American Psychological Association

Psychological Methods

1082-989X/19/$12.00

2019, Vol. 24, No. 1, 1–19
http://dx.doi.org/10.1037/met0000195

Statistical Power in Two-Level Models: A Tutorial Based on Monte Carlo Simulation

Matthias G. Arend and Thomas Schäfer
Chemnitz University of Technology

Abstract

The estimation of power in two-level models used to analyze data that are hierarchically structured is particularly complex because the outcome contains variance at two levels that is regressed on predictors at two levels. Methods for the estimation of power in two-level models have been based on formulas and Monte Carlo simulation. We provide a hands-on tutorial illustrating how a priori and post hoc power analyses for the most frequently used two-level models are conducted. We describe how a population model for the power analysis can be specified by using standardized input parameters and how the power analysis is implemented in SIMR, a very flexible power estimation method based on Monte Carlo simulation. Finally, we provide case-sensitive rules of thumb for deriving sufficient sample sizes as well as minimum detectable effect sizes that yield a power $\geq .80$ for the effects and input parameters most frequently analyzed by psychologists. For medium variance components, the results indicate that with lower level (L1) sample sizes up to 30 and higher level (L2) sample sizes up to 200, medium and large fixed effects can be detected. However, small L2 direct- or cross-level interaction effects cannot be detected with up to 200 clusters. The tutorial and guidelines should be of help to researchers dealing with multilevel study designs such as individuals clustered within groups or repeated measurements clustered within individuals.

Translational Abstract

In psychological research, two-level models are used to analyze data that are hierarchically structured. Such hierarchies in data can occur when participants are clustered within groups or repeated measurements are made for the same participants. Hierarchically structured data lead to quite complex dependencies among variances: (a) the outcome variable contains variance at two different levels, (b) predictor variables at both levels relate to outcome variance at the respective level (direct effects), (c) the size of the effect of a predictor variable on the lower level can vary between clusters at the higher level and (d) this variation can be explained by predictor variables of the higher level (so called cross-level interaction effects). All these variances and their dependencies must be specified to estimate the likelihood of obtaining statistically significant effects in a two-level model—known as the statistical power. We provide a hands-on tutorial illustrating the specification of these parameters and the implementation of a power analysis in the statistical environment R. We also provide rules of thumb for the sample sizes necessary to detect an effect of a certain size with sufficient power.

Keywords: multilevel modeling, hierarchical linear model (HLM), mixed effect models, power analysis, sample size determination

Supplemental materials: http://dx.doi.org/10.1037/met0000195.supp

When significance testing is used to generalize an effect from a sample to the population, it is of utmost interest to consider the probability of obtaining a significant result given that an effect holds true for the population—also known as the statistical power of the test (Cohen, 1988). This is particularly true for (a) identifying sufficiently large sample sizes prior to a study; (b) specifying the effect size that can be detected with sufficient power given a certain sample size; and (c) improving the replicability of psychological science in the long run (e.g., Asendorpf et al., 2013; Bakker, van Dijk, & Wicherts, 2012; Open Science Collaboration, 2015). Although the power of psychological research does not seem to have increased since Jacob Cohen's pioneering work in 1962 (Bakker et al., 2012; Cohen, 1962; Maxwell, 2004; Sedlmeier & Gigerenzer, 1989), awareness of the issue has grown in recent years. That is, many algorithms and programs have been developed that provide fast and easy power calculations for several basic statistical procedures such as *t* tests, analyses of variance, and regression analyses (e.g., Faul, Erdfelder, Buchner, & Lang,

This article was published Online First September 27, 2018.

Matthias G. Arend and Thomas Schäfer, Department of Psychology, Chemnitz University of Technology.

We gratefully thank Louisa Glawe, Andreas Kallenborn and Yvonne Pajonk for their assistance in article preparation, and Axel Mayer and Bertolt Meyer for helpful comments on an earlier version of the article.

Correspondence concerning this article should be addressed to Matthias G. Arend, who is now at Institute of Psychology, RWTH Aachen University, Jägerstraße 17-19, 52066 Aachen, Germany. E-mail: matthias.georg.arend@rwth-aachen.de

2009). Yet, many psychological disciplines have undergone a shift toward employing more complex statistical models, particularly multilevel models, that require more complex power computations (Aguinis & Culpepper, 2015; Castro, 2002; Mathieu, Aguinis, Culpepper, & Chen, 2012). Therefore, to increase the replicability of psychological studies in the long run, researchers should be supported in estimating power for more complex models, such as multilevel models.

Methods to estimate the power for multilevel models are difficult to apply for less experienced users. The estimation of power in multilevel models is quite complex because sample sizes are allocated at multiple levels, several types of effects can be tested, and, compared with single-level analyses, more parameters for which a standardized terminology for power analysis is not yet available must be specified (see, e.g., Helms, 1992; Snijders & Bosker, 1993; Stroup, 2002). Although approaches to determining power or optimal sample sizes for two-level models have been in use for a few decades (Dziak, Nahum-Shani, & Collins, 2012; Mathieu et al., 2012; Scherbaum & Ferreter, 2009), most of these methods are limited to specific two-level models or specific types of effects (see also Overview of Methods for Power Estimation in Two-Level Models section). Yet, a recently developed method based on Monte Carlo simulation, SIMR (Green & MacLeod, 2016a, 2016b), overcomes these limitations by being able to estimate the power for all types of effects and the most common types of two-level models. Thus, our goal with the present article is (a) to provide general guidelines for the specification of input parameters for power analyses of two-level models, (b) to demonstrate how power analyses for the most common types of two-level models can be implemented in SIMR, and (c) to derive case-sensitive rules of thumb for sufficient sample sizes and minimum detectable effect sizes (MDESs) from a large-scale simulation study that was performed with SIMR. Consequently, the present article constitutes a tutorial for both users who want to conduct a specific power analysis for their two-level study and users who seek rules of thumb for the design of their two-level studies.

The lack of such a tutorial has contributed to the neglect of considering power or sufficient sample sizes in two-level models to the extent that doing so has been the exception rather than the rule (see Literature Review section). With this tutorial, our objective is to make a tool for power estimation in two-level models available to a wide audience of researchers, which will help increase the focus on power for these types of models. But first, we give a concise overview of the characteristics of statistical power and two-level modeling.

Introduction to Statistical Power and Two-Level Models

Statistical Power

Statistical power is the probability of obtaining a statistically significant result (i.e., rejecting the null hypothesis; Cohen, 1988) for an effect that holds true for the population when a certain sample size is drawn from this population. The effect for which statistical power shall be estimated may be any statistic that can be derived from a model. Power is primarily determined by the size of the standard error (larger standard errors yield lower power), the population effect size (larger effects yield larger power), and the preset level of significance, α (lower αs yield lower power). Because the effect size that holds true for the population cannot be influenced and α is typically set to .05, a researcher interested in increasing power should aim to decrease the size of the standard error, which itself is influenced by the residual variance (smaller residual variance yields smaller standard errors, e.g., through higher accuracy of the parameter estimation; Cumming & Finch, 2005), the covariance with other predictors (covariances generally increase the standard error; Snijders & Bosker, 1993), and—most importantly—the sample size (a larger sample size yields lower standard errors).

In general, the primary goal of power analyses is to determine the sample size that provides an adequately high level of power (*a priori power analysis*; Cohen, 1988). Therefore, one must specify the target level of power, the population effect size, and α as input parameters, from which this adequate sample size can be derived (i.e., sample size as a function of power, effect size, and α). Yet, researchers might also be interested in other types of power analyses (Faul et al., 2009), such as *post hoc power analysis* (power as a function of sample size, effect size, and α), *criterion analysis* (α as a function of power, sample size, and effect size), or *MDES analysis* (also *sensitivity analysis*; effect size as a function of power, sample size, and α; see, e.g., Bloom, 1995). It is important to note that although these power analyses focus on different parameters, they use essentially the same information. That is, one parameter is always estimated as a function of the other parameters and the different power analyses are (theoretically) all conversions of the original power formula (Cohen, 1988). For example, if a sample size of 50, an effect size of .30, and an α of .05 yield a power of .80 (post hoc power analysis), an MDES analysis based on a power of .80, sample size of 50, and α of .05 will yield an effect size of .30.

To sum up, a power analysis always requires the specification of input parameters, on the basis of which different types of power analyses can be conducted. In the following sections, we describe (a) how the input parameters for power analyses in two-level models can be specified, and (b) how power analyses (a priori and post hoc) for two-level models can be conducted. But we begin with a short introduction on two-level modeling.

Two-Level Models

Multilevel models are a statistical method for analyzing relationships for data that are hierarchically structured (i.e., lower level units are clustered within higher level units; Hox, 1998, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). As such, by accounting for relationships at and between all levels of the data, multilevel modeling extends single-level statistical methods, in which units must be independent from each other (i.e., no such clustering occurs; Snijders & Bosker, 2012). Other terms for multilevel models are (linear) mixed effect models (e.g., Cao & Ramsay, 2010; Cudeck & Klebe, 2002), random coefficient models (Longford, 1993), multilevel covariance component models (e.g., Goldstein, 1987), hierarchical linear models (e.g., Lindenberger & Pötter, 1998; Raudenbush & Bryk, 2002), and multilevel regression models (Bauer & Curran, 2005). In the present study, we use the terminology of hierarchical linear models (HLMs; Raudenbush & Bryk, 2002; Raudenbush & Liu, 2000).

Two-level models are used when data clustered on two levels are analyzed. In psychology, there are two dominant types of data analyzed with two-level models: (a) individuals clustered within clusters such as teams or groups (e.g., organizational/applied/social psychology), schools or classes (e.g., educational psychology), or health care institutions or therapy practices (health/clinical psychology); and (b) repeated measurements clustered within individuals (i.e., longitudinal data).

Independent of the type of data, when these are hierarchically structured on two levels, it can be expected that the outcome variable also contains variance at these two levels: The lower level (L1) variance component (σ^2) resides within clusters (i.e., between units referring to one cluster) and the higher level (L2) variance component (τ_{00}) resides between clusters. Thus, when the mean of an outcome variable Y_{ij} that contains L1 units (indicated by lowered *i*) within clusters (indicated by lowered *j*) is to be estimated, a grand mean and a cluster-specific deviation can be modeled by one equation at each level (describing the *null model*):

$$Y_{ij} = \beta_{0j} + R_{ij} \quad (1)$$

$$\beta_{0j} = \gamma_{00} + U_{0j} \quad (2)$$

Equation 1 describes the estimation of a mean value (intercept β_{0j}) for each cluster and the deviation of each within-cluster unit from β_{0j} (R_{ij}) on L1. On L2, Equation 2 combines the cluster-specific intercepts into one general estimate of the mean (γ_{00}) and the deviation of each cluster-specific mean from the grand mean γ_{00} (U_{0j}). The sample-wide collection of the R_{ij} is σ^2, the L1 variance component; and the sample-wide collection of U_{0j} is τ_{00}, the L2 variance component:

$$\sigma^2 = \text{var}(R_{ij}) \quad (3)$$

$$\tau_{00} = \text{var}(U_{0j}) \quad (4)$$

The share of the L2 variance in the total (L2 and L1) variance is indicated by the *intraclass correlation coefficient* (ICC; ρ) of the null model:

$$\rho = \frac{\tau_{00}}{(\tau_{00} + \sigma^2)} \quad (5)$$

The ICC of the null model can vary between 0 and 1, with $\rho = .00$ indicating no variance between clusters and $\rho = 1.00$ indicating no variance within clusters (yet, two-level models should be used in both cases; Nezlek, 2008). Two-level models draw on these variances when the influence of the L1 predictors on the L1 outcome variance (L1 direct effect) and the influence of the L2 predictors on the L2 outcome variance (L2 direct effect) are modeled. This is (theoretically) done by regressing the outcome on the L1 predictor for each cluster. Consequently, each cluster can have its cluster-specific intercept and slope. These between-cluster differences in the intercepts and slopes are the outcomes that are regressed on L2 predictors: In a *means-* (Raudenbush & Bryk, 2002) or *intercepts-as-outcomes model* (Hofmann, 1997), the L2 predictor explains L2 variance (i.e., intercept variance) in the outcome; in a *slopes-as-outcomes model*, the L2 predictor explains differences in the cluster-specific slopes (*cross-level interaction [CLI] effect*; Kahn, 2011; Nezlek, 2008). Correspondingly, the equations for L1 and L2 direct effects and CLI effects are as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + R_{ij} \quad (6)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + U_{0j} \quad (7)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + U_{1j} \quad (8)$$

