Behav Res (2014) 46:346–356
DOI 10.3758/s13428-013-0387-1

Generalized sample size determination formulas for experimental research with hierarchical data

Satoshi Usami

Published online: 3 October 2013

Abstract Hierarchical data sets arise when the data for lower units (e.g., individuals such as students, clients, and citizens) are nested within higher units (e.g., groups such as classes, hospitals, and regions). In data collection for experimental research, estimating the required sample size beforehand is a fundamental question for obtaining sufficient statistical power and precision of the focused parameters. The present research extends previous research from Heo and Leon (2008) and Usami (2011b), by deriving closed-form formulas for determining the required sample size to test effects in experimental research with hierarchical data, and by focusing on both multisite-randomized trials (MRTs) and cluster-randomized trials (CRTs). These formulas consider both statistical power and the width of the confidence interval of a standardized effect size, on the basis of estimates from a random-intercept model for three-level data that considers both balanced and unbalanced designs. These formulas also address some important results, such as the lower bounds of the needed units at the highest levels.

Keywords Sample size · Statistical power · Hierarchical data · Multilevel data · Experimental research

A hierarchical linear model (HLM[1]) is a regression model for hierarchical data sets. Hierarchical data sets result from nesting the data for lower units (e.g., individuals such as students, clients, and citizens) within higher units (e.g., groups such as classes, hospitals, and regions). Repeated measures data and data collected through paired designs are also hierarchical data (Raudenbush & Bryk, 2002; Singer & Willett, 2003). The main advantages of using HLMs are attaining improved estimates of parameters and improved information on the residuals at different hierarchical levels. Software packages for hierarchical data analysis include HLM, MLwiN, Mplus, R, SAS, and SPSS.

In data collection for experimental research, estimating the required sample size beforehand is fundamental to obtaining sufficient statistical power and precision of the focused parameters, and sample size determination problems are often closely related to human resource and budget requirements (Chow, Shao, & Wang, 2003; Raudenbush, 1997; Usami, 2011a). Estimating the expected statistical power before beginning research by power analysis is sometimes crucial to avoiding wrong conclusions (Cohen, 1988). However, actual psychological research is often underpowered (Bezeau & Graves, 2001; Cohen, 1962; Maxwell, 2004; Maxwell, Kelley, & Rausch, 2008). Although in simpler data collection designs there are multiple ways of conveniently conducting power analysis for both nonhierarchical data (e.g., Dupont & Plummer, 1998, for PS; Faul, Erdfelder, Lang, & Buchner, 2007, for G*Power 3) and hierarchical data (e.g., Donner & Klar, 2000, for ACluster; Fosgate, 2007; Raudenbush, Spybrook, Congdon, Liu, & Martinez, 2011, for OD), a more intuitive method is strongly desired for estimating the required sample size for more general data collection designs.

The present research provides closed-form formulas that generalize sample size requirements for testing effects in experimental research with hierarchical data, focusing on both multisite-randomized trials (MRTs), in which individuals are randomized (Raudenbush & Liu, 2000), and cluster-randomized trials (CRTs), in which clusters are randomized (Heo & Leon, 2008). Although MRTs are generally preferable to CRTs, since in CRTs the dependency of individual response data within clusters (i.e., intraclass correlation) inflates the standard errors of estimates of the experimental effects (see the Comparing CRTs and MRTs section for details). However, in many cases CRTs have to be chosen, due to the research

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

S. Usami (✉)
Department of Psychology, University of Southern California, Los Angeles, CA, USA
e-mail: usami_s@p.u-tokyo.ac.jp

purposes (e.g., a difference of doctors is a focused issue of the intervention).

These formulas are derived through considering both statistical power and the width of the confidence interval for a standardized effect size, on the basis of estimates from random-intercept models for three-level data that consider both balanced and unbalanced designs. As was summarized in Usami (2011a), although several methods have been developed for sample size determination in hierarchical data (e.g., Heo & Leon, 2008; Okumura, 2007; Raudenbush, 1997; Raudenbush & Liu, 2000; Roy, Bhaumik, Aryal, & Gibbons, 2007; Usami, 2011b), these methods were developed under designs featuring restricted data collection and numbers of levels. For example, Heo and Leon derived a closed-form power function and a formula for determining the sample size required to detect a single experimental effect in three-level hierarchical CRTs. The derived formulas were restricted to CRTs under a balanced design, however, and formulas to obtain desired confidence intervals were not considered. Roy et al. devised a general method for sample size determination in a three-level HLM for longitudinal data, but it was not in a closed form and the experimental design was not directly considered. In the Japanese literature, Usami (2011b) derived formulas for MRTs and CRTs, but the derived formulas were restricted to two-level hierarchical data under a balanced design. Three-level hierarchies arise frequently in both cross-sectional studies (e.g., students are nested within classes within schools) and longitudinal studies (e.g., longitudinally obtained data are nested within patients within hospitals). In the present research, formulas were derived in a unified way, using generalized least squares estimators for experimental effects, in order to overcome the restrictions of the former research. These formulas also address additional results not derived in previous research, such as lower bounds on the number of required units in the highest (third) level and cases involving more than three levels.

This article is organized into five sections. The following section introduces a three-level HLM. The one after gives derivations of the generalized formulas and examples of estimating the required sample size on the basis of programs provided in the Appendix. Next, additional results obtained from the derived formulas are addressed, and the final section discusses prospects for the proposed method and related problems.

Statistical model

This section introduces a three-level random-intercept model that considers both balanced and unbalanced designs, referring to Heo and Leon (2008). Experimental and control groups are sometimes unbalanced due to practical considerations. For example, producing experimental drugs for clinical trials is expensive, so the experimental and control groups may be of unequal size (Ogungbenro & Aarons, 2009). For brevity, the discussion here will be confined to the case in which the numbers of Level 1 units (e.g., students) and Level 2 units (e.g., classes) are equal within the Level 3 unit (e.g., schools), and in which no attrition occurs during trials (on this point, some conventional alternatives are addressed in the Discussion section).

