial se^2(\hat{\delta}_p)}{\partial \rho_{m+1}}$ for $p = 1, 2, 3, 4$ and $m = 1, 2, 3$), due to differences in magnitudes of determination coefficients (R^2) and slope variances (ω) between different levels.

Likewise, from equation (8) it can be shown that the effect of changes in the number of units

is given by

$$\frac{\partial se^2(\hat{\delta}_4)}{\partial n_1} = \frac{\partial se^2(\hat{\delta}_3)}{\partial n_1} = \frac{\partial se^2(\hat{\delta}_2)}{\partial n_1} = \frac{\partial se^2(\hat{\delta}_1)}{\partial n_1}, \tag{35}$$

$$\frac{\partial se^2(\hat{\delta}_4)}{\partial n_2} = \frac{\partial se^2(\hat{\delta}_3)}{\partial n_2} = \frac{\partial se^2(\hat{\delta}_2)}{\partial n_2} \neq \frac{\partial se^2(\hat{\delta}_1)}{\partial n_2}, \tag{36}$$

$$\frac{\partial se^2(\hat{\delta}_4)}{\partial n_3} = \frac{\partial se^2(\hat{\delta}_3)}{\partial n_3} \neq \frac{\partial se^2(\hat{\delta}_2)}{\partial n_3} \neq \frac{\partial se^2(\hat{\delta}_1)}{\partial n_3}, \tag{37}$$

$$\frac{\partial se^2(\hat{\delta}_4)}{\partial n_4} \neq \frac{\partial se^2(\hat{\delta}_3)}{\partial n_4} \neq \frac{\partial se^2(\hat{\delta}_2)}{\partial n_4} \neq \frac{\partial se^2(\hat{\delta}_1)}{\partial n_4}. \tag{38}$$

Obviously, increasing the sizes of units always decreases standard errors of an experimental effect estimate for any randomization. From equations (35)–(38), it can be shown that the effects of increasing level-m units are equal in m'-level randomization for every m that satisfies $m' \geq m$, but are different when $m' < m$. Note that the relative effects of the sizes of different levels in each randomization cannot be simply evaluated, because the magnitude relation depends on the unit size at each level.

For the impact of proportions of experimental group size (i.e., of setting P at different values), the following relations can be derived:

$$\frac{\partial se^2(\hat{\delta}_4)}{\partial P} \leq \frac{\partial se^2(\hat{\delta}_3)}{\partial P} \leq \frac{\partial se^2(\hat{\delta}_2)}{\partial P} \leq \frac{\partial se^2(\hat{\delta}_1)}{\partial P} \leq 0 \qquad (0 \leq P \leq 0.5) \tag{39}$$

$$\frac{\partial se^2(\hat{\delta}_4)}{\partial P} \geq \frac{\partial se^2(\hat{\delta}_3)}{\partial P} \geq \frac{\partial se^2(\hat{\delta}_2)}{\partial P} \geq \frac{\partial se^2(\hat{\delta}_1)}{\partial P} \geq 0 \qquad (0.5 \leq P \leq 1). \tag{40}$$

Thus, if $0 \leq P \leq 0.5$, then increasing P has a stronger impact in designs with higher-level randomization, and the standard error decreases with each randomization. When $0.5 \leq P \leq 1$, increasing

P also has a stronger impact in designs with higher-level randomization, but under this condition the standard error increases with each randomization. This result indicates that the impact of P on standard errors is most dominant in HD.

Next, regarding the impact of changes of coefficients of determination for random intercept R^2, the following relations can be derived:

$$\frac{\partial se^2(\hat{\delta}_4)}{\partial R_1^2} = \frac{\partial se^2(\hat{\delta}_3)}{\partial R_1^2} = \frac{\partial se^2(\hat{\delta}_2)}{\partial R_1^2} = \frac{\partial se^2(\hat{\delta}_1)}{\partial R_1^2} \leq 0 \tag{41}$$

$$\frac{\partial se^2(\hat{\delta}_4)}{\partial R_2^2} = \frac{\partial se^2(\hat{\delta}_3)}{\partial R_2^2} = \frac{\partial se^2(\hat{\delta}_2)}{\partial R_2^2} \leq \frac{\partial se^2(\hat{\delta}_1)}{\partial R_2^2} = 0 \tag{42}$$

$$\frac{\partial se^2(\hat{\delta}_4)}{\partial R_3^2} = \frac{\partial se^2(\hat{\delta}_3)}{\partial R_3^2} \leq \frac{\partial se^2(\hat{\delta}_2)}{\partial R_3^2} = \frac{\partial se^2(\hat{\delta}_1)}{\partial R_3^2} = 0 \tag{43}$$

$$\frac{\partial se^2(\hat{\delta}_4)}{\partial R_4^2} \leq \frac{\partial se^2(\hat{\delta}_3)}{\partial R_4^2} = \frac{\partial se^2(\hat{\delta}_2)}{\partial R_4^2} = \frac{\partial se^2(\hat{\delta}_1)}{\partial R_4^2} = 0 \tag{44}$$

From equations (41)–(44), the effects of increasing level-m coefficients of determination are nonzero and equal between m'-level and $m' + 1$ randomizations when $m' \geq m$. However, the effect is 0 when $m' < m$, because standard errors are unrelated to this coefficient of determination under such conditions. Note that the influences of increasing level-m and level-$(m + 1)$ coefficients of determination for random intercepts are different in m'-level randomization when $m' > m$, because the magnitude relation depends on the unit size (n) and the residual variance (ρ) at each level.

Similarly, the following relations can be derived for the impact of the coefficient of determina-

tion for random slope (R_s^2):

$$\frac{\partial se^2(\hat{\delta}_4)}{\partial R_{s2}^2} = \frac{\partial se^2(\hat{\delta}_3)}{\partial R_{s2}^2} = \frac{\partial se^2(\hat{\delta}_2)}{\partial R_{s2}^2} = 0 \geq \frac{\partial se^2(\hat{\delta}_1)}{\partial R_{s2}^2} \tag{45}$$

$$\frac{\partial se^2(\hat{\delta}_4)}{\partial R_{s3}^2} = \frac{\partial se^2(\hat{\delta}_3)}{\partial R_{s3}^2} = 0 \geq \frac{\partial se^2(\hat{\delta}_2)}{\partial R_{s3}^2} = \frac{\partial se^2(\hat{\delta}_1)}{\partial R_{s3}^2} \tag{46}$$

$$\frac{\partial se^2(\hat{\delta}_4)}{\partial R_{s4}^2} = 0 \geq \frac{\partial se^2(\hat{\delta}_3)}{\partial R_{s4}^2} = \frac{\partial se^2(\hat{\delta}_2)}{\partial R_{s4}^2} = \frac{\partial se^2(\hat{\delta}_1)}{\partial R_{s4}^2} \tag{47}$$

From equations (45)–(47), the effects of increasing level-m coefficients of determination for slopes are nonzero but equal between m'-level and $m'+1$ randomizations when $m' < m-1$. However, the effect is 0 when $m' \geq m$, because standard errors are unrelated to this coefficient of determination under such conditions. Note that the influences of increasing level-m and level-$(m+1)$ coefficients of determination in slopes are different in m'-level randomization when $m' < m$, because the magnitude relation depends on the unit size (n), the slope variance (ω), and the residual variance (ρ) at each level.

Finally, regarding the impact of slope variances (ω), the following relations can be derived:

$$\frac{\partial se^2(\hat{\delta}_1)}{\partial \omega_2} \geq \frac{\partial se^2(\hat{\delta}_2)}{\partial \omega_2} = \frac{\partial se^2(\hat{\delta}_3)}{\partial \omega_2} = \frac{\partial se^2(\hat{\delta}_4)}{\partial \omega_2} = 0 \tag{48}$$

$$\frac{\partial se^2(\hat{\delta}_1)}{\partial \omega_3} = \frac{\partial se^2(\hat{\delta}_2)}{\partial \omega_3} \geq \frac{\partial se^2(\hat{\delta}_3)}{\partial \omega_3} = \frac{\partial se^2(\hat{\delta}_4)}{\partial \omega_3} = 0 \tag{49}$$

$$\frac{\partial se^2(\hat{\delta}_1)}{\partial \omega_4} = \frac{\partial se^2(\hat{\delta}_2)}{\partial \omega_4} = \frac{\partial se^2(\hat{\delta}_3)}{\partial \omega_4} \geq \frac{\partial se^2(\hat{\delta}_4)}{\partial \omega_4} = 0 \tag{50}$$

From equations (48)–(50), the effects of increasing level-m slope variances are nonzero but equal in m'-level and $m'+1$ randomizations when $m' < m-1$. However, the effect is 0 when $m' \geq m$, because standard errors are unrelated to this slope variance under such conditions. Note that the effects of increasing level-m and level-$m+1$ slope variances are different in m'-level randomization when $m' < m$, because the magnitude relation depends on the unit size (n), the coefficients of determination for slopes (R_s^2), and the residual variance (ρ) at each level.

5.3 Asymptotic standard errors

From equation (8), the following lower limits for standard errors can be obtained:

$$\lim_{n_3\to\infty}\lim_{n_2\to\infty}\lim_{n_1\to\infty} se^2(\hat{\delta}_1) = \frac{\sigma^2(1-R_{s4}^2)\rho_4\omega_4}{n_4}, \tag{51}$$

$$\lim_{n_3\to\infty}\lim_{n_2\to\infty}\lim_{n_1\to\infty} se^2(\hat{\delta}_2) = \frac{\sigma^2(1-R_{s4}^2)\rho_4\omega_4}{n_4}, \tag{52}$$

$$\lim_{n_3\to\infty}\lim_{n_2\to\infty}\lim_{n_1\to\infty} se^2(\hat{\delta}_3) = \frac{\sigma^2(1-R_{s4}^2)\rho_4\omega_4}{n_4}, \tag{53}$$

$$\lim_{n_3\to\infty}\lim_{n_2\to\infty}\lim_{n_1\to\infty} se^2(\hat{\delta}_4) = \frac{\sigma^2(1-R_4^2)\rho_4}{P(1-P)n_4}. \tag{54}$$