Each equation itself is characterized by an intercept, a slope, and a residual term. Equation 7 contains the fixed intercept (i.e., the average intercept of all clusters), the L2 direct effect (γ_{01}), and an error term, the residuals of the cluster-specific intercepts (U_{0j}). Equation 8 contains the L1 direct effect (i.e., the average slope of the L1 predictor on the outcome; γ_{10}), the CLI effect (i.e., the effect of the L2 predictor on the cluster-specific slopes; γ_{11}), and an error term, the residuals of cluster-specific slopes (U_{1j}). The whole model will be most often depicted in one equation. Substitution and conversion yields

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} W_j + \gamma_{11} W_j X_{ij} + U_{0j} + U_{1j} X_{ij} + R_{ij} \quad (9)$$

This equation is adapted so that the elements can be separated into the fixed part (γ_{00}, γ_{10}, γ_{01}, γ_{11}) and the random part (U_{0j}, U_{1j}, R_{ij}) of the model. Correspondingly, these parts also contain the fixed effects (γ_{00}, γ_{10}, γ_{01}, γ_{11}) and the residual terms on which the variance components (σ^2, τ_{00}, and the third variance component, the random slope variance $\tau_{11} = var[U_{1j}]$) are based. These are the parameters for which significance can be tested in these models. In that, fixed effects are comparable with regression coefficients for which, in general, one- or two-sided tests can be implemented (i.e., directional or nondirectional hypotheses can be tested; Kimmel, 1957). Any fixed effect (γ) can be tested for significance with a test statistic that is achieved by dividing γ by its standard error (Snijders & Bosker, 2012; Wald, 1943), which itself is influenced by factors such as the predictor's variance and covariances with other predictors (for more details see Snijders & Bosker, 1993). If significance for more than one fixed effect is to be analyzed together, this can be done with the Wald test using the *F*-distribution (Snijders & Bosker, 2012; Wald, 1943). Furthermore, when sample sizes at one or both levels are low, approximations of the degrees of freedoms or test statistic should be used (such as the Satterthwaite or Kenward–Roger approximation; Kenward & Roger, 1997; Satterthwaite, 1946). For the variance components, standard errors and significance can also be estimated. Yet, because they are variances and thus larger than zero, it has been considered unreasonable to compute standard errors for variance components (Baayen, Davidson, & Bates, 2008). Instead, the deviance test is often used for significance testing of variance components (Snijders & Bosker, 2012).

A power analysis in two-level models requires the specification of additional input parameters because predictors at two levels explain outcome variance at two levels, the effects of L1 predictors can differ between clusters, and the differences in these effects can be explained by L2 predictors. That is, one must specify not only the significance level, the sample sizes, the target level of power, and the effect sizes but also the sizes of the variance components and the covariance of the random slope and random intercept. Consequently, the power analysis in two-level models requires a standardized specification of additional parameters and is thus much more complex compared to power analyses of single-level models. Therefore, in the present article we seek to provide guidance for power analysis in two-level models by illustrating how

these input parameters can be specified for the most commonly used types of two-level-model effects.

Review of Demands on and Methods for Power Analysis in Two-Level Models

We begin with a literature review we conducted to detect the types of two-level effects that are frequently studied by psychologists. To help psychologists identify the methods that can estimate the power for these types of effects, an overview of methods developed for the power analysis in two-level models is furthermore provided below.

Literature Review: Frequently Studied Effects in Two-Level Models

The literature review was based on a search term that included the most common notations for multilevel models of articles published between 2000 and 2016 in the *Journal of Abnormal Psychology, Journal of Applied Psychology, Journal of Educational Psychology*, and *Journal of Personality and Social Psychology*. These journals contain research from a broad range of psychological disciplines, such as organizational, educational, applied, social, and personality psychology—all disciplines that tend to use multilevel modeling frequently.

The literature search yielded 269 studies, 151 of which included two-level models that were applied to a total of 185 samples (some studies examined more than one sample). Ninety-one of these samples contained longitudinal data and 90 included individuals at L1 (of the remaining, two studies modeled groups and one study each departments and teams at L1).

In general, these samples were used to analyze hypotheses referring to L1 direct effects (182), L2 direct effects (122), CLI effects (107), and variance components (27, in five of these, the random slope variance component was required to be significant for examining CLI effects). Models examining a combination of L1 direct, L2 direct, and CLI effects were by far the most frequent (43.8%), followed by a combination of L1 and L2 effects (20.5%), L1 effects alone (20.0%), and L1 direct combined with CLI effects (14.1%). Hypotheses covering variance components (11.9%) were always combined with hypotheses about fixed effects. Consequently, it seems that variance components are frequently modeled (which is important for the adequate specification of a model) yet often not specified in terms of a hypothesis. Hence, we focus on power estimation for fixed effects with models that include random effects (i.e., all types of variance components) in the following tutorial.

Furthermore, the sample sizes and ICCs of the null models—if reported in the articles—were collected. A null model for $n = 197$ ICCs nested in $N = 107$ samples was estimated. The average estimate of the ICC was .30 (*Mdn* = .24). For samples with individuals at L1, the estimate of the average ICC was .19 (*Mdn* = .17) and for samples with repeated measurements at L1 it was .42 (*Mdn* = .40). Because the sample sizes (for unbalanced data sets, the mean sample size was recorded) at L1 (range = 2–294) and at L2 (range = 5–17,385) varied greatly, only the median (plus the lower quartile Q1 and the upper quartile Q3) is reported here: The average sample sizes were *Mdn* = 11 (Q1 = 5; Q3 = 22) at L1 and *Mdn* = 100.5 (Q1 = 50.75; Q3 = 196) at L2.

Finally, we also used the analyzed studies to obtain an impression of how multilevel researchers have dealt with estimating and reporting power. In almost two out of three samples (120; 64.8%) researchers did not remark on statistical power at all. Of the remaining 65 samples, for 38 (20.5%), power considerations were made based on informed guesses (without quantitative estimation), for eight (4.3%) on analyses other than the two-level models, and for five (2.7%) on references/literature. The remaining five samples (2.7%) included specific considerations on factors influencing power (such as replicability or the significance level). Nine samples (4.8%) contained an exact determination of power (by calculation or simulation). Once again, this lack of estimating and reporting of statistical power indicates the need to provide psychologists with basic guidance regarding the estimation of statistical power in multilevel models.

In summary, the results from the literature review suggest that the types of effects most frequently studied are particularly fixed effects (yet, in models that also contain random effects). Thus, to support most psychologists, power analysis methods should be able to estimate the power and required sample sizes for all types of fixed effects in two-level models that include fixed and random effects.

Overview of Methods for Power Estimation in Two-Level Models

The question of how to choose and allocate the sample sizes in two-level research to achieve a target level of power has been addressed in prior work with two approaches: (a) approximate formulas and (b) simulation methods. The most important methods of both approaches are summarized in Table 1. In the following sections, we give a brief description of approximate formulas versus simulation methods and introduce the method used for estimating power in this tutorial.

One way to estimate statistical power in two-level models is to use approximate formulas. These formulas calculate the approximate size of the standard error of an effect based on several input parameters, particularly the two sample sizes and the size of the variance components (and/or the ICC). The major advantage of these equations is that they can be converted so that each of the parameters can be estimated by the others (e.g., sample sizes can be estimated as a function of the effect size, the power, and α). Yet, formulas often refer to specific models and, thus, have only limited flexibility (Arnold, Hogan, Colford, & Hubbard, 2011). This is because adding additional predictors or variance components complicates their deduction to the extent that a definition of such formulas has been considered impossible (Cools, Van den Noortgate, & Onghena, 2008; Snijders & Bosker, 2012). Furthermore, approximate formulas rely on several assumptions that must be met, which is the reason for their approximate nature and narrows their usability (Cools et al., 2008). As outlined in Table 1, multiple formulas for the computation of fixed effects have been identified, as well as one formula for L2 variance components. To our knowledge, there is no approximate formula for random slopes (this is also reflected in articles and chapters that have provided power calculations for multilevel models; see Scherbaum & Ferreter, 2009; Snijders & Bosker, 2012). Most of these formulas are implemented in small software packages for direct power estimation (see Table 1).

Table 1
Overview of Methods for Power Estimation in Two-Level Models

Method	Types of effects					Notes (and most important references)
	L1	L2	CLI	RI	RS	
Power estimation by approximate formulas						
PinT	X	X	X	—	—	Computer program to estimate the SE of fixed effects for continuous outcomes. PinT computes various models (all relevant input parameters can be specified) and provides optimal sample sizes under budget constraints (Bosker, Snijders, & Guldemond, 2003).
OptimalDesign	X	X	—	—	—	Computer program that estimates the power for fixed effects for several types of outcomes, numbers of covariates and, specifically, for longitudinal data. Contains MDES-analyses (Raudenbush, 1997; Raudenbush & Liu, 2000; Spybrook et al., 2011).
Longford	—	—	—	X	—	Formula for direct computation of the SE of random intercepts (Longford, 1993).
Power estimation by simulation (all executable in the statistical environment R)						
SIMR	X	X	X	X	X	SIMR runs power analyses for all types of models with different types of outcomes, input parameters, significance tests, and ranges of sample sizes. The simulation can be based on either a real or a simulated dataset. SIMR uses the packages that are conventionally used for multilevel analysis in R (Green & MacLeod, 2016a, 2016b).
MLPowSim	X	X	—	—	—	MLPowSim simulates power for different types of multilevel models (either in MLwiN or in R; Rasbash, Steele, Browne, & Goldstein, 2017). After the specification of input parameters, the necessary code is created by the program (Browne, Lahi, & Parker, 2009)
ML Power Tool	—	—	X	—	—	ML Power Tool is an online simulation tool to estimate the power of a CLI effect (see https://aguinis.shinyapps.io/ml_power/) in a model with one predictor at each level of analysis (Mathieu, Aguinis, Culpepper, & Chen, 2012).
PAMM	—	—	—	X	X	PAMM simulates power for random intercepts and slopes variance components based on a deviance test used for testing significance (Martin, 2015; Martin, Nussey, Wilson, & Re, 2011)

Note. There are other methods that are not included here because they were not sufficiently documented in widely accessible manuals (e.g., nlmeU; Gałlecki & Burzykowski, 2013, including formulas and simulation), they referred to corrected sample size determinations originally meant for single-level analyses (see, e.g., Longpower; Donohue & Edland, 2016; Liu & Liang, 1997) or to very specific cases of two-level data sets that are clustered within crossed time points and can thus be regarded as three-level models (clusterPower; Reich, Myers, Obeng, Milstone, & Perl, 2012; see also Dziak, Nahum-Shani, & Collins, 2012), or they were two-level models without random slopes and intercepts (sim.glmm, which corresponds to covariance analysis; Johnson, Barry, Ferguson, & Pie, 2015). L1 = L1 direct effect; L2 = L2 direct effect; CLI = Cross-level interaction; RI = Random intercept; RS = Random slope; MDES = minimum detectable effect sizes.

The other way to estimate statistical power of multilevel effects is to use Monte Carlo simulation. To do so, several parameters must be specified by the user (e.g., sample sizes, fixed effects, and variance components)—similar to power analysis with approximate formulas. These parameters are those that are expected to hold true for the population, on the basis of which data sets are repeatedly created. Subsequently, the significance of the effect of interest is estimated in each of these data sets and, because the parameters specified by the user count as the population parameters, the share of significant replicates equals the statistical power (Arnold et al., 2011; Cools et al., 2008; Johnson, Barry, Ferguson, & Müller, 2015). Consequently, when an a priori power analysis is conducted, those steps must be iterated with different sample sizes as input parameters until the target level of power is reached. In general, power simulations are very flexible (different models and tests can be used) and, implemented as functions in statistical packages, they are user-friendly, efficient to use, and easy to implement (Martin, Nussey, Wilson, & Re, 2011). There are many tools that can estimate the power for at least one fixed effect or random effects in two-level models by simulation (see Table 1 for an overview). Most of these tools must be used in (or in combination with) the statistical environment R (R Core Team, 2015).

In summary, in our overview of methods for two-level power estimation, SIMR was the only method able to estimate the power for all types of effects, all types of models, and all types of significance tests. Another valuable advantage of SIMR is that it relates to the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015), which is typically used to conduct multilevel analyses in R. Consequently, its flexibility, connection to data analysis in R, and user friendliness make SIMR the method psychologists will likely find the most beneficial when conducting power analyses. Hence, the implementation of a priori and post hoc power analysis in two-level models is described in SIMR.

Objectives

In general, the same input parameters must be specified for a power analysis in a two-level model regardless of whether formulas or simulations are used. Consequently, our first objective is to

provide guidelines for how to adequately specify these input parameters.