Let Y_{ijk} be the outcome for an i (= 1, 2, . . . , I)-th Level 1 unit nested within a j (= 1, 2, . . . , J)-th Level 2 unit, which is again nested within a k (= 1, 2, . . . , K)-th Level 3 unit. The following Level 1 model is assumed for expressing Y_{ijk}:

$$Y_{ijk} = \beta_{0jk} + \delta X_{ijk} + e_{ijk}. \quad (1)$$

Here, X_{ijk} is a corresponding assignment indicator variable, set to 1 when it is assigned to an experimental group and to 0 when it is assigned to a control group. Let the proportion of an experimental group size be P ($0 < P < 1$). The balanced condition is satisfied only when $P = .5$. In CRT, essentially $X_{ijk} = X_k$, since clusters are randomized, so the number of Level 1 units assigned to an experimental group per Level 2 units is $P \times I$ for MRTs, whereas it is 0 or I for CRTs. β_{0jk} is a random intercept denoting the overall control group mean in the jth Level 2 unit nested within the kth Level 3 unit. e_{ijk} is the corresponding residual, assumed to be independent of X_{ijk}. Additionally, e_{ijk} is assumed to be distributed as $e_{ijk} \sim N(0,{\sigma_1}^2)$. Here, ${\sigma_3}^2$ is the residual variance for the respective groups in each Level 1 unit.

The Level 2 model is a decomposition form of β_{0jk} as

$$\beta_{0\,jk} = \beta_{0k} + e_{jk}. \quad (2)$$

Here, β_{0k} is a random intercept denoting the overall control group mean in the kth Level 3 unit. e_{jk} is the corresponding residual and is assumed to be independent of X_{ijk} and of other residuals. Additionally, e_{jk} is assumed to be distributed as $e_{jk} \sim N(0,{\sigma_2}^2)$, and ${\sigma_2}^2$ is the residual variance for the respective groups in each Level 2 unit.

β_{0k} can be further decomposed in order to obtain the following Level 3 model:

$$\beta_{0k} = \beta_0 + e_k. \quad (3)$$

Here, β_0 is the overall control group mean. e_k is the corresponding residual, assumed to be independent of X_{ijk} and of other residuals. Additionally, e_k is assumed to be distributed as $e_k \sim N(0,{\sigma_3}^2)$, and ${\sigma_3}^2$ is the residual variance for the respective groups in each Level 3 unit. From Eqs. 2 and 3, Eq. 1 can now be written as

$$Y_{ijk} = \left(\beta_0 + \delta X_{ijk}\right) + \left(e_k + e_{jk} + e_{ijk}\right). \quad (4)$$

This combined form clarifies the fixed and residual parts of the three-level model. From Eq. 4, it is evident that the mean of Y_{ijk}, given X_{ijk}, is

$$E\left(Y_{ijk}\,|X_{ijk}\right)=\beta_0+\delta X_{ijk}, \tag{5}$$

where $E()$ denotes the mean. Additionally, the covariance of Y_{ijk} and $Y_{i'j'k'}$ can be generally expressed as

$$\begin{aligned}\mathrm{cov}\left(Y_{ijk},Y_{i'j'k'}\,|\,X_{ijk},X_{i'j'k'}\right)&=1(i=i'\&j=j'\&k=k')\sigma_1{}^2\\&+1(j=j'\&k=k')\sigma_2{}^2+1(k=k')\sigma_3{}^2,\end{aligned} \tag{6}$$

where $cov()$ denotes the covariance, and 1() is an indicator function, which has a value of 1 if the conditions in parentheses are satisfied and 0 if they are not. From Eq. 6, the variance of Y_{ijk} (namely, $i=i'$, $j=j'$, and $k=k'$, respectively) can be expressed as

$$\mathrm{Var}\left(Y_{ijk}\,|X_{ijk}\right)=\sigma_1^2+\sigma_2^2+\sigma_3^2=\sigma^2. \tag{7}$$

Here, $Var()$ denotes the variance. The standardized effect size Δ of an experimental effect δ is defined according to Cohen (1988) by using the pooled standard deviation σ. Namely,

$$\Delta=\frac{\delta}{\sigma}. \tag{8}$$

Therefore, the intraclass correlation coefficient (ICC) among the Level 2 data can now be expressed as

$$\rho_2=\mathrm{Corr}\left(Y_{ijk},Y_{i'j'k}\right)=\frac{\sigma_3^2}{\sigma_1^2+\sigma_2^2+\sigma_3^2}=\frac{\sigma_3^2}{\sigma^2}, \tag{9}$$

and the ICC among the Level 1 data can be expressed as

$$\rho_1=\mathrm{Corr}\left(Y_{ijk},Y_{i'jk}\right)=\frac{\sigma_2^2+\sigma_3^2}{\sigma_1^2+\sigma_2^2+\sigma_3^2}=\frac{\sigma_2^2+\sigma_3^2}{\sigma^2}. \tag{10}$$

Here, $Corr()$ denotes the correlation.

Derivation of generalized formulas

Standard errors of δ

Without loss of generality, we can set $\sigma^2=1$, and then from Eqs. 8–10, $\Delta=\delta$, $\rho_2=\sigma_3{}^2$, and $\rho_1=\sigma_2{}^2+\sigma_3{}^2$. If residual variances $\sigma_1{}^2$, $\sigma_2{}^2$, and $\sigma_3{}^2$ are known, the test statistic Z for the null hypothesis $H_0:\delta=0$ can be constructed as

$$Z=\frac{\widehat{\delta}}{se\left(\widehat{\delta}\right)}, \tag{11}$$

where $se\left(\widehat{\delta}\right)$ denotes a standard error of estimate $\widehat{\delta}$. Z is normally distributed as $Z\sim N(\delta,1)$. To derive formulas under a clear and unified procedure, consider a matrix notation of Eq. 4:

$$\boldsymbol{Y}=\widetilde{\boldsymbol{X}}\boldsymbol{\beta}+\widetilde{\boldsymbol{\varepsilon}}. \tag{12}$$

Here, $\boldsymbol{\beta}=(\beta_0,\delta)'$ and $\boldsymbol{Y}$ is an $(I\times J\times K)\times 1$ vector, and its elements are arranged as $\boldsymbol{Y}=(\boldsymbol{Y'_1},\ldots,\boldsymbol{Y'_k},\ldots,\boldsymbol{Y'_K})'$, where $\boldsymbol{Y_k}=(\boldsymbol{Y'_{1k}},\ldots,\boldsymbol{Y'_{jk}},\ldots,\boldsymbol{Y'_{Jk}})'$ and $\boldsymbol{Y_{jk}}=(Y_{1jk},\ldots,Y_{ijk},\ldots,Y_{Ijk})'$. $\widetilde{\boldsymbol{X}}=(\boldsymbol{1_{IJK}},\boldsymbol{X})$ is a corresponding $(I\times J\times K)\times 2$ matrix, and $\boldsymbol{X}$ is a corresponding $(I\times J\times K)\times 1$ vector including information about X_{ijk}. $\widetilde{\varepsilon}$ is also a corresponding $(I\times J\times K)\times 1$ vector including information about $\widetilde{e}_{ijk}=e_k+e_{jk}+e_{ijk}$. From Eq. 6, it can be shown that $\widetilde{e}_{ijk}$ is distributed as $\widetilde{e}_{ijk}\sim N(0,\boldsymbol{\Sigma})$, where

