Summary-statistics-based power analysis: a new and practical method to determine sample size for mixed-effects modelling

Article

Accepted Version

Murayama, Kou, Usami, Satoshi and Sakaki, Michiko ORCID logoORCID: https://orcid.org/0000-0003-1993-5765 (2022)
Summary-statistics-based power analysis: a new and practical method to determine sample size for mixed-effects modelling. Psychological Methods, 27 (6). pp. 1014-1038. ISSN 1082-989X doi: https://doi.org/10.1037/met0000330 Available at https://centaur.reading.ac.uk/100388/

Running Head: POWER AND MIXED-EFFECTS MODELLING

Summary-statistics-based power analysis: A new and practical method to determine sample size for mixed-effects modelling

Kou Murayama[1, 2, 3], Satoshi Usami[4], Michiko Sakaki[1, 3]

1. Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Germany
2. School of Psychology and Clinical Language Sciences, University of Reading, UK
3. Research Institute, Kochi University of Technology
4. Graduate School of Education, University of Tokyo, Japan

Accepted by *Psychological Methods* on 17/9/2021

Author footnotes

Correspondence concerning this article should be addressed to Kou Murayama, Hector Research Institute of Education Sciences and Psychology, Europastraße 6, 72072 Tübingen, Germany. E-mail: k.murayama@uni-tuebingen.de. This research was supported by JSPS KAKENHI (Grant Numbers 18H01102; 18K18696 to KM), F. J. McGuigan Early Career Investigator Prize from American Psychological Foundation (to KM); Jacobs Foundation Advanced Research Fellowship (to KM), the Leverhulme Trust Research Leadership Award (Grant Number RL-2016-030, to KM), and the Alexander von Humboldt Foundation (the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research; to K. M.). We thank Cristina Pascua Martin for helping literature review and Method Interest Group at the University of Tübingen for critical discussion on the manuscript. The original draft of the manuscript is posted in OSF preprints (https://osf.io/6cer3/). The script for the illustrative example is also posted to OSF (https://osf.io/d4mub/).

Abstract

This article proposes a summary-statistics-based power analysis --- a practical method for conducting power analysis for mixed-effects modelling with two-level nested data (for both binary and continuous predictors), complementing the existing formula-based and simulation-based methods. The proposed method bases its logic on conditional equivalence of the summary-statistics approach and mixed-effects modelling, paring back the power analysis for mixed-effects modelling to that for a simpler statistical analysis (e.g., one-sample *t* test). Accordingly, the proposed method allows us to conduct power analysis for mixed-effects modelling using popular software such as *G*Power* or the *pwr* package in *R* and, with minimum input from relevant prior work (e.g., *t* value). We provide analytic proof and a series of statistical simulations to show the validity and robustness of the summary-statistics-based power analysis and show illustrative examples with real published work. We also developed a web app (https://koumurayama.shinyapps.io/summary_statistics_based_power/) to facilitate the utility of the proposed method. While the proposed method has limited flexibilities compared to the existing methods in terms of the models and designs that can be appropriately handled, it provides a convenient alternative for applied researchers when there is limited information to conduct power analysis.

Key words: Mixed regression; random-effects models; hierarchical linear model; multilevel-modelling; summary-measures approach.

Translational abstract

For applied researchers, statistical power analysis with mixed-effects modelling (or multilevel modelling) poses a big challenge, because it requires substantive expertise on modelling, use of special software, and a number of input parameters which are usually not available in published work. In fact, despite the number of research papers on this topic, we found that applied researchers rarely use appropriate statistical power analysis recommended by experts. To improve the current state of practice, this article proposes an easy and practical method to conduct statistical power analysis for mixed-effects modelling, called summary-statistics-based power analysis. While the proposed method has limited flexibilities (e.g., it can be applied only to two-level nested data), it has greater advantages over traditional power-analysis methods (formula- and simulation-based power analysis) in terms of usability and practicality. In fact, the proposed method can determine appropriate level-2 sample size of a new study by using only a *t* value and level-2 sample size from a previous study. Moreover, the sample size calculation can be easily conducted using popular software such as *G*Power* or the *pwr* package in *R*. A series of statistical simulations demonstrate the validity and robustness of the summary-statistics-based power analysis. We also provide illustrative examples with real published work. To further ease the demand for applied researchers, we developed a web app (https://koumurayama.shinyapps.io/summary_statistics_based_power/) that automatically performs the proposed method.

With the proliferation of diverse research methodologies (e.g., multi-site data, intensive longitudinal data), the need for mixed-effects modelling (also called "multilevel modelling" or "hierarchical linear modelling) is ever growing in psychology. This class of methods effectively deals with dependency of data (e.g., nested data), allowing researchers to flexibly model the effects of different levels of predictors (Goldstein, 2010; Hox, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012).

One of the major practical challenges in mixed-effects modelling for applied researchers is sample size determination, which is typically carried out with a power analysis. With simpler statistical analysis such as a *t*-test, power analysis can be performed by using popular software such as *G*Power* (Faul, Erdfelder, Lang, & Buchner, 2007) or the *pwr* package in *R* (Champely, 2018). To conduct power analysis for mixed-effects modelling, on the other hand, there are generally two classes of methods. *Formula-based power analysis* applies analytic formulae to calculate statistical power (e.g., Raudenbush & Liu, 2000; Snijders & Bosker, 1993). As the formulae are normally complicated (except for very simplified situations), researchers often need to use specialized statistical packages (e.g., PINT; Bosker, Snijders, & Guldemond, 2003). The second method, *simulation-based power analysis*, computes statistical power of mixed-effects modelling based on Monte-Carlo simulation. This normally requires a specialized package (e.g., *simr* package in *R*; Green & MacLeod, 2016; MLPowSim; Browne, Lahi, & Parker, 2009; clusterPower; Reich et al., 2012), or researchers can also write their own customized code using general purpose statistical language (e.g., *R*).

Importantly, however, both formula- and simulation-based power analyses require many intricate parameters which are specific to mixed-effects modelling. In fact, one of the key challenges of power analysis in mixed-effects modelling lies in the fact that statistical power depends on many parameters that are often hard to expect. For example, Multilevel Power Tool (Mathieu, Aguinis, Culpepper, & Chen, 2012; https://aguinis.shinyapps.io/ml_power/), an online app that computes statistical power of a cross-level interaction effect in multilevel data (based on simulation), requires 12 inputs: average level 1 sample size, level 2 sample size, intraclass correlation for the independent variable, six fixed effects, and variance components of intercept, slope, and within-cluster residuals. PINT, a power calculation program of two-level multilevel models based on an analytic formula (Bosker, Snijders, & Guldemond, 2003), requires the variance/covariance matrix of predictors separately at the within- and between-cluster levels. Even in a simple situation such as a clustered randomized design with balanced data, researchers need to know variance components in addition to effect size (Raudenbush & Liu, 2000). When researchers want to conduct a power analysis based on a previous study on a similar topic, these parameters may not be reported. Even if researchers have access to the entire data (e.g., researchers use their own pilot data or open data available online), understanding, computing, and entering these parameters into software requires extensive care and expertise.

Given the complication, it is not surprising that researchers seem to avoid statistical power analysis of mixed-effects modelling in empirical literature. We performed a literature review for the papers published in *Psychological Science*, which introduced a "*Research Disclosure Statements*" section in January, 2014, mandating authors to provide the justification of their sample size, including statistical power analysis (Eich, 2013). With this policy, researchers are now required to provide justification for the sample size in the method section. We collected papers in *Psychological Science* that conducted mixed-effects modelling (hierarchical linear modelling or multilevel modelling; some researchers used different names such as random-effects regression) as the main part of the analysis, and examined whether there was any mention of statistical power

analysis. We focused on the papers published after January 2015 as we thought it would take time for Research Disclosure Statements (effective since January 2014) to consolidate in published papers. Of the 135 papers published until August 2019 that used mixed-effects modelling as the primary analysis [1], only 4.4% of the papers conducted a proper statistical power analysis considering the dependency of the data with specialized software (*simr, Mplus*, etc.; none of the studies used a customized code). 1.5% referred to past work on the power analysis of mixed-effects modelling (e.g., Brysbaert & Stevens, 2018) to justify sample size. Other 20.0% of the papers reported quantitative statistical power analysis but authors did not mention how they considered the dependency of data (e.g., reported a "standard" power analysis based on Cohen's *d*, *r*, etc.). The rest of the papers (74.1%) did not report (or refer to) any quantitative power analysis --- 17.8% of the papers mentioned past similar work to justify the sample size (e.g., the sample size is similar to or bigger than the relevant work in the past) and 56.3% of the papers provided only obscure reasons (e.g., the research group tried to recruit as many participants as possible by the end of the term) or even little justifications. These numbers are quite similar to those reported by Arend and Schäfer (2019), who examined reporting practice of statistical power analysis of mixed-effects modelling in other major journals of psychology between 2000 and 2016 (i.e. *Journal of Abnormal Psychology, Journal of Applied Psychology, Journal of Educational Psychology, and Journal of Personality and Social Psychology)*. They showed that only 4.8% of authors performed a quantitative computation of power considering the dependency of the data. 20.5% mentioned power but it was based on informed guesses without quantitative estimation. 64.8% did not mention statistical power at all.

The purpose of the current article is to provide a new, practical power analysis method for mixed-effects modelling of nested data, called *summary-statistics-based power analysis*. The proposed method is complementary to the existing methods of power analysis (i.e. formula- and simulation-based power analysis) with its own strengths and weaknesses. Importantly, the proposed method can effectively address the above-mentioned challenge for applied researchers, potentially facilitating the use of power analysis for mixed-effects modelling. Our idea is based on the fact that, under certain conditions, *t* values (or *p* values) computed in mixed-effects modelling are equivalent to or approximated to *t* (or *p*) values obtained using a one sample *t* test or correlation analysis of the aggregated data of the same dataset (called the "summary-statistics approach"). This means that researchers can conduct a power analysis of mixed-effects modelling using popular software (e.g., *G*Power*, *pwr* package in *R*) by treating these *t* values reported in a previous study using mixed-effects modelling as if they were the output of a one-sample *t* test or a correlation analysis.

In the following, we first provide an illustrative example to demonstrate the equivalence of mixed-effects modelling and the one-sample *t* test of aggregated data under certain conditions. We then delineate the basic logic of the proposed summary-statistics-based power analysis, followed by a comparison of the proposed method with formula- and simulation-based power analyses. Afterwards we evaluate the robustness of the proposed method with statistical simulations. Finally, we show empirical examples and development of a web app which makes it easier for applied researchers to implement the proposed summary-statistics-based power analysis (https://koumurayama.shinyapps.io/summary_statistics_based_power/).

Conditional Equivalence of Mixed-effects Modelling and a *t* test on Aggregated Data

[1] Some papers reported multiple studies within a paper. We included a paper if any of the studies in the paper used mixed-effects modelling as the primary analysis.

Mixed-effects modelling is often regarded as the gold standard to handle multilevel (nested) data, and the literature tends to emphasize the advantage of mixed-effects modelling over simpler statistical analysis on aggregated data. In fact, unlike statistical analysis of aggregated data, mixed-effects modelling makes full use of the information in the nested data all at once, and therefore, it is reasonable to posit that mixed-effects modelling is better informed than the analysis on aggregated data. However, it is not well recognized that mixed-effects modelling produces approximately identical results to a simpler statistical test (e.g., one-sample *t*-test) on aggregated data under certain conditions.

To illustrate this, think about a hypothetical dataset in which 10 data points (Level 1; L1) are nested within 12 participants (Level 2; L2). For each participant, 5 data points are from a control condition (coded as -0.5) and 5 data points are from an experimental condition (coded as 0.5). Overall means for the experimental and control condition are 19.2 (*SD* = 3.99) and 17.9 (*SD* = 3.89), respectively (*SDs* were computed across all data points). The data and code are available on the Open Science Framework (https://osf.io/d4mub/). This example treats participants as level-2 but the following logic applies to any type of two-level nested data such as participants nested within study sites (e.g., schools, companies).

Because the data are clustered, we apply mixed-effects modelling to the data, using the experimental condition as the independent variable (predictor). Specifically, we examine the following model.

$$y_{ij} = (\gamma_{00} + u_{0j}) + (\gamma_{10} + u_{1j})x_{ij} + e_{ij} \qquad (1)$$

where y_{ij} is the dependent variable of i th data point of j th participant, u_{0j} and u_{1j} are random intercepts and slopes of participants, and γ_{00} is the overall intercept and γ_{10} is the overall slope of experimental effect. x_{ij} is a L1 (effect coded) independent variable representing the experimental condition of the i th data point of the j th participant. We assume that the focal level-1 predictor has a random slope --- this is an important condition to prevent potential inflation of Type-1 error rate (Barr, Levy, Scheepers, & Tily, 2013; Brauer & Curtin, 2018). e_{ij} is a random error. We assume that errors are normally distributed and independent from each other: $e_{ij} \sim N(0, \sigma^2)$, where σ^2 is error (or, within-cluster) variance, meaning that random errors have a simple structure (e.g., there is no autocorrelation within participants). Random intercepts and slopes follow a multivariate normal distribution: $\mathrm{Var}(u_{0j}) = \tau_{00}$ (random intercept variance), $\mathrm{Var}(u_{1j}) = \tau_{11}$ (random slope variance) and $\mathrm{Cov}(u_{0j}, u_{1j}) = \tau_{10}$ (covariance between random intercept and random slope). The primary effect of interest is the L1 effect γ_{10} (i.e. the effect of the experimental condition). Normally, parameter estimates and standard errors are computed using an iterative procedure. However, when (1) cluster size and (2) variance of the independent variable are identical across clusters (which is the case in this example), an analytic formula to compute the standard error of the γ_{10} is available as follows (Snijders, 2001, 2005):

$$SE(\hat{\gamma}_{10}) = \sqrt{\frac{\tau_{11} + \frac{\sigma^2}{nS_x^2}}{J}}. \qquad (2)$$

Here, J = level-2 (L2) sample size, and n = level-1 (L1) sample size per cluster (or cluster size). Therefore, the total sample = Jn., S^2_x is the within-cluster variance of x_{ij} and in this example, S^2_x = 0.25. t value for γ_{10} can be computed by dividing the parameter estimates of γ_{10} by the standard

error.

When we apply the model (1) to the data (using *lme4* package in R with restricted maximum likelihood estimation), the results showed that the estimated γ_{10} (i.e. fixed effect of intervention) was 1.25, which corresponds to the raw mean difference between the conditions, with t (11) = 1.32 and p = 0.21.

Alternatively, you can aggregate the data at the participant level (i.e. compute the means of the intervention and control conditions for each participant) and conduct a paired-samples *t* test (J = 12) to examine the effect of the experimental condition. This paired-samples *t* test is equivalent to a one-sample *t* test using difference scores between the conditions. This is a common approach in experimental psychology or neuroscience literature in which trials are nested within participants --- researchers frequently compute the mean for each participant per condition and compare the conditions using a *t*-test or ANOVA. This analytic approach to aggregated data is often called by-participant analysis (when level-2 unit is participants; e.g., Murayama, Sakaki, Yan, & Smith, 2014), two-step procedure (Achen, 2005), or *summary-statistics* (or *summary-measures*) approach (Ahn, Heo, & Zhang, 2015; Frison & Pocock, 1992; Matthews, Altman, Campbell, & Royston, 1990).

Critically, the *t* test with summary-statistics approach shows exactly the same results, t (11) = 1.32 and p = 0.21, indicating that these two analytic approaches are equivalent. In fact, when the two aforementioned conditions are met (i.e. cluster size and the variance of x are constant across clusters), mixed-effects modelling and summary-statistics approach are mathematically equivalent. We show the mathematical proof in Appendix[2]. Note also that, in this specific example, the equivalence holds (i.e. the results stay the same) even if we used a method that adjusts degrees of freedom and/or standard errors for small sample size such as the Kenward-Roger correction (Kenward & Roger, 1997).

The above example focused on the case of comparing two conditions, i.e. the independent variable was binary. What if the focal independent variable is continuous? Here we assume that independent variables are all centered within clusters, which is a common and appropriate option to dissociate level-1 from level-2 effects (Enders & Tofighi, 2007). Mixed-effects modelling takes into account the fact that regression slopes are different across clusters, and estimates a single overall within-cluster regression slope as well as the variance of the slopes between clusters. If the overall regression coefficient is statistically significant, researchers will conclude that the independent variable is related to the dependent variable.

A parallel summary-statistics approach is to run a regression analysis for each cluster --- if researchers have 12 clusters, for example, they run 12 regression analyses predicting the dependent variable from the independent variable. They can then test whether the average of the regression coefficients is different from zero using a one-sample *t* test. This analysis has been common in experimental psychology, in which researchers frequently run trial-level regression or correlation analysis for each participant (Lorch & Myers, 1990; Monin & Oppenheimer, 2005; Murayama et al., 2014). Similarly, in neuroimaging analysis, researchers often apply a linear regression model (more precisely, a general linear model) to individual data to estimate regression coefficients for each participant, which are then subject to a one-sample *t* test to examine whether the coefficients are significantly different from zero across participants (Monti, 2011). In fact, this is the procedure

[2] Some studies demonstrated that mixed-effects modelling and mixed-effects analysis of variance (ANOVA) are mathematically equivalent (Barr et al., 2013). But these studies consider ANOVA applied to hierarchical (long-format) data themselves, not the aggregated data. This is not the equivalence of mixed-effects modelling and summary-statistics approach as described here.

currently adopted by Statistical Parametric Mapping (SPM), widely used neuroimaging software.

Importantly, even when x_{ij} is a continuous variable, as demonstrated mathematically in Appendix, Equation (2) is asymptotically true, and mixed-effects modelling and the summary-statistics approach produce almost identical *t* values under the same conditions (i.e., cluster size and the within-cluster variance of *x* are identical across clusters). Note that this equivalence is further generalized to a situation in which there are other (non-orthogonal) L1 predictors with random slopes (see Appendix for proof).