Second, in terms of flexibility and usability, SIMR was the only method able to estimate the power for all commonly studied types of two-level effects. Therefore, our second objective is to describe how a priori as well as post hoc power analyses can be conducted in SIMR.

Finally, to support researchers who are looking for quick guidance on how to design a two-level study, we conducted a large-scale simulation study with SIMR from which rules of thumb for determining adequate sample sizes as well as the MDESs that yield a high power (i.e., a power of .80 or more; cf. Cohen, 1992) for the typical two-level effects in psychological research can be derived. Thus, readers who do not want to delve into the details of exact power estimation might go directly to the section Rules of Thumb and MDESs for Power Analysis in Two-Level Models.

Power Analysis in Two-Level Models

Specification of Input Parameters

To specify the input parameters for a power analysis in two-level models, researchers must decide which fixed, random, and cross-level interaction effects will be included in the analysis. Table 2 gives an overview of the types of two-level models most typically studied by psychologists, as well as the input parameters that must be specified for the power analyses in these two-level models.

Once the model and type of two-level effects that will be studied are selected, the input parameters must first be specified in a standardized way. For example, a standardized indicator for the strength of the association between a predictor variable and the outcome (i.e., the fixed effects), or the size of the proportion of the L2 variance in the total variance (i.e., the ICC) must be specified. Second, a population model for the power analysis must be derived from the standardized input parameters. That is, the size of the unconditional variance components is derived from the ICC, the fixed effect estimates are derived from the standardized effect sizes and the sizes of the unconditional variance components, and the conditional variance components are deduced from the unconditional ones. Finally, more general specifications for the power analysis, such as the significance level or the target level of power, must also be made. The following sections describe these steps in detail and provide guidelines for the specification of standardized input parameters. These guidelines refer to the model from Equations 6–9.

Specification of standardized input parameters. The parameters for which standardized effect sizes must be specified are the fixed effects, the random slope variance, the ICC, and the correlation between random intercept and random slope. The following sections provide guidelines on how to choose these input parameters.

Standardized effect sizes for fixed effects. In general, one can draw on the magnitude of effects that have typically been found in previous studies in the same area of research to determine small, medium, and large standardized effect sizes, or use Cohen's conventions (Cohen, 1988, 1992). For single effects of continuous predictor variables on the outcome, one can use the conventions for product-moment correlation coefficients (γ_{std} = .10 for small, γ_{std} = .30 for medium, and γ_{std} = .50 for large effect sizes) to describe the strength of this association (e.g., Maas & Hox, 2005). It should be noted that for two-level models, these conventions apply to relationships of predictor variables to different types of variances in the outcome: L1 direct effects relate L1 predictor variables to L1 outcome variance, L2 direct effects relate L2 predictor variables to L2 outcome variance, and CLI effects relate L2 predictor variables to random slope variance. Accordingly, a small L1 direct effect ($\gamma_{10.std}$) explains 1% ($R^2 = .10^2$) of the L1 variance component, a medium L2 direct effect ($\gamma_{01.std}$) 9% ($R^2 = .30^2$) of the L2 variance component, and a large CLI effect ($\gamma_{11.std}$) 25% ($R^2 = .50^2$) of the random slope variance component (see also Aarts, Verhage, Veenvliet, Dolan, & van der Sluis, 2014).

Standardized effect size for the random slope variance component. Because variances are usually not covered by specific hypotheses in psychological research, it is quite unusual to have specific expectations about their magnitude in the population and good estimates for the random slope variance component are hard to make. Again, if no other estimates are available from prior research, one should ask to what extent the effects of the L1 predictor on the criterion can be assumed to vary between clusters (using the standardized conventions). For example, if one expects that 68% of the cluster-specific slopes lie within a standard deviation of .10 above/below the fixed effect estimate, this indicates a standardized random slope variance component of .01 (the squared

Table 2
Input Parameters That Must Be Specified for Power Analyses in Two-Level Models

	Two-level model including:					Input parameters								
	L1 predictor	L2 predictor	Random intercept	Random slope	CLI effect	$\gamma_{10.std}$	$\gamma_{01.std}$	$\gamma_{11.std}$	ρ	$\tau_{11.std}$	$Cor(U_{0j}, U_{1j})$	α	*n*	*N*
M1	No	No	Yes	No	No	—	—	—	X	—	—	X	X	X
M2	Yes	No	Yes	No	No	X	—	—	X	—	—	X	X	X
M3	No	Yes	Yes	No	No	—	X	—	X	—	—	X	X	X
M4	Yes	Yes	Yes	No	No	X	X	—	X	—	—	X	X	X
M5	Yes	No	Yes	Yes	No	X	—	—	X	X	X	X	X	X
M6	Yes	Yes	Yes	Yes	Yes	X	X	X	X	X	X	X	X	X

Note. α = significance level; *n* = L1 sample size; *N* = L2 sample size; $\gamma_{10.std}$ = L1 direct effect; $\gamma_{01.std}$ = L2 direct effect; $\gamma_{11.std}$ = CLI effect; ρ = ICC; $\tau_{11.std}$ = random slope variance component; $Cor(U_{0j}, U_{1j})$ = slope-intercept correlation; CLI = Cross-level interaction. M1 refers to a null model, M3 to an intercepts-as-outcomes model, M5 to a random-slopes model, and M6 to the full model (Equations 6–9). M1–M6 cover the models that are typically studied by psychologists (see Literature Review section).

value of this standard deviation). This can, in keeping with the standardized size of the standard deviation, be regarded as a small random slope variance component. In this way, one can use the squared values of Cohen's suggestions for the product-moment correlation coefficient as estimates for the magnitude of variances. Consequently, it additionally follows that if 68% of the cluster-specific slopes lie within a standard deviation of .30 around the L1 direct effect (cf. Spybrook et al., 2011), the standardized random slope variance is medium ($\tau_{11.std} = .09$); and if 68% of the cluster-specific slopes lie within a standard deviation of .50 around the L1 direct effect, it is large ($\tau_{11.std} = .25$). These values include the range of random slope variance components typically chosen for power estimation in two-level models (e.g., Martin et al., 2011; Mathieu et al., 2012; Raudenbush & Liu, 2000). When no sufficiently informed estimator for the size of the random slope is available, choosing a medium effect size ($\tau_{11.std} = .09$) is a good option, because small random slope variances would yield anti-conservative estimates and large slope variances would yield over-conservative estimates of the power for CLI effects.

Intraclass correlation coefficient. In general, the ICC is a standardized indicator for the relation of the L2 outcome variance to the total outcome variance. In the literature, different values that generally lie between .05 (small) and .30 (large) have been chosen to reflect small, medium, and large ICCs (Bliese, 2000; Dziak et al., 2012; James, 1982; Maas & Hox, 2005; Mathieu et al., 2012; Snijders & Bosker, 1999). Yet, the mean ICC of the studies in the literature review presented above was $\rho = .30$. Consequently, we propose regarding this as the medium size of the ICC. Furthermore, if the L1-units are individuals, one would expect a rather small ICC and if they are repeated-measurement occasions, one would expect rather large ICCs (see Literature Review section). Therefore, an ICC of $\rho = .10$ can be considered small, and an ICC of $\rho = .50$ large.

Correlation between random slope and random intercept. If the model includes a random slope, its correlation with the random intercept, $Cor(U_{0j}, U_{1j})$, should be specified. To achieve a generally conservative power estimate one should set this correlation to .00 (Martin et al., 2011). This will not greatly influence the results, because the standard errors of fixed effects are not sensitive to the correlation (Bosker, Snijders, & Guldemond, 2003; Snijders & Bosker, 1993). Yet, if necessary, one can also specify a correlation other than zero.

Derivation of a population model for the power analysis. Before implementing the power analysis, a population model for the power analysis must be derived from the standardized input parameters. Therefore, one must make assumptions about the scale of the outcome and predictor variables and adjust the standardized input parameters to their scale. How these adjustments are made is illustrated in the following sections.

Specification of predictor variables. Basically, it is possible to specify predictor variables with all kinds of scales. Nonetheless, by taking a closer look at the literature on power analysis in two-level models, it becomes evident that predictor variables are often specified as being *z*-standardized (based on their total variance, between-cluster variance, or within-cluster variance), implying $M = 0$ and $SD = 1$ (consequently $Var = 1$; e.g., Bosker et al., 2003; Dziak et al., 2012; Hox, 2010). The L2 predictor (W_j; standardization based on its total variance) thus has a variance of 1. For the L1 predictor (X_{ij}), cluster-mean centering (i.e., a *z*-standardization for each cluster based on the within-cluster variance) is adequate for examining L1 and L2 direct effects as well as CLI effects (Enders & Tofighi, 2007). Consequently, we recommend specifying the L2 predictor with total variance equal to 1 and the L1 predictor with within-cluster variance equal to 1.

Specification of the outcome variable. In two-level models, the total outcome variance can be partitioned into the L1 and the L2 variance component, and the relation of the L2 variance component to the total outcome variance equals the ICC. To analyze power in two-level models, one must first derive the outcome variance components from the ICC. A common practice in this respect is to fix the L1 variance component at a specific value and to subsequently derive the L2 variance component from the ICC (e.g., Dziak et al., 2012; Maas & Hox, 2005; Mathieu et al., 2012). In line with this, we recommend fixing the L1 variance component of the null model to $\sigma^2 = 1.00$. Converting Equation 5, one can then derive the size of the L2 variance component by applying the formula from Equation 10:

$$\tau_{00} = \frac{\rho_{std}}{(1 - \rho_{std})} \tag{10}$$

For instance, from $\rho_{std} = .30$ and $\sigma^2 = 1.00$ it follows that $\tau_{00} = .43$. Furthermore, from these unconditional variance components ($Var_{unconditional}$) that refer to the null model (which does not account for the effects of predictors), the conditional variance components ($Var_{condiditional}$; i.e., the unexplained variance that remains after accounting for the effects of the predictors) must now be derived for the population model. Therefore, each $Var_{condiditional}$ is derived from $Var_{unconditional}$ and the respective *standardized* effect size γ_{std} on each level (see Equation 11; cf. Bosker et al., 2003; note that the formula would be different for two or more *correlated* predictors at either level).

$$Var_{conditional} = Var_{unconditional}(1 - \gamma^2_{std}) \tag{11}$$

For example, $Var_{unconditional} = .43$ and $\gamma_{std} = .30$ yield $Var_{conditional} = .39$. Consequently, the population model conditional L1 variance component $\sigma^2_{Y|X}$ is derived from σ^2 and the standardized effect size of the L1 predictor ($\gamma_{10.std}$), and the population model conditional L2 variance component $\tau_{00Y/W}$ is derived from τ_{00} and the standardized effect size of the L2 predictor ($\gamma_{01.std}$).

Note that one must also specify the fixed intercept of the null model (i.e., the mean of the outcome variable Y_{ij}). Consistent with other research in this area (e.g., Bosker et al., 2003; Mathieu et al., 2012), we propose here to set the fixed intercept to 0 if its size cannot be derived from hypotheses.

Adjustment of the random slope variance. Because random slopes are cluster-specific effects of a L1 predictor on the L1 outcome variance, the random slope variance component must refer to the scale of the L1 variance component. Therefore, the standardized random slope variance component ($\tau_{11.std}$) must often be adjusted, yielding the unconditional random slope variance component (τ_{11}).

As suggested by an anonymous expert reviewer, this can be understood by considering the index of slope reliability (ISR; Willett, 1989). The ISR is a measure of the reliability of slopes (Rast & Hofer, 2014), comparable with the ICC2 (a measure of reliability of the intercepts which accounts for the sample size at L1; Bliese, 2000; James, 1982). The ISR (ρ_{slope}) is based on the random slope variance component (τ_{11}), the L1 variance compo-

nent (σ^2), and the sum of squared deviations of each L1-unit from the cluster mean, which, when the L1 predictor variable is *z*-standardized within clusters, equals the cluster size *n*. Consequently, Equation 12 can be derived from Willett (1989; see also Rast & Hofer, 2014):

$$\rho_{slope} = \frac{\tau_{11}}{\tau_{11} + \left(\frac{\sigma^2}{n}\right)} \quad (12)$$

Because the same size of the random slope would yield higher ISRs for higher L1 variance components, and a higher ISR is related to higher power (Rast & Hofer, 2014), the standardized value of the random slope must be adjusted to the size of the L1 variance component. This leads to Equation 13:

$$\tau_{11} = \tau_{11.\text{std}} \times \sigma^2 \quad (13)$$

For example, with $n = 30$, a medium value for the standardized random slope of $\tau_{11.\text{std}} = .09$ yields the same ISR of .73 for an L1 variance component of .90 as for an L1 variance component of .50. Furthermore, Equation 13 also implies that when the size of the unconditional L1 variance component has been fixed at 1 (as described above), the adjusted unconditional random slope variance component equals the standardized one specified above. This illustrates the advantage of fixing the L1 variance component at 1.