From equations (51)–(54), the minimum required numbers in the highest units (n_4) become

$$n_4 \geq \frac{4\sigma^2(1-R_{s4}^2)\rho_4\omega_4}{L^2} \tag{55}$$

in one-, two-, and level-three RBD, and

$$n_4 \geq \frac{4\sigma^2(1-R_4^2)\rho_4}{L^2P(1-P)} \tag{56}$$

in HD. Specifically, if the numbers of level-4 units (e.g., districts, as in the example of the previous section) do not satisfy the above relations, then the confidence interval will not have the desired width L on average, even when the total amount of data from lower units (i.e., $n_1n_2n_3$; total number of students in each district) becomes infinite. From the right-hand sides of equations (55)–(56), these minimum required sizes become easier to achieve when L and the coefficients of determination are larger (or, residual variances and slope variances are sufficiently small). In HD, although minimum required numbers are unrelated to ω_4, the proportion of the experimental group size P does have an influence and the minimum number becomes smallest when $P = 0.5$ (i.e., balanced design). When the number of levels is 3, similar results can be obtained. Namely, the minimum required numbers in the highest units (n_3) are

$$n_3 \geq \frac{4\sigma^2(1-R_{s3}^2)\rho_3\omega_3}{L^2} \tag{57}$$

in one- and level-two RBD, respectively, and

$$n_3 \geq \frac{4\sigma^2(1-R_3^2)\rho_3}{L^2P(1-P)} \tag{58}$$

in HD. Similarly, when the number of levels is 2, the minimum required numbers in the highest units (n_2) are

$$n_2 \geq \frac{4\sigma^2(1-R^2_{s2})\rho_2\omega_2}{L^2} \tag{59}$$

in level-one RBD, and

$$n_2 \geq \frac{4\sigma^2(1-R^2_2)\rho_2}{L^2P(1-P)} \tag{60}$$

in HD. These relations will help researchers to minimize the risk of conducting experiments that are statistically unlikely to show the presence of an experimental effect. Table 2 provides minimum required values of level-M units (n_M, $M = 2, 3, 4$) in RBD when the number of levels is M under various specifications of ρ_M, R^2_{sM}, ω_M, and L (the two-sided significance level is $\alpha = .05$). Table 3 provides similar minimum required values for level-M units in HD when the number of levels is M under various specifications of ρ_M, R^2_M, P and L. Note that the same tables can be used whether the number of levels is three or two because equations (55), (57), and (59), and equations (56), (58), and (60), are equivalent except for the indices of levels. Thus, if researchers conduct an RBD study when the number of levels is $M = 4$, to calculate minimum required values from Table 2, specifying values of ρ_4, R^2_{s4}, ω_4, and L is sufficient, while setting ρ_3, R^2_{s3}, ω_3, and L is required if researchers conduct an RBD study when the number of levels is $M = 3$.

From Table 2, it can be observed that a smaller desired width of confidence interval L, such

as 0.1 and 0.2, might require unrealistically large minimum required values, so including covariates that can reduce the magnitude of random intercept or slope variances is especially important. Note that in RBD minimum required values are always 1 (i.e., there are essentially no lower limits) when a random intercept model can be assumed, because $\omega_M = 0$ in such cases (e.g., Usami, 2014). The magnitude of the heterogeneity of the experimental effect (random slopes) is thus related to minimum required values, indicating that researchers should correctly specify the analysis model (e.g., whether random slopes should be assumed or not) to obtain precise estimates of required sizes in using formulas.

One notable difference between Table 2 and Table 3 (i.e., RBD or HD) is that HD demands much larger required values on average. In addition, in HD, minimum required values are a function of R^2_m (rather than R^2_{sm}) and P, and for extreme values of P ($P = 0$ or $P = 1$) the minimum required values become larger. Thus, including useful covariates and choosing a balanced design ($P = 0.5$) are especially important in HD to achieve minimum required values that are realistic if researchers demand high accuracy in estimating experimental effects (i.e., a narrower width of confidence interval L). Note that unlike RBD minimum required values are related to random intercept variance (i.e., $\sigma^2(1 - R^2_M)\rho_M$) rather than random slope variance (i.e., $\sigma^2(1 - R^2_{sM})\rho_M\omega_M$) in HD, and the right sides of equations (56), (58) and (60) do not become 0 when random intercept model is data generation model (i.e., $\omega_M = 0$). Therefore, when HD is chosen for experimental design, regardless of the choice of analysis model (i.e., random intercept model or random intercept and slope model), researchers should always take care to satisfy this minimum required value before starting

experiments. The choice of RBD or HD thus yields different consequences in various aspects of sample size determination, and in many cases RBD is preferable to HD in consideration of required sizes (i.e., smaller standard errors) unless slope variances at the highest level (e.g., ω_4 in a four-level design) are very large (see equation (8) regarding this point).

5.4 Cases with more than four levels

As indicated by the results provided in Appendix A, the standard error for an experimental effect estimate $se(\delta_{m|M})$ in a design with level-m ($m = 1, 2, \ldots, M$; M is the number of levels) randomization can be derived as

$$se(\delta_{m|M}) = \sigma \sqrt{\frac{f_m}{N^* P(1-P)}}, \tag{61}$$

where $N^* = \Pi_{m^*=1}^{M} n_m^*$, and

$$f_m = \sum_{m^*=m+1}^{M} \left[P(1-P)(\Pi_{m^{**}=1}^{m^*-1} n_{m^{**}}) \rho_{m^*} \omega_{m^*} (1 - R^2_{sm^*}) \right] + \sum_{m^*=2}^{m} \left[(\Pi_{m^{**}=1}^{m^*-1} n_{m^{**}}) \rho_{m^*} (1 - R^2_{m^*}) \right] + \rho_1 (1 - R_1^2). \tag{62}$$

Naturally, when $M = 4$ this corresponds to equation (8). By applying the same procedure discussed in the previous section, some generalized results can be derived for designs when the number of levels is more than four. For example, for the influence of residual variances (random intercept

variances) on standard errors, the relations

$$\frac{\partial se(\delta_{1|M})}{\partial \rho_{m^*}} = \cdots = \frac{\partial se(\delta_{m^*-2|M})}{\partial \rho_{m^*}} = \frac{\partial se(\delta_{m^*-1|M})}{\partial \rho_{m^*}} \geq \frac{\partial se(\delta_{m^*|M})}{\partial \rho_{m^*}} = \cdots = \frac{\partial se(\delta_{M-1|M})}{\partial \rho_{m^*}} = \frac{\partial se(\delta_{M|M})}{\partial \rho_{m^*}} \geq 0 \tag{63}$$

can be obtained. Additionally, the minimum required values n_M given a number of levels M can be expressed through equations similar to (55)–(60), as

$$n_M \geq \frac{4\sigma^2(1 - R^2_{sM})\rho_M\omega_M}{L^2} \tag{64}$$

in level-m RBD ($m = 1, 2, \ldots, M-1$), and

$$n_M \geq \frac{4\sigma^2(1 - R^2_{M})\rho_M}{L^2P(1-P)} \tag{65}$$

in HD.

6 Discussion

The present research has provided closed-form and iterative sample size determination formulas that can be used to ensure the desired width of a confidence interval for hierarchical data. These formulas have been derived for a four-level hierarchical model that assumes random slopes and covariates, considering both balanced and unbalanced designs. Examples of estimating the required sample size are also shown using R functions provided in the Online Supporting Materials. We have

also addressed several mathematical properties of required sample sizes via standard errors of experimental effect estimates: the relative impact of several indices (e.g., random intercept/slope variance at each level) on standard errors, asymptotic standard errors, and generalized expressions of standard errors for designs with any-level randomization under any number of levels.

We have seen many differences in RBD and HD. For example, the factors that are influential on standard errors are different according to research designs: In RBD, coefficients of determination for random slope at the highest level (R^2_{s4}), proportion of residual variances (random intercept variances) at the highest level (ρ_4), the ratio of the variance of the experimental effect (i.e., random slope variances) at the highest level (ω_4), and P are more influential on standard errors, while coefficients of determination for random intercept at the highest level (R^2_4), ρ_4, and P are more influential in HD. Because standard errors of experimental effect estimates in HD are unrelated to slope variances, the impact of ω can be ignored in this design. It was also addressed in Section 5 that the influences of P on standard errors are largest in HD, and that asymptotic standard errors and minimum required values at the highest level when the total number of units at lower levels are infinite are different between RBD and HD. On this point, Tables 2 and 3 will help researchers to minimize the risk of conducting experiments that are statistically unlikely to show the presence of an experimental effect. Table 4 summarizes the differences between RBD and HD. In addition to finding better research designs, the present investigation also helps to clarify the relative importance of accurate specification of each index (or parameter), contributing to making the whole procedure of sample size determination more efficient.

Although the factors that are influential on standard errors are different according to research designs, it is common that precise specifications of indices are important in estimating required sizes. As illustrated in the example, we believe that one of the most useful and convenient approaches is to calculate the required sizes under the various conditions of these indices, because this can reduce the risk of obtaining standard errors that are unexpectedly large. In addition, including covariates that can explain the variance of outcomes (i.e. reducing the magnitude of random intercept or slope variances) is useful. We also believe that reporting the estimates of parameters and indices is highly desirable, and the accumulation of such information in each research area (see, e.g., Spybrook, 2013, Spybrook & Kelcey, 2016 and Westine, Spybrook, and Taylor, 2013 for examples and discussion) would aid researchers to better evaluate required sample sizes for future research.

There remain important areas for future development. First, although we assume continuous outcomes in a two-group comparison, the extension to formulas for an arbitrary number of groups, and to cases in which non-continuous outcomes (e.g., binary, ordered, count, rate, time-to-event data) can also be straightforwardly derived (e.g., Usami, 2011a; Ahn, Heo & Zhang, 2014; Rutterford, Copas & Eldridge, 2015 and references therein), and such extensions must be an intriguing topic. Another important research topic is the extension of this work to other multilevel modelling approaches, such as cross-classified models (e.g., Rasbash & Browne, 2001; Raudenbush & Bryk, 2002) and contextual models (e.g., Lüdtke et al., 2008), because the issues of sample size determination, bias, and Type-1 error rates of experimental effects or contextual effects have recently attracted attention in multiple fields of behavioural and psychological science (e.g., Baayen, Davidson, &

Bates, 2008; Judd, Westfall, & Kenny, 2012, 2017; Lüdtke et al., 2008; Murayama, Sakaki, Yan, & Smith, 2014, Usami, 2017).

7 References

Ahn, C., Heo, M., & Zhang S. (2014). *Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research*. CRC Press: Boca Raton, FL.

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6^{th} ed.). Washington, D.C.: Author.

American Statistical Association (2016). http://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf.

Baayen, R.H., Davidson, D.J., & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412. doi:10.1016/j.jml.2007.12.005

Balluerka, N., Gomez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology, 1*, 55-70.

Bloom, H.S. (2005). Randomizing groups to evaluate place-based programs. In H.S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches (pp.115-172).* New York: Russell Sage Foundation.