$$\widetilde{\boldsymbol{\Sigma}}=I_K\otimes\boldsymbol{\Sigma}, \tag{13}$$

$$\begin{aligned}\boldsymbol{\Sigma}&=\sigma_3^2\boldsymbol{1}_{IJ}\boldsymbol{1}'_{IJ}+\boldsymbol{I}_J\otimes\left(\sigma_2^2\boldsymbol{1}_I\boldsymbol{1}'_I\right)+\sigma_1^2\boldsymbol{I}_{IJ}\\&=\rho_2\boldsymbol{1}_{IJ}\boldsymbol{1}'_{IJ}+\boldsymbol{I}_J\otimes[(\rho_1-\rho_2)\boldsymbol{1}_I\boldsymbol{1}'_I]+(1-\rho_1)\boldsymbol{I}_{IJ}.\end{aligned} \tag{14}$$

Here we assume that $\sigma_1{}^2\geq 0$, $\sigma_2{}^2\geq 0$, and $\sigma_3{}^2\geq 0$, and that the inverse matrix of $\boldsymbol{\Sigma}$ (denoted as $\boldsymbol{\Sigma}^{-1}$) exists. Let the diagonal elements of $\boldsymbol{\Sigma}^{-1}$ be a, the off-diagonal elements denoting the same Level 2 and Level 3 units in $\boldsymbol{\Sigma}^{-1}$ be b, and the off-block diagonal elements denoting the same Level 3 unit in $\boldsymbol{\Sigma}^{-1}$ be c. Comparing the left and right sides of the identity $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}=\boldsymbol{I}$, the following equations are obtained:

$$\begin{aligned}&a+(I-1)\,\rho_1 b+I\,(J-1)\,\rho_2 c=1,\\&b+\rho_1 a+(I-2)\,\rho_1 b+I\,(J-1)\,\rho_2 c=0,\\&[1+(I-1)\rho_1]c+\rho_2[a+(I-1)b]+(J-2)I\rho_2 c=0.\end{aligned} \tag{15}$$

These equations can be rewritten as

$$\begin{aligned}a&=b+\frac{1}{1-\rho_1},\\b&=\frac{(f-I\rho_2)\rho_1-I(J-1)\rho_2^2}{I^2(J-1)\rho_2^2+(I\rho_2-f)[(I-1)\rho_1+1]}\left[\frac{1}{1-\rho_1}\right],\\c&=\frac{\rho_2}{I\rho_2-f}\left[Ib+\frac{1}{1-\rho_1}\right],\end{aligned} \tag{16}$$

where $f=1+I(J-1)\rho_2+(I-1)\rho_1$ is a variance inflation factor or design effect (Heo & Leon, 2008). Simple calculation shows that

$$\frac{1}{a+(I-1)b+I(J-1)c}=f. \tag{17}$$

Using the generalized least squares estimators, a sample distribution of $\widehat{\boldsymbol{\beta}}$ can be expressed as $\widehat{\boldsymbol{\beta}} \sim N\left[\left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y}, \left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{X}}\right)^{-1}\right]$, and then $se\left(\widehat{\delta}\right)$ can be evaluated by (2, 2) elements of $\left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{X}}\right)^{-1/2} = \left[\widetilde{\boldsymbol{X}}'\left(\boldsymbol{I}_K \otimes \boldsymbol{\Sigma}^{-1}\right)\widetilde{\boldsymbol{X}}\right]^{-1/2}$.

Let $\boldsymbol{x_m}$ and $\boldsymbol{x_c}$ be $\boldsymbol{x_m} = (\boldsymbol{1}'_{\boldsymbol{PI}}, \boldsymbol{0}'_{\boldsymbol{(1-P)I}})'$ and $\boldsymbol{x_c} = (\boldsymbol{1}'_{\boldsymbol{PK}}, \boldsymbol{0}'_{\boldsymbol{(1-P)K}})'$, respectively. Now $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{X} = \begin{cases} \boldsymbol{1_{JK}} \otimes \boldsymbol{x_m} & (MRTs) \\ \boldsymbol{x_c} \otimes \boldsymbol{1_{IJ}} & (CRTs). \end{cases} \tag{18}$$

for the respective randomized trials. Then, from Eqs. 17 and 18, $se\left(\widehat{\delta}\right)$ can be calculated as

$$se\left(\widehat{\delta}\right) = \begin{cases} se\left(\widehat{\delta}_M\right) = \sqrt{\dfrac{1}{IJKP(1-P)(a-b)}} = \sqrt{\dfrac{1-\rho_1}{IJKP(1-P)}}, & (MRTs) \\ se\left(\widehat{\delta}_C\right) = \sqrt{\dfrac{1}{IJKP(1-P)[a+(I-1)b+I(J-1)c]}} = \sqrt{\dfrac{f}{IJKP(1-P)}}, & (CRTs). \end{cases} \tag{19}$$

for the respective randomized trials. As for $se\left(\widehat{\delta}_C\right)$ in Eq. 19, this completely corresponds to the results of Heo and Leon (2008) when a balanced design is used (i.e., $P = 1/2$). From Eq. 19, it is evident that larger ρ_2 and ρ_1 lead to an $se\left(\widehat{\delta}\right)$ that is smaller in MRTs and larger in CRTs, and $se\left(\widehat{\delta}_M\right) = se\left(\widehat{\delta}_C\right)$ when $\rho_2 = \rho_1 = 0$.

Generalized formulas for desired statistical power

Let α be a two-sided significance level for the test of the null hypothesis $H_0 : \delta = 0$. Under test statistic Z in Eq. 11, statistical power ϕ can be evaluated as

$$\phi = \Phi\left[z_{\alpha/2} - E(Z)\right] + \Phi\left[E(Z) - z_{1-\alpha/2}\right]. \tag{20}$$

Here, Φ is a cumulative density function of the standard normal distribution, and z_α denotes the $100\alpha\%$ point of a standard normal distribution.