Like the other variance components, the *conditional* instead of the unconditional random slope variance must be used for the population model. The population model conditional random slope variance component τ_{11YlXW} can consequently be derived from τ_{11} and the standardized effect size of the CLI effects ($\gamma_{11.\text{std}}$) by applying Equation 11.

Covariance of random intercept and slope. Furthermore, one must deduce the covariance of the random intercept and random slope from the correlation coefficient specified above. If, as advised, the correlation coefficient is set to .00, the covariance also equals .00. Yet, if one has specified another size for this correlation, for example, $Cor(U_{0j}, U_{1j}) = .30$, the covariance of these values can be estimated applying Equation 14:

$$Cov(U_{0j}, U_{1j}) = Cor(U_{0j}, U_{1j}) \times \sqrt{\tau_{00}} \times \sqrt{\tau_{11}} \quad (14)$$

Adjustment of fixed effects. Furthermore, it is also necessary to adjust the standardized effect sizes to the respective outcome variances of the population model (the standardized effects hold only for the case when predictor and outcome variances at the respective level are 1.0), which yields the fixed effects for the population model. This can be done with the following equation (derived from Hox, 2010):

$$\gamma = \frac{\gamma_{\text{std}} \times SD_{\text{outcome}}}{SD_{\text{predictor}}} \quad (15)$$

In Equation 15, γ refers to the resulting adjusted fixed effect size (L1 or L2 direct effect or CLI effect) and γ_{std} to the respective standardized effect size. The $SD_{\text{predictor}}$ is always 1.0 (for L1 as well as L2 predictors). The SD_{outcome} is the square root of the respective *unconditional* variance component (i.e., the L1 and L2 variance components of the null model and the random slope variance component without CLI effect). Consequently, the equation for L1 direct effects is $\gamma_{10} = \gamma_{10.\text{std}}\sqrt{\sigma^2}$, for L2 direct effects $\gamma_{01} = \gamma_{01.\text{std}}\sqrt{\tau_{00}}$, and for CLI effects $\gamma_{11} = \gamma_{11.\text{std}}\sqrt{\tau_{11}}$.

Further specifications for the power analysis in two-level models. The sections above described how the input parameters specific to power analyses in two-level models can be specified and adjusted to yield a population model for the power analysis. Yet, there are further specifications that must be made for all power analyses, such as the significance level used, the sample sizes, and the target level of power.

Significance level. Power increases with a larger significance level α. Nonetheless, the decision to apply higher values for α to increase power should be based on careful considerations described elsewhere (Cohen, 1988). For the present approach we recommend setting α to .05, the level typically used in psychological research.

Minimum L1 and L2 sample sizes. When determining a possible range of sample sizes, it is also necessary to consider what the minimum sample sizes to achieve unbiased parameter (and standard error) estimates are. When restricted maximum likelihood (REML) is used to estimate the model parameters, a relatively small number of clusters (even down to 10; Maas & Hox, 2005) with a relatively small cluster size (down to five; cf. Maas & Hox, 2005) should be sufficient to provide unbiased estimates of fixed effects (see also Afshartous, 1995; McNeish & Stapleton, 2016a). Although REML does not completely account for the downward bias in standard errors for small samples, the nominal Type I error rate can be achieved for small samples by applying the Kenward–Roger approximation (Kenward & Roger, 1997) to the test statistic (McNeish & Stapleton, 2016b). In summary, a minimum of 10 clusters with a minimum cluster size of five can yield unbiased parameter estimates if the REML method is used for parameter estimation and the Kenward–Roger approximation is used for correcting the standard errors (for simulations, this can already be specified in the power analysis).

Target level of power. The target level of power plays a pivotal role: Researchers must decide what power they are trying to achieve. Cohen (1988, 1992) defined .80 as the lower limit of acceptable statistical power (high statistical power). Therefore, one should seek to achieve a target level of power of .80 or more (Cohen, 1988).

Specifications for power simulation. When conducting a power analysis via simulation it is also necessary to define the number of replications (or reiterations) and the parameter estimation method that shall be used. The higher the number of replications, the more accurate the power estimate will be (SIMR provides confidence intervals for the power estimates that can be used for assessing its accuracy). It is often recommended to use 1,000 or more replications (e.g., Green & MacLeod, 2016a, 2016b; Mathieu et al., 2012).

The method that is used to estimate the model parameters also has an impact on the parameter accuracy: In multilevel models, maximum likelihood (ML) and REML methods yield the most efficient and accurate parameter estimates (Kreft, 1996, as cited in Maas & Hox, 2004; van der Leeden, Busing, & Meijer, 1997), but only REML provides unbiased estimates of all standard errors for a small number of clusters (Browne & Draper, 2000; McNeish & Stapleton, 2016b). Therefore, when available, we recommend using REML for power estimation.

In summary, to conduct a power analysis with one of the methods provided in the section Overview of Methods for Power Estimation, one must first have specified and adjusted the input parameters described in the sections above. Using these input parameters, one can conduct the power analysis with any of the methods. Because we have identified SIMR as the most flexible method, the following section illustrates how a power analysis can be implemented in SIMR.

Implementation of a Power Analysis in a Two-Level Model in SIMR

This section illustrates how a power analysis is generally implemented in SIMR. Furthermore, we describe how this general implementation can be used for estimating a priori and post hoc power.

Power analysis in SIMR. This section illustrates how a power analysis is implemented in SIMR. First, Table 3 provides a general description of how to do this in R. Furthermore, two specific examples are provided, one for individuals within clusters and one for longitudinal data (measurements within individuals). These will be very basic examples but should give the reader an impression of how SIMR can be used.

As can be seen in Table 3, a few lines of R code are necessary for conducting the power analysis in SIMR. The code provided in this table refers to a model with random slopes and intercepts, in which the power for one L1 direct effect, one L2 direct effect, and one CLI effect is to be estimated (i.e., the model from Equation 9). For conducting the power analysis, R must be installed on the computer and the package SIMR must be downloaded and in-

Table 3
How to Estimate Power in Two-Level Models With SIMR in R

Steps	Description	R Code
Installing and loading SIMR	Download and installation of SIMR Load SIMR	install.packages("simr", dependencies = TRUE) library(simr)
Specifying standardized input parameters	Significance level $\alpha = .05$ L1 sample size $n = 20$ L2 sample size $N = 40$ L1 direct effect $\gamma_{10.std} = .30$ L2 direct effect $\gamma_{01.std} = .30$ CLI effect $\gamma_{11.std} = .50$ ICC $\rho = .30$ Stand. random slope $\tau_{11.std} = .09$	alpha.S <− .05 Size.clus <− 20 N.clus <− 40 L1_DE_standardized <− .30 L2_DE_standardized <− .30 CLI_E_standardized <− .50 ICC <− .30 rand.sl <− .09
Creating variables for power simulation	Creates a dataset with one L1-predictor *x* (standardized based on within-cluster variance) and one L2-predictor *Z* (standardized based on total variance)	x <− scale(rep(1:Size.clus)) g <− as.factor(1:N.clus) X <− cbind(expand.grid("x"=x, "g"=g)) X <− data.frame(X, Z = as.numeric(X$g)) X$Z <− scale(X$Z)
Adapting the standardized parameters	L1 variance component (σ^2) L2 variance component (τ_{00}) Random slope variance (τ_{11}) L1 direct effect (γ_{10}) L2 direct effect (γ_{01}) CLI effect (γ_{11})	varL1 <− 1 varL2 <− ICC/(1-ICC) varRS <− rand.sl*varL1 L1_DE <− L1_DE_standardized*sqrt(varL1) L2_DE <− L2_DE_standardized*sqrt(varL2) CLI_E <− CLI_E_standardized*sqrt(varRS)
Creating conditional variances	L1 variance ($\sigma^2_{Y\|X}$) L2 variance ($\tau_{00Y\|W}$) Random slope variance ($\tau_{11Y\|XW}$)	s <− sqrt((varL1)*(1-(L1_DE_standardized^2))) V1 <− varL2*(1-(L2_DE_standardized^2)) rand_sl.con <− varRS*(1-(CLI_E_standardized^2))
Creating a population model for simulation	Vector of fixed effects (fixed intercept, L1-, L2- direct and CLI effect). Random effects covariance matrix (with covariances set to 0). Function that creates a typical lme4-model (function: lmer). The model can have a fixed (1|g) or random (x|g) slope (when fixed, "VarCorr = V1" must be used). $x = \gamma_{00}$; $Z = \gamma_{01}$; $x{:}Z = \gamma_{11}$	b <− c(0, L1_DE, L2_DE, CLI_E) V2 <− matrix(c(V1,0,0,rand_sl.con), 2) model <− makeLmer(y ~ x + Z + x:Z + (x|g), fixef = b, VarCorr = V2, sigma = s, data = X) print(model)
Implementing the power analysis	Estimates power for *x* based on 1,000 repetitions (for more repetitions change nsim). For estimating power for *Z* or *x*:*Z*, use 'fixed("Z")' or 'fixed ("x:Z")' respectively. "kr" = Kenward Roger test (for other tests see Green & MacLeod, 2016a, 2016b)	sim.ef <− powerSim(model, fixed("x","kr"), alpha = alpha.S, nsim = 1,000) print(sim.ef)

Notes. CLI = Cross-level interaction; ICC = intraclass correlation coefficient. For less advanced users we recommend installing RStudio. All considerations are based on Green and MacLeod (2016a, 2016b).

stalled. Then, the input parameters for the power analysis are specified as described in the section Specification of Input Parameters. With these parameters, a population model is set up with the SIMR function makeLmer. Finally, the power analysis for each of the effects specified in this model can be conducted. When Supplementary Material A that contains the code from Table 3 (and "set.seed(123)") is used, the simulation yields a power of 1.00 for the L1 direct effect, a power of .47 for the L2 direct effect, and a power of .78 for the CLI effect. In addition to this general example, two practical examples are given below.

Example 1: Power for a study with individuals within clusters. Consider a researcher who is planning a study on team members within work teams in which individual behavior (e.g., performance) is determined by individual-level (e.g., tenure, a continuous variable, assumed to be cluster-mean centered) as well as team-level (e.g., team faultlines, a continuous variable, assumed to be grand-mean centered, see, e.g., Meyer & Glenz, 2013) predictor variables. It is necessary here to determine the power for the L1-direct and the L2 direct effect. Say the researcher has access to a sample of $N = 150$ teams, with average size $n = 5$. For individuals-within-clusters studies, the ICC tends to be rather low and could thus be set to .10. Tenure is expected to have a small standardized L1 direct effect of .10, and the team-level faultlines to have a standardized L2 direct effect of .50 on performance. The power estimation is based on 1,000 simulations. Because an L2 direct effect is expected, one would treat the intercept as random (the slope is fixed here because tenure is expected to have the same effect on performance in all teams). The final model, depicted in Equation 16, would correspond to M4 in Table 2 and looks like this:

$$Performance_{ij} = \gamma_{00} + \gamma_{10}\, Tenure_{ij} + \gamma_{01}\, Faultlines_j + U_{0j} + R_{ij} \quad (16)$$

After the standardized input parameters for the simulation are set, it is necessary to convert these values into parameters for the population model. This can be done by applying all the steps described in the section Specification of Input Parameters (or correspondingly with the R code from Supplementary Material B). Afterward, the simulation can be run. This process can be based on the description given in Table 3. The simulation reveals a power of .70 for the L1 direct effect and a power of .85 for the L2 direct effect.

Example 2: Power for a study with repeated measures within individuals. Researchers often use designs where the L1-units are repeated-measurement occasions (Bryk & Raudenbush, 1987). In this case, as outlined in the section Intraclass Correlation Coefficient, a high ICC of .50 is expected. One could, for instance, hypothesize that the competence in interacting with a certain technical system increases over time (within-person centered; L1 direct effect of time) but that this happens more quickly for younger and more slowly for older (L2 variable: age) individuals (CLI effect). One could assume a small size for the L1 direct effect and large size for the CLI effect (i.e., older people develop the competence *much more slowly* than younger people). Suppose that competence is measured for $N = 100$ participants every day for two successive weeks, yielding $n = 14$. The size of the standardized random slope is set to medium and the power estimation is based on 1,000 simulations. The final model would correspond to M6 in Table 2 and is depicted in Equation 17:

$$Competence_{ij} = \gamma_{00} + \gamma_{10}\, Time_{ij} + \gamma_{11}\, Age_j\, Time_{ij} + U_{0j} + U_{1j}\, Time_{ij} + R_{ij} \quad (17)$$

Again, the R code from Supplementary Material C can be used to specify the population model according to the input parameters. For this simulation model, a power of .78 can be found for the L1 direct effect and a power of .99 for the CLI effect.