Summary-statistics-based Power Analysis for Sample-size Determination with Mixed-effects Modelling

The close relation of the test statistics between the mixed-effects modelling and summary-statistics approaches provides us with a simple approach to compute statistical power: Summary-statistics-based power analysis. Imagine that there is a published study or a pilot study that researchers conducted, and they want to determine a L2 sample size (e.g., number of participants in case of a multi-trial experimental study or a diary study) for their new study based on that information. They are interested in a level-1 fixed-effect (γ_{10}) in Equation (1). With our proposed approach, the researchers simply need the *t* value (or *p* value) associated with this fixed effect (γ_{10}) and the L2 sample size *J* from the previous study to compute an approximate statistical power. This information is so basic as to be readily available from most published papers[3] and yet sufficient to conduct a power analysis for that fixed-effect.

The summary-statistics-based power analysis is summarized in Table 1. Specifically, once researchers obtain the information on the *t*-value and the L2 sample size, they would need to convert the *t* value to Cohen's *d as if* the *t* value were obtained from a one-sample *t* test with the sample size = *J* (Step 1). This Cohen's *d* can be then used to conduct a power analysis of one-sample *t* test with *G*Power*, *pwr*, or other software (Step 2; it is also possible to use an available formula to determine a sample size with one-sample *t* test). Put simply, they can run a power analysis by regarding the L1 statistical test in mixed-effects modelling as a one-sample *t* test. The fundamental idea is that statistical power analysis for mixed-effects modelling can be substituted with that of the summary-statistics approach, because the two approaches provide equivalent statistical test results when the conditions are met[4]. This procedure is effective even when there is more than one L1 predictor.

The proposed summary-statistics-based power analysis is primarily meant to conduct a power analysis based on the information from relevant prior work. It is possible to conduct a power analysis based on a pre-defined effect size (in that case, researchers should simply skip Step 1 in Table 1), such as minimal clinically important difference (Jaeschke et al., 1989), but this approach should be used with caution, which we will detail in the General Discussion.

Importantly, while using prior work is one of the common procedures for power analysis,

---

[3] We can also compute the *t* value even if the original study only reported the beta value and its standard error; we simply need to divide the beta by the standard error to obtain a *t* value.

[4] Mathematically speaking, the sufficiency of the *t* value to conduct a power analysis for L2 sample size stems from the fact that the *t* value is completely proportional to $\sqrt{J}$ as shown in Equation (2). This means that, once we obtain a *t* value from prior work, we can update standard error estimates and the non-centrality parameter by changing L2 sample size *J* without knowing the other parameters (i.e. γ_{10}, τ_{10}, σ^2, *n*, and s_x^2). Then we can use an existing formula (e.g., Cohen, 1988) to calculate a desired sample size to achieve a specific power given a pre-defined Type-1 error rate α. This alternative method allows researchers to directly conduct power analysis, bypassing the summary-statistics approach as described in Table 1. However, we prefer taking the proposed steps in Table 1 because these steps provide applied researchers with (1) a clearer understanding of what they are doing, and (2) more accessibility through the use of existing easy-to-use software for power analysis.

applied researchers should keep in mind its potential problems and measures to remedy them. First, power analysis based on prior work assumes that there is a high degree of similarity between the previous study and future study. When researchers aim to do a replication study or a pilot study, this is not a big issue. But when researchers plan sample size by looking at the past relevant work, the degree of similarity between the past work and planned work (design, population, etc.) has influence on the validity of sample size estimates. There are many sources of divergence (e.g., differences in study design, included predictors, study population, etc.) and to the extent that the prior work is dissimilar with the planned study, power estimates would become inaccurate. If researchers feel that previous studies are not similar enough, it is best to conduct a pilot study with the exact procedure. This sounds costly but it can save researchers from potential interpretive difficulty when the main study does not produce a statistically significant effect.

The second issue is that the results from the past work include uncertainty (i.e. sampling errors). To address the issue, various approaches have been proposed (e.g., Du & Wang 2016; McShane & Böckenholt, 2016; Pek & Park, 2019). One simple method is a safeguard power, in which researchers are encouraged to use the lower boundary of the effect size confidence interval (e.g., 60% CI) as a protection against the potential underestimation of the effect size (Perugini et al., 2014). Summary-statistics-based power analysis can easily incorporate this idea. Specifically, sampling error of the *t* value computed from a past study always follows a non-central *t* distribution regardless of the sample size and sample variance of the study. As such, instead of computing the effect size confidence interval, researchers can conduct a safeguard power analysis by deriving a confidence interval of the *t* value. For example, if a previous study reported a *t* value of 2.40 (df = 60), we can suppose that the sampling distribution of the *t* value follows a non-central *t* distribution with the non-centrality parameter $\delta = 2.40$ ($df = 60$), without knowing any other outputs from the study. Its 60% CI is [1.56, 3.29], and as such, researchers can use $t\ (60) = 1.56$ as input to calculate appropriate sample size based on the proposed method.

The third issue is publication bias. Publication bias generally overestimates effect size (and associated test statistics) and therefore, conducting a power analysis based on prior work may risk underestimating sample size. To address the issue, others provided a likelihood function of the non-centrality parameter (i.e. *t* value) when publication bias exists (Anderson et al., 2017; Taylor & Muller, 1996). By applying the formula to the non-centrality parameter obtained in a prior study, researchers can obtain an adjusted non-centrality parameter that accounts for the uncertainty and publication bias. There is a Web App (https://designingexperiments.com/shiny-r-web-apps/) and an *R* package (*BUCSS*; Anderson & Kelley, 2020) that automatically compute the adjusted non-centrality parameter. This correction method fits well with the proposed summary-statistics-based power analysis in mixed-effects modelling. Specifically, researchers can simply enter the non-centrality parameter (i.e. *t* value) of mixed-effects modelling results from a previous study to compute an adjusted non-centrality parameter using the software, and then use the obtained value as the input of the summary-statistics-based power analysis to plan an appropriate sample size. To our knowledge, the issue of publication bias in power analysis has not been sufficiently discussed in the literature of mixed-effects modelling but the method described here provides a possible solution.

More Complex Models

We have thus far focused on a simple model in which there are only L1 predictors. Here we extend the idea, considering a more complex model in which there are Level-1 predictors, Level-2 predictors, and cross-level interactions. Let's first consider the following full model:

$$y_{ij} = (\gamma_{00} + \gamma_{01} w_j + u_{0j}) + (\gamma_{10} + \gamma_{11} w_j + u_{1j}) x_{ij} + e_{ij} \quad (3)$$

γ_{10} is the overall slope of the level-1 predictor x_{ij} (L1 effect), γ_{01} is the main effect of the level-2 predictor w_j (L2 effect), and γ_{11} is the cross-level interaction between x_{ij} and w_j (L12 effect). Again, as customary, we assume that L1 predictor x_{ij} is centered within clusters and has random slopes u_{1j}.

In this full model, L2 effect is similar to a regression predicting cluster-averaged scores of the dependent variable from the L2 predictor. Likewise, L12 effect is conceptually similar to a regression predicting within-cluster slopes from the L2 predictor. As such, we may treat the statistical test of L2 and L12 effects in mixed-effects modelling as similar to the test of regression coefficients using clusters as the unit of analysis. This logic suggests that, for L1 as well as L2 and L12 effects, the summary-statistics approach is still possible with a full model specified in Equation (3). Note that the same conceptual idea applies even if there is more than one L2 and L12 effect --- in this case, the focal L2 effect or L12 effect can be seen as a partial regression coefficient after controlling for the other L2 or L12 effects. The statistical test for the partial regression coefficient is essentially a test of the regression coefficient with residuals after explaining the other variables. As such, we can still consider the focal L2 or L12 effect as stemming from regression, except that we need to adjust the degrees of freedom based on the number of the other L2 or L12 effects (see Algina & Olejnik, 2003).

What about the L1 effect? One important difference from the previous simple case is that, the interpretation of L1 effect γ_{10} is dependent on the L2 predictor W_j. That is, as Equation (3) indicated, γ_{10} is the effect of L1 predictor when L2 predictor W_j is zero. Therefore, with the summary statistics approach, a statistical test for L1 effect is essentially equivalent to the statistical test of the intercept term obtained from the regression analysis predicting within-cluster slopes from the L2 predictor (i.e. the regression to estimate L12 effect). The same logic applies when there is more than one L1/L2 predictor. This means that we can still treat the statistical test of L1 effect as similar to a one-sample *t* test across clusters, although we need to adjust for the degrees of freedom based on the number of the cross-level interactions related to this L1 effect.

Under the same conditions of invariant cluster size and within-cluster variance of the L1 predictor, we can analytically derive approximate formulae to compute the standard errors of these fixed effects (L1 effect, L2 effect, and L12 effect), and we can prove that the summary-statistics approach described above would give approximately identical test statistics to those obtained from mixed-effects modelling. This is the case even when there is more than one L1 and/or L2 predictor (under the condition of invariant cluster size and within-cluster variance/covariance matrix of the L1 predictors). Appendix shows these formulae and the proof of equivalence.

Summary-statistics-based power analysis with complex models

Given the equivalence of mixed-effects modelling and the summary-statistics approach in the complex model, we can extend the proposed method to plan L2 sample size based on past work for such complex models (Table 1). Again, the fundamental idea is that we can substitute the power analysis of mixed-effects modelling with that of the summary-statistics approach based on well-known effect size metrics (i.e. Cohen's *d*, *r*). For the power analysis of L1 effect, the procedure is the same: We should simply regard the *t* value from the past work as that from a one-sample *t* test. Then we can convert the *t* value into Cohen's *d* (Step 1), with which we can conduct a power analysis using a standard software such as *G*Power* or *pwr* or available formula (Step 2). The single difference is that we need to adjust for the number of the cross-level interactions that involve the L1 predictor of interest (Step 1 and Step 3; see Table 1 for precise procedure).

For the power analysis of L2 effect, we first consider the *t* value of the mixed-effects modelling from the past work *as if* it were obtained from a test of simple correlation (or partial correlation in case of multiple L2 effects) between the L2 predictor and cluster-average scores of the dependent variable. Note that we use the correlation metric, rather than the regression coefficient, to quantify the relationship. Power analysis for correlation and that for regression produce identical or almost identical results and as such, choice of the metric has little impact on the actual power calculation. However, correlation has more intuitive metric than unstandardized regression coefficient, and is easier to conduct power analysis with available software. In fact, based on the available formula, we can then easily convert the *t* value (along with the L2 sample size information) back to correlation *r* (Step 1). Then we can simply use the *r* value as effect size input to conduct a power analysis of correlation using standard software or available formula to determine L2 sample size for a new study (Step 2). Similarly, power analysis of L12 effect starts with the idea that the *t* value of the mixed-effects modelling can be regarded *as if* it were obtained from a test of simple correlation between the L2 predictor and within-cluster slopes. Using the same formula, we can easily convert the *t* value (along with the L2 sample size information) to correlation *r* (Step 1). Then we can use *r* as effect size input to conduct a power analysis of correlation using standard software or the available formula (Step 2). In both cases, it is more accurate to adjust the number of the L2 or L12 predictors (Step 1 and Step 3; see Table 1 for precise procedure).

When Determining Cluster Size in Addition to L2 Sample Size

The proposed approach provides researchers with a relatively easy and accessible way to determine L2 sample size based on past work. However, the estimated L2 sample size with this approach is deemed valid only when cluster size is the same between the past study and the planned study (or when the planned cluster size is larger than the past study; in such a case, the proposed method will provide a conservative sample size). This is because changing cluster size is likely to change the stability of within-cluster regression coefficients, resulting in the change of the effect size Cohen's *d*. This limitation compromises the potential value of the summary-statistics-based power analysis.

However, there is a way to expand the summary-statistics based power analysis to address this issue. Consider a simple model as described in Equation (1) in the first place. As indicated in Equation (2), *t* value is the function of (1) fixed effect γ_{10}, (2) cluster size (*n*), (3) number of clusters (*J*), (4) random slope variance (τ_{11}), and (5) the ratio of the within-cluster variance and the observed variance of the independent variable (σ^2/s_x^2). When there is more than one L1 predictor, the last element (σ^2/s_x^2) is substituted by $\sigma^2/s_x^2(1-R_x^2)$, where R_x^2 is the R^2 when the focal L1 predictor is regressed upon the other L1 predictors (see Appendix for the formula). Therefore, in addition to *t* value and L2 sample sizes, if estimated random slope variance (τ_{11}) and some basic information (estimated L1 fixed-effect γ_{10} and cluster size *n*) are available, we can compute σ^2/s_x^2 (or $\sigma^2/s_x^2(1-R_x^2)$, when there is more than one level-1 predictor). Then, we can investigate the impact of changing cluster size (*n*) on the *t* value using the same equation. This procedure allows us to estimate a new adjusted *t* value and associated Cohen's *d* *if* cluster size had been different in the prior work. Based on this new Cohen's *d*, we can conduct a power analysis using the available software or formula. In essence, with the additional information (e.g., τ_{11}), we can estimate the most inaccessible information in published work (σ^2/s_x^2 or $\sigma^2/s_x^2(1-R_x^2)$) from the *t* value, allowing us to conduct statistical power analysis in a relatively easy manner[5].

[5] Once we obtain the formula to compute effect size from the combination of L2 sample size and cluster size, it would

The same strategy applies to a complex model as described in Equation (3), and expanded models that have additional level-1 and/or level-2 predictors to Equation (3). Even with such complex models, we can estimate the most inaccessible information σ^2/s_x^2 or $\sigma^2(1-R_x^2)/s_x^2$ based on some additional information. According to the analytic formula derived in Appendix, if the focal effect is the L1 effect, like the simple model we discussed above, the additional information we need is only random slope variance (τ_{11}) as well as some basic information (i.e. estimated focal fixed effect γ_{10} and cluster size n) from the prior work. When the focal effect is the L2 effect or L12 effect, we (unfortunately) need further information: the variance of the focal level-2 predictor (S_W^2), and in case where there is more than one level-2 predictor, the proportion of the variance in the focal level-2 predictor explained by the other level-2 predictors (R_W^2). The information is obtainable if previous work reported descriptive statistics and correlation matrix of the level-2 predictors.

For the purpose of illustration, we arbitrarily generated a single dummy dataset based on Equation (3) (cluster size = 32; L2 sample size = 30), which we regarded as data obtained from hypothetical past work, and drew power curves for L1, L2, and L12 effects using both summary-statistics-based and simulation-based power analysis. For simulation-based power-analysis, power curves were drawn by simulating the statistical power for each combination of the cluster size and L2 sample size (with *simr* package in R; replication = 5,000 for each combination). The bold green line in Figure 1 (cluster size = 32) shows a power curve based on the proposed approach (solid line) and simulation-based approach (dotted line) when cluster size is the same as the hypothetical prior work (32). Other power curves concern the situation when the cluster size is different from the hypothetical prior work (thus summary-statistics-based power analysis requires more information to draw these curves as noted above). The graph indicates that both summary-statistics-based and simulation-based power analysis produce very similar power curves.

Although summary-statistics-based power analysis allows us to determine cluster size with additional information, the workaround described here is cumbersome or error-prone given the additional computations. To ease the use of the summary-statistics-based power analysis in such cases, we developed an online app. We will illustrate the app with real data examples later (see "Online App and Illustrative Examples" section).

Strengths and Weaknesses: Comparison of Three Power analysis Methods

Summary-statistics-based power analysis is not meant to replace the existing formula- and simulation-based power analysis. Rather, they have complementary strengths and weakness, and different statistical assumptions. Table 2 summarizes the comparison of the three classes of power analysis methods for mixed-effects modelling.

Practical aspects

All of the three power analysis methods (i.e. formula-, simulation-, and summary-statistics-based power analyses) can be used to determine sample size using the information from similar previous work. On the other hand, some researchers may want to conduct power analysis with a pre-defined effect size. In this situation, the summary-statistics-based method can still compute statistical power. However, its capacity is limited to the determination of L2 sample size only (see General Discussion), while traditional methods would be more comprehensive (although they require more input).

Formula- and simulation-based power analyses require researchers to use software

---

also be possible to work out an optimal design which maximizes the statistical power given a budgetary constraint (Rutterford et al., 2015; van Breukelen & Candel, 2012).

specialized for power analysis in multi-level modelling (e.g., *PINT* or *simr*). Alternatively researchers can write a customized code to run simulations. On the other hand, summary-statistics-based power analysis can be conducted by standard power analysis software that many researchers are familiar with such as *G*Power* or *pwr* package in *R*. When researchers are interested in determining cluster size, computational steps before using these software are cumbersome (see the section "When Determining Cluster Size in Addition to L2 Sample Size") but we developed a web app to automate the process (introduced later). In general, summary-statistics-based power analysis can be conducted with a smaller number of input values, in comparison to the other methods.

One critical limitation of the summary-statistics-based power analysis is that they can only handle a relatively simple mixed-effects model: two-level linear model with nested data. For example, the summary-statistics-based power analysis cannot be applied to a model with categorical or ordinal outcomes (i.e. generalized linear mixed-effects modelling). Although we believe that the proposed method can still give a rough estimate of statistical power, its performance should be empirically evaluated. Summary-statistics-based power analysis cannot handle the case in which there is more than one crossed random effect (cross-classified model). In fact, in such a case, *t* value from prior work does not contain sufficient information to calculate statistical power with varying sample size. Formula-based power analysis has slightly better flexibility. For example, when independent variables are binary and data are balanced, researchers can conduct formula-based power analysis for cross-classification models (e.g., Westfall, Kenny, & Judd, 2014) and three-level models (e.g., Dong & Maynard, 2013). Simulation-based power analysis is most flexible in this regard, and can run power analysis for all of the existing mixed-effects models in theory, although researchers may need to write a customized simulation script in some situations.

There is a hidden practical cost of the simulation-based power analysis: computational time. To perform a simulation-based power analysis, researchers need to run numerous replications to obtain a power value. To draw a power curve to find an appropriate sample size, this process needs to be repeated by systematically changing sample sizes, each of which could take a few hours. If researchers want to test different combinations of parameters, the same simulation needs to be repeated. On the other hand, both formula- and summary-statistics-based power analyses can produce a power curve within a second.