Without loss of generality, a positive experimental effect (i.e., $\delta \geq 0$) is assumed here. Then, the probability that $z_{\alpha/2}$ exceeds $E(Z)$ (i.e., $\Phi[z_{\alpha/2} - E(Z)]$) is generally very low, and the first term in Eq. 20 can be pragmatically ignored, unless sample size or effect size is too small (e.g., Usami, 2011b).[2] The above equation therefore becomes

$$\phi \approx \Phi\left[E(Z) - z_{1-\alpha/2}\right]. \tag{21}$$

[2] For example, even when $\alpha = .05$ and a small value of $E(Z) = .2$ are assumed, $\Phi[z_{a/2} - E(Z)] = \Phi[z_{.025} - .2] = \Phi[-1.9599 - .2] = .0154$, and this probability can be pragmatically ignored. In cases of situations such as $E(Z) > .2$, this probability becomes much lower.

For a desired statistical power ψ, we can obtain the following relations (Usami, 2011b):

$$\begin{aligned} &\Phi\left[E(Z) - z_{1-a/2}\right] \geq \psi, \\ &\leftrightarrow E(Z) - z_{1-a/2} \geq z_\psi, \\ &\leftrightarrow E(Z) \geq z_{1-a/2} + z_\psi. \end{aligned} \tag{22}$$

Additionally, when ${\sigma_1}^2$, ${\sigma_2}^2$, and ${\sigma_3}^2$ (namely, ρ_2 and ρ_1) are known, $E\left(\widehat{\delta}\right) = \delta$, since $E\left(\widehat{\boldsymbol{\beta}}\right) = E\left[\left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y}\right] = (\widetilde{\boldsymbol{X}}'\ \boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{\Sigma}}^{-1}E(\boldsymbol{Y}) = \left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{\Sigma}}^{-1}\left(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}\right) = \boldsymbol{\beta}$. Then, this relation and Eqs. 11, 19, and 22, give the following required sample size (IJK) for an MRT design:

$$IJK \geq \frac{\left(z_{1-\alpha/2} + z_\psi\right)^2(1-\rho_1)}{P(1-P)\Delta^2}. \tag{23}$$

Note that $\delta = \Delta$ because σ^2 is assumed to be 1. From this formula, under fixed α and P, larger ρ_1 and Δ lead to a smaller sample size requirement for a desired statistical power ψ. Additionally, a balanced design in which $P = 1/2$ provides the least demands on sample size. Likewise, in a CRT design, the formula for required sample size can be obtained from Eqs. 11, 19, and 22. Since $se\left(\widehat{\delta}_C\right)$ includes f in its numerator, we get the following sample size determination formulas for the respective units:

$$I \geq \frac{(1-\rho_1)\left(z_{1-\alpha/2} + z_\psi\right)^2}{JKP(1-P)\Delta^2 - \left(z_{1-\alpha/2} + z_\psi\right)^2[(J-1)\rho_2 + \rho_1]} \tag{24}$$

$$J \geq \frac{[I(\rho_1 - \rho_2) + 1 - \rho_1]\left(z_{1-\alpha/2} + z_\psi\right)^2}{IKP(1-P)\Delta^2 - \rho_2 I\left(z_{1-\alpha/2} + z_\psi\right)^2} \quad (25)$$

$$K \geq \frac{f\left(z_{1-\alpha/2} + z_\psi\right)^2}{IJP(1-P)\Delta^2} = \frac{[1 + I(J-1)\rho_2 + (I-1)\rho_1]\left(z_{1-\alpha/2} + z_\psi\right)^2}{IJP(1-P)\Delta^2}. \quad (26)$$

Note that Eqs. 24–26 reduce to the formulas obtained in Heo and Leon (2008) when a balanced design is used (i.e., $P = 1/2$). In each equation, for a fixed α and P, a larger ρ_2 and ρ_1 and a smaller Δ lead to larger sample size requirements for a desired statistical power ψ. Additionally, as in an MRT design setting, a balanced design ($P = 1/2$) provides the least demands on sample size.

Generalized formulas for confidence intervals

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

$$\hat{\delta} - z_{1-\alpha/2} se\left(\hat{\delta}\right) \leq \delta \leq \hat{\delta} + z_{1-\alpha/2} se\left(\hat{\delta}\right), \quad (27)$$

so the width of a confidence interval L can be evaluated as

$$L = 2z_{1-\alpha/2} se\left(\hat{\delta}\right). \quad (28)$$

Note that this is also the width of the confidence interval for Δ because σ^2 is assumed to be 1. When a desired width of the confidence interval is specified as L', using Eqs. 19 and 28, the relation $L \leq L'$ can be reexpressed in MRTs as follows:

$$IJK \geq \frac{4z_{1-\alpha/2}^2(1-\rho_1)}{P(1-P)L'^2}. \quad (29)$$

Naturally, as ρ_1 and L' become smaller, the required sample size becomes larger. In CRTs, as with the formulas for desired statistical power (Eqs. 24–26), determination formulas can be derived for the respective units as

$$I \geq \frac{4z_{1-\alpha/2}^2(1-\rho_1)}{JKP(1-P)L'^2 - 4z_{1-\alpha/2}^2[(J-1)\rho_2 + \rho_1]} \quad (30)$$

$$J \geq \frac{4z_{1-\alpha/2}^2[1 - I\rho_2 + (I-1)\rho_1]}{IKP(1-P)L'^2 - 4z_{1-\alpha/2}^2 I\rho_2} \quad (31)$$

$$K \geq \frac{4z_{1-\alpha/2}^2 f}{IJP(1-P)L'^2} = \frac{4z_{1-\alpha/2}^2[1 + I(J-1)\rho_2 + (I-1)\rho_1]}{IJP(1-P)L'^2} \quad (32)$$

using Eqs. 19 and 28. Naturally, these equations reduce to the results obtained in Usami (2011b) when the number of levels and P are restricted to two ($K = 1$ and $\rho_2 = 0$) and 1/2 (balanced design), respectively.

Examples

To facilitate the use of the derived formulas, R programs are provided in the Appendices to estimate the minimum required sample size in MRTs and CRTs for statistical power (Eqs. 23–26) and the width of the confidence intervals for different standardized effect sizes of the experimental effects (Eqs. 29–32). Here, we consider a hypothetical situation in which students from different classes and schools are assigned to either an experimental or a control group in order to evaluate the experimental effect on test scores of new learning programs for English conversation. From previous research results, the variance of test scores and the size of the experimental effect δ are assumed to be $\sigma^2 = 20^2$ and $\delta = 16$, respectively, so that $\Delta = 16/20 = .80$. As for ICC, the variance of the means of English conversation ability is assumed to be small among schools, but large among classes in each school, so that ${\sigma_3}^2$ is small and ${\sigma_2}^2$ is large. Therefore, ρ_1 and ρ_2 are set as .15 and .03, respectively.