A priori power analysis in SIMR. In most cases, researchers perform a power analysis to identify the sample sizes that are sufficiently large to identify an effect that is expected to hold true for the population (i.e., a priori power analysis). To determine the optimal sample size, one must repeat the power analysis by adapting the sample sizes until the target level of power is achieved. How this can be done is illustrated with the example for the two-level studies with individuals nested within teams described above. If the researcher wants to achieve a power $\geq .80$ for the medium L1 direct effect (for 150 teams with five members per team, the power equals .70), he or she should increase the sample size until the target level of power is reached. Sampling additional teams is likely very difficult for the researcher, so he or she would try to increase power by increasing the number of members per team. A subsequent power analysis with 150 teams and six members per team reveals that a power of .81 for the L1 direct effect could be achieved with this sample size. Consequently, the researcher should try to sample additional team members to achieve a power $\geq .80$ for the L1 direct effect.

Post hoc power analysis in SIMR. Because the a priori power estimates for effects from a certain multilevel model strongly depend on the input parameters chosen, it is advised here (and elsewhere, e.g., Mathieu et al., 2012) to also estimate post hoc power. For instance, consider that a high ICC was used as input parameter for an a priori power estimation for an L2 direct effect but in the sample data, the ICC was low. This would result in significantly decreased power for the same standardized effect size. To avoid misinterpretation based on the a priori power analysis (e.g., the interpretation that a certain effect likely does *not* hold true for the population in the case of nonsignificance), a post hoc power analysis that is based on the input parameters that are derived from the sample and the effect that was hypothesized to hold true for the population before the study should be conducted (Cohen, 1988). For instance, let us assume that an expected medium sized effect was not significant but a priori power was .80 for a given set of input parameters. The nonsignificance could be a result of misspecification of input parameters other than the effect size. In this respect, a post hoc power analysis can help avoid fallacies (Onwuegbuzie & Leech, 2004): If the post hoc power was markedly smaller than the a priori power, this difference could be due to differences in the specification of input parameters other than the effect size (e.g., the ICC or random slope variance). In this case, one should estimate the effect size that yields a power of .80 (i.e., the MDES; O'Keefe, 2007). Consequently, one can argue that the power to detect an effect of this size would have been adequate and that it thus seems

unlikely that an effect of this size or larger holds true for the population.

To estimate the post hoc power, one must specify a null model with the empirical data set. From this null model, all input parameters can be derived except for the standardized effect sizes. Instead, the standardized effect sizes that were used for the a priori power analysis can be adapted to the scale of the sampled variables using Equation 12. For the resulting unstandardized effect size, one can now implement a power analysis as described above in this section to estimate post hoc power. It should be reported to what extent the post hoc power differs from the a priori power, and the MDES for which post hoc power is $\geq .80$.

Rules of Thumb and MDESs for Power Analysis in Two-Level Models

Designing multilevel studies has long been based on rules of thumb, such as the 30/30 rule (i.e., 30 clusters on L2, each with cluster size 30 on L1; Kreft & De Leeuw, 1998) or the 50/20 rule by Hox (1998). The validity and credibility of these rules has been disputed and it has been shown that they do not hold for most situations (Mathieu et al., 2012). However, for many researchers striving for acceptable power, such rules of thumb can constitute important benchmarks in the process of designing a study, in particular when they simply do not have the resources to delve into power estimation as described in the previous section. Thus, the purpose of the present section is to provide guidelines for fast and frugal power estimation while avoiding the problems of all-too-simple rules of thumb. More precisely, case-sensitive rules of thumb for each combination of effect size (small, medium, large) and size of the ICC or random slope variance component (small, medium, large) are given that cover most of the sample sizes typically used in two-level research (i.e., L1 sample sizes between three and 30 and L2 sample sizes between 30 and 200).

Therefore, we performed a large-scale simulation-based MDES analysis, which provides the standardized effect size that can be detected with a power of .80 for each combination of input parameters. The MDES approach should be particularly helpful for two-level modeling, because it might not always be possible to sample sufficient units to achieve the necessary power to detect, for example, small effects (e.g., when the number of accessible participants is limited, or the budget is constrained). In these cases, researchers can estimate the size of the effect that can still be detected with sufficient power and decide whether it is still worth conducting their study.

From the results of the simulation study, three guidelines are provided: The first guideline comprises general recommendations for designing a two-level study with sufficient power. The second guideline comprises rules of thumb from which sample sizes necessary to detect certain effects with a power of .80 can be derived. The third guideline comprises the results of the MDES analysis which can be used for a priori as well as post hoc power estimation.

Simulation Design

The model that was used for the MDES analysis contained one predictor at each level and all types of fixed effects (one L1 direct effect, one L2 direct effect, and one CLI effect). Consequently, the intercept and slope were set to be random (the model corresponds to the one depicted in Equation 9 and to the one for which power was analyzed in Table 3). The L1 predictor did not contain between-cluster variance, so the guidelines also account for longitudinal data in which the L1 predictor has similar values for all participants (e.g., when participants are measured at similar points in time and time is the L1 predictor). The target level of power was .80 for all analyses. The input parameters, from which the respective population models were subsequently created, were set as described in Table 4. The sample sizes were chosen to reflect those identified in our literature review as closely as possible (for the L1 sample size, 72% lay between three and 30, and for the L2 sample size, 67% lay between 30 and 200). Because it can be anticipated that longitudinal data might contain a rather low number of L1-units, all possible sample sizes between three (the lowest possible sample size) and 10 were included in the simulation. Because of these small sample sizes, the Kenward–Roger approximation (Halekoh & Højsgaard, 2014; Kenward & Roger, 1997) was used for testing significance to avoid false negatives (i.e., anticonservative power estimates) due to biased standard errors (McNeish & Stapleton, 2016b).

For deriving the MDES, the standardized effect sizes for each effect were varied between .01 and .99 for each combination of sample sizes and variance components. To estimate the MDES, the

Table 4
Input Parameters for the MDES Analysis

Input parameter		Values
Significance level	α	.05
L1 sample size	n	3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25, 30
L2 sample size	N	30, 40, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200
Standardized L1 direct effect	$\gamma_{10.std}$	.10, .30., .50 (fixed at .10 when power for other effects is estimated)
Standardized L2 direct effect	$\gamma_{01.std}$	.10, .30., .50 (fixed at .10 when power for other effects is estimated)
Standardized CLI effect	$\gamma_{11.std}$	.10, .30., .50 (fixed at .10 when power for other effects is estimated)
Standardized random slope	$\tau_{11.std}$	.01, .09, .25 (fixed at .09 when power for effects other than CLI is estimated)
ICC	ρ	.10, .30, .50 (fixed at .30 when power for CLI effects is estimated)
Slope-intercept covariance	$Cor(U_{0j}, U_{1j})$	.00
Number of replicates	—	5,000

Note. CLI = Cross-level interaction; ICC = intraclass correlation coefficient; MDES = minimum detectable effect sizes. If only one value is specified, it will be fixed throughout the whole simulation procedure.

following decision algorithm was applied: The standardized effect size that yielded the power estimate closest to and above .80 was used as the MDES (i.e., the MDES is the lowest standardized effect size for which the target level of power was ≥ .80). Results are depicted in the *power figures* (see Figures 1, 2, and 3), for which the range of L2 sample sizes (N = 50, 100, 150) was chosen to reflect those typical for two-level studies as close as possible (see Literature Review section). The Tables 5–7 give all MDES values for all combinations of input parameters depicted in Table 4.

One major limitation of the simulation design is that it does not account for residual autocorrelation (because SIMR and lme4 do not provide direct functions to model residual autocorrelation; Bates, Mächler, Bolker, & Walker, 2015). Yet, in repeated-measurement designs, residual autocorrelation among repeated measurements might occur (e.g., higher correlations with nearer and lower correlations with more distant measurement occasions; Field, Miles, & Field, 2012). However, the necessity of modeling autocorrelated errors in power analyses has been questioned in the literature (for general linear models, see Huitema & McKean, 1998), and most articles reported in the literature review that were based on longitudinal data did not report modeling residual autocorrelation. Consequently, the results of the present power analysis can also be used for longitudinal data (if there is no assumption suggesting that substantial residual autocorrelation might occur).

Second, the simulation is based on balanced designs (i.e., the same sizes for all clusters). This might not cover the situation most two-levels researchers will face. Nonetheless, it has been argued that the negative effect of unbalanced designs is rather negligible (Cools, Van den Noortgate, & Onghena, 2009). Hence, the results of the present power analysis should be valid when the design is not strongly unbalanced.

Finally, in the present model, only one predictor per level is considered (and each explains variance only at its respective level). If more predictors at either level that are substantively correlated with each other are considered, standard errors of the fixed effects will be increased due to multicollinearity, and thus power will be decreased (Clark, 2013; Kubitschek & Hallinan, 1999). Nonetheless, because of the standardization of variables (within-cluster z-standardization of the predictor variable; see Kreft & de Leeuw, 1998), the power estimates provided here should be reasonable approximations in most cases even if multiple predictors per level are considered (see also Shieh & Fouladi, 2003). Only if the predictors are strongly related ($R^2 > .50$; variance inflation factors (VIF) > 2; Clark, 2013) will multicollinearity substantially decrease the power of small effects. For medium effects, a strong impact on power would require even larger interrelations of the predictors ($R^2 > .80$; $VIF > 5$; Clark, 2013).

General Recommendations for Designing Two-Level Studies

First, Figures 1–3 indicate that one should consider several general mechanisms when designing a two-level study. One can see that independent of the size of the ICC and sample-size combinations, the L1 direct effects yielded on average the smallest MDES, followed by the L2 direct effects and the CLI effects. This means that if an effect will be studied in combination with the L1 direct effect and both standardized effects are assumed to have similar size, it is recommended to first try to achieve sufficient sample sizes for the other effect, because this sample size will probably also suffice for the L1 direct effect. Furthermore, with the values for the random slope chosen here, the MDES for the L2 direct effect of a certain level (i.e., small, medium, or large) of the ICC is slightly smaller than the MDES for the CLI effect for the same level of the random slope. This means that if both L2 direct effects and CLI effects are studied in combination and are assumed to have similar standardized effect sizes, one should focus on the rules of thumb for the CLI effect to achieve sufficient sample sizes for both effects.

In addition, for the CLI effects, larger random slope variance components are related to higher power. This is because the adjusted CLI effect sizes are based on the size of these variance components, and thus, larger variance components yield larger values for the effects (in relation to the total variance; cf. Dziak et al., 2012; Mathieu et al., 2012). Consequently, the larger the

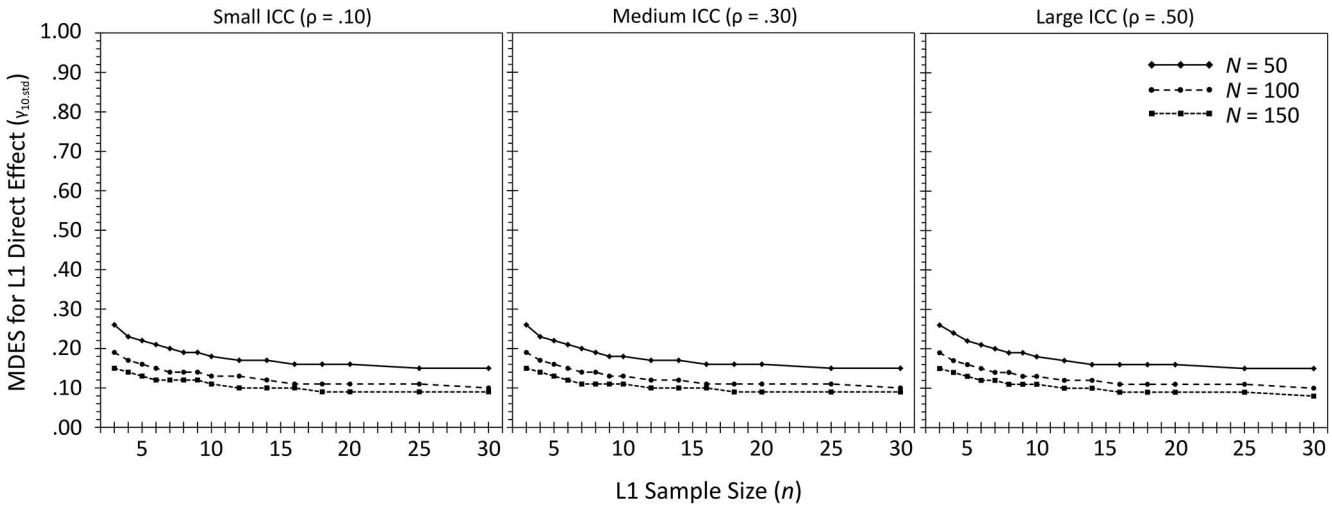

Figure 1. MDESs for L1 direct effects. N = L2 sample size; ICC = intraclass correlation coefficient.