Statistical Assumptions

Summary-statistics-based power analysis is essentially a version of formula-based power analysis. Therefore, these two methods have similar statistical assumptions, and if the assumptions are met, they provide precise power estimates. We already elucidated some of these statistical assumptions when explaining the equivalence of the summary-statistics approach and mixed-effects modelling, but here we elaborate on the assumptions in more detail.

First, both formula-based and summary-statistics-based power analyses assume constant cluster size across clusters (although some work on formula-based power analysis provides formulae to adjust sample size in case of unequal cluster size; e.g., Manatunga et al., 2001). When researchers expect heterogeneous cluster size across clusters, researchers may use the expected average or harmonic-mean of cluster sizes as a proxy. Previous simulation studies showed that formula-based power analysis is reasonably robust for the violation of this assumption, unless cluster size is extremely heterogeneous (Candel et al., 2008; van Breukelen et al., 2007). In a similar manner, our simulation in a later section also showed that the summary-statistics-based power analysis is quite robust for such a violation.

Second, formula- and summary-statistics-based power analyses can deal with continuous

predictors but assume that predictors have equal variance across clusters. To our knowledge, we are not aware of statistical simulation studies that directly addressed the robustness of the formula-based power analysis against the violation of this assumption when predictors are continuous. However, in the simulation we will conduct in a later section, we showed that summary-statistics-based power analysis is reasonably robust for the moderate violation of this assumption (and therefore we can expect that formula-based power analysis is also reasonably robust for the violation).

There are three related assumptions that are specific to summary-statistics-based power analysis. First, summary-statistics-based power analysis has an implicit assumption that the variance of independent variables is not directly related to cluster size. Although this is a reasonable assumption in most applied studies, there is one exception: research that includes time as a predictor variable, such as growth curve modelling (i.e. adding time points naturally changes the variance of the time predictor). In such models, if researchers want to determine an optimal number of time points (i.e. cluster size) rather than L2 sample size only, we can still apply the method but we should modify the procedure to take into account the change in variance of the time variable when adding or subtracting time points. Second, summary-statistics-based power analysis assumes that predictors are cluster-mean centered (or have equal means across clusters). Finally, summary-statistics-based power analysis is based on a formula that assumes that all L1 predictors have random slopes. That said, when random slopes are dropped due to the fact that slopes do not vary across clusters (most common reason for not including random slopes), summary-statistics-based power analysis should still give accurate power estimates.

On top of these assumptions, of course, standard assumptions for mixed-effects modelling apply to summary-statistics-based power analysis (e.g., normality of random effects and L1 errors, homoscedasticity, independence of L2 random effects and L1 errors). L1 errors are assumed to be independent (this is the assumption for *lme4* package in *R*). This is often violated when data are longitudinal and measurements are taken close together in time. Formula-based power analysis basically has the same standard assumptions especially for the models that can be dealt with using specialized software packages. Finally, both formula- and summary-statistics-based power analysis methods assume that researchers will analyze data using restricted maximum-likelihood (REML) estimation, which is normally preferrable to maximum-likelihood (ML) estimation due to the fact that ML potentially has larger underestimation bias in the standard errors estimates (McNeish & Stapleton, 2016).

Simulation-based power analysis, on the other hand, can loosen most of these assumptions in principle. This flexibility is a big advantage of simulation-based power analysis (Arend & Schäfer, 2019). On the other hand, this considerable flexibility also means that researchers need to make many educated decisions (with certain assumptions) --- researchers need to deliberately specify various information to run power analysis with such a complicated design or data. When researchers expect that cluster size will be different across clusters, for example, they need to decide how to generate (or which distribution should be used to generate) cluster sizes for each simulation. Similarly, if researchers plan to use continuous predictors, they need to decide the way to generate these predictors for each simulation. It should also be noted that, for some type of complex models, software availability may be limited. For example, *simr* in R cannot handle a model with a complex covariance structure of L1 errors.

With rare exceptions, mixed-effects models are known to underestimate standard errors when L2 sample size is small (McNeish & Stapleton, 2016). Therefore, both formula- and summary-statistics-based power analyses generally give somewhat overestimated statistical power,

when researchers plan to take into account small-sample bias using an existing correction method (e.g., Kenward-Roger correction method) in the main data analysis. In such cases, researchers may wish to oversample slightly to make up for inflated statistical power. Alternatively, researchers can use simulation-based-power analysis with a correction method. However, the Kenward-Roger correction method (one of the most popular and well-performing correction methods) sometimes requires considerable computation time (Kuznetsova, Brockhoff, & Christensen, 2017); therefore, this option may not be practical.

Evaluating the Robustness of the Summary-statistics-based Power Analysis

Like formula-based power analysis, the summary-statistics-based power analysis is asymptotically accurate when the statistical assumptions described above are met. In reality, however, these assumptions are rarely met. Here we evaluate the robustness of the proposed power analysis by focusing on the two assumptions --- constant within-cluster variance of predictor and constant cluster size. Although other statistical assumptions are also important, we focused on these two assumptions because these are the major assumptions that are not routinely assumed in mixed-effects modelling and are often violated in real data. To evaluate the robustness, we compared the summary-statistics-based power analysis with simulation-based power analysis which took into account the violations of these assumptions. We used a complex model in Equation (3) so that we can evaluate the robustness of the proposed method in a relatively broad context.

Simulation 1

Simulation 1 Method. We first simulated 500 datasets from a data generation model based on Equation (3) with cluster size n = 32 and L2 sample size J = 30, and regarded the simulated datasets as 500 separate pilot test datasets. We then applied mixed-effects modelling to each of the (hypothetical) pilot datasets. Next, based on the obtained results, we estimated a power curve for L1, L2, and L12 effects respectively, once using the summary-statistics-based method and again using a simulation-based method (with the *simr* package in *R*, replication = 1,000; Satterthwaite approximation was used to define degrees of freedom for the mixed-effects model). For the summary-statistics-based power analysis, we only used the t value obtained from the mixed-effects modelling and L2 sample size J = 30 as input to obtain a power curve when cluster size n = 32. To obtain power estimates with different cluster sizes, the summary-statistics-based power analysis used further information specified in the previous section.

Given the excessive computation time of repeating the simulation method, in this simulation, we only looked at the estimated power of three different L2 sample sizes (J = 20, 30, and 40) combined with three different cluster sizes (n = 22, 32, and 42) for each hypothetical data, and used only one set of parameters to generate hypothetical datasets. Specifically, for the simulation to examine the comparability of the L1 effect: standardized L1 effect = 0.20; standardized L2 effect = 0.20; standardized L12 effect = 0.20; random intercept variance = 0.6; and random slope variance = 0.3. For the simulation to examine the comparability of L2 effect: standardized L1 effect = 0.20; standardized L2 effect = 0.45; standardized L12 effect = 0.20; random intercept variance = 0.3, random slope variance 0.6. For the simulation to examine the comparability of L12 effect: standardized L1 effect = 0.20; standardized L2 effect = 0.20; standardized L12 effect = 0.45; random intercept variance = 0.3; random slope variance = 0.3. For all models, error variance was set to 1.0 and correlation between the random intercept and slope was set to 0.3. These parameter values were deliberately chosen in order not to have a floor or ceiling effect of power estimates and to have visible change in power estimates as a function of sample size.

There were three different conditions for the data generation model. The first condition

generated data under the restrictions that cluster size and predictor variance are constant across clusters (n = 32, and predictor variance = 1). This condition serves as a baseline to evaluate the performance of the other conditions --- this condition theoretically produces the same power estimates across the two methods of power analysis. In the second condition, the independent variable x_{ij} in the hypothetical pilot dataset was generated from a normal distribution with mean = 0 and standard deviation = 1, meaning that the observed (not the population) variance of the independent variable is different across clusters. In the simulation-based power analysis, we need to specify the data points of the independent variable in the simulation model that evaluates statistical power. In this condition, simulation-based power analysis used the same normal distribution to define the data points of x_{ij} in the simulation model. In the third condition, cluster size was varied across clusters while the mean cluster size was still the same as the other conditions (i.e. n = 32). Specifically, cluster size for each cluster was varied --- 3 repeats of the following 10 numbers: 14, 18, 22, 26, 30, 34, 38, 42, 46, and 50. Again, simulation-based power analysis requires us to specify the individual cluster sizes in the simulation model when they are heterogeneous across clusters. By arbitrarily setting up cluster sizes of the hypothetical pilot dataset as 3 repeats of the same 10 numbers, we circumvented the issue. Specifically, when we estimated the power with L2 sample size = 20 or 40, we simply omitted or added these 10 numbers to specify the simulation model in the simulation-based power analysis. This means that we used an idealistic scenario to increase the precision of the simulation-based power analysis.

In sum, this simulation produced 500 (hypothetical pilot datasets) x 3 (effect in focus: L1, L2, or L12) x 3 (conditions) = 4,500 hypothetical pilot datasets, which were analyzed both by the summary-statistics- and simulation-based (replication = 1,000) power analyses to compare the results. For each hypothetical pilot dataset, both methods produced nine power estimates with three different levels of L2 sample sizes (J = 20, 30, and 40) combined with three different cluster sizes (n = 22, 32, and 42). Therefore, the simulation method ran 4500 x 1000 x 3 x 3 = 40,500,000 models to obtain power estimates. The simulation was run on the Laboratory of Neuroimaging (LONI) Pipeline system provided by the University of Southern California Mark and Mary Stevens Neuroimaging and Informatics Institute (Rex, Ma, & Toga, 2003).

To evaluate the similarity of power estimates from these two methods, we computed (1) the average power for each method and (2) the root-mean square error (RMSE) defined as follows:

$$RMSE = \sqrt{\frac{\sum^{R} (Power_{summary} - Power_{simulation})^2}{R}}$$

R denotes the number of pilot data generated (i.e. 500) and $Power_{summary}$ and $Power_{simulation}$ are the power estimates based on the summary-statistics- and simulation-based power analyses.

Simulation 1 Results. The results are summarized in Table 3. Overall, summary-statistics-based power analysis produced comparable results with the simulation-based power analysis. In the baseline condition, where cluster size and predictor variance are constant across clusters, summary-statistics-based power analysis should show very similar power estimates with simulation-based power analysis. In fact, we saw little discrepancy between the proposed and simulation-based methods in the average power and RMSEs (0.009 – 0.019). Given the general equivalence of the two approaches, the observed discrepancy is likely due to sampling errors in the simulation method. In fact, with 1,000 replications of the simulation-based power analysis, the standard error of a proportion estimate is maximally 0.016 (when true proportion is 0.50).

Critically, the results were robust even under conditions where predictor variance or cluster size was varied across clusters. On average, both methods showed very similar power estimates, indicating that the proposed method is not biased in a systematic manner. RMSEs are mostly below 0.025, although the discrepancy was slightly larger when both L2 sample size cluster sizes were small and cluster size was varied (maximum = 0.030). In comparison to the baseline condition, RMSEs became larger only by 0 – 0.012. The robustness of the summary-statistics-based power analysis is consistent with the previous observations that cluster size heterogeneity does not have much impact on the statistical power estimates (e.g., van Breukelen et al., 2007).

Simulation 2

Simulation 2 Method. The first set of simulations showed that summary-statistics-based power analysis is robust to violation of the two critical assumptions: constant within-cluster variance of predictors and constant cluster size. However, the simulation was limited in that it focused only on a fixed set of parameters to generate hypothetical pilot datasets. This is due to the considerable computational demands of simulation-based power analysis, which make it practically infeasible to run a similar simulation by factorially combining different parameter values.

To examine the effects of the different parameter values on power accuracy while minimizing the computational demand, we conducted the second simulation using the following procedure. Specifically, for each hypothetical pilot data generation, we randomly sampled the parameters of the data generation model from the following range (using a uniform distribution): For the models testing L1 effect, L2 sample size J = [20, 48]; cluster size n = [20, 60]; standardized L1 effect = [0.10, 0.40]; standardized L2 effect = [0.20, 0.50]; standardized L12 effect = [0.20, 0.50]; random intercept variance = [0.6, 1.2]; random slope variance = [0.5, 1.0]; correlation between random intercepts and random slopes = [0, 0.5]. Error variance was fixed to 1.0. For the models testing L2 effect and those testing L12 effect, the ranges were almost the same except for the random slope variance = [0.3, 0.6]. These ranges were chosen deliberately so that overall statistical power would not hit the floor or ceiling in the majority of cases. We generated 1,000 hypothetical pilot datasets and then applied summary-statistics- and simulation-based (replication = 1,000) power analyses, estimating the power of three different L2 sample sizes (J - 10, J, and J + 10, where J is the L2 sample size of the hypothetical data) combined with three different cluster sizes (n – 10, n, and n + 10, where n is the cluster size of the hypothetical data). This procedure resulted in 9,000 (1,000 x 3 x 3) pairs of power estimates for L1 effect from the two methods with varying parameter values. An additional 9,000 pairs of power estimates were obtained for L2 effect, and another 9,000 pairs of power estimates were obtained for L12 effect.

We then regressed the difference of the power estimates between summary-statistics- and simulation-based power analyses onto the parameter values. This analysis reveals the impacts of parameters in the data generation model on the discrepancy of power estimates in the two power-analysis methods. We examined two types of power estimate differences. The first type is the relative difference ($Power_{summary}$ - $Power_{simulation}$), which allows us to investigate the factors that contribute to a systematic overestimation or underestimation of power. The second type is absolute difference ($|Power_{summary} - Power_{simulation}|$), which allows us to examine the factors that contribute to absolute deviation of the two power estimates. This index is conceptually similar to RMSE.

The independent variables of the regression analysis were (1) L2 sample size of the hypothetical (original) dataset (J), (2) cluster size of the hypothetical (original) dataset (n), (3) random intercept variance, (4) random slope variance, (5) standardized L1 effect, (6) standardized L2 effect, (7) standardized L12 effect, (8) correlation between random intercepts and random

slopes, (9) L2 sample size to estimate power *relative to* the original sample size (-10: J – 10, 0: J, +10: J + 10), (10) cluster size to estimate power *relative to* the original sample size (-10: n-10, 0: n, +10: n +10). All the independent variables were centered so that the intercept represented the overall discrepancy of power estimates between summary-statistics- and simulation-based power analyses.

Simulation 2 had two conditions: the condition of heterogeneous within-cluster variance of the predictor, and the condition of heterogeneous cluster size. The within-cluster variance of the predictor was determined in the same way as the previous simulation. For the heterogeneity of cluster size, we tried to increase the variation of cluster size in comparison to the first simulation. Variation of cluster size is often quantified using a coefficient of cluster size variation (CsV), which is computed by dividing the standard deviation of cluster sizes by mean cluster size. In the Simulation 1, heterogeneous cluster size condition had the CsV of 0.32 in the generated pilot data. In the current simulation, we further tried to increase the CsV by sampling individual cluster sizes from a uniform distribution with a relatively wide range (minimum cluster size = 4): In the simulated hypothetical pilot dataset, the average CsV was 0.56. To determine the individual cluster size to estimate power in the simulation-based method, we used the same approach so that the distribution of the cluster size would become as similar as possible between the hypothetical pilot dataset and simulated models.

Simulation 2 Results. The regression results are summarized in Table 4. To quantify the stability of parameter estimates, we also showed 95% CI (we corrected for the standard errors to take into account the fact that the same hypothetical data produced 9 pairs of power estimates). In short, the results confirmed the robustness of the summary-statistics-based power analysis for both conditions. For L1 effect, overall discrepancy was very small (for relative power difference, -0.0002 and -0.0009; for absolute power difference, 0.0095 and 0.0098). For both relative difference and absolute difference, none of the predictors had strong effects on the discrepancy. For example, L1 effect size showed a negative effect of -0.0039 on the absolute difference in power in the variant cluster size condition. This means that the absolute power difference between the two power analysis methods increased by 0.00039 when (standardized) L1 effect size is smaller by 0.1.

For L2 effect, overall discrepancy was very small (for relative difference, -0.0008 and -0.0013; for absolute difference, 0.0117 and 0.0124). None of the predictors had strong effects, although the effects seem to be slightly larger than the L1 effect overall. For example, random intercept SD had a positive effect of β = 0.0125 on relative power difference in the varied cluster variance condition. This means that summary-statistics-based power analysis tended to overestimate the statistical power by 0.00125 when random intercept SD increased by 0.1 (i.e. as large as the SD of errors).

For L12 effect, again, overall discrepancy was very small (for relative power difference, 0.0004 and -0.0004; for absolute power difference, 0.0119 and 0.0146). Most of the predictors did not have strong effects. The biggest effect seems to be that of the standardized L12 effect on relative power difference, which showed β = -0.0543 in the varied cluster size condition. This suggests that summary-statistics-based power analysis tends to underestimate power by 0.00543 in comparison to the simulation-based method when standardized L12 effect size increases by 0.1.

Simulation 3

Simulation 3 Method. In the last simulation, we evaluated the performance of the proposed method in a similar manner to Simulation 1 (i.e. fixed set of parameters), but with a relatively adverse condition. Specifically, we ran a set of simulations under a condition in which cluster sizes

were vastly different, predictor variance was varied across clusters, and L2 sample size to estimate power was set to low values (10, 15, and 20).

The simulation procedure was almost identical to Simulation 1 with the following changes. First, we simulated 1,000 (instead of 500) hypothetical datasets to evaluate the performance of the proposed method more accurately (the number of replications for the simulation-based power analysis stayed the same, i.e. 1,000). Second, we reduced the L2 sample size of the hypothetical dataset from 30 to 15, and estimated power for three different L2 sample sizes: 10, 15 and 20. Third, we created and focused only on a condition in which both cluster size and within-cluster predictor variance were varied across clusters. Fourth, we used the same procedure as Simulation 2 to determine cluster sizes of the hypothetical datasets, and cluster sizes to estimate power, with the aim to increase the variation of cluster size. The average CsV in the hypothetical pilot datasets was 0.55.