The desired statistical power and the two-sided significance level for testing the null hypothesis $H_0 : \delta = 0$ are set as $\psi = .80$ and $\alpha = .05$, respectively. If MRTs are conducted, Eq. 23 indicates that the required minimum sample sizes of IJK to achieve $\phi \geq .80$ are calculated as being 42 and 50 for different proportions of experimental group sizes $P = .5$ and .7, respectively. When using the provided programs, the same results can be obtained:

```
MRTpower(alpha = 0.05, psi = 0.80, rho1 = 0.15, Delta = 0.80, P = 0.50)
42
MRTpower(alpha = 0.05, psi = 0.80, rho1 = 0.15, Delta = 0.80, P = 0.70)
50
```

When the required sample size is determined on the basis of the desired width of a confidence interval L' so that $L' = .30$, from Eq. 29 the required minimum sample sizes IJK to achieve $L' \leq .30$ are calculated as being 581 and 692 for $P = .5$ and .7, respectively. When using the provided programs, the same results can be obtained:

```
MRTconfidenceinterval (alpha = 0.05, rho1 = 0.15, L = 0.30, P = 0.50)
581
MRTconfidenceinterval (alpha = 0.05, rho1 = 0.15, L = 0.30, P = 0.70)
692
```

If CRTs are conducted and J and K are fixed at $J = 3$ and $K = 10$, from Eq. 24 the required minimum number of units I to achieve $\phi \geq .80$ is calculated as 3 for $P = .5$. When using the provided programs, the same result can be obtained:

```
CRTpowerI (J = 3, K = 10, alpha = 0.05, psi = 0.80,
rho1 = 0.15, rho2 = 0.03, Delta = 0.80, P = 0.5)
3
```

When the required sample size is determined on the basis of a desired width of the confidence interval of $L' = .70$, from Eq. 30 the required minimum number of units I to achieve $L' \leq .70$ is calculated as being 30 for $P = .5$. When using the provided programs, the same result can be obtained:

```
CRTconfidenceintervalI (J = 3, K = 10, alpha = 0.05, rho1 = 0.15,
rho2 = 0.03, L = 0.70, P = 0.5)
30
```

Some results relating the derived formulas

This section addresses several useful and important results relating the formulas derived above.

Comparing CRTs and MRTs

Comparing the numerators in the square root of $se\left(\widehat{\delta}_C\right)$ and $se\left(\widehat{\delta}_M\right)$ in Eq. 19 shows that $f-(1-\rho_1) = 1 + I(J-1)\rho_2 + (I-1)\rho_1 - (1-\rho_1) = I(J-1)\rho_2 + I\rho_1 = I[(J-1)\rho_2 + \rho_1] \geq 0$, since $I \geq 1$, $J \geq 1$, $\rho_2 \geq 0$, and $\rho_1 \geq 0$. Therefore, $se\left(\widehat{\delta}_C\right) \geq se\left(\widehat{\delta}_M\right)$, and MRTs are always preferable to CRTs. This relation indirectly indicates that $\phi_c \leq \phi_m$ and $L_c \geq L_m$ for any combination of I, J, K, ρ_2, and ρ_1. Here, ϕ_c (or ϕ_m) and L_c (or L_m) are the statistical power and the width of the confidence intervals for CRTs (or MRTs). From Eq. 19, increasing I, J, and K and conducting a balanced design ($P = 1/2$) both lead to a smaller $se\left(\widehat{\delta}\right)$, since $\partial se(\delta)/\partial I < 0$, $\partial se(\delta)/\partial J < 0$, $\partial se(\delta)/\partial K < 0$, and $\partial se(\delta)/\partial P < 0$ ($P \leq 1/2$) in MRTs and CRTs. However, the strengths of the effect of improving I, J, K, and P on $se(\delta)$ are different between MRTs and CRTs. For example, it can be shown that

$$\partial se^2(\delta_m)/\partial J - \partial se^2(\delta_c)/\partial J = I^2KP(1-P)(\rho_1-\rho_2)/W \geq 0, \tag{33}$$

$$\partial se^2(\delta_m)/\partial K - \partial se^2(\delta_c)/\partial K = I^2JP(1-P)[(J-1)\rho_2 + \rho_1]/W \geq 0, \tag{34}$$

$$\partial se^2(\delta_m)/\partial P - \partial se^2(\delta_c)/\partial P = I^2JK(1-2P)[(J-1)\rho_2 + \rho_1]/W \geq 0, \tag{35}$$

where $W = [IJKP(1-P)]^2 \geq 0$. Namely, the effect of improving the values J, K, and P are always more dominant in MRTs than in CRTs. Interestingly, the similar result for I becomes $\partial se^2(\delta_m)/\partial I - \partial se^2(\delta_c)/\partial I = 0$. Namely, the strengths of the effect of improving I are the same between MRTs and CRTs.

As for the influences of ρ_1, an opposite relation holds between MRTs and CRTs, since $\partial se^2(\delta_m)/\partial\rho_1 = -1/IJKP(1-P) < 0$ and $\partial se^2(\delta_c)/\partial\rho_1 = (I-1)/IJKP(1-P) \geq 0$. Namely, a larger ρ_1 always leads to a smaller $se\left(\widehat{\delta}_M\right)$ and a larger $se\left(\widehat{\delta}_C\right)$, and when $I \geq 2$, the absolute strength of ρ_1 is larger in CRTs than in MRTs.

Relative influences of ρ_2 and ρ_1 in CRTs

In CRTs, both ρ_2 and ρ_1 are included in $se(\delta_C)$, and the influences of these ICCs on $se(\delta_C)$ differ. Namely, as Heo and Leon (2008) briefly derived in the case of a balanced design, although larger ρ_2 and ρ_1 lead to a larger $se(\delta)$, the influence of ρ_2 is greater than that of ρ_1, because $\partial f/\partial\rho_2 = I(J-1) > \partial f/\partial\rho_1 = (I-1) \geq 0$ when $J \geq 2$, indicating the dominance of ${\sigma_3}^2$ over ${\sigma_2}^2$.