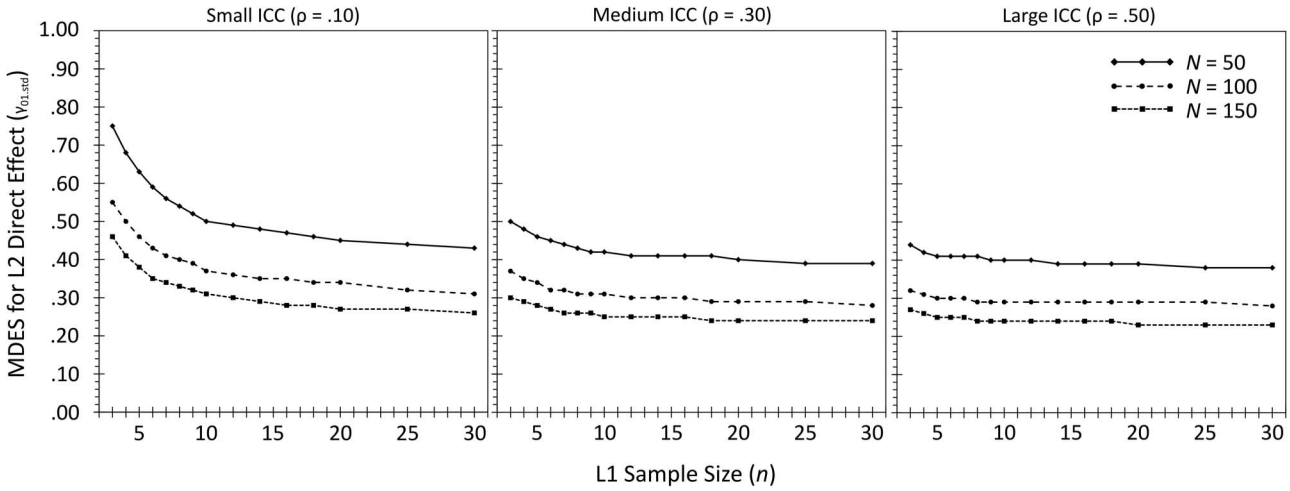

Figure 2. MDESs for L2 direct effects. N = L2 sample size; ICC = intraclass correlation coefficient.

variance, the larger the adjusted CLI effect size of the population model will be. This covers what one would expect mathematically, because a reduction (or an increase) in the variance of one or both of two variables is related to their reduced (or increased) covariance (e.g., Schmidt, Hunter, & Urry, 1976).

Finally, all the figures indicate that it is particularly worthwhile to use L1 sample sizes larger than three. For nearly all conditions, increasing the L1 sample sizes that are in the low single figures is related to a significant increase in power (except for the L2 direct effect with a high ICC). For L2 direct effects and CLI effects, this positive impact is stronger for smaller variance components. Consequently, if one assumes low variance components as population parameters (e.g., for studying members within teams) and one wants to increase power for these effects, it is best to increase the L1 sample size if possible (e.g., sample teams with six instead of four team members). In contrast, if one assumes high variance components as population parameters (e.g., for studying repeated-measurement occasions within individuals) and power should be increased, one should first choose a L1 sample size slightly larger than three and then focus on increasing the L2 sample size (e.g., increase the number of individuals). It should be noted that for CLI effects, increasing the sample size at L1 has a particularly strong positive impact on power for all sizes of the random slope, whereas for L2 direct effects, this positive impact is strong only for small ICCs.

Rules of Thumb for Deriving Sufficient Sample Sizes in Two-Level Models

The derivation of rules of thumb for sufficient sample sizes was based on all MDES values that resulted from the simulation (see Tables 5–7). Because the L1 (n) and L2 (N) sample sizes that were used for the simulation cover the samples that are most frequently used in two-level research (see literature review section), each rule of thumb provides a sample-size combination for which the MDES equals (or is smaller than and closest to) a certain standardized effect size (small, medium, or large). To account for the different

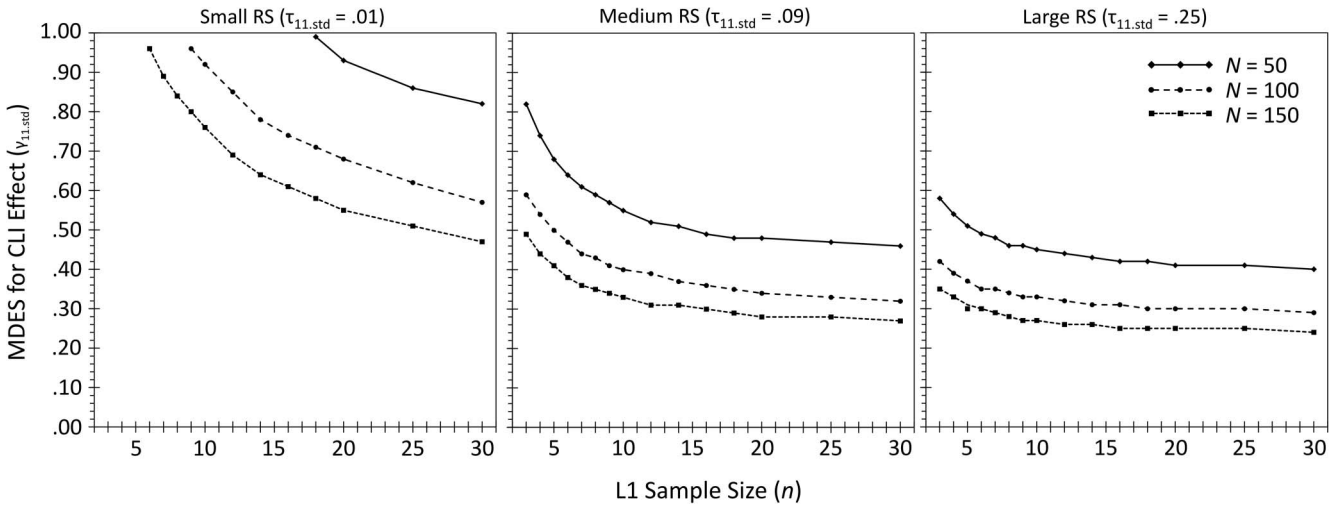

Figure 3. MDESs for CLI effects. N = L2 sample size; RS = random slope variance component.

Table 5
MDESs for L1 Direct Effects (Target Level of Power ≥ .80)

ICC	L2 sample size (*N*)	L1 sample size (*n*)														
		3	4	5	6	7	8	9	10	12	14	16	18	20	25	30
Small	30	.33	.31	.29	.27	.26	.24	.24	.23	.22	.22	.21	.20	.20	.20	.19
Small	40	.29	.27	.25	.23	.22	.21	.21	.20	.19	.19	.18	.18	.17	.17	.16
Small	50	.26	.23	.22	.21	.20	.19	.19	.18	.17	.17	.16	.16	.16	.15	.15
Small	60	.24	.22	.20	.19	.18	.17	.17	.16	.16	.15	.15	.14	.14	.14	.14
Small	70	.22	.20	.18	.18	.17	.16	.16	.15	.15	.14	.14	.13	.13	.13	.13
Small	80	.21	.19	.17	.16	.16	.15	.15	.14	.14	.13	.13	.13	.12	.12	.12
Small	90	.20	.18	.16	.15	.15	.14	.14	.13	.13	.12	.12	.12	.12	.12	.11
Small	100	.19	.17	.16	.15	.14	.14	.14	.13	.13	.12	.11	.11	.11	.11	.10
Small	125	.16	.15	.14	.13	.13	.13	.13	.12	.12	.11	.10	.10	.10	.10	.09
Small	150	.15	.14	.13	.12	.12	.12	.12	.11	.10	.10	.10	.09	.09	.09	.09
Small	175	.14	.13	.12	.11	.11	.11	.11	.10	.10	.09	.09	.09	.09	.08	.08
Small	200	.13	.12	.11	.11	.10	.10	.10	.09	.09	.09	.08	.08	.08	.08	.08
Medium	30	.34	.31	.29	.27	.26	.25	.24	.23	.22	.22	.21	.21	.20	.19	.19
Medium	40	.30	.26	.25	.23	.22	.21	.20	.20	.19	.19	.18	.18	.17	.17	.16
Medium	50	.26	.23	.22	.21	.20	.19	.18	.18	.17	.17	.16	.16	.16	.15	.15
Medium	60	.24	.21	.20	.19	.18	.17	.17	.16	.16	.15	.15	.14	.14	.13	.13
Medium	70	.22	.20	.19	.18	.17	.16	.16	.15	.15	.14	.14	.13	.13	.12	.12
Medium	80	.21	.19	.18	.17	.16	.15	.15	.14	.14	.13	.13	.13	.12	.12	.12
Medium	90	.20	.18	.17	.16	.15	.14	.14	.13	.13	.12	.12	.12	.12	.11	.11
Medium	100	.19	.17	.16	.15	.14	.14	.13	.13	.12	.12	.11	.11	.11	.11	.10
Medium	125	.17	.15	.14	.13	.13	.12	.12	.11	.11	.11	.10	.10	.10	.09	.09
Medium	150	.15	.14	.13	.12	.11	.11	.11	.11	.10	.10	.10	.09	.09	.09	.09
Medium	175	.14	.13	.12	.11	.11	.10	.10	.10	.09	.09	.09	.09	.08	.08	.08
Medium	200	.13	.12	.11	.11	.10	.10	.09	.09	.09	.09	.09	.08	.08	.08	.08
Large	30	.34	.31	.28	.26	.26	.25	.24	.23	.22	.21	.21	.20	.20	.19	.19
Large	40	.30	.26	.25	.23	.22	.21	.21	.20	.19	.19	.18	.17	.17	.17	.16
Large	50	.26	.24	.22	.21	.20	.19	.19	.18	.17	.16	.16	.16	.16	.15	.15
Large	60	.24	.22	.20	.19	.18	.17	.17	.16	.16	.15	.15	.14	.14	.14	.13
Large	70	.22	.20	.19	.18	.17	.16	.16	.15	.14	.14	.14	.13	.13	.13	.12
Large	80	.21	.19	.17	.16	.16	.15	.15	.14	.14	.13	.13	.12	.12	.12	.11
Large	90	.20	.18	.16	.15	.15	.14	.14	.13	.13	.13	.12	.12	.12	.11	.11
Large	100	.19	.17	.16	.15	.14	.14	.13	.13	.12	.12	.11	.11	.11	.11	.10
Large	125	.17	.15	.14	.13	.13	.12	.12	.11	.11	.11	.10	.10	.10	.10	.09
Large	150	.15	.14	.13	.12	.12	.11	.11	.11	.10	.10	.09	.09	.09	.09	.08
Large	175	.14	.13	.12	.11	.11	.10	.10	.10	.09	.09	.09	.09	.08	.08	.08
Large	200	.13	.12	.11	.11	.10	.10	.09	.09	.09	.09	.08	.08	.08	.08	.08

Note. ICC = intraclass correlation coefficient; MDESs = minimum detectable effect size.

capabilities in extending the L1 (e.g., sampling additional members in work teams) or L2 (e.g., sampling more countries) sample size, all resulting sample-size combinations are reported in Table 8 (L1 sample sizes in ascending order). Each rule of thumb is depicted as *N*/*n*.

Table 8 provides those sample-size combinations of L1 sample sizes between three and 30 and L2 sample sizes between 30 and 200 that yielded a power of .80 for small, medium, and large effect sizes as well as small, medium, and large ICCs/random slope variance components. It becomes evident that L1 sample sizes of up to 30 and L2 sample sizes of up to 200 are not large enough to study small standardized L2 direct effects and CLI effects as well as medium CLI effects with a small variance component. All other standardized effects for all types of fixed effects can be detected with these sample sizes. In the following paragraphs, we present the sample-size combinations that yield the smallest total sample size with an L1 sample size equal to or larger than five (because five is the minimum number of L1-units that provides unbiased estimates of fixed effects; Maas & Hox, 2005) for a power of .80.