Finally, we slightly changed the parameters of the data generation model (for hypothetical datasets) to obtain statistical power estimates that were not close to the boundary (i.e. 0.00 and 1.00). This was necessary, as the reduction of L2 sample size inevitably decreases overall statistical power. More specifically, the following values were used. For the simulation to examine the comparability of the L1 effect: standardized L1 effect = 0.30; standardized L2 effect = 0.20; standardized L12 effect = 0.20; random intercept variance = 0.6; and random slope variance = 0.3. For the simulation to examine the comparability of L2 effect: standardized L1 effect = 0.20; standardized L2 effect = 0.6; standardized L12 effect = 0.20; random intercept variance = 0.6; random slope variance = 0.3. For the simulation to examine the comparability of L12 effect: standardized L1 effect = 0.20; standardized L2 effect = 0.20; standardized L12 effect = 0.6; random intercept variance = 0.3; random slope variance = 0.3. Other parameter values were identical to those used in Simulation 1.

Simulation 3 Results. Results are reported in Table 5. Overall, despite the adverse condition, the proposed method showed reasonably good performance across different combinations of L2 sample size and cluster size, generally indicating the robustness of the summary-statistics-based power analysis. Of particular interest was the case when L2 sample size was small (i.e. 10). Our two major observations were as follows. First, on average, the proposed method showed very similar statistical power to the simulation-based power analysis. Second, when L2 sample size was small, RMSE increased up to 0.034 (for L1 effect) to 0.048 (for L12 effect); This was a ~0.02-0.03 increase from the baseline model in Simulation 1. Together, these results demonstrate the robustness of the summary-statistics-based power analysis, even with strong violations of the assumption of equal cluster size, and with the modest violation of the assumption of equal predictor variance. The results also suggest that, when dealing with small sample size data, researchers need to bear in mind the possibility that a power estimate based on the proposed method may not be very accurate (i.e. it is better to oversample to be on a safe side).

In sum, these simulation studies demonstrate the reasonable robustness of the summary-statistics-based power analysis even when critical statistical assumptions are violated. The discrepancy between the proposed and simulation-based methods do not seem to be strongly impacted by the nature of the prior data (i.e., true parameters that define hypothetical pilot data). Even under the conditions of unequal within-cluster variance of predictor and cluster size (up to CsV = 0.55-0.56), power estimates by the proposed method did not systematically differ from those by the simulation-based method on average, although RMSE became slightly higher when sample size was small.

Online App and Illustrative Examples

Applied researchers can use summary-statistics-based power analysis just by following Table 1. When determining cluster size as well as L2 sample size, the computation is cumbersome. To ease the use of the summary-statistics-based power analysis, we developed an online app (Figure 2; https://koumurayama.shinyapps.io/summary_statistics_based_power/) by which researchers can easily conduct a priori power analysis by entering necessary information taken from pilot or prior work. The app allows researchers to conduct L2 sample size determination for L1, L2, and L12 effects by entering a *t* value and L2 sample size. The app also automatically adjusts for degrees of freedom when there is more than one predictor (Table 1). Critically, if additional input is available (e.g., τ_{11}), the app can conduct power analysis with different cluster sizes (not varying across clusters). Below, we illustrate how we can conduct a priori power analysis of mixed-effects modelling using the app based on two empirical papers.

Hackel and Zaki (2018)

The study examined whether reciprocity was influenced by the wealth of the target using an economic game. The procedure is rather complicated but in short, at the final stage of the experiment, participants decided whether they wanted to share their endowment with other players who interacted with the participants ("givers"). Some of the givers were wealthier than others. Eighty-seven participants made a decision for 20 trials, 10 times for wealthy and 10 times for less wealthy givers (these givers were different across participants). The givers' players also exhibited different levels of generosity at an early stage of the experiment, which served as another continuous predictor variable. To test their hypothesis, they fitted a mixed-effects model with trials as level-1 and participants as level-2, predicting percentage shared with the giver on each trial as a function of giver wealth (higher wealth = 1, lower wealth = −1), giver generosity (centred within clusters), and their interaction. Consistent with their hypothesis, giver wealth significantly influenced the shared percentage, $b = 0.036$, $t\ (77.17) = 5.40$, $p < .01$, suggesting that wealthy givers were more likely to be reciprocated.

Given the lack of the detailed information in the paper, such as intraclass correlation, it is challenging to perform power analysis using existing software. However, the summary-statistics-based power analysis can easily compute a desired sample size if researchers were to conduct a similar study with the same number of trials. The computed Cohen's *d* for the L1 effect of the giver wealth based on summary statistics is relatively large, $d = 0.58$. Based on this effect size, the app indicated an appropriate sample size of 26 to achieve 80% statistical power. It is also easy to do the power analysis that takes into account the uncertainty of the *t* value reported in the study and publication bias by additionally using BUCSS package in *R* (Anderson & Kelley, 2020), which we mentioned earlier. Here, we conducted another power analysis with assurance = 80% and with the assumption that there was a publication bias (i.e. studies with $p > .05$ had not been published). Assurance is the percentage of times the sample size planning approach will be successful in reaching or exceeding the intended level of statistical power, taking into consideration of the uncertainty (i.e. sampling error) of previous results. Entering $t = 5.40$ and $N = 87$ as input (into the function to compute power based on a one-sample *t* test), BUCSS showed that the *t* value should be corrected to 4.469. Using this corrected value as an input, our app produces power curve and indicated an appropriate sample size of 37 to achieve 80% statistical power (Figure 2).

Eckerlein et al. (2019)

The study administered a learning diary to university students in which they responded to a set of questions about the upcoming psychology exam every day during the 14-day exam preparation period. Overall, data from 115 participants were analysed, with the average assessments per participant = 10.5 ($SD = 3.0$, range = 3 -14). The main dependent variable was the

daily report of invested effort ("I tried especially hard today"; ICC = 0.18) and the main independent variable was daily report of motivational difficulties in the learning process ("Today I struggled to keep my study motivation on a high level"). They also assessed the quantity of motivation regulation and quality of motivation regulation as individual differences (i.e. L2 variables) in a pre-test. One of the main hypotheses was that the negative within-person relationship between motivational difficulties and invested effort would be moderated by the quantity and quality of motivation regulation: the negative within-person relationship would be weaker for those who had high quantity and/or quality of motivation regulation (i.e. a positive cross-level interaction effect).

They conducted mixed-effects modelling in which daily invested effort was predicted by daily motivational difficulties and time of assessment at the within-person level. Random slopes were specified for both predictors. They further included the quantity of motivation regulation and quality of motivation regulation as two simultaneous L2 predictors, both to predict the random intercept (L2 effect) and the random slopes of the daily motivational difficulties (L12 effect). All the variables were *z*-standardized before the analysis. Partially supporting their hypothesis, there was a significant positive cross-level interaction effect between motivational difficulties and quality of motivational regulation, $\beta = 0.07$ (SE = 0.03), $p < .05$. On the other hand, quantity of motivation regulation did not significantly predict the random slope, $\beta = 0.01$ (SE = 0.03).

The information is sufficient to conduct summary-statistics-based power analysis for a new study that aims to test a similar cross-level interaction effect. Specifically, using the L2 sample size (115), the estimated *t* value from SE (0.07/0.03 = 2.33), and the number of cross-level interactions related to the focal L1 effect (2), the web app indicated that L2 sample size = 168 is required to achieve the statistical power of 80% (effect size $r = 0.22$), provided that average cluster size stays the same (i.e. 10.5).

More importantly, the study also reported the random slope variance of the effects of motivational difficulties (0.05) and the correlation between the two L2 variables ($r = 0.49$). Entering the information (i.e. $\tau_{11} = 0.05$, $R_W^2 = 0.49^2 = 0.2401$, and $S_W^2 = 1$ as all variables were standardised) as well as the average cluster size (10.5) as additional input, summary-statistics-based power analysis now allows us to compute statistical power as a function of both L2 sample size and cluster size. For example, if one were to force all students to complete the learning diary every day (i.e. cluster size = 14), the required sample size would reduce to 153 to achieve the statistical power of 80%. A reproduced power curve as a function of L2 sample size and cluster size based on the entered information is presented in Figure 3.

Note that the output from the last analysis includes an *adjusted t* value = 2.445 and its degrees of freedom = 112. The adjusted *t* value is a hypothetical *t* value if the cluster size were 14. This value is useful when a researcher wants to conduct power analysis for a different cluster size that takes into account the uncertainty of prior data or publication bias. For example, if a researcher wants to do a safeguard power analysis (Perugini et al., 2014) using the lower bound of the 60% confidence interval (this is conceptually similar to doing power analysis with assurance = 80%), the researcher should first compute the lower bound of the 60% confidence interval using the (non-central) *t* distribution with $df = 112$ and non-centrality parameter = 2.445. This is equal to $t = 1.602$. Then using the new *t* value, the researcher can draw a power curve with varying L2 sample size using the app again. In the current example, the required L2 sample size (with cluster size = 14) is 349 to achieve 80% of statistical power --- much higher than when not considering the uncertainty.

General Discussion

For applied researchers, currently-available methods for power analysis of mixed-effects modelling (i.e. formula- and simulation-based power analyses) can be complicated as they require substantial expertise and a lot of input information, which may not be readily available. The present manuscript proposed a simple and practically-useful alternative method, called summary-statistics-based power analysis for mixed-effects modelling of nested data. The proposed method only needs a *t* value and L2 sample size from a previous study or pilot data to plan L2 sample size for a new study. Even if one were to plan cluster size and L2 sample size altogether, summary-statistics-based power analysis requires minimal additional information from the previous study (e.g., τ_{11}). The proposed method can be implemented in popular power-analysis software (e.g., *G*Power*, *pwr*), and we also provided a web app to further ease the usage of the method (https://koumurayama.shinyapps.io/summary_statistics_based_power/). Our simulation results demonstrated that the summary-statistics-based power analysis is generally robust to violations of critical underlying assumptions (i.e., homogeneity of within-cluster variance and homogeneity of cluster size). While summary-statistics-based power analysis has less flexibility (e.g., only L2 sample size can be determined based on a pre-defined effect size) than formula- and simulation-based power analyses, it has a complementary benefit of usability and practicality, providing an attractive power analysis option for applied researchers.

Summary-statistics-based power analysis and the summary-statistics approach

As the name indicates, summary-statistics-based power analysis takes advantage of the fact that the summary-statistics approach (i.e. aggregating the 1st level by summary statistics before the 2nd level analysis) is mathematically equivalent to mixed-effects modelling under certain conditions. Summary-statistics approach has been recurrently discussed as an alternative to mixed-effects modelling (Achen, 2005; Austin, 2007; Dowding & Haufe, 2018; Feldman, 1988; Frison & Pocock, 1992; Lorch & Myers, 1990; Saxonhouse, 1976; Wishart, 1938). Random-effect meta-analysis can also be considered as a version of this summary-statistics approach (Borenstein, Hedges, Higgins, & Rothstein, 2008). However, there has been limited discussion on how this approach can be applied to statistical power analysis for mixed-effects modelling (for a similar idea in a limited context, see Lefante, 1990). The current manuscript is an attempt to make an explicit connection between these seemingly disjointed issues.

Considering the proposed power analysis method in the context of the summary-statistics approach further highlights its strengths and weaknesses. For example, recent studies have made note that applied researchers using mixed-effects modelling frequently face the issue of non-positive definite covariance matrices of random effects (Brauer & Curtin, 2018; McNeish & Bauer, 2020). This issue is often ignored in simulation-based power analysis --- in the simr package in R, for example, the program checks the statistical significance for each replicate regardless of whether estimates lie inside the parameter space. Mixed-effects models also run the risk of failing to obtain appropriate parameter estimates due to non-convergence (this is especially the case when sample size is small). On the other hand, the summary-statistics approach does not suffer from such issues because it does not directly estimate random effect components and does not require complicated estimation procedures. As such, in situations where simulation-based power analysis produces an overwhelming number of non-positive definite covariance matrices and/or non-convergence errors, summary-statistics-based power analysis may be a good alternative for researchers. Also, in the context of big data analysis, mixed-effects modelling may not be practically useful given the considerable computation time. Here again, the summary-statistics approach can be an attractive and efficient alternative given the little computational demand (see also Beckmann, Jenkinson, & Smith, 2003, for the utility of this method in the context of neuroimaging analaysis),

and the proposed method would be a practical solution to compute statistical power.

Of course, the summary-statistics approach has its own limitations, and these limitations reflect the limitations of the proposed power analysis method. One notable limitation is that the summary-statistics approach is accurate only when certain assumptions are met. As such, the proposed power analysis method is also accurate only when its assumptions are met. We already laid out the assumptions earlier and highlighted two critical assumptions --- constant within-cluster variance of predictors and constant cluster size. Our simulation studies, however, showed that summary-statistics-based power analysis is reasonably robust for the violation of these assumptions. Future studies should examine the robustness of the proposed method when there is an even more extreme violation of the assumptions. It is also worth noting that research on the summary-statistics approach has provided a solution to the violation of these two assumptions in the proposed power analysis. When cluster size is different across clusters, for example, researchers can integrate summary statistics by taking into account the sampling variability of the clusters (Achen, 2005; see also Goldberg et al., 2005 for "weighted *t*-statistics"). Dowding and Haufe (2018) called this method the *sufficient* summary-statistics approach. This fact suggests that it is theoretically possible to extend the proposed summary-statistics-based power analysis to analytically derive appropriate sample size when these assumptions are not met. We have not explored this possibility in the current manuscript, but future studies should benefit from considering such an extension.

Using the proposed power analysis with a pre-defined effect size

There are situations in which researchers want to use a pre-defined effect size to conduct power analysis, rather than relying on prior relevant data. Although summary-statistics-based power analysis is primarily developed to conduct power analysis with prior work, it also allows researchers to use a pre-defined effect size as input when researchers are interested in L2 sample size. This is because the proposed method computes Cohen's *d* (defined in relation to one-sample *t* test) or *r* in the course of estimating statistical power. Specifically, instead of estimating Cohen's *d* or *r* by using the output from prior work (e.g., *t* value), researchers can simply set a pre-defined Cohen's *d* or *r* based on a theoretical or practical consideration (new Step 1), and then conduct a priori power analysis as if planning to conduct a one-sample *t* test or correlation analysis based on these effect sizes (Steps 2 and 3 in Table 1) using *G*Power* or *pwr* package in *R*. Note that a similar procedure can be used when researchers want to show the minimum effect size that can be detected given specific power and L2 sample size values (sensitivity analysis).

This approach seems simple and attractive. However, there are two (related) critical limitations that researchers should bear in mind. Specifically, (1) this approach can only provide an L2 sample size to achieve a specified power, and (2) the effect sizes Cohen's *d* or *r* used in the proposed approach are the effect sizes of the *observed* (not the true) relationship, which can be influenced by cluster size. To see the reason why, and to appropriately use this approach, we need to precisely understand the meaning of effect size measures used in summary-statistics-based power analysis. For example, Cohen's *d* that we use in the summary-statistics-based power analysis is the effect size of the averaged regression slopes considering the observed between-cluster differences of the slopes (e.g., individual differences in the slopes among participants). Both γ_{10} (L1 fixed-effect slope) and *d* are effect size estimates, but the meaning is substantially different: While γ_{10} represents the effect size of *within*-cluster relations (i.e. expected within-cluster change when the independent variable increases by one), Cohen's *d* evaluates the magnitude of γ_{10} against observed *between*-cluster differences. Even when the overall within-cluster regression slope γ_{10} is small, if there are little observed between-cluster differences in the

slopes, Cohen's *d* can be substantially large. Similarly, even when the overall within-cluster regression slope γ_{10} is large, if there are large observed between-cluster differences in the slopes, Cohen's *d* can be small. In other words, Cohen's *d* in the summary-statistics-based power analysis can be interpreted as representing the generalizability of the observed effect across clusters (e.g., generalizability across participants), rather than the strength of the association between the level-1 predictor and the dependent variable. Therefore, when defining a Cohen's *d*, researchers should ask "how stable the observed L1 effect would be across clusters?" rather than "how strong L1 effect is within a cluster?".

Importantly, Cohen's *d* defined in this context concerns the between-cluster stability of the *observed* L1 effect, which is a function of both true between-cluster and within-cluster variances. For example, when there are only a few trials in an experiment, even if the *true* within-person regression slope is invariant across participants (i.e. no true between-cluster variance: large stability across clusters), we still observe substantial individual differences of within-person regression slopes due to trial-level sampling errors. In that case, Cohen's *d* becomes smaller. In other words, Cohen's *d* in the summary-statistics-based power analysis implicitly takes into account the cluster size, and that is why this approach can be applied only to determine L2 sample size. This issue is the same for the power analysis of L2 and L12 effects. Unlike L1 effect, the interpretation of these effects is relatively straightforward and consistent with the corresponding effect sizes (γ_{01} and γ_{11}): the magnitude of the association between the L2 predictor and the observed intercepts or slopes. However, the magnitude of the association is dependent on the cluster size. In fact, when cluster size is small, the magnitude of the association would be attenuated due to L1 sampling errors. Therefore, *r* in the summary-statistics-based power analysis also implicitly accounts for the cluster size of data. Therefore, we recommend that researchers use this approach only when cluster size is large, such that it is relatively easy to calibrate an effect size of the *observed* summary statistics (i.e. cluster average scores)[6].

Practical Guideline

Summarizing all the discussions made in the manuscript, we provide the following practical guidelines on when we should (and should not) use summary-statistics-based power analysis.