Asymptotic power and confidence intervals and minimum requirement for K in CRTs

Since $se(\delta_C)$ includes I and J in its numerator and denominator, $se(\delta_C)$ does not take a value near 0, but rather has a lower limit even when I and J become infinite under a fixed number of highest units (K). A lower limit of $se(\delta_C)$ when $I \to \infty$, $J \to \infty$ under a fixed K can be derived as follows:

$$\begin{aligned} &\lim_{J\to\infty}\lim_{I\to\infty}\sqrt{\frac{f}{IJKP(1-P)}} \\ &= \lim_{J\to\infty}\left[\lim_{I\to\infty}\sqrt{\frac{1+I(J-1)\rho_2+(I-1)\rho_1}{IJKP(1-P)}}\right] \\ &= \lim_{J\to\infty}\sqrt{\frac{(J-1)\rho_2+\rho_1}{JKP(1-P)}} \\ &= \sqrt{\frac{\rho_2}{KP(1-P)}} > 0\,. \end{aligned} \tag{36}$$

Combining this result and the relation of Eq. 11, a limit value of $E(Z)$ can be evaluated as $\left(\sqrt{KP(1-P)}\Delta\right)/\sqrt{\rho_2}$. Then, from Eq. 22, the following relation is obtained, indicating the minimum requirement for K:

$$K \geq \frac{\rho_2\left(z_{1-\alpha/2}+z_{\psi}\right)^2}{P(1-P)\Delta^2}. \tag{37}$$

Namely, if K does not satisfy the relation above, the actual statistical power ϕ does not exceed the desired statistical power ψ, even when I and J become infinite. Additionally, from the right side of Eq. 37, this

minimum required K becomes trivial when $\rho_2 = 0$ or Δ is sufficiently large.

A similar equation can be derived for a desired width of the confidence interval, using Eq. 28:

$$K \geq \frac{4z_{1-\alpha/2}^2 \rho_2}{P(1-P)L'^2}. \tag{38}$$

Naturally, Eqs. 37 and 38 reduce to the results obtained in Usami (2011b) when the number of levels and P are restricted to two ($K = 1$ and $\rho_2 = 0$) and 1/2 (balanced design). Minimum integer values of K under $\alpha = .05$ for both a desired statistical power and a desired width of the confidence interval are summarized in Table 1.

Interestingly, these results can also be derived from Eqs. 25 and 31. Namely, the denominators on the right sides of these equations should be positive [$IKP(1-P)\Delta^2 - \rho_2 I(z_{1-\alpha/2} + z_\psi)^2 \geq 0$ and $IKP(1-P)L'^2 - 4z_{1-\alpha/2}^2 I\rho_2 \geq 0$] since both numerators are always positive, and the restriction $J \geq 1$ should be satisfied. Then, Eqs. 37 and 38 can be directly derived from these relations.

Cases with more than three levels

Through the same procedure discussed in the previous section, more generalized formulas can be derived for more than three levels. Let D and $N_1, N_2, \ldots, N_D$ be the number of levels and the number of units for each level, respectively. Let $\sigma_d^2 (d = 1,2\ldots,D)$ and $\rho_d = (\sum_{d+1}^{D} \sigma_d^2)/(\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_D^2)(d; = 1,2,\ldots,D-1)$ be the residual variances and ICCs among the Level d units under the similar d-level models discussed in the Statistical Model section. Thus, the more generalized form of the standard errors of $\widehat{\delta}$ can be derived as

$$se\left(\widehat{\delta}_G\right) = \begin{cases} se\left(\widehat{\delta}_{GM}\right) = \sqrt{\dfrac{1-\rho_1}{\left(\Pi_d^D N_d\right)P(1-P)}}, (MRTs) \\ se\left(\widehat{\delta}_{GC}\right) = \sqrt{\dfrac{1+\sum_{q=1}^{D-2}\left[\left(\Pi_{d=1}^{D-q-1} N_d\right)\left(N_{D-q}-1\right)\rho_{D-q}\right] + (N_1-1)\rho_1}{\left(\Pi_d^D N_d\right)P(1-P)}}, (CRTs) \end{cases} \tag{39}$$

for the respective trials when $D \geq 3$. Naturally, when $D = 3$ (i.e., $N_1 = I$, $N_2 = J$, and $N_3 = K$), this formula corresponds to Eq. 19. In MRTs, more generalized formulas for required sample sizes $\Pi_{d=1}^{D} N_d$ to achieve a desired statistical power and width of the confidence intervals can be obtained through the same equations—namely, Eqs. 23 and 29. In CRTs, sample size determination formulas can be obtained for the respective units as well, and the details are omitted here.

The results discussed in the previous subsections also hold even when $D \geq 3$. For example, the strengths of the effect of improving the values N_2, N_3, and N_D toward $se(\delta)$ are always more dominant in MRTs than in CRTs, whereas the strengths of improving N_1 are the same between MRTs and CRTs. Additionally, the minimum required number of the highest units N_D to achieve a desired statistical power and width of the confidence intervals can be expressed through equations similar to Eqs. 37 and 38, as $N_D \geq \rho_{D-1}(z_{1-\alpha/2} + z_\psi)^2/(P(1-P)\Delta^2)$ and $N_D \geq 4z_{1-\alpha/2}^2 \rho_{D-1}/(P(1-P)L'^2)$, respectively.

Discussion

The present research provides closed-form generalized sample size determination formulas to use when testing effects in experimental research with hierarchical data, focusing on MRTs and CRTs, and these formulas are derived considering both statistical power and the width of the confidence interval of a standardized effect size, on the basis of estimates from a random-intercept model for three-level data that considers both balanced and unbalanced designs. In the present research, as in Usami (2011b), formulas have been derived in a unified way that uses generalized least squared estimators for an experimental effect to overcome the restrictions of previous research. Some additional useful results not derived in the previous research, such as lower bounds on the needed units in the highest (third) level and equations for cases of more than three levels, are also addressed by these formulas. As was noted in the introduction, repeated measures data, paired data, and pre–post data can be analyzed through HLM, and these data are also within the scope of applying the formulas derived here. Additionally, R programs for calculating needed sample sizes are provided in the Appendices to facilitate the use of the derived formulas. Developing a more flexible program is an important topic for future research that will also provide various outputs, including numerical tables. The present and improved programs will be available on the author's website (http://satoshiusami.com/).