For a *small ICC* or *small random slope variance component*, a large L1 direct effect can be detected with 30/5 (although, from a power perspective, every other combination with $N \geq 30$ and $n \geq 3$ is already sufficient), a large L2 direct effect requires 90/5, and a large CLI effect 200/20. A medium L1 direct effect can be detected with 30/5, a medium L2 direct effect with 200/7, and a medium CLI effect cannot be detected with sample sizes up to 200 (L2) and 30 (L1). Finally, in contrast to a small L1 direct effect that can be detected with 200/6, small L2 direct and small CLI effects cannot be detected with the sample sizes examined here.

When no information on their size is available, we recommend using a *medium ICC and medium random slope variance component* for deriving sufficient sample sizes. For these, detecting a large L1 direct effect requires 30/5 (although, from a power perspective, every other combination with $N \geq 30$ and $n \geq 3$ is already sufficient), a large L2 direct effect requires 40/5, and a large CLI effect requires 100/5. For detecting a medium direct effect size, sample sizes of 30/5 for L1 direct effects, 125/5 for L2 direct effects, and 200/9 for CLI effects are sufficient. Again, beyond the small L1 direct effect requir-

Table 6
MDESs for L2 Direct Effects (Target Level of Power ≥ .80)

ICC	L2 sample size (*N*)	L1 sample size (*n*)														
		3	4	5	6	7	8	9	10	12	14	16	18	20	25	30
Small	30	.95	.87	.79	.74	.71	.68	.66	.64	.62	.59	.58	.57	.56	.55	.54
Small	40	.84	.76	.70	.65	.62	.60	.58	.57	.54	.51	.51	.50	.50	.48	.47
Small	50	.75	.68	.63	.59	.56	.54	.52	.50	.49	.48	.47	.46	.45	.44	.43
Small	60	.70	.64	.58	.54	.52	.50	.48	.47	.45	.44	.43	.42	.42	.40	.40
Small	70	.65	.59	.55	.51	.48	.47	.45	.44	.43	.41	.40	.39	.39	.38	.38
Small	80	.61	.55	.51	.48	.45	.44	.43	.42	.40	.39	.38	.37	.37	.36	.35
Small	90	.58	.52	.48	.45	.43	.41	.40	.40	.38	.37	.36	.35	.35	.34	.34
Small	100	.55	.50	.46	.43	.41	.40	.39	.37	.36	.35	.35	.34	.34	.32	.31
Small	125	.50	.45	.41	.39	.37	.36	.35	.34	.33	.32	.31	.30	.30	.29	.29
Small	150	.46	.41	.38	.35	.34	.33	.32	.31	.30	.29	.28	.28	.27	.27	.26
Small	175	.42	.38	.35	.33	.32	.31	.30	.29	.28	.27	.26	.26	.25	.25	.25
Small	200	.40	.36	.33	.31	.30	.29	.28	.28	.27	.25	.25	.24	.24	.23	.23
Medium	30	.63	.59	.57	.56	.54	.54	.53	.52	.52	.51	.51	.51	.51	.50	.50
Medium	40	.56	.52	.50	.49	.48	.47	.47	.46	.46	.45	.45	.44	.44	.43	.43
Medium	50	.50	.48	.46	.45	.44	.43	.42	.42	.41	.41	.41	.41	.40	.39	.39
Medium	60	.47	.44	.43	.41	.40	.40	.39	.39	.38	.38	.38	.37	.37	.37	.36
Medium	70	.44	.41	.39	.38	.38	.37	.37	.36	.36	.35	.35	.35	.35	.34	.34
Medium	80	.41	.39	.37	.36	.35	.34	.34	.34	.33	.33	.33	.32	.32	.32	.31
Medium	90	.39	.36	.36	.34	.34	.33	.33	.32	.32	.32	.31	.31	.31	.30	.30
Medium	100	.37	.35	.34	.32	.32	.31	.31	.31	.30	.30	.30	.29	.29	.29	.28
Medium	125	.33	.31	.30	.29	.28	.28	.28	.28	.27	.27	.27	.27	.27	.26	.26
Medium	150	.30	.29	.28	.27	.26	.26	.26	.25	.25	.25	.25	.24	.24	.24	.24
Medium	175	.28	.27	.26	.25	.24	.24	.24	.23	.23	.23	.23	.23	.22	.22	.22
Medium	200	.26	.25	.24	.24	.23	.22	.22	.22	.22	.21	.21	.21	.21	.21	.20
Large	30	.55	.53	.52	.51	.51	.50	.50	.50	.50	.49	.49	.49	.49	.48	.48
Large	40	.48	.47	.46	.45	.45	.45	.44	.44	.44	.43	.43	.43	.43	.43	.43
Large	50	.44	.42	.41	.41	.41	.41	.40	.40	.40	.39	.39	.39	.39	.38	.38
Large	60	.40	.39	.38	.38	.37	.37	.37	.37	.37	.36	.36	.36	.36	.36	.36
Large	70	.38	.37	.36	.35	.35	.34	.34	.34	.34	.34	.34	.34	.33	.33	.33
Large	80	.35	.34	.34	.33	.33	.32	.32	.32	.32	.32	.32	.32	.32	.31	.31
Large	90	.34	.32	.32	.32	.31	.31	.30	.30	.30	.30	.30	.30	.30	.29	.29
Large	100	.32	.31	.30	.30	.30	.29	.29	.29	.29	.29	.29	.29	.29	.29	.28
Large	125	.29	.28	.27	.27	.27	.26	.26	.26	.26	.26	.26	.26	.26	.25	.25
Large	150	.27	.26	.25	.25	.25	.24	.24	.24	.24	.24	.24	.24	.23	.23	.23
Large	175	.24	.24	.23	.23	.23	.23	.22	.22	.22	.22	.22	.22	.22	.21	.21
Large	200	.23	.22	.22	.22	.22	.21	.21	.21	.21	.21	.21	.21	.20	.20	.20

Note. ICC = intraclass correlation coefficient; MDESs = minimum detectable effect size.

ing 200/6, small L2 direct and CLI effects cannot be detected with the sample sizes analyzed here.

Finally, for a *large ICC or random slope variance component*, detecting a large L1 direct effect can already be achieved with 30/5 (although, from a power perspective, every other combination with $N \geq 30$ and $n \geq 3$ is already sufficient), a large L2 direct effect requires 40/5, and a large CLI effect 60/5. A medium L1 direct effect can be detected with 30/5, a medium L2 direct effect with 100/5, and a medium CLI effect with 175/5. For detecting a small L1 direct effect, 200/7 are required. Again, small L2 direct or small CLI effects cannot be detected with up to 200 L2- and up to 30 L1-units.

MDESs for A Priori and Post Hoc Power Estimation in Two-Level Models

Whereas the previous section was concerned with a priori power considerations and giving advice on how to choose sample sizes when the study design can be easily adapted, the present section provides guidelines for those researchers who are either seeking to assess power post hoc or working under substantial budget or sample-size constraints and, thus, cannot increase the sample sizes (any more). In this case, the MDES approach can be used by providing the standardized effect size that could be detected with a power of .80 given a specific sample size at each of the two levels. This MDES can be used as an a priori estimate as well as a post hoc estimate for the effect size that would be significant in at least 80% of the cases if it holds true for the population.

First, consider a researcher who is planning a study but, because of financial (e.g., the budget for the study does not allow sampling more units at either level) or sample-size (e.g., all potential organizations have already been contacted and accepted or declined the request to participate in the study) constraints, cannot adapt the sample sizes any more. It is assumed that this researcher has a L2 sample size of 80 organizations, each providing 10 members for the study. The researcher expects a medium standardized L2 direct effect with a medium ICC to hold true for the population (i.e., .30). The resulting power for this sample size is smaller than .80. With the MDES approach, it can be determined which effect size can still be detected given a power ≥ .80 and the sample sizes, revealing a MDES of .34 (see Table 6). Because this is not too

Table 7
MDESs for CLI Effects (Target Level of Power ≥ .80)

RS	L2 sample size (*N*)	L1 sample size (*n*)														
		3	4	5	6	7	8	9	10	12	14	16	18	20	25	30
Small	30	—	—	—	—	—	—	—	—	—	—	—	—	—	—	.99
Small	40	—	—	—	—	—	—	—	—	—	—	—	—	—	.94	.91
Small	50	—	—	—	—	—	—	—	—	—	—	—	.99	.93	.86	.82
Small	60	—	—	—	—	—	—	—	—	—	—	.95	.90	.86	.79	.73
Small	70	—	—	—	—	—	—	—	—	—	.93	.88	.84	.80	.72	.66
Small	80	—	—	—	—	—	—	—	—	.94	.89	.83	.79	.75	.69	.64
Small	90	—	—	—	—	—	—	—	.98	.89	.83	.78	.74	.71	.65	.61
Small	100	—	—	—	—	—	—	.96	.92	.85	.78	.74	.71	.68	.62	.57
Small	125	—	—	—	—	.97	.91	.87	.83	.76	.71	.67	.63	.61	.55	.50
Small	150	—	—	—	.96	.89	.84	.80	.76	.69	.64	.61	.58	.55	.51	.47
Small	175	—	—	.97	.89	.83	.78	.74	.70	.66	.60	.56	.54	.52	.48	.44
Small	200	—	—	.90	.83	.77	.74	.68	.65	.61	.56	.53	.51	.49	.44	.41
Medium	30	—	.94	.86	.81	.76	.73	.71	.69	.65	.63	.62	.60	.59	.57	.56
Medium	40	.92	.82	.76	.70	.68	.65	.62	.60	.58	.56	.54	.53	.52	.50	.49
Medium	50	.82	.74	.68	.64	.61	.59	.57	.55	.52	.51	.49	.48	.48	.47	.46
Medium	60	.76	.68	.62	.59	.56	.54	.52	.50	.48	.47	.45	.44	.43	.42	.40
Medium	70	.71	.63	.58	.55	.52	.50	.48	.47	.45	.44	.42	.41	.40	.39	.38
Medium	80	.66	.59	.54	.52	.49	.48	.46	.44	.43	.41	.40	.38	.37	.36	.35
Medium	90	.63	.56	.52	.48	.46	.45	.43	.42	.41	.39	.38	.37	.36	.34	.33
Medium	100	.59	.54	.50	.47	.44	.43	.41	.40	.39	.37	.36	.35	.34	.33	.32
Medium	125	.54	.48	.44	.42	.40	.38	.37	.36	.35	.33	.32	.32	.31	.30	.29
Medium	150	.49	.44	.41	.38	.36	.35	.34	.33	.31	.31	.30	.29	.28	.28	.27
Medium	175	.46	.41	.38	.36	.34	.33	.32	.31	.29	.28	.28	.27	.27	.26	.25
Medium	200	.43	.38	.36	.34	.32	.31	.30	.29	.28	.27	.26	.26	.25	.25	.24
Large	30	.73	.67	.63	.61	.59	.58	.57	.56	.54	.53	.53	.52	.51	.50	.50
Large	40	.64	.59	.56	.55	.52	.51	.51	.49	.48	.47	.47	.47	.46	.45	.44
Large	50	.58	.54	.51	.49	.48	.46	.46	.45	.44	.43	.42	.42	.41	.41	.40
Large	60	.53	.49	.46	.45	.44	.43	.42	.41	.41	.40	.40	.38	.38	.37	.37
Large	70	.50	.46	.43	.42	.41	.40	.39	.38	.37	.37	.36	.36	.35	.35	.35
Large	80	.47	.43	.41	.39	.38	.38	.37	.37	.36	.35	.34	.34	.33	.33	.33
Large	90	.44	.41	.39	.37	.36	.36	.35	.35	.33	.33	.33	.32	.32	.31	.31
Large	100	.42	.39	.37	.35	.35	.34	.33	.33	.32	.31	.31	.30	.30	.30	.29
Large	125	.38	.35	.33	.32	.31	.30	.30	.29	.29	.28	.28	.27	.27	.27	.27
Large	150	.35	.33	.31	.30	.29	.28	.27	.27	.26	.26	.25	.25	.25	.25	.24
Large	175	.32	.30	.28	.28	.26	.26	.26	.25	.25	.24	.24	.24	.23	.23	.23
Large	200	.30	.28	.27	.26	.25	.24	.24	.24	.23	.23	.22	.22	.22	.21	.21

Note. RS = random slope variance; CLI = Cross-level interaction; MDESs = minimum detectable effect size.

far from the effect size expected to hold true for the population, the researcher might conclude that the problem is not too big and it is still reasonable to conduct the study.