Borenstein, M., Hedges, L.V., & Rothstein, H. (2012). CRT Power. Teaneck, NJ: Biostat, Inc.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Mahwah, New Jersey: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Dong, N., & Maynard, R.A. (2013). PowerUp!: A tool for calculating minimum detectable effect

sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*, 24-67.

Fazzari, M.J., Kim, M.Y., & Heo, M. (2014). Sample size determination for three-level randomized clinical trials with randomization at the first or second level. *Journal of Biopharmaceutical Statistics, 24,* 579-599.

Gangwisch, J.E., Babiss, L.A., Malaspina, D., et al. (2010). Earlier parental set bedtimes as a protective factor against depression and suicidal ideation. *Sleep, 33,* 97-106.

Goldberg, D.P., Rickels, K., Downing, R., et al. (1976). A comparison of two psychiatric screening tests. *British Journal of Psychiatry, 129,* 61-67.

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

Hedges, L.V. & Rhoads, C. (2010). Statistical power analyses in education research (NCSER 2010-2006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, US Department of Education.

Hedges, L.V. & Borenstein, M. (2014). Conditional optimal design in three- and four-level experiments. *Journal of Educational and Behavioral Statistics, 39*, 257-281.

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

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Mahwah, New Jersey: Erlbaum.

Judd, C.M., Westfall, J., & Kenny, D.A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54-69. doi: 10.1037/a0028347

Judd, C.M., Westfall, J., & Kenny, D.A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology, 68*, 601-625.

Konstantopoulos, S. (2008a). The Power of the test for treatment effects in three-Level cluster randomized designs. *Journal of Research on Educational Effectiveness, 1*, 66-88.

Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness, 1*, 265-288.

Konstantopoulos, S. (2013). Optimal design in three-level block randomized designs with two levels of nesting: An ANOVA framework with random effects. *Educational and Psychological Measurement, 73*, 784-802.

Laird, N.M., & Ware, H. (1982). Random-effects model for longitudinal data. *Biometrics, 38*, 963-974.

Liang, K.Y., & Pulver, A.E. (1996). Analysis of case-control/family sampling design. *Genetic Epidemiology, 13,* 253-270.

Lüdtke, O., Marsh, H., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods, 13*, 203-229.

Maas, C.J.M., & Hox, J.J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1,* 85-91.

Matamura, M., Tochigi, M, Usami, S., Yonehara, H., Fukushima, M., Nishida, A., Togo, F., & Sasaki, T. (2014). Associations between sleep habits and mental health status and suicidality in the longitudinal survey of monozygotic-twin adolescents. *Journal of Sleep Research, 23*, 290-294.

Merikanto, I., Lahti, T., Puusniekka, R., et al. (2013). Late bedtimes weaken school performance and predispose adolescents to health hazards. *Sleep Medicine, 14,* 1105-1111.

Moerbeek, M., van Breukelen, G.J.P., & Berger, M.P.F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics, 25*, 271-284.

Moerbeek, M. (2005). Randomization of clusters versus randomization of persons within clusters: Which is preferable? *American Statistician, 59,* 173-179.

Murayama, K., Sakaki, M., Yan, V.X., & Smith, G.M. (2014). Type 1 error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 40,* 1287-1306. http://dx.doi.org/10.1037/a0036914

Muthén, L.K., & Muthén, B.O. (1998-2010). *Mplus user's guide (6th ed.)*. Los Angeles: Muthén & Muthén.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika, 69*, 167-190.

Rasbash, J., & Browne, W. J. (2001). Modeling non-hierarchical structures. In A. H. Leyland & H. Goldstein (Eds.), *Multilevel modeling of health statistics* (pp. 93-105). Chichester, England: John Wiley and Sons.

Raudenbush, S.W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173-185.

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

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods*. (2nd ed.). London: Sage.

Raudenbush, S.W., Spybrook, J., Congdon, R., Liu, X., Martinez, A., Bloom, H., & Hill, C. (2011). Optimal Design Plus empirical evidence (Version 3.0).

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.

Rutterford, C., Copas, A., & Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology, 44*, 1051-1067.

Sedlmeier, P. (2009). Beyond the significance test ritual: What is there? *Journal of Psychology, 217*, 1-5.

Singer, J.D., & Willett, J.B. (2003). *Applied longitudinal data analysis*. New York: Oxford.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling. Multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.

Snijders, T.A.B., & Bosker, R.J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*, 237-260.

Spybrook, J. (2013). Introduction to a special issue on design parameters for cluster randomized trials in education. *Evaluation Review, 37*, 435-444.

Spybrook, J., Hedges, L., & Borenstein, M. (2014). Understanding statistical power in cluster randomized trials: Challenges posed by differences in notation and terminology. *Journal of Research on Educational Effectiveness, 7*, 384-406.

Spybrook, J., & Kelcey, B. (2016). Introduction to three special issues on design parameter values for planning cluster randomized trials in the social sciences. *Evaluation Review, 40*, 491-499.

Schochet, P. (2008). Statistical power for random assignment evaluations of educational programs. *Journal of Educational and Behavioral Statistics, 33*, 62-87.

Tochigi, M., Usami, S., Matamura, M., Kitagawa, Y., Fukushima, M., Yonehara, H., Togo, F., Nishida, A., & Sasaki, T. (2015). Annual longitudinal survey at up to five time points reveals reciprocal effects of bedtime delay and depression/anxiety in adolescents. *Sleep Medicine, 17*, 81-86.

Usami, S. (2011). Statistical power of experimental research with hierarchical data. *Behaviormetrika, 38*, 63-84.

Usami, S. (2014). Generalized sample size determination formulas for experimental research with hierarchical data. *Behavior Research Methods, 46*, 346-356.

Usami, S. (2017). Generalized sample size determination formulas for investigating contextual effects by a three-level random intercept model. *Psychometrika, 82*, 133-157.

Wasserstein, R.L. & Lazar, N.A. (2016). The ASA's statement on p-values: Context, process, and purpose, *The American Statistician, 70*, 129-133.

Westine, C.D., Spybrook, J., & Taylor. J.A. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review, 37,* 490-519.

8 Appendix A: Standard errors of experimental effect $\hat{\delta}$

To derive the standard errors of experimental effect $\hat{\delta}$ in a four-level random intercept model, we consider the matrix form of equation (1):

$$\boldsymbol{Y} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}. \tag{66}$$

Here, $\boldsymbol{\beta} = (\beta_0, \delta)'$ and $\boldsymbol{Y}$ is an $(n_1 \times n_2 \times n_3 \times n_4) \times 1$ vector with its elements arranged as $\boldsymbol{Y} = (\boldsymbol{Y'_1}, \ldots, \boldsymbol{Y'_l}, \ldots, \boldsymbol{Y'_{n_4}})'$, where $\boldsymbol{Y_l} = (\boldsymbol{Y'_{1l}}, \ldots, \boldsymbol{Y'_{kl}}, \ldots, \boldsymbol{Y'_{n_3 l}})'$ and $\boldsymbol{Y_{kl}} = (\boldsymbol{Y'_{1kl}}, \ldots, \boldsymbol{Y'_{jkl}}, \ldots, \boldsymbol{Y'_{n_2 kl}})'$, giving $\boldsymbol{Y_{jkl}} = (Y_{1jkl}, \ldots, Y_{ijkl}, \ldots, Y_{n_1 jkl})'$. $\tilde{\boldsymbol{X}} = (\boldsymbol{1_N}, \boldsymbol{X})$ is an $(n_1 \times n_2 \times n_3 \times n_4) \times 2$ matrix and $\boldsymbol{X}$ is an $(n_1 \times n_2 \times n_3 \times n_4) \times 1$ vector that includes the information of X_{ijkl}. The error term $\tilde{\boldsymbol{\epsilon}}$ is also an $(n_1 \times n_2 \times n_3 \times n_4) \times 1$ vector that includes information of $\tilde{e}_{ijkl} = e_l + e_{kl} + e_{jkl} + e_{ijkl}$.

From equation (3) and the relations $\sigma^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2$ and $\rho_m = \sigma_m^2/\sigma^2$ $(m = 1, 2, 3, 4)$, it can be shown that $\tilde{e}_{ijkl}$ is distributed as $\tilde{e}_{ijkl} \sim N(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}})$, where

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{I_{n_4}} \otimes \boldsymbol{\Sigma}, \tag{67}$$

$$\begin{aligned}\boldsymbol{\Sigma} &= \sigma_4^2 \boldsymbol{1_{n_1 n_2 n_3}} \boldsymbol{1'_{n_1 n_2 n_3}} + \boldsymbol{I_{n_3}} \otimes (\sigma_3^2 \boldsymbol{1_{n_1 n_2}} \boldsymbol{1'_{n_1 n_2}}) + \boldsymbol{I_{n_2 n_3}} \otimes (\sigma_2^2 \boldsymbol{1_{n_1}} \boldsymbol{1'_{n_1}}) + \sigma_1^2 \boldsymbol{I_{n_1 n_2 n_3}} \\ &= \sigma^2 [\rho_4 \boldsymbol{1_{n_1 n_2 n_3}} \boldsymbol{1'_{n_1 n_2 n_3}} + \boldsymbol{I_{n_3}} \otimes (\rho_3 \boldsymbol{1_{n_1 n_2}} \boldsymbol{1'_{n_1 n_2}}) + \boldsymbol{I_{n_2 n_3}} \otimes (\rho_2 \boldsymbol{1_{n_1}} \boldsymbol{1'_{n_1}}) + \rho_1 \boldsymbol{I_{n_1 n_2 n_3}})].\end{aligned} \tag{68}$$