As Usami (2011b) noted, almost no previous research focusing on sample size determination for hierarchical data has provided closed formulas and numerical tables that consider the desired width of the confidence intervals that would be usable by applied researchers. Null hypothesis significance

Table 1 Minimum required values of Level 3 units (K) at two-sided significance level of α=.05 in CRTs

(a) Statistical power (desired statistical power φ=.80)

	ρ_2/Δ^*	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	3.00	5.00
P = .5	0.01	32	8	4	2	2	1	1	1	1	1	1	1
	0.05	157	40	18	10	7	5	4	3	2	2	1	1
	0.10	314	79	35	20	13	9	7	5	4	4	1	1
	0.20	628	157	70	40	26	18	13	10	8	7	1	1
	0.30	942	236	105	59	38	27	20	15	12	10	2	1
	0.50	1,570	393	175	99	63	44	33	25	20	16	2	1
P = .7	0.01	38	10	5	3	2	2	1	1	1	1	1	1
	0.05	187	47	21	12	8	6	4	3	3	2	1	1
	0.10	374	94	42	24	15	11	8	6	5	4	1	1
	0.20	748	187	84	47	30	21	16	12	10	8	1	1
	0.30	1,122	281	125	71	45	32	23	18	14	12	2	1
	0.50	1,869	468	208	117	75	52	39	30	24	19	3	1
P = .9	0.01	88	22	10	6	4	3	2	2	2	1	1	1
	0.05	437	110	49	28	18	13	9	7	6	5	1	1
	0.10	873	219	97	55	35	25	18	14	11	9	1	1
	0.20	1,745	437	194	110	70	49	36	28	22	18	2	1
	0.30	2,617	655	291	164	105	73	54	41	33	27	3	2
	0.50	4,361	1,091	485	273	175	122	89	69	54	44	5	2

(b) Confidence interval

	ρ_2/L'^*	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	3.00	5.00
P = .5	0.01	62	16	7	4	3	2	2	1	1	1	1	1
	0.05	308	77	35	20	13	9	7	5	4	4	1	1
	0.10	615	154	69	39	25	18	13	10	8	7	1	1
	0.20	1,230	308	137	77	50	35	26	20	16	13	2	1
	0.30	1,844	461	205	116	74	52	38	29	23	19	3	1
	0.50	3,073	769	342	193	123	86	63	49	38	31	4	2
P = .7	0.01	74	19	9	5	3	3	2	2	1	1	1	1
	0.05	366	92	41	23	15	11	8	6	5	4	1	1
	0.10	732	183	82	46	30	21	15	12	10	8	1	1
	0.20	1,464	366	163	92	59	41	30	23	19	15	2	1
	0.30	2,195	549	244	138	88	61	45	35	28	22	3	1
	0.50	3,659	915	407	229	147	102	75	58	46	37	5	2
P = .9	0.01	171	43	19	11	7	5	4	3	3	2	1	1
	0.05	854	214	95	54	35	24	18	14	11	9	1	1
	0.10	1,708	427	190	107	69	48	35	27	22	18	2	1
	0.20	3,415	854	380	214	137	95	70	54	43	35	4	2
	0.30	5,122	1,281	570	321	205	143	105	81	64	52	6	3
	0.50	8,537	2,135	949	534	342	238	175	134	106	86	10	4

* ρ_2, intraclass correlation among Level 2 units; Δ, effect size of an intervention; L', desired width of confidence interval

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 (Balluerka, Gómez, & Hidalgo, 2005; Cohen, 1994, Sedlmeier, 2009). The American Psychological Association has therefore recommended that researchers report confidence intervals (American Psychological Association, 2009). The derived Formulas 29–32 are simple and have closed forms, and thus seem to be effective tools to encourage applied researchers to collect data and interpret the obtained results on the basis of confidence intervals.

For simplicity, the explicit development of the method proposed here has been confined to a single factor and two levels. However, it will be straightforward to extend the proposed formulas to an arbitrary number of levels and factors. On this point, Usami (2011a) illustrated a simple, unified method of estimating the statistical power of various types of contrasts to be evaluated regarding main effects and interactions for two-factor between-subjects designs, using multiparameter tests based on Wald statistics.

Cases in which the outcome is binary or ordered are also intriguing topics for future research. An important disadvantage of the derived formulas comes from the assumption that no units will be missing for all levels, although attrition does often occur, especially in Level 1 and 2 units. However, as Heo and Leon (2008) discussed, if variation of the numbers of respective units is completely random, in the sense of missing data, then the fixed sample sizes J and I may be replaced by $\tilde{J} = (1/K)\sum_{k=1}^{K} n_k$ and $\tilde{I} = \left(1/\tilde{J}K\right)\sum_{k=1}^{K} \sum_{j=1}^{n_k} n_{jk}$, respectively. Here, n_{jk} denotes the number of Level 2 units in the kth Level 3 unit, and n_{jk} denotes the number of Level 1 units in jth Level 2 unit nested within the kth Level 3 unit.

One possible major limitation of the present research regards the fact that the formulas were derived on the basis of the random-intercept model. There are merits to considering the random-intercepts model, since this model provides direct information about intraclass correlations, which are helpful in determining whether multilevel models are required in the first place. However, the values of slopes can vary significantly among clusters, and the random-intercepts model may not be realistic in actual data. Several researchers (Maas & Hox, 2005; Raudenbush & Liu, 2000; Usami 2011a) have loosened this assumption and discussed ways for evaluating the needed sample size on the basis of a random-intercepts-and-slopes model. Although the relevant parameters and indices for evaluating statistical power would be more complex, the formulas proposed here could be directly extended to the case of a random-intercepts-and-slopes model, and this could be an intriguing topic for future research.

Another limitation is the unrealistic assumption that the residual variances are already known, leading to ignoring asymptotic features of the sample distribution and to the use of a normal distribution to conduct the statistical test of Eq. 11. Therefore, sample sizes calculated from the derived formulas will generally be optimistic and negatively biased. As Heo and Leon (2008) noted, more accurate formulas could be evaluated under noncentral t distributions. However, differences between these distributions are trivial—as simulations performed by Heo and Leon (2008) and Usami (2011a) showed—because when the degrees of freedom are more than 20, the t distribution approaches a normal distribution. This point seems to be important only when the estimated required sample size becomes small (i.e., less than 20)—for example, when a large effect size is assumed.

In estimating the required sample size for hierarchical data, one major problem facing all researchers designing CRTs is the need to specify ICCs (Smeeth & Ng, 2002). In CRTs, the specification of ICCs becomes problematic in behavioral research (for clinical trials, see Hedges & Hedberg, 2007; Murray, Varnell, & Biltstein, 2004; Shoukri, Asyali, & Donner, 2004), since slight misspecifications of the ICCs may cause seriously biased estimation of required sample sizes. As Smeeth and Ng (2002) and Usami (2011b) pointed out, the ideal solution would be to have ICCs available from previous studies that were large enough and that had sufficient clusters to generate reasonably accurate estimates of the ICC for the variable of interest, although this is generally impossible in practice. As Usami (2011b) noted, although the conventional criteria provided in the literature, such as Raudenbush and Bryk (2002, where ICCs of .05, .10, and .15 are small, medium, and large, respectively) and Hox (2010, where ICCs of .10, .20, and .30 are small, medium, and large, respectively), are useful when no informative data are available, actual ICCs depend heavily on the features of the variables of interest and the units. Presenting estimated ICCs for a range of outcomes through a review, as Smeeth and Ng have done for clinical trial research, will be a very useful to aid for the specification of ICCs, and such reviews will be strongly desired for various research areas. As another strategy, constructing models that include covariates to explain the variance of outcomes Y would also be a useful approach to excluding the influence of ICCs (Hedges & Hedberg, 2007; Murray & Blitstein, 2003). However, note that when such covariates correlate highly not only with the outcomes Y but also with the assignment indicator variable X, estimates for an experimental effect δ may be strongly biased and more difficult to interpret (Usami, 2011b).