1. Summary-statistics-based power analysis should only be used when the planned model is within the scope of the method as described in Table 2. For example, the proposed method cannot be used when one plans to use a three-level multilevel model or a model with time being a predictor (e.g., growth curve models). Although the proposed method is based on the assumption of equal within-cluster predictor variance and cluster size across clusters, our statistical simulation showed that the method is reasonably robust to violations of these assumptions.
2. When using a pre-defined effect size to do a power analysis (or a sensitivity analysis), the proposed power analysis should be used only when it is relatively easy to interpret effect size based on cluster-average scores (e.g., cluster size is sufficiently large).
3. When using prior relevant work as input for a power analysis, researchers should ensure that the prior work has sufficient similarity with the planned new study in the first place. When necessary, researchers should also use the previously-developed correction

[6] It is also important to emphasize that this does not mean that Cohen's *d* or *r* used in the proposed method is meaningless --- our argument is that we need to accurately interpret the metric. For example, in the illustrative example we discussed at the beginning, it is natural for researchers to use summary-statistics approach to analyze the data, and then Cohen's *d* is the effect size metric that should be reported. Indeed, it is common that researchers compute Cohen's *d* for aggregated data; our discussion reveals the nuances for this commonly-reported effect size metric.

methods to counter potential uncertainty and publication bias in the prior data.

Concluding remark

Applied researchers often consider mixed-effects modelling as overly complicated and difficult to understand. This class of analysis is indeed complicated and requires substantial expertise to appropriately analyze complex data. Accordingly, power analysis for mixed-effects modelling is also often regarded as complicated. At the same time, researchers tend to overlook the fact that such advanced statistical methods are actually built upon, and sometimes even reduced to, a set of simpler analyses that we are more familiar with. We believe that conceptualizing mixed-effects modelling this way (i.e. summary-statistics approach) would considerably help applied researchers understand the complicated modelling at a deeper level, and can even provide insights into various issues that are commonly deemed challenging. We are hoping that the current article provides one good case for such an argument --- a demonstration of how this "simple" perspective promotes our understanding and usage of power analysis in mixed-effects modelling.

References

Achen, C. H. (2005). Two-Step Hierarchical Estimation: Beyond Regression Analysis. Political Analysis, 13(4), 447–456. https://doi.org/10.1093/pan/mpi033

Ahn, C., Heo, M., & Zhang, S. (2015). *Sample size calculations for clustered and longitudinal outcomes in clinical research.* New York: Chapman and Hall/CRC.

Algina, J., & Olejnik, S. (2003). Sample Size Tables for Correlation Analysis with Applications in Partial Correlation and Multiple Regression Analysis. Multivariate Behavioral Research, 38(3), 309-323. doi:10.1207/S15327906MBR3803_02

Anderson, S. F. & Kelley, K. (2020). BUCSS: Bias and Uncertainty Corrected Sample Size. R package version 1.2.1. https://CRAN.R-project.org/package=BUCSS

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychological Science*, *28*(11), 1547–1562. https://doi.org/10.1177/0956797617723724

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods, 24*(1), 1-19. doi:10.1037/met0000195

Austin, P. C. (2007). A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. *Statistics in Medicine*, *26*(19), 3550–3565. https://doi.org/10.1002/sim.2813

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 10.1016/j.jml.2012.1011.1001. doi:10.1016/j.jml.2012.11.001

Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *Neuroimage, 20*(2), 1052-1063. doi:https://doi.org/10.1016/S1053-8119(03)00435-X

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2008). Comprehensive Meta-analysis (Version 2.2.048) [Computer software].

Bosker, R. J., Snijders, T. A. B., & Guldemond, H. (2003). *PINT version 2.1*. Retrieved from https://www.stats.ox.ac.uk/~snijders/multilevel.htm#progPINT

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods, 23*(3), 389-411. doi:10.1037/met0000159

Browne, W. J., Lahi, M. G., & Parker, R. M. A. (2009). *A guide to sample size calculations for random effect models via simulation and the MLPowSim software package*. Retrieved from https://seis.bristol.ac.uk/~frwjb/esrc/MLPOWSIMmanual.pdf

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of cognition, 1*(1), 9-9. doi:10.5334/joc.10

Candel, M. J. J. M., Van Breukelen, G. J. P., Kotova, L., & Berger, M. P. F. (2008). Optimality of equal vs. unequal cluster sizes in multilevel intervention studies: A Monte Carlo study for small sample sizes. *Communications in Statistics - Simulation and Computation, 37(1),* 222–239. https://doi.org/10.1080/03610910701724052

Champely, S. (2018). *pwr: Basic Functions for Power Analysis*. R. package version 1.2-2. https://CRAN.R-project.org/package=pwr

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

Dong, N., & Maynard, R. (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*(1), 24–67. https://doi.org/10.1080/19345747.2012.673143

Dowding, I., & Haufe, S. (2018). Powerful Statistical Inference for Nested Data Using Sufficient Summary Statistics. *Frontiers in Human Neuroscience, 12*(103). doi:10.3389/fnhum.2018.00103

Du, H., & Wang, L. (2016). A Bayesian Power Analysis Procedure Considering Uncertainty in Effect Size Estimates from a Meta-analysis. *Multivariate Behavioral Research, 51(5)*, 589–605. https://doi.org/10.1080/00273171.2016.1191324

Eckerlein, N., Roth, A., Engelschalk, T., Steuer, G., Schmitz, B., & Dresel, M. (2019) The role of motivational regulation in exam preparation: Results from a standardized diary study. *Frontiers in Psychology*, 10:81. doi: 10.3389/fpsyg.2019.00081

Eich, E. (2013). Business Not as Usual. *Psychological Science*, *25*(1), 3-6. doi:10.1177/0956797613512465

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121-138. doi:10.1037/1082-989x.12.2.121

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*(2), 175-191. doi:10.3758/BF03193146

Feldman, H. A. (1988). Families of lines: Random effects in linear regression analysis. *Journal of Applied Physiology*, *64*(4), 1721–1732. https://doi.org/10.1152/jappl.1988.64.4.1721

Frison, L., & Pocock, S. J. (1992). Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Statistics in medicine, 11*(13), 1685-1704. doi:10.1002/sim.4780111304

Goldberg, L. R., Kercheval, A. N., & Lee, K. (2005). T ‐ statistics for weighted means in credit risk modeling. *The Journal of Risk Finance*, *6*(4), 349–365. https://doi.org/10.1108/15265940510613688

Goldstein, H. (2010). *Multilevel Statistical Models, 4th Edition*. Chichester, UK: Wiley.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493-498. doi:10.1111/2041-210X.12504

Hackel, L. M., & Zaki, J. (2018). Propagation of Economic Inequality Through Reciprocity and Reputation. *Psychological Science*, *29*(4), 604-613. doi:10.1177/0956797617741720

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications, 2nd ed*. New York, NY, US: Routledge/Taylor & Francis Group.

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials, 10(4)*, 407–415. https://doi.org/10.1016/0197-2456(89)90005-6

Kenward, M. G., & Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics, 53(3)*, 983–997. https://doi.org/10.2307/2533558

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). *lmerTest* Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82(13)*. https://doi.org/10.18637/jss.v082.i13

Lefante, J. J. (1990). The power to detect differences in average rates of change in longitudinal studies. *Statistics in Medicine, 9*(4), 437–446. https://doi.org/10.1002/sim.4780090414

Littell, R. C. (2002). Analysis of unbalanced mixed model data: A case study comparison of ANOVA versus REML/GLS. *Journal of Agricultural, Biological, and Environmental Statistics, 7*(4), 472. https://doi.org/10.1198/108571102816

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 149-157. doi:http://dx.doi.org/10.1037/0278-7393.16.1.149

Manatunga, A. K., Hudgens, M. G., & Chen, S. (2001). Sample Size Estimation in Cluster Randomized Studies with Varying Cluster Size. *Biometrical Journal*, *43*(1), 75–86. https://doi.org/10.1002/1521-4036(200102)43:1<75::AID-BIMJ75>3.0.CO;2-N

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology, 97*(5), 951-966. doi:10.1037/a0028380

Matthews, J. N., Altman, D. G., Campbell, M. J., & Royston, P. (1990). Analysis of serial measurements in medical research. *BMJ (Clinical research ed.), 300*(6719), 230-235. doi:10.1136/bmj.300.6719.230

McNeish, D., & Bauer, D. J. (2020). Reducing Incidence of Nonpositive Definite Covariance Matrices in Mixed Effect Models. *Multivariate Behavioral Research*, 1–23. https://doi.org/10.1080/00273171.2020.1830019

McNeish, D., & Stapleton, L. (2016). The Effect of Small Sample Size on Two Level Model Estimates: A Review and Illustration. *Educational Psychology Review, 28*. https://doi.org/10.1007/s10648-014-9287-x

McShane, B. B., & Böckenholt, U. (2016). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods, 21(1),* 47–60. https://doi.org/10.1037/met0000036

Monin, B., & Oppenheimer, D. M. (2005). Correlated Averages vs. Averaged Correlations: Demonstrating the Warm Glow Heuristic Beyond Aggregation. *Social Cognition, 23*(3), 257-278. doi:10.1037/0022-3514.64.3.431

Monti, M. (2011). Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach. *Frontiers in Human Neuroscience, 5*(28). doi:10.3389/fnhum.2011.00028

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.

Pek, J., & Park, J. (2019). Complexities in power analysis: Quantifying uncertainties with a Bayesian-classical hybrid approach. *Psychological Methods, 24(5),* 590–605. https://doi.org/10.1037/met0000208

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard Power as a Protection Against Imprecise Power Estimates. *Perspectives on Psychological Science, 9*(3), 319–332. https://doi.org/10.1177/1745691614528519

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

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199-213. doi:10.1037/1082-989X.5.2.199

Rex, D. E., Ma, J. Q., & Toga, A. W. (2003). The LONI Pipeline Processing Environment. *Neuroimage*, *19*(3), 1033-1048. doi:https://doi.org/10.1016/S1053-8119(03)00185-X

Reich, N. G., Myers, J. A., Obeng, D., Milstone, A. M., & Perl, T. M. (2012). Empirical Power and Sample Size Calculations for Cluster-Randomized and Cluster-Randomized Crossover Studies. *PLOS ONE*, *7*(4), e35564. https://doi.org/10.1371/journal.pone.0035564

Rutterford, C., Copas, A., & Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, *44*(3), 1051–1067. https://doi.org/10.1093/ije/dyv113

Saxonhouse, G. R. (1976). Estimated Parameters as Dependent Variables. *The American Economic Review*, *66*(1), 178–183.

Snijders, T. A., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*(3), 237-259. doi:10.2307/1165134

Snijders, T. A. B. (2001). Sampling. In A. Leyland & H. Goldstein (Eds.), *Multilevel modelling of health statistics.* (pp. 159-174). New York: Wiley.

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (pp. 1570-1573). Chicester: Wiley.

Snijders, T. A., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*(3), 237–259. https://doi.org/10.2307/1165134

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and applied multilevel analysis (2nd edition)*. London: Sage.

Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics - Theory and Methods*, *25*(7), 1595–1610. https://doi.org/10.1080/03610929608831787

van Breukelen, G. J. P., & Candel, M. J. J. M. (2012). Calculating sample sizes for cluster randomized trials: We can keep it simple and efficient! *Journal of Clinical Epidemiology*, *65*(11), 1212–1218. https://doi.org/10.1016/j.jclinepi.2012.06.002

van Breukelen, G. J. P., Candel, M. J. J. M., & Berger, M. P. F. (2007). Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in medicine*, *26*(13), 2589-2603. doi:10.1002/sim.2740

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020-2045. doi:10.1037/xge0000014

Wishart, J. (1938). Growth-Rate Determinations in Nutrition Studies with the Bacon Pig, and Their Analysis. *Biometrika*, *30*(1/2), 16–28. https://doi.org/10.2307/2332221

Table 1 Summary-statistics-based power analysis to determine L2 sample size based on prior work or pilot study

	L1 effect	L2 effect	L12 cross-level interaction
Input from prior work/pilot	*t*, *J*, p_{L12}	*t*, *J*, p_{L2}	*t*, *J*, p_{L12}
Core idea	Regard *t* as being from a one-sample t test (with sample size = J - p_{L12})	Regard *t* as being from a correlation test (with sample size = $J - [p_{L2} - 1]$)	Regard *t* as being from a correlation test (with sample size = $J - [p_{L12} - 1]$)
Step 1	Convert *t* to Cohen's *d* $d = \frac{t}{\sqrt{J - p_{L12}}}$	Convert *t* to Pearson's correlation *r* $r = \sqrt{\frac{t^2}{J - p_{L2} - 1 + t^2}}$	Convert *t* to Pearson's correlation *r* $r = \sqrt{\frac{t^2}{J - p_{L12} - 1 + t^2}}$
Step 2 (with *G*Power*, *pwr*, etc.)	Input effect size *d* to conduct a power analysis of one-sample *t* test.	Input effect size *r* to conduct a power analysis of correlation.	Input effect size *r* to conduct a power analysis of correlation.
Step 3	Add p_{L12} to the required sample size	Add p_{L2} - 1 to the required sample size	Add p_{L12} - 1 to the required sample size
Description (Information obtained from the prior work/pilot)	*t* = *t* value of the effect of interest *J* = Number of L2 sample size p_{L2} = Number of L2 main effects (including the L2 predictor of interest) p_{L12} = Number of L12 cross-level interaction terms related to the L1 effect of interest (including the cross-level interaction effect of interest)		

Table 2 Comparison of three classes of power-analysis methods

	Formula based	Simulation based	Summary-statistics based
Power analysis based on previous similar work	Yes	Yes	Yes
Power analysis based on pre-defined effect size	Yes	Yes	Yes but only for L2 sample size determination.
Main Packages	Specialized software: e.g., *PINT* (Boskers et al., 2003), *Optimal Design* (Raudenbush, 1997; Spybrook et al., 2011)	Specialized software: e.g., *simr* (Green & MacLeod, 2016), *MLPowSim* (Browne et al., 2009), *clusterPower* (Reich et al., 2012)	Standard power analysis software: *G*Power* (Faul et al., 2007) *pwr* (Champely, 2018)
Required input	Extensive (e.g., all variance components, within-cluster predictor variance) and often not reported in previous work.	Extensive (e.g., all variance components, individual data points of predictors) and often not reported in previous work.	Minimum and normally reported in previous work (e.g., a *t* value). But additional input (e.g., τ_{11}) is needed when planning a cluster size.
Continuous predictors	Yes but limited software availability.	Yes but limited software availability.	Yes.
Assumptions of L1 predictors	In theory the method is not constrained by the assumptions of L1 predictors. But available software typically assumes constant within-cluster variance and mean-centering within clusters. The within-cluster variance needs to be specified.	There is no assumption of L1 predictor. Predictor values need to be simulated from a certain distribution that researchers find reasonable.	The method assumes constant within-cluster variance and cluster means of L1 predictor but the variance does not need to be specified. The model should include random slopes of L1 predictors when they exist.
Homogeneous cluster size	Assumed. If violated, average or harmonic mean may be used.	Not assumed. Individual cluster sizes need to be simulated from a certain distribution that researchers find reasonable.	Assumed. If violated, average or harmonic mean may be used.
Complex L1 error structure	Difficult to address with the existing software	Can be modelled, but software availability is limited.	Difficult to address.
Complex models (e.g., cross-classified model, three-level model)	Yes but only for limited designs	Yes but limited software availability.	No
Computational time	Little	Extensive	Little

Table 3 Power estimates computed by the summary-statistics- and simulation-based power analyses, when pilot datasets are simulated by a model specified by Equation (3) with L2 sample size = 30 and cluster size = 32 (Simulation 1).

		Invariant cluster size/predictor variance			Varied predictor variance			Varied cluster size		
		Average power			Average power			Average power		
L2 size	cl. size	Summary	Simulation	RMSE	Summary	Simulation	RMSE	Summary	Simulation	RMSE
Level-1 effect										
20	22	0.609	0.613	0.014	0.618	0.620	0.017	0.597	0.582	0.021
20	32	0.647	0.653	0.013	0.657	0.661	0.015	0.638	0.644	0.015
20	42	0.669	0.675	0.014	0.678	0.684	0.015	0.662	0.669	0.015
30	22	0.753	0.751	0.012	0.758	0.756	0.015	0.741	0.723	0.022
30	32	0.784	0.784	0.011	0.788	0.788	0.012	0.774	0.773	0.010
30	42	0.800	0.800	0.010	0.804	0.805	0.011	0.792	0.795	0.011
40	22	0.833	0.830	0.010	0.834	0.830	0.013	0.820	0.805	0.021
40	32	0.856	0.854	0.009	0.856	0.855	0.010	0.845	0.844	0.009
40	42	0.868	0.867	0.009	0.868	0.867	0.009	0.858	0.859	0.009
Level-2 effect										
20	22	0.422	0.427	0.019	0.436	0.436	0.017	0.425	0.417	0.023
20	32	0.457	0.460	0.018	0.472	0.474	0.019	0.462	0.466	0.021
20	42	0.478	0.480	0.018	0.494	0.496	0.018	0.483	0.489	0.020
30	22	0.557	0.562	0.016	0.574	0.576	0.016	0.560	0.552	0.023
30	32	0.593	0.595	0.016	0.611	0.610	0.016	0.597	0.601	0.018
30	42	0.614	0.614	0.016	0.631	0.631	0.016	0.618	0.622	0.017
40	22	0.649	0.652	0.014	0.667	0.667	0.014	0.651	0.642	0.022
40	32	0.683	0.685	0.013	0.700	0.700	0.014	0.685	0.686	0.016
40	42	0.701	0.700	0.014	0.718	0.717	0.013	0.703	0.705	0.014
Cross-level effect										
20	22	0.418	0.420	0.018	0.412	0.414	0.020	0.413	0.397	0.030
20	32	0.454	0.456	0.019	0.448	0.450	0.021	0.451	0.453	0.021
20	42	0.475	0.477	0.019	0.470	0.474	0.019	0.474	0.479	0.020
30	22	0.551	0.552	0.017	0.547	0.549	0.020	0.546	0.533	0.027

		Invariant cluster size/predictor variance			Varied predictor variance			Varied cluster size		
		Average power			Average power			Average power		
L2 size	cl. size	Summary	Simulation	RMSE	Summary	Simulation	RMSE	Summary	Simulation	RMSE
30	32	0.587	0.589	0.016	0.585	0.587	0.018	0.586	0.588	0.018
30	42	0.607	0.608	0.016	0.606	0.609	0.018	0.608	0.612	0.018
40	22	0.640	0.642	0.013	0.640	0.643	0.018	0.638	0.624	0.025
40	32	0.674	0.676	0.013	0.675	0.678	0.015	0.674	0.675	0.016
40	42	0.692	0.691	0.013	0.695	0.697	0.014	0.694	0.697	0.015

Note: cl. size = cluster size; Summary = Summary-statistics-based power analysis; Simulation = Simulation-based power analysis; RMSE = root-mean square error

Table 4 Relative and absolute difference of power estimates between the summary-statistics- and simulation-based power analyses regressed on true parameters to generate hypothetical pilot datasets (Simulation 2).