Here, we assume that $\sigma_1^2 \geq 0$, $\sigma_2^2 \geq 0$, $\sigma_3^2 \geq 0$, and $\sigma_4^2 \geq 0$, and that the inverse matrix of $\boldsymbol{\Sigma}$ (denoted by $\boldsymbol{\Sigma^{-1}}$) exists. Let the diagonal elements of $\boldsymbol{\Sigma^{-1}}$ be $\sigma^{(1)}$, the off-diagonal elements denoting the

same level-2 unit in $\mathbf{\Sigma^{-1}}$ be $\sigma^{(2)}$, and the off-block diagonal elements denoting the same level-3 and level-4 units in $\mathbf{\Sigma^{-1}}$ be $\sigma^{(3)}$ and $\sigma^{(4)}$, respectively. Comparing the left- and right-hand sides of the identity $\mathbf{\Sigma\Sigma^{-1}} = \boldsymbol{I}$, the following equations are obtained:

$$\begin{aligned}
&\sigma^2[\sigma^{(1)} + (n_1 - 1)(\rho_2 + \rho_3 + \rho_4)\sigma^{(2)} + [n_1(n_2 - 1)](\rho_3 + \rho_4)\sigma^{(3)} + [n_1 n_2(n_3 - 1)]\rho_4\sigma^{(4)}] = 1,\\
&\sigma^2[\sigma^{(2)} + (\rho_2 + \rho_3 + \rho_4)\sigma^{(1)} + (n_1 - 2)(\rho_2 + \rho_3 + \rho_4)\sigma^{(2)} + [n_1(n_2 - 1)](\rho_3 + \rho_4)\sigma^{(3)} + [n_1 n_2(n_3 - 1)]\rho_4\sigma^{(4)}] = 0,\\
&\sigma^2[\sigma^{(3)} + (n_1 - 1)(\rho_2 + \rho_3 + \rho_4)\sigma^{(3)} + n_1(n_2 - 2)(\rho_3 + \rho_4)\sigma^{(3)} + (\rho_3 + \rho_4)[\sigma^{(1)} + (n_1 - 1)\sigma^{(2)}] + [n_1 n_2(n_3 - 1)]\rho_4\sigma^{(4)}] = 0,\\
&\sigma^2[\sigma^{(4)} + (n_1 - 1)(\rho_2 + \rho_3 + \rho_4)\sigma^{(4)} + n_1(n_2 - 1)(\rho_3 + \rho_4)\sigma^{(4)} + n_1 n_2(n_3 - 2)\rho_4\sigma^{(4)} + \rho_4[\sigma^{(1)} + (n_1 - 1)\sigma^{(2)} + n_1(n_2 - 1)\sigma^{(3)}]] = 0.
\end{aligned} \tag{69}$$

These equations can be rewritten as

$$\sigma^{(1)} = \sigma^{(2)} + \frac{1}{f_1}, \qquad \sigma^{(2)} = \sigma^{(3)} - \frac{\rho_2}{f_1 f_2}, \qquad \sigma^{(3)} = \sigma^{(4)} - \frac{\rho_3}{f_2 f_3}, \qquad \sigma^{(4)} = -\frac{\rho_4}{f_3 f_4}, \tag{70}$$

where

$$\begin{aligned}
f_1 &= \sigma^2\rho_1\\
f_2 &= \sigma^2[n_1\rho_2 + \rho_1]\\
f_3 &= \sigma^2[n_1 n_2\rho_3 + n_1\rho_2 + \rho_1]\\
f_4 &= \sigma^2[n_1 n_2 n_3\rho_4 + n_1 n_2\rho_3 + n_1\rho_2 + \rho_1].
\end{aligned} \tag{71}$$

Simple calculation shows that f_1, f_2, f_3, and f_4 can also be expressed as functions of $\sigma^{(1)}$, $\sigma^{(2)}$, $\sigma^{(3)}$, and $\sigma^{(4)}$, as follows.

$$
\begin{aligned}
f_1 &= \frac{1}{\sigma^{(1)} - \sigma^{(2)}} \\
f_2 &= \frac{1}{\sigma^{(1)} + (n_1 - 1)\sigma^{(2)} - n_1\sigma^{(3)}} \\
f_3 &= \frac{1}{\sigma^{(1)} + (n_1 - 1)\sigma^{(2)} + n_1(n_2 - 1)\sigma^{(3)} - n_1 n_2\sigma^{(4)}} \\
f_4 &= \frac{1}{\sigma^{(1)} + (n_1 - 1)\sigma^{(2)} + n_1(n_2 - 1)\sigma^{(3)} + n_1 n_2(n_3 - 1)\sigma^{(4)}}
\end{aligned} \tag{72}
$$

In this, f_1, f_2, f_3, and f_4 can be regarded as variance inflation factors or design effects, as will be shown soon.

Using the generalized least squares estimators, sample distributions of $\hat{\boldsymbol{\beta}}$ can be characterized as $\hat{\boldsymbol{\beta}} \sim N((\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y}, (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{X}})^{-1})$. Then, $se(\hat{\delta})$ is given by the square root of the $(2,2)$ element of

$$
(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{X}})^{-1} = (\tilde{\boldsymbol{X}}'[\boldsymbol{I}_K \otimes \boldsymbol{\Sigma}^{-1}]\tilde{\boldsymbol{X}})^{-1}. \tag{73}
$$

Let $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, $\boldsymbol{x_3}$, and $\boldsymbol{x_4}$ be

$$
\boldsymbol{x_1} = (\boldsymbol{1}'_{Pn_1}, \boldsymbol{0}'_{(1-P)n_1})', \boldsymbol{x_2} = (\boldsymbol{1}'_{Pn_2}, \boldsymbol{0}'_{(1-P)n_2})', \boldsymbol{x_3} = (\boldsymbol{1}'_{Pn_3}, \boldsymbol{0}'_{(1-P)n_3})', \boldsymbol{x_4} = (\boldsymbol{1}'_{Pn_4}, \boldsymbol{0}'_{(1-P)n_4})', \tag{74}
$$

respectively. Now, $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{X} = \begin{cases} \mathbf{1}_{n_2 n_3 n_4} \otimes \boldsymbol{x_1} & \text{(RBD with level-one randomization)} \\ \mathbf{1}_{n_3 n_4} \otimes \boldsymbol{x_2} \otimes \mathbf{1}_{n_1} & \text{(RBD with level-two randomization)} \\ \mathbf{1}_{n_4} \otimes \boldsymbol{x_3} \otimes \mathbf{1}_{n_1 n_2} & \text{(RBD with level-three randomization)} \\ \boldsymbol{x_4} \otimes \mathbf{1}_{n_1 n_2 n_3}, & \text{(HD)} \end{cases} \tag{75}$$

for the respective randomized trials. Then, $se(\hat{\delta})$ can be calculated as

$$se(\hat{\delta}) = \begin{cases} se(\hat{\delta}_1) = \sigma \sqrt{\frac{1}{NP(1-P)(\sigma^{(1)}-\sigma^{(2)})}} = \sigma \sqrt{\frac{f_1}{NP(1-P)}} = \sigma \sqrt{\frac{\rho_1}{NP(1-P)}}, \\ se(\hat{\delta}_2) = \sigma \sqrt{\frac{1}{NP(1-P)[\sigma^{(1)}+(n_1-1)\sigma^{(2)}-n_1\sigma^{(3)}]}} = \sigma \sqrt{\frac{f_2}{NP(1-P)}} = \sigma \sqrt{\frac{(n_1\rho_2+\rho_1)}{NP(1-P)}}, \\ se(\hat{\delta}_3) = \sigma \sqrt{\frac{1}{NP(1-P)[\sigma^{(1)}+(n_1-1)\sigma^{(2)}+n_1(n_2-1)\sigma^{(3)}-n_1n_2\sigma^{(4)}]}} = \sigma \sqrt{\frac{f_3}{NP(1-P)}} = \sigma \sqrt{\frac{(n_1n_2\rho_3+n_1\rho_2+\rho_1)}{NP(1-P)}}, \\ se(\hat{\delta}_4) = \sigma \sqrt{\frac{1}{NP(1-P)[\sigma^{(1)}+(n_1-1)\sigma^{(2)}+n_1(n_2-1)\sigma^{(3)}+n_1n_2(n_3-1)\sigma^{(4)}]}} = \sigma \sqrt{\frac{f_4}{NP(1-P)}} = \sigma \sqrt{\frac{(n_1n_2n_3\rho_4+n_1n_2\rho_3+n_1\rho_2+\rho_1)}{NP(1-P)}}, \end{cases} \tag{76}$$

for the respective randomized trials, leading to the same results as shown in (6). In addition, the relation $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ can be easily derived, since

$$E(\hat{\boldsymbol{\beta}}) = E((\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y}) = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}E(\boldsymbol{Y}) = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\boldsymbol{X}}\boldsymbol{\beta}) = \boldsymbol{\beta}. \tag{77}$$

In our case, this result indicates the relation $E(\hat{\delta}) = \delta$.

When heterogeneity of the experimental effect (random slopes) is assumed at each level (i.e., replacing δ with δ_{jkl}, δ_{kl}, or δ_l; see footnote 4 in the main manuscript), substituting $\tilde{\boldsymbol{\Sigma}}_s = \tilde{\boldsymbol{\Sigma}} + \boldsymbol{\Sigma}_s$ into

$\tilde{\boldsymbol{\Sigma}}$ of equation (73) can provide the standard error of the experimental effect estimate ($se(\hat{\delta})$). Here,

$$\boldsymbol{\Sigma}_s = \boldsymbol{X_0} \circ (\boldsymbol{I}_{n_4} \otimes \boldsymbol{\Sigma_0}), \tag{78}$$

$$\boldsymbol{\Sigma_0} = \sigma^2[\rho_4\omega_4 \mathbf{1}_{n_1n_2n_3}\mathbf{1}'_{n_1n_2n_3} + \boldsymbol{I}_{n_3} \otimes (\rho_3\omega_3 \mathbf{1}_{n_1n_2}\mathbf{1}'_{n_1n_2}) + \boldsymbol{I}_{n_2n_3} \otimes (\rho_2\omega_2 \mathbf{1}_{n_1}\mathbf{1}'_{n_1})], \tag{79}$$

and

$$\boldsymbol{X_0} = \begin{cases} \boldsymbol{I}_{n_2n_3n_4} \otimes (\boldsymbol{x_1}\boldsymbol{x}'_{\boldsymbol{1}}) & \text{(RBD with level-one randomization)} \\ \boldsymbol{I}_{n_3n_4} \otimes (\boldsymbol{x_2} \otimes \mathbf{1}_{n_1}) \otimes (\boldsymbol{x_2} \otimes \mathbf{1}_{n_1})' & \text{(RBD with level-two randomization)} \\ \boldsymbol{I}_{n_4} \otimes (\boldsymbol{x_3} \otimes \mathbf{1}_{n_1n_2}) \otimes (\boldsymbol{x_3} \otimes \mathbf{1}_{n_1n_2})' & \text{(RBD with level-three randomization)} \\ (\boldsymbol{x_4} \otimes \mathbf{1}_{n_1n_2n_3}) \otimes (\boldsymbol{x_4} \otimes \mathbf{1}_{n_1n_2n_3})', & \text{(HD)} \end{cases} \tag{80}$$

Specifically, $se(\hat{\delta})$ is given by the square root of the $(2, 2)$ element of

$$(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{s}}^{-\mathbf{1}}\tilde{\boldsymbol{X}})^{-1}, \tag{81}$$

leading to the same results as shown in equation (7). Note that $\sigma^2_{sm} = \sigma^2_m\omega_m$ ($m = 2, 3, 4$) indicates random slope variances.