Although many issues are left to be investigated in future research, in designing an experiment with hierarchical data based on either MRTs or CRTs in order to evaluate an experimental effect, the derived formulas and related results here will be of great help in estimating the required sample size to achieve a desired statistical power and width of the confidence intervals in actual research.

Appendix A

Table 2 Programs for estimating minimum required sample size in MRTs and CRTs (Eqs. 23–26) when desired statistical power is ψ

```
MRTpower<-function(alpha,psi,rho1,Delta,P){
IJK<-floor((qnorm(1-alpha/2) + qnorm(psi))^2*(1-rho1)/(P*(1-P)*Delta^2)) + 1
return(IJK)
}
CRTpowerI<-function(J,K,alpha,psi,rho1,rho2,Delta,P){
I<-floor(((1-rho1)*(qnorm(1-alpha/2) + qnorm(psi))^2)/(J*K*P*(1-P)*Delta^2-
  (qnorm(1-alpha/2) + qnor m(psi))^2*((J-1)*rho2 + rho1))) + 1
return(I)
}
CRTpowerJ<-function(I,K,alpha,psi,rho1,rho2,Delta,P){
J<-floor(((I*(rho1-rho2) + 1-rho1)*(qnorm(1-alpha/2) + qnorm(psi))^2)/
  (I*K*P*(1-P)*Delta^2-rho2*I*(q norm(1-alpha/2) + qnorm(psi))^2)) + 1
return(J)
}
CRTpowerK<-function(I,J,alpha,psi,rho1,rho2,Delta,P){
K<-floor(((1 + I*(J-1)*rho2 + (I-1)*rho1)(qnorm(1-alpha/2) + qnorm(psi))^2)/
  (I*J*P*(1-P)*Delta^2)) + 1
return(K)
}
```

* I, number of Level 1 units; J, number of Level 2 units; K, number of Level 3 units; alpha, two-sided significance level; psi, desired statistical power; rho1, intraclass correlation coefficient (ICC) for Level 1 (Eq. 10); rho2, ICC for Level 2 (Eq. 9); Delta, standardized effect size (Eq. 8); P, proportion of experimental group size

Appendix B

Table 3 Programs for estimating minimum required sample size in MRTs and CRTs (Eqs. 29–32) when desired width of the confidence interval is L

```
MRTconfidenceinterval<-function(alpha,rho1,L,P){
IJK<-floor((4*qnorm(1-alpha/2)^2*(1-rho1))/(P*(1-P)*L^2)) + 1
return(IJK)
}
CRTconfidenceintervalI<-function(J,K,alpha,rho1,rho2,L,P){
I<-floor((4*qnorm(1-alpha/2)^2*(1-rho1))/(J*K*P*(1-P)*L^2-4*qnorm
  (1-alpha/2)^2*((J-1)*rho2 + rho1 ))) + 1
return(I)
}
CRTconfidenceintervalJ<-function(I,K,alpha,rho1,rho2,L,P){
J<-floor((4*qnorm(1-alpha/2)^2(1-I*rho2 + (I-1)*rho1))/(I*K*P*(1-P)*L^2-
  4*qnorm(1-alpha/2)^2*I*rho2)) + 1
return(J)
}
CRTconfidenceintervalK<-function(I,J,alpha,rho1,rho2,L,P){
K<-floor((4*qnorm(1-alpha/2)^2*(1 + I*(J-1)*rho2 + (I-1)*rho1))/
  (I*J*P*(1-P)*L^2)) + 1
return(K)
}
```

* I, number of Level 1 units; J, number of Level 2 units; K, number of Level 3 units; alpha, two-sided significance level; L, desired width of the confidence interval; rho1, intraclass correlation coefficient (ICC) for Level 1 (Eq. 10); rho2, ICC for Level 2 (Eq. 9); P, proportion of experimental group size

References

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Balluerka, N., Gómez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology, 1,* 55–70. doi:10.1027/1614-1881.1.2.55

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology, 23,* 399–406.

Chow, S. C., Shao, J., & Wang, H. (2003). *Sample size calculation in clinical research* (2nd ed.). New York, NY: Chapman & Hall.

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153. doi:10.1037/h0045186

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. doi:10.1037/0003-066X.49.12.997

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London, UK: Arnold.

Dupont, W. D., & Plummer, W. D. (1998). Power and sample size calculations for studies involving linear regression. *Controlled Clinical Trials, 19,* 589–601.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. doi:10.3758/BF03193146

Fosgate, G. T. (2007). A cluster-adjusted sample size algorithm for proportions was developed using a beta-binomial model. *Journal of Clinical Epidemiology, 60,* 250–255.

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

Hedges, L. V., & Hedberg, E. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis, 29,* 60–87.

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, NJ: Erlbaum.

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

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1,* 86–92. doi:10.1027/1614-2241.1.3.86

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59,* 537–563.

Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review, 27,* 79–103.

Murray, D. M., Varnell, S. P., & Biltstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health, 94,* 423–432.

Ogungbenro, K., & Aarons, L. (2009). Sample size/power calculations for repeated ordinal measurements in population pharmacodynamic experiments. *Journal of Pharmacokinetics and Pharmacodynamics, 37,* 67–83.

Okumura, T. (2007). Sample size determination for hierarchical linear models considering uncertainty in parameter estimates. *Behaviormetrika, 34,* 79–94.

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

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

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

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X., & Martinez, A. (2011). Optimal design software for multi-level and longitudinal research (Version 3.01) [Software]. Available from http://www.wtgrantfoundation.org.

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.

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

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: Review and new results. *Statistical Methods in Medical Research, 13,* 251–271.

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

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

Smeeth, L., & Ng, E. S.-W. (2002). Intra-class correlation coefficients for cluster randomized trials in primary care: Data from the MRC trial of the assessment and management of older people in the community. *Control Clinical Trials, 23,* 409–421.

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

Usami, S. (2011b). A unified method for determining sample size needed to evaluate mean difference in hierarchical research design and construction of numerical table—Focusing on statistical power and confidence interval of effect size. *Japanese Journal of Educational Psychology, 59,* 385–401.