	Varied predictor variance				Varied cluster size			
	relative difference		absolute difference		relative difference		absolute difference	
Predictors	Beta	95% CI	Beta	95% CI	Beta	95% CI	Beta	95% CI
Level-1 effect								
Intercept	-0.0002	[-0.0005, 0.0000]	0.0095	[0.0093, 0.0097]	-0.0009	[-0.0012, -0.0006]	0.0098	[0.0096, 0.0101]
Original L2 N	0.0000	[0.0000, 0.0001]	0.0000	[-0.0001, 0.0000]	0.0000	[0.0000, 0.0001]	-0.0001	[-0.0001, 0.0000]
Original cluster size	0.0000	[-0.0001, 0.0000]	0.0000	[0.0000, 0.0000]	0.0000	[0.0000, 0.0001]	0.0000	[0.0000, 0.0000]
Random intercept SD	-0.0019	[-0.0036, -0.0001]	-0.0002	[-0.0019, 0.0015]	0.0003	[-0.0018, 0.0024]	-0.0006	[-0.0024, 0.0011]
Random slope SD	0.0010	[-0.0011, 0.0031]	-0.0001	[-0.0021, 0.0019]	0.0020	[-0.0007, 0.0047]	-0.0013	[-0.0034, 0.0009]
L1 effect size	-0.0028	[-0.0057, 0.0002]	-0.0019	[-0.0049, 0.0011]	-0.0039	[-0.0074, -0.0005]	-0.0024	[-0.0055, 0.0007]
L2 effect size	0.0026	[0.0001, 0.0052]	-0.0006	[-0.0032, 0.0019]	-0.0017	[-0.0048, 0.0013]	0.0018	[-0.0009, 0.0044]
L12 effect size	0.0015	[-0.0016, 0.0046]	0.0009	[-0.0021, 0.0039]	0.0002	[-0.0037, 0.0040]	0.0002	[-0.0029, 0.0034]
random effects cor.	0.0003	[-0.0015, 0.0020]	0.0004	[-0.0013, 0.0021]	0.0009	[-0.0012, 0.0030]	-0.0005	[-0.0023, 0.0013]
Relative L2 N in power analysis	0.0003	[0.0002, 0.0003]	-0.0001	[-0.0001, 0.0000]	0.0003	[0.0002, 0.0003]	-0.0001	[-0.0001, -0.0001]
Relative cluster size in power analysis	0.0000	[-0.0001, 0.0000]	0.0000	[0.0000, 0.0000]	0.0001	[0.0001, 0.0002]	0.0000	[-0.0001, 0.0000]
Level-2 effect								
Intercept	-0.0008	[-0.0014, -0.0003]	0.0117	[0.0115, 0.0120]	-0.0013	[-0.0019, -0.0007]	0.0124	[0.0121, 0.0127]
Original L2 N	-0.0001	[-0.0002, -0.0001]	-0.0001	[-0.0001, -0.0001]	-0.0001	[-0.0002, -0.0001]	-0.0001	[-0.0001, -0.0001]
Original cluster size	0.0000	[0.0000, 0.0001]	0.0000	[0.0000, 0.0000]	0.0000	[0.0000, 0.0001]	0.0000	[0.0000, 0.0000]
Random intercept SD	0.0125	[0.0090, 0.0159]	-0.0023	[-0.0040, -0.0006]	0.0089	[0.0048, 0.0130]	-0.0049	[-0.0070, -0.0028]
Random slope SD	0.0014	[-0.0057, 0.0084]	0.0003	[-0.0030, 0.0037]	-0.0023	[-0.0103, 0.0058]	-0.0013	[-0.0053, 0.0028]
L1 effect size	-0.0017	[-0.0077, 0.0044]	-0.0003	[-0.0031, 0.0025]	-0.0067	[-0.0133, 0.0000]	-0.0029	[-0.0063, 0.0005]
L2 effect size	-0.0345	[-0.0396, -0.0294]	0.0071	[0.0046, 0.0097]	-0.0250	[-0.0306, -0.0195]	0.0077	[0.0047, 0.0106]
L12 effect size	-0.0053	[-0.0157, 0.0050]	-0.0042	[-0.0091, 0.0008]	0.0081	[-0.0039, 0.0200]	-0.0006	[-0.0066, 0.0055]

	Varied predictor variance				Varied cluster size			
	relative difference		absolute difference		relative difference		absolute difference	
Predictors	Beta	95% CI	Beta	95% CI	Beta	95% CI	Beta	95% CI
random effects cor.	-0.0003	[-0.0038, 0.0031]	0.0014	[-0.0003, 0.0032]	-0.0019	[-0.0058, 0.0020]	-0.0007	[-0.0027, 0.0013]
Relative L2 N in power analysis	0.0001	[0.0001, 0.0001]	-0.0001	[-0.0002, -0.0001]	0.0001	[0.0000, 0.0001]	-0.0001	[-0.0001, -0.0001]
Relative cluster size in power analysis	0.0002	[0.0002, 0.0002]	0.0000	[0.0000, 0.0000]	0.0003	[0.0003, 0.0003]	0.0000	[-0.0001, 0.0000]
Cross-level effect								
Intercept	0.0004	[-0.0002, 0.0009]	0.0119	[0.0117, 0.0122]	-0.0004	[-0.0012, 0.0003]	0.0146	[0.0142, 0.0150]
Original L2 N	-0.0001	[-0.0002, -0.0001]	-0.0001	[-0.0001, -0.0001]	-0.0002	[-0.0003, -0.0001]	-0.0001	[-0.0001, 0.0000]
Original cluster size	-0.0001	[-0.0001, 0.0000]	0.0000	[-0.0001, 0.0000]	0.0000	[0.0000, 0.0001]	-0.0001	[-0.0001, 0.0000]
Random intercept SD	0.0015	[-0.0021, 0.0052]	0.0002	[-0.0017, 0.0020]	0.0030	[-0.0022, 0.0082]	0.0001	[-0.0030, 0.0031]
Random slope SD	0.0123	[0.0045, 0.0200]	-0.0106	[-0.0144, -0.0068]	0.0131	[0.0026, 0.0237]	-0.0098	[-0.0162, -0.0033]
L1 effect size	0.0003	[-0.0062, 0.0068]	0.0007	[-0.0024, 0.0038]	0.0021	[-0.0063, 0.0105]	0.0005	[-0.0044, 0.0055]
L2 effect size	0.0001	[-0.0056, 0.0058]	-0.0009	[-0.0037, 0.0019]	-0.0020	[-0.0097, 0.0057]	-0.0007	[-0.0052, 0.0039]
L12 effect size	-0.0438	[-0.0548, -0.0327]	0.0156	[0.0100, 0.0213]	-0.0543	[-0.0700, -0.0385]	0.0095	[-0.0001, 0.0190]
random effects cor.	0.0026	[-0.0014, 0.0065]	-0.0002	[-0.0022, 0.0017]	-0.0005	[-0.0058, 0.0049]	-0.0009	[-0.0041, 0.0023]
Relative L2 N in power analysis	0.0001	[0.0000, 0.0001]	-0.0001	[-0.0002, -0.0001]	0.0000	[0.0000, 0.0001]	-0.0001	[-0.0001, 0.0000]
Relative cluster size in power analysis	0.0000	[0.0000, 0.0001]	0.0000	[-0.0001, 0.0000]	0.0004	[0.0003, 0.0004]	-0.0001	[-0.0001, -0.0001]

Table 5 Power estimates in Simulation 3 (small sample size, varied cluster-size/predictor variance) computed by the summary-statistics- and simulation-based power analysis. Pilot data are generated from a model specified by Equation (3) with L2 sample size = 15 and cluster size = 32.

		Varied cluster size/predictor variance		
		Average power		
L2 size	cl. size	Summary	Simulation	RMSE
Level-1 effect				
10	22	0.593	0.593	0.034
10	32	0.642	0.639	0.030
10	42	0.669	0.668	0.025
15	22	0.758	0.759	0.025
15	32	0.795	0.793	0.020
15	42	0.815	0.810	0.019
20	22	0.842	0.840	0.019
20	32	0.869	0.865	0.016
20	42	0.882	0.877	0.015
Level-2 effect				
10	22	0.468	0.485	0.038
10	32	0.490	0.498	0.034
10	42	0.502	0.506	0.032
15	22	0.628	0.643	0.031
15	32	0.649	0.656	0.026
15	42	0.660	0.662	0.025
20	22	0.725	0.736	0.024
20	32	0.743	0.746	0.020
20	42	0.753	0.751	0.020
Cross-level effect				
10	22	0.379	0.374	0.048
10	32	0.420	0.417	0.042
10	42	0.445	0.440	0.042
15	22	0.523	0.531	0.045
15	32	0.567	0.567	0.039
15	42	0.593	0.590	0.035
20	22	0.620	0.628	0.040
20	32	0.661	0.661	0.035
20	42	0.684	0.681	0.031

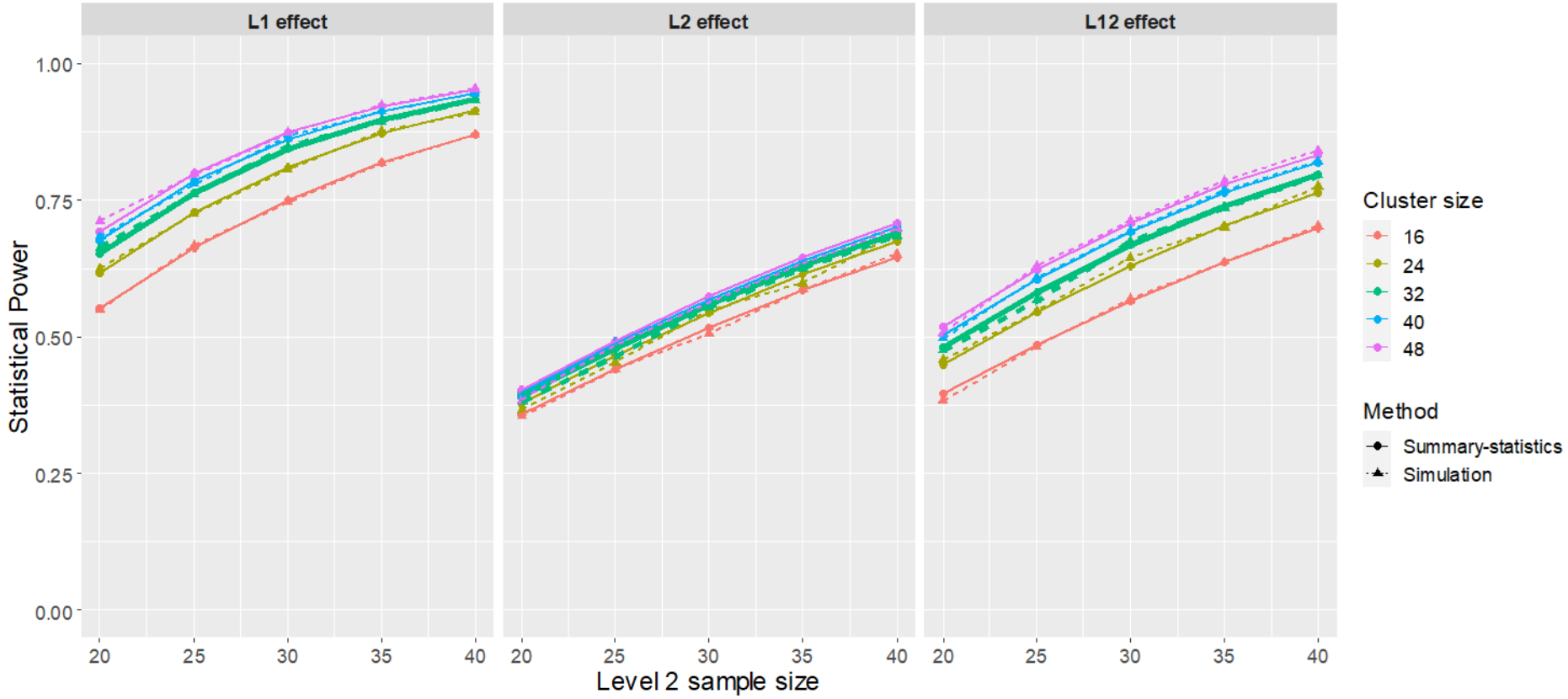

Figure 1. Power curves of L1, L2, and L12 effects with a complex model obtained from simulation-based power analysis (dotted line) and summary-statistics-based power analysis (solid line) using a simulated dataset (cluster size = 32; L2 sample size = 30). Model: $y_{ij} = (\gamma_{00} + \gamma_{01} wj + u_{0j}) + (\gamma_{10} + \gamma_{11} wj + u_{1j}) x_{ij} + e_{ij}$.

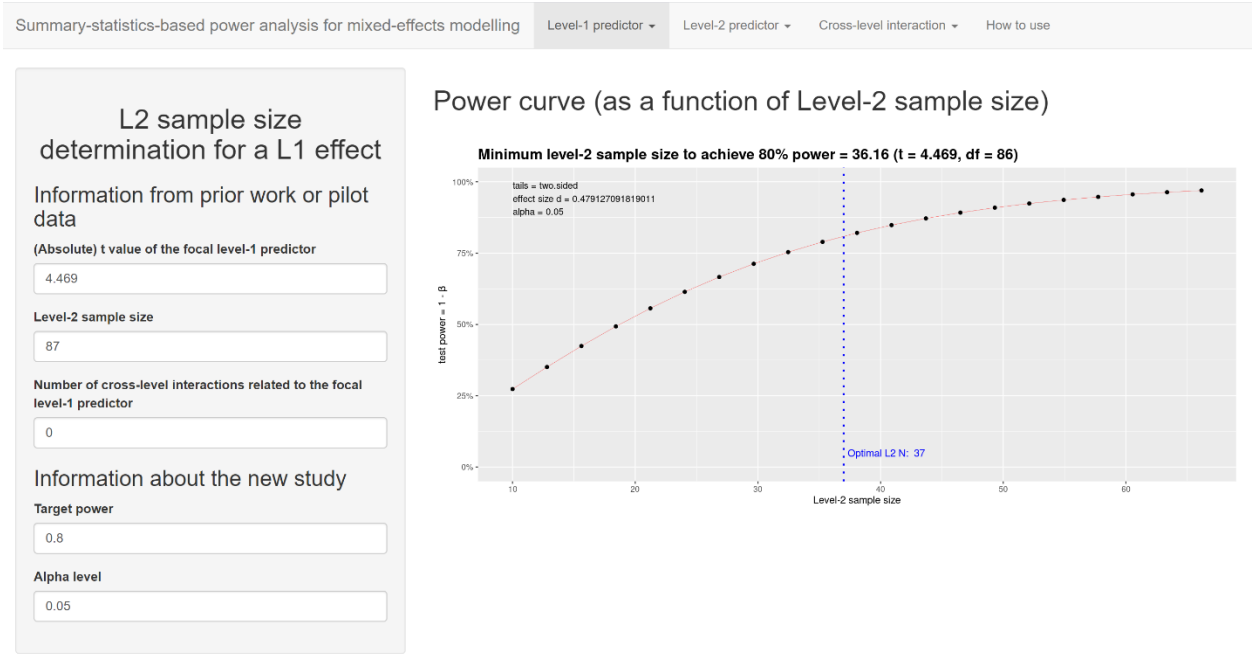

Figure 2. Web app to conduct a priori power analysis based on the summary-statistics-based power analysis (https://koumurayama.shinyapps.io/summary_statistics_based_power/).

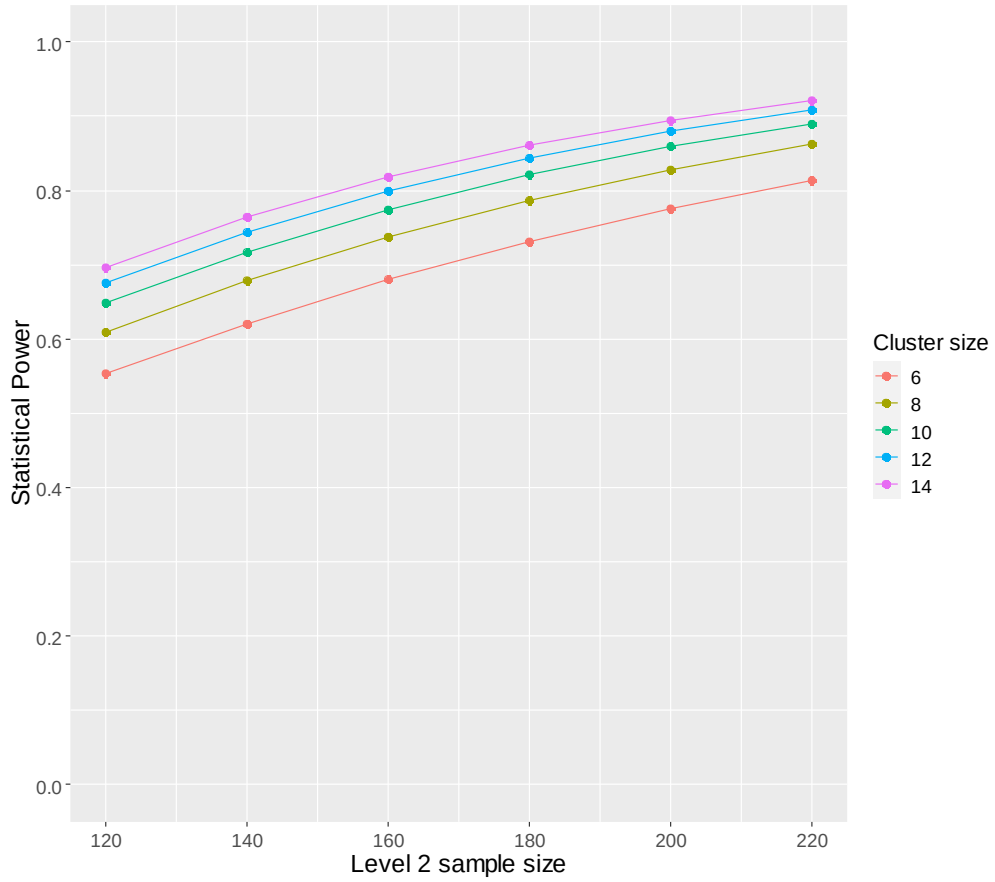

Figure 3. Power curve for a cross-level interaction as a function of L2 sample size (number of participants) and cluster size (assessments of learning diary), based on the information reported by Eckerlein et al. (2019).