Likewise, when covariates are included to reduce the magnitudes of residual variances at each level, using conditional variances after including covariates (i.e., σ^{*2}_m and σ^{*2}_{sm}) rather than unconditional variances (i.e., σ^2_m and σ^2_{sm}) in equation (79) can provide a standard error of the experimental effect estimate that is equivalent to equation (8).

9 Appendix B: Standard error of experimental effect $\hat{\delta}$ when the number of levels is two or three

As a special case of the results shown in (8), standard errors of the experimental effect $\hat{\delta}$ (i.e., $se(\hat{\delta})$) when the number of levels is three can be expressed as

$$se(\hat{\delta}) = \begin{cases} se(\hat{\delta}_1) = \sigma\sqrt{\frac{P(1-P)n_1n_2\rho_3\omega_3(1-R^2_{s3})+P(1-P)n_1\rho_2\omega_2(1-R^2_{s2})+\rho_1(1-R^2_1)}{NP(1-P)}} \\ se(\hat{\delta}_2) = \sigma\sqrt{\frac{P(1-P)n_1n_2\rho_3\omega_3(1-R^2_{s3})+n_1\rho_2(1-R^2_2)+\rho_1(1-R^2_1)}{NP(1-P)}} \\ se(\hat{\delta}_3) = \sigma\sqrt{\frac{n_1n_2\rho_3(1-R^2_3)+n_1\rho_2(1-R^2_2)+\rho_1(1-R^2_1)}{NP(1-P)}} \end{cases} \tag{82}$$

where $N = n_1n_2n_3$, $\rho_m = \sigma^2_m/(\sigma^2_1+\sigma^2_2+\sigma^2_3)$ $(m = 1, 2, 3)$ and $\sigma^2 = \sigma^2_1+\sigma^2_2+\sigma^2_3$. The values of df differ according to the levels of randomization: $df = n_3 - g_3 - 1$ for one- and level-two RBD, and $df = n_3 - g_3 - 2$ for HD, where g_3 denotes the number of covariates at level three. When the desired width of the confidence interval is specified as L, the relation $2t_{1-\alpha/2,df}\,se(\hat{\delta}) \le L$ can be re-expressed for each level unit in level-one RBD as

$$n_1 > \frac{4\sigma^2(1-R^2_1)\rho_1 t^2_{1-\alpha/2,df}}{P(1-P)[L^2n_2n_3 - 4(1-R^2_{s3})n_2\rho_3\omega_3 t^2_{1-\alpha/2,df} - 4(1-R^2_{s2})\rho_2\omega_2 t^2_{1-\alpha/2,df}]}, \tag{83}$$

$$n_2 > \frac{4\sigma^2[P(1-P)(1-R^2_{s2})n_1\rho_2\omega_2 + (1-R^2_1)\rho_1]t^2_{1-\alpha/2,df}}{P(1-P)n_1[L^2n_3 - 4(1-R^2_{s3})\rho_3\omega_3 t^2_{1-\alpha/2,df}]}, \tag{84}$$

$$n_3 > \frac{4\sigma^2[P(1-P)(1-R^2_{s3})n_1n_2\rho_3\omega_3 + P(1-P)(1-R^2_{s2})n_1\rho_2\omega_2 + (1-R^2_1)\rho_1]t^2_{1-\alpha/2,df}}{P(1-P)n_1n_2L^2}. \tag{85}$$

Similar results

$$n_1 > \frac{4\sigma^2(1-R_1^2)\rho_1 t_{1-\alpha/2,df}^2}{L^2P(1-P)n_2n_3 - 4P(1-P)(1-R_{s3}^2)n_2\rho_3\omega_3 t_{1-\alpha/2,df}^2 - 4(1-R_2^2)\rho_2 t_{1-\alpha/2,df}^2}, \tag{86}$$

$$n_2 > \frac{4\sigma^2[(1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{P(1-P)n_1[L^2n_3 - 4(1-R_{s3}^2)\rho_3\omega_3 t_{1-\alpha/2,df}^2]}, \tag{87}$$

$$n_3 > \frac{4\sigma^2[P(1-P)(1-R_{s3}^2)n_1n_2\rho_3\omega_3 + (1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{P(1-P)n_1n_2L^2}. \tag{88}$$

for level-two RBD, and

$$n_1 > \frac{4\sigma^2(1-R_1^2)\rho_1 t_{1-\alpha/2,df}^2}{L^2P(1-P)n_2n_3 - 4(1-R_3^2)n_2\rho_3 t_{1-\alpha/2,df}^2 - 4(1-R_2^2)\rho_2 t_{1-\alpha/2,df}^2}, \tag{89}$$

$$n_2 > \frac{4\sigma^2[(1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{n_1[L^2P(1-P)n_3 - 4(1-R_3^2)\rho_3 t_{1-\alpha/2,df}^2]}, \tag{90}$$

$$n_3 > \frac{4\sigma^2[(1-R_3^2)n_1n_2\rho_3 + (1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{P(1-P)n_1n_2L^2}. \tag{91}$$

for HD can also be derived. If the number of levels is two, $se(\hat{\delta})$) can be expressed as

$$se(\hat{\delta}) = \begin{cases} se(\hat{\delta}_1) = \sigma\sqrt{\frac{P(1-P)n_1\rho_2\omega_2(1-R_{s2}^2)+\rho_1(1-R_1^2)}{NP(1-P)}} \\ se(\hat{\delta}_2) = \sigma\sqrt{\frac{n_1\rho_2(1-R_2^2)+\rho_1(1-R_1^2)}{NP(1-P)}} \end{cases} \tag{92}$$

where $N = n_1n_2$, $\rho_m = \sigma_m^2/(\sigma_1^2+\sigma_2^2)$ $(m = 1, 2)$, and $\sigma^2 = \sigma_1^2+\sigma_2^2$. The values of df differ according to the levels of randomization: $df = n_2 - g_2 - 1$ for level-one RBD and $df = n_2 - g_2 - 2$ for HD, where g_2 denotes the number of covariates in level two. The relation $2t_{1-\alpha/2,df}se(\hat{\delta}) \leq L$ can be re-expressed for each level unit in level-one RBD as

$$n_1 > \frac{4\sigma^2(1-R_1^2)\rho_1 t_{1-\alpha/2,df}^2}{P(1-P)[L^2n_2 - 4(1-R_{s2}^2)\rho_2\omega_2 t_{1-\alpha/2,df}^2]}, \tag{93}$$

$$n_2 > \frac{4\sigma^2[P(1-P)(1-R_{s2}^2)n_1\rho_2\omega_2 + (1-R_1^2)\rho_1]t_{1-\alpha/2,df}^2}{P(1-P)n_1L^2}. \tag{94}$$

Similar results

$$n_1 > \frac{4\sigma^2(1-R_1^2)\rho_1 t^2_{1-\alpha/2,df}}{L^2P(1-P)n_2 - 4(1-R_2^2)\rho_2 t^2_{1-\alpha/2,df}}, \tag{95}$$

$$n_2 > \frac{4\sigma^2[(1-R_2^2)n_1\rho_2 + (1-R_1^2)\rho_1]t^2_{1-\alpha/2,df}}{P(1-P)n_1L^2}. \tag{96}$$

can also be given for HD.

Table 1. Example of an assignment indicator variable for each design for four-level data ($n_1 = n_2 = n_3 = n_4 = 2$).

	level-one RBD	level-two RBD	level-three RBD	HD
x_{1111}	1	1	1	1
x_{2111}	0	1	1	1
x_{1211}	1	0	1	1
x_{2211}	0	0	1	1
x_{1121}	1	1	0	1
x_{2121}	0	1	0	1
x_{1221}	1	0	0	1
x_{2221}	0	0	0	1
x_{1112}	1	1	1	0
x_{2112}	0	1	1	0
x_{1212}	1	0	1	0
x_{2212}	0	0	1	0
x_{1122}	1	1	0	0
x_{2122}	0	1	0	0
x_{1222}	1	0	0	0
x_{2222}	0	0	0	0

*RBD…randomized blocked design, HD…hierarchical design

Table 2. Minimum required values of level-$\boldsymbol{M}$ units ($\boldsymbol{n_M}$, $\boldsymbol{M}$= total number of levels = 2,3,4) under a randomized blocked design when the total amount of data from lower units becomes infinite (two-sided significance level is $\boldsymbol{\alpha = 0.05}$).

R^2_{sM}	ρ_M	ω_M	Desired width of confidence interval L for standardized experimental effects (Δ)									
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	0.1	0.1	4	1	1	1	1	1	1	1	1	1
0.1	0.3	0.1	11	3	2	1	1	1	1	1	1	1
0.1	0.5	0.1	19	5	3	2	1	1	1	1	1	1
0.1	0.7	0.1	26	7	3	2	2	1	1	1	1	1
0.1	0.9	0.1	33	9	4	3	2	1	1	1	1	1
0.1	0.1	0.5	19	5	3	2	1	1	1	1	1	1
0.1	0.3	0.5	55	14	7	4	3	2	2	1	1	1
0.1	0.5	0.5	91	23	11	6	4	3	2	2	2	1
0.1	0.7	0.5	127	32	15	8	6	4	3	2	2	2
0.1	0.9	0.5	163	41	19	11	7	5	4	3	3	2
0.1	0.1	1	37	10	5	3	2	2	1	1	1	1
0.1	0.3	1	109	28	13	7	5	4	3	2	2	2
0.1	0.5	1	181	46	21	12	8	6	4	3	3	2
0.1	0.7	1	253	64	29	16	11	8	6	4	4	3
0.1	0.9	1	325	82	37	21	13	10	7	6	5	4
0.3	0.1	0.1	3	1	1	1	1	1	1	1	1	1
0.3	0.3	0.1	9	3	1	1	1	1	1	1	1	1
0.3	0.5	0.1	15	4	2	1	1	1	1	1	1	1
0.3	0.7	0.1	20	5	3	2	1	1	1	1	1	1
0.3	0.9	0.1	26	7	3	2	2	1	1	1	1	1
0.3	0.1	0.5	15	4	2	1	1	1	1	1	1	1
0.3	0.3	0.5	43	11	5	3	2	2	1	1	1	1
0.3	0.5	0.5	71	18	8	5	3	2	2	2	1	1
0.3	0.7	0.5	99	25	11	7	4	3	3	2	2	1
0.3	0.9	0.5	127	32	15	8	6	4	3	2	2	2
0.3	0.1	1	29	8	4	2	2	1	1	1	1	1
0.3	0.3	1	85	22	10	6	4	3	2	2	2	1
0.3	0.5	1	141	36	16	9	6	4	3	3	2	2
0.3	0.7	1	197	50	22	13	8	6	5	4	3	2
0.3	0.9	1	253	64	29	16	11	8	6	4	4	3
0.5	0.1	0.1	3	1	1	1	1	1	1	1	1	1
0.5	0.3	0.1	7	2	1	1	1	1	1	1	1	1
0.5	0.5	0.1	11	3	2	1	1	1	1	1	1	1
0.5	0.7	0.1	15	4	2	1	1	1	1	1	1	1
0.5	0.9	0.1	19	5	3	2	1	1	1	1	1	1
0.5	0.1	0.5	11	3	2	1	1	1	1	1	1	1
0.5	0.3	0.5	31	8	4	2	2	1	1	1	1	1
0.5	0.5	0.5	51	13	6	4	3	2	2	1	1	1
0.5	0.7	0.5	71	18	8	5	3	2	2	2	1	1
0.5	0.9	0.5	91	23	11	6	4	3	2	2	2	1
0.5	0.1	1	21	6	3	2	1	1	1	1	1	1
0.5	0.3	1	61	16	7	4	3	2	2	1	1	1
0.5	0.5	1	101	26	12	7	5	3	3	2	2	2
0.5	0.7	1	141	36	16	9	6	4	3	3	2	2
0.5	0.9	1	181	46	21	12	8	6	4	3	3	2

*$\boldsymbol{R^2_{sM}}$…coefficient of determination for slopes at level-$\boldsymbol{M}$, $\boldsymbol{\rho_M}$…(proportion of) residual variance for intercepts at level-$\boldsymbol{M}$, $\boldsymbol{\omega_M}$…ratio of slope variances to intercept variances at level-$\boldsymbol{M}$.

Table 3. Minimum required values of level-M units (n_M, M= total number of levels =2,3,4) under a hierarchical design when the total amount of data from lower units becomes infinite (two-sided significance level is $\alpha = 0.05$).

R_M^2	ρ_M	P	Desired width of confidence interval L for standardized experimental effects (Δ)									
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	0.1	0.5	145	37	17	10	6	5	3	3	2	2
0.1	0.3	0.5	433	109	49	28	18	13	9	7	6	5
0.1	0.5	0.5	721	181	81	46	29	21	15	12	9	8
0.1	0.7	0.5	1009	253	113	64	41	29	21	16	13	11
0.1	0.9	0.5	1297	325	145	82	52	37	27	21	17	13
0.1	0.1	0.7	172	43	20	11	7	5	4	3	3	2
0.1	0.3	0.7	515	129	58	33	21	15	11	9	7	6
0.1	0.5	0.7	858	215	96	54	35	24	18	14	11	9
0.1	0.7	0.7	1201	301	134	76	49	34	25	19	15	13
0.1	0.9	0.7	1543	386	172	97	62	43	32	25	20	16
0.1	0.1	0.9	401	101	45	26	17	12	9	7	5	5
0.1	0.3	0.9	1201	301	134	76	49	34	25	19	15	13
0.1	0.5	0.9	2001	501	223	126	81	56	41	32	25	21
0.1	0.7	0.9	2801	701	312	176	113	78	58	44	35	29
0.1	0.9	0.9	3601	901	401	226	145	101	74	57	45	37
0.3	0.1	0.5	113	29	13	8	5	4	3	2	2	2
0.3	0.3	0.5	337	85	38	22	14	10	7	6	5	4
0.3	0.5	0.5	561	141	63	36	23	16	12	9	7	6
0.3	0.7	0.5	785	197	88	50	32	22	17	13	10	8
0.3	0.9	0.5	1009	253	113	64	41	29	21	16	13	11
0.3	0.1	0.7	134	34	15	9	6	4	3	3	2	2
0.3	0.3	0.7	401	101	45	26	17	12	9	7	5	5
0.3	0.5	0.7	667	167	75	42	27	19	14	11	9	7
0.3	0.7	0.7	934	234	104	59	38	26	20	15	12	10
0.3	0.9	0.7	1201	301	134	76	49	34	25	19	15	13
0.3	0.1	0.9	312	78	35	20	13	9	7	5	4	4
0.3	0.3	0.9	934	234	104	59	38	26	20	15	12	10
0.3	0.5	0.9	1556	389	173	98	63	44	32	25	20	16
0.3	0.7	0.9	2178	545	242	137	88	61	45	35	27	22
0.3	0.9	0.9	2801	701	312	176	113	78	58	44	35	29
0.5	0.1	0.5	81	21	9	6	4	3	2	2	1	1
0.5	0.3	0.5	241	61	27	16	10	7	5	4	3	3
0.5	0.5	0.5	401	101	45	26	17	12	9	7	5	5
0.5	0.7	0.5	561	141	63	36	23	16	12	9	7	6
0.5	0.9	0.5	721	181	81	46	29	21	15	12	9	8
0.5	0.1	0.7	96	24	11	6	4	3	2	2	2	1
0.5	0.3	0.7	286	72	32	18	12	8	6	5	4	3
0.5	0.5	0.7	477	120	53	30	20	14	10	8	6	5
0.5	0.7	0.7	667	167	75	42	27	19	14	11	9	7
0.5	0.9	0.7	858	215	96	54	35	24	18	14	11	9
0.5	0.1	0.9	223	56	25	14	9	7	5	4	3	3
0.5	0.3	0.9	667	167	75	42	27	19	14	11	9	7
0.5	0.5	0.9	1112	278	124	70	45	31	23	18	14	12
0.5	0.7	0.9	1556	389	173	98	63	44	32	25	20	16
0.5	0.9	0.9	2001	501	223	126	81	56	41	32	25	21

*R_M^2…coefficient of determination for intercepts at level-M, ρ_M…(proportion of) residual variance for intercepts at level-M, P…proportion of experimental group size.

Table 4. Summary of results for different experimental designs

	RBD			HD
	level-one RBD	level-two RBD	level-three RBD	
Assignment indicator variable	X_{ijkl}	X_{jkl}	X_{kl}	X_{l}
Sample size determination formulas	Equations 11–14	Equations 15–18	Equations 19–22	Equations 23–26
Dependence of heterogeneity of experimental effect (random slope variances: ω) in formulas	Yes			No
Degree of freedom in testing $H_0 : \delta = 0$	$n_4 - g_4 - 1$			$n_4 - g_4 - 2$
Asymptotic standard errors (Equations 51–54)	$\sqrt{\frac{\sigma(1-R_{s4}^2)\rho_4\,\omega_4}{n_4}}$			$\sqrt{\frac{\sigma(1-R_4^2)\rho_4}{P(1-P)n_4}}$
Minimum required numbers in the highest units n_4 when the total amount of data from lower units ($n_1\ n_2\ n_3$) becomes infinite	$n_4 \geq \frac{4\sigma^2(1-R_{s4}^2)\rho_4\,\omega_4}{L^2}$ (Equation 55)			$n_4 \geq \frac{4\sigma^2(1-R_4^2)\rho_4}{L^2P(1-P)}$ (Equation 56)
Standard errors of experimental effect estimates (when $\omega_2 = \omega_3 = \omega_4 = 0$)	$se(\hat{\delta}_1) \leq se(\hat{\delta}_2) \leq se(\hat{\delta}_3) \leq se(\hat{\delta}_4)$			
Relative influences of the proportion of units in the experimental group (P) on standard errors (equations 39–40)	small ⟷ large			

*RBD…randomized blocked design, HD…hierarchical design, n_4 …unit size at the fourth level, g_4 … **the number of covariates at the fourth level**, R_{s4}^2 ⋯coefficient of determination for slopes at the **four**th level, R_4^2…coefficient of determination for intercepts at the **four**th level, ρ_4…(proportion of) residual variance for intercepts at the **four**th level, ω_4 …ratio of slope variances to intercept variances at the **four**th level, P…proportion of experimental group size, L…desired width of confidence intervals.

Confidence Interval-Based Sample Size Determination Formulas and Some Mathematical Properties for Hierarchical Data

Online Supporting Materials

level-1 randomization in four-level design (equations 11-14 in the main manuscript) …p.2
level-2 randomization in four-level design (equations 15-18 in the main manuscript) …p.3
level-3 randomization in four-level design (equations 19-22 in the main manuscript) …p.4
level-4 randomization in four-level design (equations 23-26 in the main manuscript) …p.5
level-1 randomization in three-level design …p.6
level-2 randomization in three-level design …p.7
level-3 randomization in three-level design …p.8
level-1 randomization in two-level design …p.9
level-2 randomization in two-level design …p.10

level-1 randomization in four-level design (rho1+rho2+rho3+rho4=1)

#Equation 11 in main manuscript

```
L4random1n1<-function(L,alpha,g4,n2,n3,n4,rho1,rho2,rho3,rho4,R1sq,Rs2sq,Rs3sq,Rs4sq,w2,w3,w4,P,sigma){
n1<-floor((4*sigma^2*(1-R1sq)*rho1*qt(1-alpha/2,n4-g4-1)^2)/(P*(1-P)*(L^2*n2*n3*n4-4*(1-
Rs4sq)*n2*n3*rho4*w4*qt(1-alpha/2,n4-g4-1)^2-4*(1-Rs3sq)*n2*rho3*w3*qt(1-alpha/2,n4-g4-1)^2-4*(1-
Rs2sq)*rho2*w2*qt(1-alpha/2,n4-g4-1)^2)))+1
return(n1)
}
L4random1n1(L,alpha,g4,n2,n3,n4,rho1,rho2,rho3,rho4,R1sq,Rs2sq,Rs3sq,Rs4sq,w2,w3,w4,P,sigma)
```

#Equation 12 in main manuscript

```
L4random1n2<-function(L,alpha,g4,n1,n3,n4,rho1,rho2,rho3,rho4,R1sq,Rs2sq,Rs3sq,Rs4sq,w2,w3,w4,P,sigma){
n2<-floor((4*sigma^2*(P*(1-P)*(1-Rs2sq)*n1*rho2*w2+(1-R1sq)*rho1)*qt(1-alpha/2,n4-g4-1)^2)/(P*(1-
P)*n1*(L^2*n3*n4-4*(1-Rs4sq)*n3*rho4*w4*qt(1-alpha/2,n4-g4-1)^2-4*(1-Rs3sq)*rho3*w3*qt(1-alpha/2,n4-g4-
1)^2)))+1
return(n2)
}
L4random1n2(L,alpha,g4,n1,n3,n4,rho1,rho2,rho3,rho4,R1sq,Rs2sq,Rs3sq,Rs4sq,w2,w3,w4,P,sigma)
```