Appendix

Snijders & Bosker (1993) derived general formulae for the standard errors of estimated regression coefficients in two-level designs using the generalized least squares (GLS) estimator. However, as far as we are aware, Snijders & Bosker (1993) (and other studies) did not provide a specific closed-form expressions to evaluate standard errors for specific models like the ones we discussed in the current article.

In this Appendix, we first explain how the standard errors of estimated regression coefficients can be generally expressed based on Snijders & Bosker (1993), and then discuss how the closed-form formulae of standard errors can be derived for several specific models. Specifically, we first derived the closed-form formulae for a mixed-effects model that includes only one L1 predictor (i.e. Equation 1). We then extend it to a model that includes two L1 predictors, and a model that includes two L1 predictors and one L2 predictor. Importantly, for each of the models, based on the closed-form expressions of the standard errors, we also demonstrate the mathematical equivalence between mixed-effects modelling and the summary-statistics approach. Finally, to discuss the generality of the argument, we further extend these models to the ones that include more than one predictor at each level. Note that the formulae we derived to compute standard errors are only asymptotically correct. It is well known that the GLS estimator underestimates SEs when sample size is small (Kenward & Roger, 1997; Littell, 2002). The purpose of the appendix is not to derive the accurate formula for SEs, but to show the mathematical equivalence when certain conditions are met. The discussion for the issue associated with small sample size can be found in the main text.

As noted in the article, here we assume that variance/covariance matrix of the L1 independent variables (number of L1 independent variables = p_1) is the same across clusters, and cluster size (n) is equal across clusters. However, we will also discuss the case when these restrictions are not imposed toward the end. For the purpose of simplicity, it is also assumed that the same set of L2 independent variables (number of L2 independent variables = p_2) are used to explain random slopes of each of the L1 predictors. In addition, without loss of generality of discussion, we assume that L1 independent variable x_{ij} is within-cluster deviation score (i.e. *centering* within *clusters*) and that L2 independent variable w_j is cluster-mean centered. In other words, mean of x_{ij} is zero for jth cluster and mean of w_j is also zero. Let $\boldsymbol{X_j^*} = (\mathbf{1}, \boldsymbol{X_j})$ be $n \times (p_1 + 1)$ data matrix for L1 independent variables in jth cluster that include the focal variable, and let $\boldsymbol{W}^* = (\mathbf{1}, \boldsymbol{W})$ be $J \times (p_2 + 1)$ data matrix for L2 independent variables including the focal variable. For both $\boldsymbol{X_j^*}$ and $\boldsymbol{W}^*$, elements in the first column are all ones ($= \mathbf{1}$) to denote intercepts.

Synthesizing different equations (i.e. Equations 22, 25, 30 and 31) provided by Snijders & Bosker (1993) and assuming random effects (co)variances are known, a general form of expected variances-covariances matrix of estimated regression coefficients by GLS estimator (denoted as $cov(\hat{\boldsymbol{\gamma}}_{GLS})$) can be expressed as

$$cov(\hat{\boldsymbol{\gamma}}_{GLS}) = \frac{1}{J}\left(\boldsymbol{T} + \frac{\sigma^2}{n}[\boldsymbol{e}\boldsymbol{e}' + \boldsymbol{\Sigma_w}]^{-1}\right) \otimes (\mu\mu' + \boldsymbol{\Sigma_B})^{-1}, \quad \text{(A1)}$$

where $\boldsymbol{T}$ is the random effects variance-covariance matrix, n is cluster size, J is L2 sample size, σ^2 is an error (or, within-cluster) variance, $\boldsymbol{e}$ is a vector that includes mean of L1 independent variables for jth cluster (i.e. $\boldsymbol{e} = (1, \bar{x}_{1.j}, \cdots, \bar{x}_{p_1.j})'$), $\boldsymbol{\Sigma_w} = cov\left(\boldsymbol{X_j^*}\right)$ within jth cluster (i.e. $(p_1 + 1) \times (p_1 + 1)$ within cluster sample variance-covariance), μ is a vector that includes mean of L2 independent variables (i.e. $\mu = (1, \bar{w}_{1j}, \cdots, \bar{w}_{p_2 j})'$), and $\boldsymbol{\Sigma_B} = cov(\boldsymbol{W}^*)$ (i.e. $(p_2 + 1) \times (p_2 + 1)$ between cluster sample variance-covariance). $\otimes$ denotes a Kronecker product. Recall that mean of x_{ij} is zero for jth cluster and mean of w_j is also zero, indicating $\boldsymbol{e} = (1, \mathbf{0}')'$ and $\mu =$

$(1, \mathbf{0}')'$.

Because random effects (co)variances (i.e. $\boldsymbol{T}$ and σ^2) are generally unknown, they are usually estimated by maximum likelihood methods under the normality assumption. Estimated GLS estimator (denoted as $cov(\hat{\boldsymbol{\gamma}})$) is approximately equivalent to the GLS estimator as cluster size *J* tends to infinity (i.e. $cov(\hat{\boldsymbol{\gamma}}) \approx cov(\hat{\boldsymbol{\gamma}}_{GLS})$).

A Model That Includes Only One L1 Independent Variable (i.e. Equation 1)

Closed Form Expression for Standard Errors

With Equation A1 we can obtain the closed form expression of standard errors for a model that includes only one L1 independent variable (i.e. Equation (1)). In this case, $p_1 = 1$ (for a vector of the focal L1 variable $\boldsymbol{X_j}$) , $p_2 = 0$, $\boldsymbol{T} = \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{pmatrix}$, $\boldsymbol{e} = (1,0)'$, $\boldsymbol{\Sigma_w} = cov(\mathbf{1}, \boldsymbol{X_j}) = \begin{pmatrix} 0 & 0 \\ 0 & s_x^2 \end{pmatrix}$ (where s_x^2 is a within-cluster sample variance of the focal L1 independent variable), $\mu = 1$, $\boldsymbol{\Sigma_B} = 0$. Substituting these values into Equation A1 gives

$$cov(\hat{\boldsymbol{\gamma}}_{GLS}) = \frac{1}{J}\left(\begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{pmatrix} + \frac{\sigma^2}{n}\begin{pmatrix} 1 & 0 \\ 0 & s_x^2 \end{pmatrix}^{-1}\right) \otimes (1+0)^{-1} = \frac{1}{J}\begin{pmatrix} \tau_{00} + \frac{\sigma^2}{n} & \tau_{01} \\ \tau_{01} & \tau_{11} + \frac{\sigma^2}{n s_x^2} \end{pmatrix}. \quad \text{(A2)}$$

Therefore, the standard errors of estimated regression coefficient (γ_{10}) can be evaluated by the square root of the (2,2) element of $cov(\hat{\boldsymbol{\gamma}}_{GLS})$:

$$SE(\hat{\gamma}_{10}) = \sqrt{\frac{\tau_{11} + \frac{\sigma^2}{n s_x^2}}{J}}. \quad \text{(A3)}$$

This formula is the same as Equation 2, which is also shown in Snijders (2005).

Equivalence with Summary-Statistics Approach

We are now ready to prove the mathematical equivalence between mixed-effects modelling and the summary-statistics approach. For the purpose of simplicity, we assume that the focal L1 variable x_{ij} is binary like the example we discussed in the article (i.e. mindfulness intervention) but as noted later, the results can be easily generalised to a continuous L1 independent variable (which is briefly discussed later). Let $\mu_{jE} - \mu_{jC}$ be a population mean difference between intervention and control conditions for jth cluster, and let β_{1j} be jth population regression coefficient when outcome y_{ij} is regressed on the intervention variable x_{ij} in jth cluster. If the cluster size is the same between two conditions, β_{1j} is mathematically equivalent to the $\mu_{jE} - \mu_{jC}$. Therefore, the summary-statistics approach, which conducts a paired-samples *t* test to examine the effect of intervention, can be viewed as an approach that tests the null hypothesis for population mean of β_{1j} (denoted as $E(\beta_{1j}) = a_{10}$), i.e. H_0: $a_{10} = 0$. With this point in mind, the relation between β_{1j} and a_{10} can be modelled as the form of the level-2 equation in the standard mixed-effects model as

$$\beta_{1j} = \mu_{jE} - \mu_{jC} = a_{10} + u_{1j}, \quad \text{(A4)}$$

where u_{1j} is a deviation term and $var(u_{1j}) = var(\beta_{1j}) = \tau_{11}$. However, we cannot actually observe the population value of β_{1j} and usually estimate it by mean difference of observed outcomes (i.e. $\bar{y}_{jE} - \bar{y}_{jC}$). The relation between β_{1j} and its estimates (denoted as $\hat{\beta}_{1j} = \bar{y}_{jE} - \bar{y}_{jC}$) can be formulated by introducing a sampling error (denoted as e_j) as

$$\hat{\beta}_{1j} = \bar{y}_{jE} - \bar{y}_{jC} = \beta_{1j} + e_j, \quad \text{(A5)}$$

where $var(e_j) = var(\hat{\beta}_{1j}|x_{ij}) = \frac{var(y_{ij}|x_{ij})}{\sum_{i=1}^{n} x_{ij}^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_{ij}^2} = \frac{\sigma^2}{ns_x^2}$, from the standard formula of variance-covariance of estimated regression coefficients by ordinary least squares (OLS) (i.e. in the situation where outcome y_{ij} is regressed on x_{ij} to estimate β_{1j}). Note that s_x^2 is the observed variance of the focal predictor, and is 0.25 in the current case where an intervention variable is binary.

Combining Equation A4 and Equation A5 we obtain:

$$\hat{\beta}_{1j} = a_{10} + u_{1j} + e_j. \qquad \text{(A6)}$$

Because of the standard assumption that u_{1j} and e_j are independent (i.e. $cov(u_{1j}, e_j) = 0$), $var(\hat{\beta}_{1j}) = var(u_{1j} + e_j) = var(\mu_{1j}) + var(e_j) = \tau_{11} + \sigma^2/[ns_x^2]$. Therefore, in the summary-statistics approach with a paired-samples *t* test, a test statistic can now be expressed as

$$t = \frac{\hat{a}_{10}}{SE(\hat{a}_{10})} = \frac{\frac{1}{J}\sum_{i=1}^{J}\hat{\beta}_{1j}}{\sqrt{\frac{V(\hat{\beta}_{1j})}{J}}} = \frac{\frac{1}{J}\sum_{i=1}^{J}(\bar{y}_{jE} - \bar{y}_{jC})}{\sqrt{\frac{\tau_{11} + \sigma^2/ns_x^2}{J}}}. \qquad \text{(A7)}$$

Obviously, the denominator of this equation is mathematically equivalent to the one provided in Equation A3. Considering the relation $\hat{\gamma}_{10} = \hat{a}_{10} = \frac{1}{J}\sum_{i=1}^{J}\hat{\beta}_{1j} = \frac{1}{J}\sum_{i=1}^{J}(\bar{y}_{jE} - \bar{y}_{jC})$, *t*-statistics from two different approaches thus become equivalent. Even if x_{ij} is continuous, because the condition $\hat{\gamma}_{10} = \hat{a}_{10} = \frac{1}{J}\sum_{i=1}^{J}\hat{\beta}_{1j}$ is unchanged, both standard errors and *t*-statistics for testing the null-hypothesis are mathematically equivalent between $\hat{\gamma}_{10}$ (i.e. mixed-effects modelling) and $\hat{a}_{10}$ (i.e. summary-statistics approach).

A Model That Includes Two L1 Independent Variables (i.e. Equation 1 with a Covariate) Closed Form Expression for Standard Errors

A mixed-effects model for this case can be expressed as

$$y_{ij} = (\gamma_{00} + \mu_{0j}) + (\gamma_{10} + \mu_{1j})\, x_{1ij} + (\gamma_{20} + \mu_{2j})\, x_{2ij} + e_{ij}, \qquad (A8)$$

where x_{1ij} denotes the focal L1 variable and x_{2ij} is a L1 covariate. γ_{00}, γ_{10} and γ_{20} are overall intercept and slopes. $\boldsymbol{U} = (\mu_{0j}, \mu_{1j}, \mu_{2j})$ is a vector of random effects that has the moment assumption

$$E(\boldsymbol{U}) = \mathbf{0} \quad cov(\boldsymbol{U}) = \boldsymbol{T} = \begin{pmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{01} & \tau_{11} & \tau_{12} \\ \tau_{02} & \tau_{12} & \tau_{22} \end{pmatrix}. \qquad \text{(A9)}$$

In this case, $p_1 = 2, p_2 = 0, \boldsymbol{e} = (1,0,0)'$, $\boldsymbol{\Sigma_w} = cov(\mathbf{1}, \boldsymbol{X_{1j}}, \boldsymbol{X_{2j}}) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & s_{x1}^2 & s_{x12} \\ 0 & s_{x12} & s_{x2}^2 \end{pmatrix}$, $\mu = 1$, $\boldsymbol{\Sigma_B} = 0$. Therefore, according to Equation A1, $cov(\hat{\boldsymbol{\gamma}}_{GLS})$ becomes

$$cov(\boldsymbol{\gamma}_{GLS}) = \frac{1}{J}\left(\begin{pmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{01} & \tau_{11} & \tau_{12} \\ \tau_{02} & \tau_{12} & \tau_{22} \end{pmatrix} + \frac{\sigma^2}{n} \begin{pmatrix} 1 & 0 & 0 \\ 0 & s_{x1}^2 & s_{x12} \\ 0 & s_{x12} & s_{x2}^2 \end{pmatrix}^{-1} \right) \otimes (1+0)^{-1}$$

$$= \frac{1}{J}\left(\begin{pmatrix}\tau_{00} & \tau_{01} & \tau_{02}\\ \tau_{01} & \tau_{11} & \tau_{12}\\ \tau_{02} & \tau_{12} & \tau_{22}\end{pmatrix} + \frac{\sigma^2}{n}\begin{pmatrix}1 & 0 & 0\\ 0 & s_{x2}^2/(s_{x1}^2 s_{x2}^2 - s_{x12}^2) & -s_{x12}/(s_{x1}^2 s_{x2}^2 - s_{x12}^2)\\ 0 & -s_{x12}/(s_{x1}^2 s_{x2}^2 - s_{x12}^2) & s_{x1}^2/(s_{x1}^2 s_{x2}^2 - s_{x12}^2)\end{pmatrix}\right)$$

$$= \frac{1}{J}\begin{pmatrix}\tau_{00} + \frac{\sigma^2}{n} & \tau_{01} & \tau_{02}\\ \tau_{01} & \tau_{11} + \frac{\sigma^2}{n s_{x1}^2(1-r_{x12}^2)} & \tau_{12} - \frac{\sigma^2}{n}\left(\frac{r_{x12}^2}{s_{x12}(1-r_{x12}^2)}\right)\\ \tau_{02} & \tau_{12} - \frac{\sigma^2}{n}\left(\frac{r_{x12}^2}{s_{x12}(1-r_{x12}^2)}\right) & \tau_{22} + \frac{\sigma^2}{n s_{x2}^2(1-r_{x12}^2)}\end{pmatrix}, \quad \text{(A10)}$$

where $r_{x12} = s_{x12}/s_{x1}s_{x2}$ denotes sample correlation, and r_{x12}^2 is equivalent to the proportion of variance explained when x_2 is regressed on x_1 and vice versa. Standard errors of estimated overall intercept (γ_{00}) and intercepts (γ_{10} and γ_{20}) can now be evaluated as the square root of the diagonal elements of $cov(\hat{\boldsymbol{\gamma}}_{GLS})$ and therefore, the standard errors of the focal variable is:

$$SE(\hat{\gamma}_{10}) = \sqrt{\frac{\tau_{11} + \frac{\sigma^2}{n s_{x1}^2(1-r_{x12}^2)}}{J}}. \quad \text{(A11)}$$

Equivalence with Summary-Statistics Approach

Let β_{0j}, β_{1j}, and β_{2j}, be the jth population intercept and regression coefficients when outcome y_{ij} is regressed on the two L1 independent variables in the jth cluster. They can be expressed using their overall means (a_{00}, a_{10}, a_{20}) and corresponding deviations $\beta_{0j} = a_{00} + u_{0j}$, $\beta_{1j} = a_{10} + u_{1j}$ and $\beta_{2j} = a_{20} + u_{2j}$, where $\boldsymbol{U} = (u_{0j}, u_{1j}, u_{2j})'$ and $cov(\boldsymbol{U}) = \begin{pmatrix}\tau_{00} & \tau_{01} & \tau_{02}\\ \tau_{01} & \tau_{11} & \tau_{12}\\ \tau_{02} & \tau_{12} & \tau_{22}\end{pmatrix}$. On the other hand, estimates $\hat{\beta}_{0j}, \hat{\beta}_{1j}$ and $\hat{\beta}_{2j}$ can be expressed as the sum of population values and sampling errors as $\hat{\beta}_{0j} = \beta_{0j} + e_{0j}$, $\hat{\beta}_{1j} = \beta_{1j} + e_{1j}$ and $\hat{\beta}_{2j} = \beta_{2j} + e_{2j}$, where

$$cov(\boldsymbol{e_j}) = cov(\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}|x_j) = var(y_{ij}|x_{1ij}, x_{2ij})(\boldsymbol{x_j^{*\prime}}\boldsymbol{x_j^*})^{-1} = \sigma^2\begin{pmatrix}n & 0 & 0\\ 0 & ns_{x1}^2 & ns_{x12}\\ 0 & ns_{x12} & ns_{x2}^2\end{pmatrix}^{-1}$$

$$= \frac{\sigma^2}{n}\begin{pmatrix}1 & 0 & 0\\ 0 & \frac{1}{s_{x1}^2(1-r_{x12}^2)} & -\frac{r_{x12}^2}{s_{x12}(1-r_{x12}^2)}\\ 0 & -\frac{r_{x12}^2}{s_{x12}(1-r_{x12}^2)} & \frac{1}{s_{x2}^2(1-r_{x12}^2)}\end{pmatrix}, \quad \text{(A12)}$$

for $\boldsymbol{e_j} = (e_{0j}, e_{1j}, e_{2j})'$, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}} = (\hat{\beta}_{0j}\ \hat{\beta}_{1j}, \hat{\beta}_{2j})'$, $\boldsymbol{x_j^*} = (\boldsymbol{1}, \boldsymbol{x_{1j}}, \boldsymbol{x_{2j}})$.