#Equation 13 in main manuscript

```
L4random1n3<-function(L,alpha,g4,n1,n2,n4,rho1,rho2,rho3,rho4,R1sq,Rs2sq,Rs3sq,Rs4sq,w2,w3,w4,P,sigma){
n3<-floor((4*sigma^2*(P*(1-P)*(1-Rs3sq)*n1*n2*rho3*w3+P*(1-P)*(1-Rs2sq)*n1*rho2*w2+(1-R1sq)*rho1)*qt(1-
alpha/2,n4-g4-1)^2)/(P*(1-P)*n1*n2*(L^2*n4-4*(1-Rs4sq)*rho4*w4*qt(1-alpha/2,n4-g4-1)^2)))+1
return(n3)
}
L4random1n3(L,alpha,g4,n1,n2,n4,rho1,rho2,rho3,rho4,R1sq,Rs2sq,Rs3sq,Rs4sq,w2,w3,w4,P,sigma)
```

#Equation 14 in main manuscript

```
L4random1n4<-function(L,alpha,g4,n1,n2,n3,rho1,rho2,rho3,rho4,R1sq,Rs2sq,Rs3sq,Rs4sq,w2,w3,w4,P,sigma){
n4<-g4+1+1; STOP<-0
while(STOP==0){
Diff<-n4-(4*sigma^2*(P*(1-P)*(1-Rs4sq)*n1*n2*n3*rho4*w4+P*(1-P)*(1-Rs3sq)*n1*n2*rho3*w3+P*(1-P)*(1-
Rs2sq)*n1*rho2*w2+(1-R1sq)*rho1)*qt(1-alpha/2,n4-g4-1)^2)/(L^2*P*(1-P)*n1*n2*n3)
if(Diff>0){
STOP<-1; n4<-n4
}else{
STOP<-0; n4<-n4+1
};}
return(n4)
}
L4random1n4(L,alpha,g4,n1,n2,n3,rho1,rho2,rho3,rho4,R1sq,Rs2sq,Rs3sq,Rs4sq,w2,w3,w4,P,sigma)
```

level-2 randomization in four-level design (rho1+rho2+rho3+rho4=1)

#Equation 15 in main manuscript

```
L4random2n1<-function(L,alpha,g4,n2,n3,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,Rs3sq,Rs4sq,w3,w4,P,sigma){
n1<-floor((4*sigma^2*(1-R1sq)*rho1*qt(1-alpha/2,n4-g4-1)^2)/(L^2*P*(1-P)*n2*n3*n4-4*P*(1-P)*(1-
Rs4sq)*n2*n3*rho4*w4*qt(1-alpha/2,n4-g4-1)^2-4*P*(1-P)*(1-Rs3sq)*n2*rho3*w3*qt(1-alpha/2,n4-g4-1)^2-4*(1-
R2sq)*rho2*qt(1-alpha/2,n4-g4-1)^2))+1
return(n1)
}
L4random2n1(L,alpha,g4,n2,n3,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,Rs3sq,Rs4sq,w3,w4,P,sigma)
```

#Equation 16 in main manuscript

```
L4random2n2<-function(L,alpha,g4,n1,n3,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,Rs3sq,Rs4sq,w3,w4,P,sigma){
n2<-floor((4*sigma^2*((1-R2sq)*n1*rho2+(1-R1sq)*rho1)*qt(1-alpha/2,n4-g4-1)^2)/(P*(1-P)*n1*(L^2*n3*n4-4*(1-
Rs4sq)*n3*rho4*w4*qt(1-alpha/2,n4-g4-1)^2-4*(1-Rs3sq)*rho3*w3*qt(1-alpha/2,n4-g4-1)^2)))+1
return(n2)
}
L4random2n2(L,alpha,g4,n1,n3,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,Rs3sq,Rs4sq,w3,w4,P,sigma)
```

#Equation 17 in main manuscript

```
L4random2n3<-function(L,alpha,g4,n1,n2,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,Rs3sq,Rs4sq,w3,w4,P,sigma){
n3<-floor((4*sigma^2*(P*(1-P)*(1-Rs3sq)*n1*n2*rho3*w3+(1-R2sq)*n1*rho2+(1-R1sq)*rho1)*qt(1-alpha/2,n4-g4-
1)^2)/(P*(1-P)*n1*n2*(L^2*n4-4*(1-Rs4sq)*rho4*w4*qt(1-alpha/2,n4-g4-1)^2)))+1
return(n3)
}
L4random2n3(L,alpha,g4,n1,n2,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,Rs3sq,Rs4sq,w3,w4,P,sigma)
```

#Equation 18 in main manuscript

```
L4random2n4<-function(L,alpha,g4,n1,n2,n3,rho1,rho2,rho3,rho4,R1sq,R2sq,Rs3sq,Rs4sq,w3,w4,P,sigma){
n4<-g4+1+1; STOP<-0
while(STOP==0){
Diff<-n4-(4*sigma^2*(P*(1-P)*(1-Rs4sq)*n1*n2*n3*rho4*w4+P*(1-P)*(1-Rs3sq)*n1*n2*rho3*w3+(1-
R2sq)*n1*rho2+(1-R1sq)*rho1)*qt(1-alpha/2,n4-g4-1)^2)/(L^2*P*(1-P)*n1*n2*n3)
if(Diff>0){
STOP<-1; n4<-n4
}else{
STOP<-0; n4<-n4+1
};}
return(n4)
}
L4random2n4(L,alpha,g4,n1,n2,n3,rho1,rho2,rho3,rho4,R1sq,R2sq,Rs3sq,Rs4sq,w3,w4,P,sigma)
```

level-3 randomization in four-level design (rho1+rho2+rho3+rho4=1)

#Equation 19 in main manuscript

```
L4random3n1<-function(L,alpha,g4,n2,n3,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,Rs4sq,w4,P,sigma){
n1<-floor((4*sigma^2*(1-R1sq)*rho1*qt(1-alpha/2,n4-g4-1)^2)/(L^2*P*(1-P)*n2*n3*n4-4*P*(1-P)*(1-
Rs4sq)*n2*n3*rho4*w4*qt(1-alpha/2,n4-g4-1)^2-4*(1-R3sq)*n2*rho3*qt(1-alpha/2,n4-g4-1)^2-4*(1-
R2sq)*rho2*qt(1-alpha/2,n4-g4-1)^2))+1
return(n1)
}
L4random3n1(L,alpha,g4,n2,n3,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,Rs4sq,w4,P,sigma)
```

#Equation 20 in main manuscript

```
L4random3n2<-function(L,alpha,g4,n1,n3,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,Rs4sq,w4,P,sigma){
n2<-floor((4*sigma^2*((1-R2sq)*n1*rho2+(1-R1sq)*rho1)*qt(1-alpha/2,n4-g4-1)^2)/(n1*(L^2*P*(1-P)*n3*n4-
4*P*(1-P)*(1-Rs4sq)*n3*rho4*w4*qt(1-alpha/2,n4-g4-1)^2-4*(1-R3sq)*rho3*qt(1-alpha/2,n4-g4-1)^2)))+1
return(n2)
}
L4random3n2(L,alpha,g4,n1,n3,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,Rs4sq,w4,P,sigma)
```

#Equation 21 in main manuscript

```
L4random3n3<-function(L,alpha,g4,n1,n2,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,Rs4sq,w4,P,sigma){
n3<-floor((4*sigma^2*((1-R3sq)*n1*n2*rho3+(1-R2sq)*n1*rho2+(1-R1sq)*rho1)*qt(1-alpha/2,n4-g4-1)^2)/(P*(1-
P)*n1*n2*(L^2*n4-4*(1-Rs4sq)*rho4*w4*qt(1-alpha/2,n4-g4-1)^2)))+1
return(n3)
}
L4random3n3(L,alpha,g4,n1,n2,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,Rs4sq,w4,P,sigma)
```

#Equation 22 in main manuscript

```
L4random3n4<-function(L,alpha,g4,n1,n2,n3,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,Rs4sq,w4,P,sigma){
n4<-g4+1+1; STOP<-0
while(STOP==0){
Diff<-n4-(4*sigma^2*(P*(1-P)*(1-Rs4sq)*n1*n2*n3*rho4*w4+(1-R3sq)*n1*n2*rho3+(1-R2sq)*n1*rho2+(1-
R1sq)*rho1)*qt(1-alpha/2,n4-g4-1)^2)/(L^2*P*(1-P)*n1*n2*n3)
if(Diff>0){
STOP<-1; n4<-n4
}else{
STOP<-0; n4<-n4+1
};}
return(n4)
}
L4random3n4(L,alpha,g4,n1,n2,n3,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,Rs4sq,w4,P,sigma)
```

level-4 randomization in four-level design (rho1+rho2+rho3+rho4=1)

#Equation 23 in main manuscript

```
L4random4n1<-function(L,alpha,g4,n2,n3,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,R4sq,P,sigma){
n1<-floor((4*sigma^2*(1-R1sq)*rho1*qt(1-alpha/2,n4-g4-2)^2)/(L^2*P*(1-P)*n2*n3*n4-4*(1-
R4sq)*n2*n3*rho4*qt(1-alpha/2,n4-g4-2)^2-4*(1-R3sq)*n2*rho3*qt(1-alpha/2,n4-g4-2)^2-4*(1-R2sq)*rho2*qt(1-
alpha/2,n4-g4-2)^2))+1
return(n1)
}
L4random4n1(L,alpha,g4,n2,n3,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,R4sq, P,sigma)
```

#Equation 24 in main manuscript

```
L4random4n2<-function(L,alpha,g4,n1,n3,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,R4sq,P,sigma){
n2<-floor((4*sigma^2*((1-R2sq)*n1*rho2+(1-R1sq)*rho1)*qt(1-alpha/2,n4-g4-2)^2)/(n1*(L^2*P*(1-P)*n3*n4-4*(1-
R4sq)*n3*rho4*qt(1-alpha/2,n4-g4-2)^2-4*(1-R3sq)*rho3*qt(1-alpha/2,n4-g4-2)^2)))+1
return(n2)
}
L4random4n2(L,alpha,g4,n1,n3,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,R4sq, P,sigma)
```

#Equation 25 in main manuscript

```
L4random4n3<-function(L,alpha,g4,n1,n2,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,R4sq,P,sigma){
n3<-floor((4*sigma^2*((1-R3sq)*n1*n2*rho3+(1-R2sq)*n1*rho2+(1-R1sq)*rho1)*qt(1-alpha/2,n4-g4-2)
^2)/(n1*n2*(L^2*P*(1-P)*n4-4*(1-R4sq)*rho4*qt(1-alpha/2,n4-g4-2)^2)))+1
return(n3)
}
L4random4n3(L,alpha,g4,n1,n2,n4,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,R4sq ,P,sigma)
```

#Equation 26 in main manuscript

```
L4random4n4<-function(L,alpha,g4,n1,n2,n3,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,R4sq,P,sigma){
n4<-g4+2+1; STOP<-0
while(STOP==0){
Diff<-n4-((4*sigma^2*((1-R4sq)*n1*n2*n3*rho4+(1-R3sq)*n1*n2*rho3+(1-R2sq)*n1*rho2+(1-R1sq)*rho1)*qt(1-
alpha/2,n4-g4-2)^2)/(L^2*P*(1-P)*n1*n2*n3))
if(Diff>0){
STOP<-1; n4<-n4
}else{
STOP<-0; n4<-n4+1
};}
return(n4)
}
L4random4n4(L,alpha,g4,n1,n2,n3,rho1,rho2,rho3,rho4,R1sq,R2sq,R3sq,R4sq ,P,sigma)
```

level-1 randomization in three-level design (rho1+rho2+rho3=1)

```
L3random1n1<-function(L,alpha,g3,n2,n3,rho1,rho2,rho3,R1sq,Rs2sq,Rs3sq,w2,w3,P,sigma){
n1<-floor((4*sigma^2*(1-R1sq)*rho1*qt(1-alpha/2,n3-g3-1)^2)/(P*(1-P)*(L^2*n2*n3-4*(1-
Rs3sq)*n2*rho3*w3*qt(1-alpha/2,n3-g3-1)^2-4*(1-Rs2sq)*rho2*w2*qt(1-alpha/2,n3-g3-1)^2)))+1
return(n1)
}
L3random1n1(L,alpha,g3,n2,n3,rho1,rho2,rho3,R1sq,Rs2sq,Rs3sq,w2,w3,P,sigma)


L3random1n2<-function(L,alpha,g3,n1,n3,rho1,rho2,rho3,R1sq,Rs2sq,Rs3sq,w2,w3,P,sigma){
n2<-floor((4*sigma^2*(P*(1-P)*(1-Rs2sq)*n1*rho2*w2+(1-R1sq)*rho1)*qt(1-alpha/2,n3-g3-1)^2)/(P*(1-
P)*n1*(L^2*n3-4*(1-Rs3sq)*rho3*w3*qt(1-alpha/2,n3-g3-1)^2)))+1
return(n2)
}
L3random1n2(L,alpha,g3,n1,n3,rho1,rho2,rho3,R1sq,Rs2sq,Rs3sq,w2,w3,P,sigma)


L3random1n3<-function(L,alpha,g3,n1,n2,rho1,rho2,rho3,R1sq,Rs2sq,Rs3sq,w2,w3,P,sigma){
n3<-g3+1+1; STOP<-0
while(STOP==0){
Diff<-n3-(4*sigma^2*(P*(1-P)*(1-Rs3sq)*n1*n2*rho3*w3+P*(1-P)*(1-Rs2sq)*n1*rho2*w2+(1-R1sq)*rho1)*qt(1-
alpha/2,n3-g3-1)^2)/(P*(1-P)*n1*n2*L^2)
if(Diff>0){
STOP<-1; n3<-n3
}else{
STOP<-0; n3<-n3+1
};}
return(n3)
}
L3random1n3(L,alpha,g3,n1,n2,rho1,rho2,rho3,R1sq,Rs2sq,Rs3sq,w2,w3,P,sigma)
```

level-2 randomization in three-level design (rho1+rho2+rho3=1)

```
L3random2n1<-function(L,alpha,g3,n2,n3,rho1,rho2,rho3,R1sq,R2sq,Rs3sq,w3,P,sigma){
n1<-floor((4*sigma^2*(1-R1sq)*rho1*qt(1-alpha/2,n3-g3-1)^2)/(L^2*P*(1-P)*n2*n3-4*P*(1-P)*(1-
Rs3sq)*n2*rho3*w3*qt(1-alpha/2,n3-g3-1)^2-4*(1-R2sq)*rho2*qt(1-alpha/2,n3-g3-1)^2))+1
return(n1)
}
L3random2n1(L,alpha,g3,n2,n3,rho1,rho2,rho3,R1sq,R2sq,Rs3sq,w3,P,sigma)


L3random2n2<-function(L,alpha,g3,n1,n3,rho1,rho2,rho3,R1sq,R2sq,Rs3sq,w3,P,sigma) {
n2<-floor((4*sigma^2*((1-R2sq)*n1*rho2+(1-R1sq)*rho1)*qt(1-alpha/2,n3-g3-1)^2)/(P*(1-P)*n1*(L^2*n3-4*(1-
Rs3sq)*rho3*w3*qt(1-alpha/2,n3-g3-1)^2)))+1
return(n2)
}
L3random2n2(L,alpha,g3,n1,n3,rho1,rho2,rho3,R1sq,R2sq,Rs3sq,w3,P,sigma)


L3random2n3<-function(L,alpha,g3,n1,n2,rho1,rho2,rho3,R1sq,R2sq,Rs3sq,w3,P,sigma){
n3<-g3+1+1; STOP<-0
while(STOP==0){
Diff<-n3-(4*sigma^2*(P*(1-P)*(1-Rs3sq)*n1*n2*rho3*w3+(1-R2sq)*n1*rho2+(1-R1sq)*rho1)*qt(1-alpha/2,n3-g3-
1)^2)/(P*(1-P)*n1*n2*L^2)
if(Diff>0){
STOP<-1; n3<-n3
}else{
STOP<-0; n3<-n3+1
};}
return(n3)
}
L3random2n3(L,alpha,g3,n1,n2,rho1,rho2,rho3,R1sq,R2sq,Rs3sq,w3,P,sigma)
```

level-3 randomization in three-level design (rho1+rho2+rho3=1)

```
L3random3n1<-function(L,alpha,g3,n2,n3,rho1,rho2,rho3,R1sq,R2sq,R3sq,P,sigma){
n1<-floor((4*sigma^2*(1-R1sq)*rho1*qt(1-alpha/2,n3-g3-2)^2)/(L^2*P*(1-P)*n2*n3-4*(1-R3sq)*n2*rho3*qt(1-alpha/2,n3-g3-2)^2-4*(1-R2sq)*rho2*qt(1-alpha/2,n3-g3-2)^2))+1
return(n1)
}
L3random3n1(L,alpha,g3,n2,n3,rho1,rho2,rho3,R1sq,R2sq,R3sq,P,sigma)


L3random3n2<-function(L,alpha,g3,n1,n3,rho1,rho2,rho3,R1sq,R2sq,R3sq,P,sigma){
n2<-floor((4*sigma^2*((1-R2sq)*n1*rho2+(1-R1sq)*rho1)*qt(1-alpha/2,n3-g3-2)^2)/(n1*(L^2*P*(1-P)*n3-4*(1-R3sq)*rho3*qt(1-alpha/2,n3-g3-2)^2)))+1
return(n2)
}
L3random3n2(L,alpha,g3,n1,n3,rho1,rho2,rho3,R1sq,R2sq,R3sq,P,sigma)


L3random3n3<-function(L,alpha,g3,n1,n2,rho1,rho2,rho3,R1sq,R2sq,R3sq,P,sigma){
N3<-g3+2+1; STOP<-0
while(STOP==0){
Diff<-n3-(4*sigma^2*((1-R3sq)*n1*n2*rho3+(1-R2sq)*n1*rho2+(1-R1sq)*rho1)*qt(1-alpha/2,n3-g3-2)^2)/(P*(1-P)*n1*n2*L^2)
if(Diff>0){
STOP<-1; n3<-n3
}else{
STOP<-0; n3<-n3+1
};}
return(n3)
}
L3random3n3(L,alpha,g3,n1,n2,rho1,rho2,rho3,R1sq,R2sq,R3sq,P,sigma)
```

level-1 randomization in two-level design (rho1+rho2=1)

```
L2random1n1<-function(L,alpha,g2,n2,rho1,rho2,R1sq,Rs2sq,w2,P,sigma){
n1<-floor((4*sigma^2*(1-R1sq)*rho1*qt(1-alpha/2,n2-g2-1)^2)/(P*(1-P)*(L^2*n2-4*(1-Rs2sq)*rho2*w2*qt(1-alpha/2,n2-g2-1)^2)))+1
return(n1)
}
L2random1n1(L,alpha,g2,n2,rho1,rho2,R1sq,Rs2sq,w2,P,sigma)


L2random1n2<-function(L,alpha,g2,n1,rho1,rho2,R1sq,Rs2sq,w2,P,sigma){
N2<-g2+1+1; STOP<-0
while(STOP==0){
Diff<-n2-(4*sigma^2*(P*(1-P)*(1-Rs2sq)*n1*rho2*w2+(1-R1sq)*rho1)*qt(1-alpha/2,n2-g2-1)^2)/(P*(1-P)*n1*L^2)
if(Diff>0){
STOP<-1; n2<-n2
}else{
STOP<-0; n2<-n2+1
};}
return(n2)
}
L2random1n2(L,alpha,g2,n1,rho1,rho2,R1sq,Rs2sq,w2,P,sigma)
```

level-2 randomization in two-level design (rho1+rho2=1)

```
L2random2n1<-function(L,alpha,g2,n2,rho1,rho2,R1sq,R2sq,P,sigma){
n1<-floor((4*sigma^2*(1-R1sq)*rho1*qt(1-alpha/2,n2-g2-2)^2)/(L^2*P*(1-P)*n2-4*(1-R2sq)*rho2*qt(1-alpha/2,n2-g2-2)^2))+1
return(n1)
}
L2random2n1(L,alpha,g2,n2,rho1,rho2,R1sq,R2sq,P,sigma)


L2random2n2<-function(L,alpha,g2,n1,rho1,rho2,R1sq,R2sq,P,sigma){
n3<-g2+2+1; STOP<-0
while(STOP==0){
Diff<-n2-(4*sigma^2*((1-R2sq)*n1*rho2+(1-R1sq)*rho1)*qt(1-alpha/2,n2-g2-2)^2)/(P*(1-P)*n1*L^2)
if(Diff>0){
STOP<-1; n2<-n2
}else{
STOP<-0; n2<-n2+1
};}
return(n2)
}
L2random2n2(L,alpha,g2,n1,rho1,rho2,R1sq,R2sq,P,sigma)
```