Combining the relations described above, the focal regression coefficient $\beta_{1j} = a_{10} + u_{1j}$ becomes $\hat{\beta}_{1j} = a_{10} + u_{1j} + e_{1j}$. Because of the assumption that u_{pj} and e_{pj} ($p = 0,1,2$) are independent, $var(\hat{\beta}_{1j}) = var(\mu_{1j}) + var(e_{1j}) = \tau_{11} + \sigma^2/[ns_{x1}^2(1 - r_{x12}^2)]$. Therefore, in the summary-statistics approach averaging $\hat{\beta}_{1j}$ over clusters to evaluate an intervention effect, a test statistic for $H_0: a_{10} = 0$ can now be expressed as

$$t = \frac{\hat{a}_{10}}{SE(\hat{a}_{10})} = \frac{\frac{1}{J}\sum_{i=1}^{J}\hat{\beta}_{1j}}{\sqrt{\frac{\tau_{11}+\frac{\sigma^2}{ns_{x1}^2(1-r_{x12}^2)}}{J}}} \quad . \quad \text{(A13)}$$

Obviously, a denominator of this equation is equivalent to the one provided in the Equation A11. Considering the relation $\hat{\gamma}_{10} = \hat{a}_{10} = \frac{1}{J}\sum_{i=1}^{J}\hat{\beta}_{1j}$, *t*-statistics are equivalent between $\hat{\gamma}_{10}$ (i.e. mixed-effects modelling) and $\hat{a}_{10}$ (i.e. summary-statistics approach).

A Model That Includes Two L1 Independent Variables and One L2 Independent variable

Closed Form Expression for Standard Errors

In this case, a mixed-effects model can be expressed as

$$y_{ij} = (\gamma_{00} + \gamma_{01}w_j + \mu_{0j}) + (\gamma_{10} + \gamma_{11}w_j + \mu_{1j})\,x_{1ij} + (\gamma_{20} + \gamma_{21}w_j + \mu_{2j})\,x_{2ij} + e_{ij} \quad \text{(A14)}$$

where x_{1ij} denotes the focal L1 independent variable, x_{2ij} is a L1 covariate, and w_j is a (cluster mean centered) L2 variable for *j*th cluster. γ_{00}, γ_{10} and γ_{20} are overall intercept and slopes (γ_{10} is the L1 effect of the focal variable) and γ_{01}, γ_{11} and γ_{21} are regression coefficients from w_j in each level-2 equation. More specifically, γ_{01} is the main effect of the L2 variable (i.e. L2 effect) and γ_{11} is the cross-level interaction between the focal L1 predictor and the L2 variable (i.e. L1 effect). $\boldsymbol{U} = (\mu_{0j}, \mu_{1j}, \mu_{2j})$ is a vector of random effects that has the same moment assumption as Equation A9. If there is no level-1 covariate (x_{2ij}), this version of the model is equivalent to the Equation 3.

In this case, $p_1 = 2, p_2 = 1$, $\boldsymbol{e} = (1,0,0)'$, $\boldsymbol{\Sigma_w} = cov(\mathbf{1}, \boldsymbol{X_{1j}}, \boldsymbol{X_{2j}}) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & s_{x1}^2 & s_{x12} \\ 0 & s_{x12} & s_{x2}^2 \end{pmatrix}$, $\mu = (1,0)'$, $\boldsymbol{\Sigma_B} = cov((\mathbf{1}, \boldsymbol{W})) = \begin{pmatrix} 0 & 0 \\ 0 & s_w^2 \end{pmatrix}$ (s_w^2 is a between cluster sample variance of w). Therefore, according to Equation A1, $cov(\hat{\boldsymbol{\gamma}}_{GLS})$ becomes

$$cov(\hat{\boldsymbol{\gamma}}_{GLS}) = \frac{1}{J}\left(\begin{pmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{01} & \tau_{11} & \tau_{12} \\ \tau_{02} & \tau_{12} & \tau_{22} \end{pmatrix} + \frac{\sigma^2}{n}\begin{pmatrix} 1 & 0 & 0 \\ 0 & s_{x1}^2 & s_{x12} \\ 0 & s_{x12} & s_{x2}^2 \end{pmatrix}^{-1}\right) \otimes \left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & s_w^2 \end{pmatrix}\right)^{-1}$$

$$= \frac{1}{J}\begin{pmatrix} \tau_{00} + \frac{\sigma^2}{n} & \tau_{01} & \tau_{02} \\ \tau_{01} & \tau_{11} + \frac{\sigma^2}{ns_{x1}^2(1-r_{x12}^2)} & \tau_{12} - \frac{\sigma^2}{n}\left(\frac{r_{x12}^2}{s_{x12}(1-r_{x12}^2)}\right) \\ \tau_{02} & \tau_{12} - \frac{\sigma^2}{n}\left(\frac{r_{x12}^2}{s_{x12}(1-r_{x12}^2)}\right) & \tau_{22} + \frac{\sigma^2}{ns_{x2}^2(1-r_{x12}^2)} \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{s_w^2} \end{pmatrix}$$

=

$$\frac{1}{J}\begin{pmatrix} \tau_{00}+\frac{\sigma^2}{n} & 0 & \tau_{01} & 0 & \tau_{02} & 0 \\ 0 & \frac{1}{s_w^2}(\tau_{00}+\frac{\sigma^2}{n}) & 0 & \tau_{01}/s_w^2 & 0 & \tau_{02}/s_w^2 \\ \tau_{01} & 0 & \tau_{11}+\frac{\sigma^2}{ns_{x1}^2(1-r_{x12}^2)} & 0 & \tau_{12}-\frac{\sigma^2}{n}\left(\frac{r_{x12}^2}{s_{x12}(1-r_{x12}^2)}\right) & 0 \\ 0 & \tau_{01}/s_w^2 & 0 & \frac{1}{s_w^2}(\tau_{11}+\frac{\sigma^2}{ns_{x1}^2(1-r_{x12}^2)}) & 0 & \frac{1}{s_w^2}(\tau_{12}-\frac{\sigma^2}{n}\left(\frac{r_{x12}^2}{s_{x12}(1-r_{x12}^2)}\right)) \\ \tau_{02} & 0 & \tau_{12}-\frac{\sigma^2}{n}\left(\frac{r_{x12}^2}{s_{x12}(1-r_{x12}^2)}\right) & 0 & \tau_{22}+\frac{\sigma^2}{ns_{x2}^2(1-r_{x12}^2)} & 0 \\ 0 & \tau_{02}/s_w^2 & 0 & \frac{1}{s_w^2}(\tau_{12}-\frac{\sigma^2}{n}\left(\frac{r_{x12}^2}{s_{x12}(1-r_{x12}^2)}\right)) & 0 & \frac{1}{s_w^2}(\tau_{22}+\frac{\sigma^2}{ns_{x2}^2(1-r_{x12}^2)}) \end{pmatrix} \quad \text{(A15)}$$

Standard errors of estimated fixed effects can be evaluated as the square root of the diagonal elements of $cov(\hat{\boldsymbol{\gamma}}_{GLS})$. Therefore,

$$SE(\hat{\gamma}_{01}) = \sqrt{\frac{\frac{1}{s_w^2}(\tau_{00}+\frac{\sigma^2}{n})}{J}},\quad SE(\hat{\gamma}_{10}) = \sqrt{\frac{\tau_{11}+\frac{\sigma^2}{ns_{x1}^2(1-r_{x12}^2)}}{J}}, SE(\hat{\gamma}_{11}) = \sqrt{\frac{\frac{1}{s_w^2}(\tau_{11}+\frac{\sigma^2}{ns_{x1}^2(1-r_{x12}^2)})}{J}} \quad \text{(A16)}$$

Again, standard errors for L2, L1 and L12 effects correspond to $SE(\hat{\gamma}_{01})$, $SE(\hat{\gamma}_{10})$ and $SE(\hat{\gamma}_{11})$. If substituting r_{x12} to 0 in these formulae, we can obtain formulae when there is no level-1 covariate (x_{2ij}) in the model (i.e. Equation 3).

Equivalence with Summary-Statistics Approach

Let β_{0j}, β_{1j}, and β_{2j},be jth population intercept and regression coefficients when outcome y_{ij} is regressed on the two L1 independent variables in the jth cluster. By incorporating L2 predictors, population linear intercept (β_{0j}) and slopes for the focal L1 independent variable and a L1 covariate (β_{1j}, β_{2j}) can be expressed using a linear combination of intercept terms (a_{00}, a_{10}, a_{20}), slope terms ($a_{01}w_j, a_{11}w_j, a_{21}w_j$) with L2 variable w_j, and deviation terms (u_{0j}, u_{1j}, u_{2j}): $\beta_{0j} = a_{00} + a_{01}w_j + u_{0j}$, $\beta_{1j} = a_{10} + a_{11}w_j + u_{1j}$ and $\beta_{2j} = a_{20} + a_{21}w_j + u_{2j}$, where $\boldsymbol{U} = (u_{0j}, u_{1j}, u_{2j})'$ and $cov(\boldsymbol{U}) = \begin{pmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{01} & \tau_{11} & \tau_{12} \\ \tau_{02} & \tau_{12} & \tau_{22} \end{pmatrix}$. On the other hand, estimates $\hat{\beta}_{0j}, \hat{\beta}_{1j}$ and $\hat{\beta}_{2j}$ can be expressed as the sum of population values and sampling errors as $\hat{\beta}_{0j} = \beta_{0j} + e_{0j}$, $\hat{\beta}_{1j} = \beta_{1j} + e_{1j}$ and $\hat{\beta}_{2j} = \beta_{2j} + e_{2j}$, where $cov(\boldsymbol{e_j})$ is expressed as the Equation (A12) for $\boldsymbol{e_j} = (e_{0j}, e_{1j}, e_{2j})'$.

Combining the relations $\beta_{0j} = a_{00} + a_{01}w_j + u_{0j}$ and $\beta_{1j} = a_{10} + a_{11}w_j + u_{1j}$ leads to $\hat{\beta}_{0j} = a_{00} + a_{01}w_j + u_{0j} + e_{0j}$ and $\hat{\beta}_{1j} = a_{10} + a_{11}w_j + u_{1j} + e_{1j}$. Because of the assumption that u_{pj} and e_{pj} ($p = 0,1,2$) are independent, $var(\hat{\beta}_{0j}|w_j) = var(\mu_{0j}) + var(e_{0j}) = \tau_{00} + \frac{\sigma^2}{n}$ and $var(\hat{\beta}_{1j}) = var(\mu_{1j}) + var(e_{1j}) = \tau_{11} + \sigma^2/[ns_{x1}^2(1 - r_{x12}^2)]$. In the summary statistics approach, as noted in the article, L2 effect is the regression coefficient when the regressor $\hat{\beta}_{0j}$ is regressed on w_j; L1 effect is the intercept when the regressor $\hat{\beta}_{1j}$ is regressed on w_j; L12 effect (cross-level interaction) is the regression coefficient in the same regression model. With this in mind, test statistics for $H_0: a_{01} = 0$ (L2 effect), $H_0: a_{10} = 0$ (L1 effect) and $H_0: a_{11} = 0$ (L12 effect) can now be expressed as

$$t = \frac{\hat{a}_{01}}{SE(\hat{a}_{01})} = \frac{cov(\hat{\beta}_{0j}, w_j)/s_w^2}{\sqrt{var(\hat{\beta}_{0j}|w_j)(\boldsymbol{w_j^*}'\ \boldsymbol{w_j^*})^{-1}_{(2,2)}}} = \frac{cov(\hat{\beta}_{0j}, w_j)/s_w^2}{\sqrt{\frac{\tau_{00}+\frac{\sigma^2}{n}}{Js_w^2}}}, \quad \text{(A17)}$$

$$t = \frac{\hat{a}_{10}}{SE(\hat{a}_{10})} = \frac{\frac{1}{J}\sum_{i=1}^{J}\hat{\beta}_{1j}}{\sqrt{var(\hat{\beta}_{1j}|w_j)(\boldsymbol{w_j^*}'\ \boldsymbol{w_j^*})^{-1}_{(1,1)}}} = \frac{\frac{1}{J}\sum_{i=1}^{J}\hat{\beta}_{1j}}{\sqrt{\frac{\tau_{11}+\frac{\sigma^2}{ns_{x1}^2(1-r_{x12}^2)}}{J}}}, \quad \text{(A18)}$$

$$t = \frac{\hat{a}_{11}}{SE(\hat{a}_{11})} = \frac{cov(\hat{\beta}_{1j}, w_j)/S_w^2}{\sqrt{var(\hat{\beta}_{1j}|w_j)(\boldsymbol{w_j^*}'\ \boldsymbol{w_j^*})^{-1}_{(2,2)}}} = \frac{cov(\hat{\beta}_{1j}, w_j)/S_w^2}{\sqrt{\frac{\tau_{11}+\frac{\sigma^2}{ns_{x1}^2(1-r_{x12}^2)}}{Js_w^2}}}, \quad \text{(A19)}$$

Where $\boldsymbol{A}_{(r,s)}$ denotes (r, s) element of an arbitrary matrix $\boldsymbol{A}$. Obviously, denominators of these equations are equivalent to the ones provided in the Equation A16. Considering the relations $\hat{\gamma}_{01} = \hat{a}_{01} = cov(\hat{\beta}_{0j}, w_j)/s_w^2, \hat{\gamma}_{10} = \hat{a}_{10} = \frac{1}{J}\sum_{i=1}^{J}\hat{\beta}_{1j}$, and $\hat{\gamma}_{11} = \hat{a}_{11} = cov(\hat{\beta}_{1j}, w_j)/s_w^2$, *t*-statistics for these three tests are equivalent between $\hat{\gamma}$ (i.e. mixed-effects modelling) and $\hat{a}$ (i.e. summary-statistics approach).

When There Are Two or More L1 and L2 Variables

When two or more independent variable are included at both L1 and L2 equations, taking the same steps with the previous cases, standard errors of L2 ($SE(\hat{\gamma}_{01})$: for cluster variable w_1), L1 ($SE(\hat{\gamma}_{10})$: for intervention variable x_1) and L12 effects ($SE(\hat{\gamma}_{11})$: $x_1 \times w_1$) can be generally expressed as

$$SE(\hat{\gamma}_{01}) = \sqrt{\frac{\frac{1}{s_{w1}^2(1-R_{w1}^2)}(\tau_{00}+\frac{\sigma^2}{n})}{J}} \quad SE(\hat{\gamma}_{10}) = \sqrt{\frac{\tau_{11}+\frac{\sigma^2}{ns_{x1}^2(1-R_{x1}^2)}}{J}} \quad SE(\hat{\gamma}_{11}) = \sqrt{\frac{\frac{1}{s_{w1}^2(1-R_{w1}^2)}(\tau_{11}+\frac{\sigma^2}{ns_{x1}^2(1-R_{x1}^2)})}{J}} \quad \text{(A20)}$$

where s_{x1}^2 is the sample variance of intervention variable x_1, and R_{x1}^2 is the proportion of variance explained when intervention variable x_1 is regressed on other level-1 covariates. Likewise, s_{w1}^2 is the sample variance of focal level-2 cluster variable w_1, and R_{w1}^2 is the proportion of variance explained when focal level-2 cluster variable w_1 is regressed on other level-2 covariates. With the closed expressions for standard errors, we can also prove that *t*-statistics obtained from the mixed-effects modelling and summary statistics approach are equivalent in the same manner (mathematical proof omitted here).

Variable Cluster Size and Predictor Variance

When cluster size and variance-covariance structure of the L1 predictors are not invariant across clusters, $cov(\hat{\boldsymbol{\gamma}}_{GLS})$ can be expressed as:

$$cov(\hat{\boldsymbol{\gamma}}_{GLS}) = \frac{1}{J}\left(\frac{1}{J}\sum_{j=1}^{N}\left(T + \frac{\sigma^2}{n_j}\left[ee' + \Sigma_{wj}\right]^{-1}\right) \otimes (\mu\mu' + \boldsymbol{\Sigma_B})^{-1}\right), \quad \text{(A21)}$$

where Σ_{wj} is within cluster variance for the jth cluster and n_j is the cluster size for the jth cluster (complete derivation omitted here). Importantly, $cov(\hat{\boldsymbol{\gamma}}_{GLS})$ (and standard errors) has the form in which all elements include J in the denominator. This suggests that the proposed summary-statistics-based method to estimate statistical power is still reasonably valid even when predictor variance and cluster size are invariant across clusters, as long as researchers are interested in determining the L2 sample size only (which was demonstrated in the simulation in the article).