Now, consider a researcher who has conducted a study with 100 individuals who provided their data on five repeated measurement occasions, which yielded a power of .80 in an a priori power analysis. This analysis was based on a large standardized CLI effect (i.e., .50) expected to hold true for the population and, because no other assumptions were available, on a medium random slope. Assume that the analysis did not yield a significant result, and the sample showed a large random slope. Consequently, because the observed random slope differs greatly from the random slope assumed for the a priori power analysis, the researcher wants to examine the standardized CLI effect that could have been detected given a power of .80. Inspecting Table 7 indicates that for a large random slope even a standardized effect of .37 could have been detected. Consequently, the researcher can argue that the nonsignificant result was probably not the result of an underpowered study.

Discussion

The present article was meant to provide a tutorial on how to estimate statistical power for the types of two-level-model effects most frequently used by psychologists. With a literature review, it was shown that fixed effects (i.e., L1 and L2 direct effects and CLI effects) were the types of effects that have most frequently been analyzed with two-level models. An overview of methods for power estimation revealed that the R package SIMR, using Monte Carlo simulation, provides the benefit of being able to estimate the power for all types of effects in two-level models. Then, we described how to specify standardized input parameters and how to derive the population model for a power analysis in two-level models. Furthermore, we described the implementation of a priori and post hoc power analysis with SIMR. Using this procedure, we performed a MDES analysis. Rules of thumbs for the rapid and economical identification of sufficient sample sizes to obtain a power of .80 were given for various effect sizes (small, medium, large) and sizes of the variance components (small, medium, large). The results indicate that with L1 sample sizes of up to 30 and L2 sample sizes of up to 200, large and medium

Table 8
Rules of Thumb for Sufficient Sample Sizes for Two-Level-Model Fixed Effects (Target Level of Power ≥ .80)

Effect	Size of the ICC (ρ)/RS ($\tau_{11.std}$)	Effect size		
		Small	Medium	Large
L1 Direct effect ($\gamma_{10.std}$)	ICC: Small (ρ = .10)	200/7; 175/10; 150/12; 125/16; 100/30	40/3; 30/5	Every combination of $30 \leq N \leq 200$ and $3 \leq n \leq 30$
	ICC: Medium (ρ = .30)	200/7; 175/8; 150/12; 125/16; 100/30	40/3; 30/5	Every combination of $30 \leq N \leq 200$ and $3 \leq n \leq 30$
	ICC: Large (ρ = .50)	200/7; 175/8; 150/12; 125/16; 100/30	40/3; 30/5	Every combination of $30 \leq N \leq 200$ and $3 \leq n \leq 30$
L2 Direct effect ($\gamma_{01.std}$)	ICC: Small (ρ = .10)	No combination of $30 \leq N \leq 200$ and $3 \leq n \leq 30$	200/7; 175/9; 150/12; 125/18	125/3; 100/4; 90/5; 80/6; 70/7; 60/8; 50/10; 40/18
	ICC: Medium (ρ = .30)	No combination of $30 \leq N \leq 200$ and $3 \leq n \leq 30$	150/3; 125/4; 100/12; 90/25	50/3; 40/5; 30/25
	ICC: Large (ρ = .50)	No combination of $30 \leq N \leq 200$ and $3 \leq n \leq 30$	125/3; 100/5; 90/9	40/3; 30/8
CLI effect ($\gamma_{11.std}$)	RS: Small ($\tau_{11.std}$ = .01)	No combination of $30 \leq N \leq 200$ and $3 \leq n \leq 30$	No combination of $30 \leq N \leq 200$ and $3 \leq n \leq 30$	200/20; 175/25; 125/30
	RS: Medium ($\tau_{11.std}$ = .09)	No combination of $30 \leq N \leq 200$ and $3 \leq n \leq 30$	200/9; 175/12; 150/16; 125/25	150/3; 125/4; 100/5; 90/6; 80/7; 70/8; 60/10; 50/16; 40/25
	RS: Large ($\tau_{11.std}$ = .25)	No combination of $30 \leq N \leq 200$ and $3 \leq n \leq 30$	200/3; 175/4; 150/6; 125/8; 100/18	70/3; 60/4; 50/6; 40/10; 30/25

Note. ICC = intraclass correlation coefficient; RS = random slope; CLI = Cross-level interaction. Each rule of thumb is given in the form L2 sample size *N*/L1 sample size *n*.

standardized effect sizes for nearly all (except for a medium CLI effect with a small size of the random slope variance component) types of fixed effects and small L1 direct effects can be detected (for more specific advice see Rules of Thumb and MDESs for Power Analysis in Two-Level Models section). All results should apply to individuals within clusters as well as most types of longitudinal two-level data.

The present article primarily offers implications for research practice. That is, following the guidelines on how to estimate power as well as the suggestions for choosing sample sizes, psychologists from different subdisciplines and with a wide range of research questions can derive sample sizes that yield sufficient power for their two-level studies. Furthermore, the validity of two-level psychological research can be improved. All too often, assumptions made by researchers about the power of two-level models identified in the present literature (see Literature Review section) have been based on guesses. Quantifying the power of one's study as well as being able to use quantitative power information from this article should improve the understanding of how the power of two-level models varies under different conditions (types of effects, effect sizes, variance components) and, thus, can be increased. This has been illustrated by two examples for individuals within teams and for longitudinal data (see Implementation of a Power Analysis in a Two-Level Model in SIMR section).

We hope that this tutorial and the guidelines given here can enhance the design of two-levels studies, the interpretability of their results and, in the long run, the replicability of multilevel research.

References

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., & van der Sluis, S. (2014). A solution to dependency: Using multilevel analysis to accommodate nested data. *Nature Neuroscience, 17,* 491–496. http://dx.doi.org/10.1038/nn.3648

Afshartous, D. (1995). *Determination of sample size for multilevel model design*. Retrieved from http://escholarship.org/uc/item/4cg0k6g0

Aguinis, H., & Culpepper, S. A. (2015). An expanded decision-making procedure for examining cross-level interaction effects with multilevel modeling. *Organizational Research Methods, 18,* 155–176. http://dx.doi.org/10.1177/1094428114563618

Arnold, B. F., Hogan, D. R., Colford, J. M., Jr., & Hubbard, A. E. (2011). Simulation methods to estimate design power: An overview for applied research. *BMC Medical Research Methodology*. Advance online publication. http://dx.doi.org/10.1186/1471-2288-11-94

Asendorpf, J. B., Conner, M., de Fruyt, F. D. E., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Vanaken, M. A. G. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108–119. http://dx.doi.org/10.1002/per.1919

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

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

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research, 40,* 373–400. http://dx.doi.org/10.1207/s15327906mbr4003_5

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Bloom, H. S. (1995). Minimum detectable effects. A simple way to report the statistical power of experimental designs. *Evaluation Review, 19,* 547–556. http://dx.doi.org/10.1177/0193841X9501900504

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

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics, 15,* 391–420. http://dx.doi.org/10.1007/s001800000041

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

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147–158.

Cao, J., & Ramsay, J. O. (2010). Linear mixed-effects modeling by parameter cascading. *Journal of the American Statistical Association, 105,* 365–374. http://dx.doi.org/10.1198/jasa.2009.tm09124

Castro, S. L. (2002). Data analytic methods for the analysis of multilevel questions. *The Leadership Quarterly, 13,* 69–93. http://dx.doi.org/10.1016/S1048-9843(01)00105-9

Clark, P. C. (2013). *The effects of multicollinearity in multilevel models* (Doctoral dissertation). Retrieved from https://etd.ohiolink.edu/pg_10?0:NO:10:P10_ACCESSION_NUM:wright1375956788

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

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155

Cools, W., Van den Noortgate, W., & Onghena, P. (2008). ML-DEs: A program for designing efficient multilevel studies. *Behavior Research Methods, 40,* 236–249. http://dx.doi.org/10.3758/BRM.40.1.236

Cools, W., Van den Noortgate, W., & Onghena, P. (2009). Design efficiency for imbalanced multilevel data. *Behavior Research Methods, 41,* 192–203. http://dx.doi.org/10.3758/BRM.41.1.192

Cudeck, R., & Klebe, K. J. (2002). Multiphase mixed-effects models for repeated measures data. *Psychological Methods, 7,* 41–63. http://dx.doi.org/10.1037/1082-989X.7.1.41

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170–180. http://dx.doi.org/10.1037/0003-066X.60.2.170

Donohue, M. C., & Edland, S. D. (2016). *longpower: Power and sample size calculators for longitudinal data* (R package version 1.0–16). Retrieved from https://cran.r-project.org/web/packages/longpower/index.html

Dziak, J. J., Nahum-Shani, I., & Collins, L. M. (2012). Multilevel factorial experiments for developing behavioral interventions: Power, sample size, and resource considerations. *Psychological Methods, 17,* 153–175. http://dx.doi.org/10.1037/a0026972

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

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149–1160. http://dx.doi.org/10.3758/BRM.41.4.1149

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London, UK: SAGE.

Gałecki, A., & Burzykowski, T. (2013). *Linear mixed-effects models using R: A step-by-step approach*. New York, NY: Springer Science & Business Media. http://dx.doi.org/10.1007/978-1-4614-3900-4

Goldstein, H. (1987). Multilevel covariance component models. *Biometrika, 74,* 430–431. http://dx.doi.org/10.1093/biomet/74.2.430

Green, P., & Macleod, C. J. (2016a). *Package "SIMR"*. Retrieved from https://cran.r-project.org/web/packages/simr/index.html

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

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models—The R package pbkrtest. *Journal of Statistical Software, 59,* 1–30. http://dx.doi.org/10.18637/jss.v059.i09

Helms, R. W. (1992). Intentionally incomplete longitudinal designs: I. Methodology and comparison of some full span designs. *Statistics in Medicine, 11,* 1889–1913. http://dx.doi.org/10.1002/sim.4780111411

Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management, 23,* 723–744. http://dx.doi.org/10.1177/014920639702300602

Hox, J. J. (1998). Multilevel modeling: When and why. In R. M. Balderjahn & M. Schader (Eds.), *Classification, data analysis, and data highways: Studies in classification, data analysis, and knowledge organization* (pp. 147–154). New York, NY: Springer. http://dx.doi.org/10.1007/978-3-642-72087-1_17

Hox, J. J. (2010). *Multilevel analysis; techniques and applications*. New York, NY: Routledge.

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods, 3,* 104–116. http://dx.doi.org/10.1037/1082-989X.3.1.104

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67,* 219–229. http://dx.doi.org/10.1037/0021-9010.67.2.219

Johnson, P. C. D., Barry, S. J. E., Ferguson, H. M., & Müller, P. (2015). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution, 6,* 133–142. http://dx.doi.org/10.1111/2041-210X.12306

Johnson, P. C. D., Barry, S. J. E., Ferguson, H. M., & Pie, M. (2015). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution, 6,* 133–142. http://dx.doi.org/10.1111/2041-210X.12306

Kahn, J. H. (2011). Multilevel modeling: Overview and applications to research in counseling psychology. *Journal of Counseling Psychology, 58,* 257–271. http://dx.doi.org/10.1037/a0022680

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53,* 983–997. http://dx.doi.org/10.2307/2533558

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin, 54,* 351–353. http://dx.doi.org/10.1037/h0046737

Kreft, I. (1996). *Are multilevel techniques necessary? An overview including simulation studies*. Unpublished manuscript, California State University, Los Angeles.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: SAGE. http://dx.doi.org/10.4135/9781849209366

Kubitschek, W. N., & Hallinan, M. T. (1999). Collinearity, bias, and effect size: Modeling "the" effect of track on achievement. *Social Science Research, 402,* 380–402. http://dx.doi.org/10.1006/ssre.1999.0653

Lindenberger, U., & Pötter, U. (1998). The complex nature of unique and shared effects in hierarchical linear regression: Implications for developmental psychology. *Psychological Methods, 3,* 218–230. http://dx.doi.org/10.1037/1082-989X.3.2.218

Liu, G., & Liang, K.-Y. (1997). Sample size calculations for studies with correlated observations. *Biometrics, 53,* 937–947. http://dx.doi.org/10.2307/2533554

Longford, N. T. (1993). *Random coefficient models*. New York, NY: Oxford University Press.

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58,* 127–137. http://dx.doi.org/10.1046/j.0039-0402.2003.00252.x

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1,* 86–92. http://dx.doi.org/10.1027/1614-2241.1.3.86

Martin, J. G. A. (2015). *Package "pamm."* Retrieved from https://cran.r-project.org/web/packages/pamm/index.html

Martin, J. G. A., Nussey, D. H., Wilson, A. J., & Re, D. (2011). Measuring individual differences in reaction norms in field and experimental studies: A power analysis of random regression models. *Methods in Ecology and Evolution, 2,* 362–374. http://dx.doi.org/10.1111/j.2041-210X.2010.00084.x

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,* 951–966. http://dx.doi.org/10.1037/a0028380

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163. http://dx.doi.org/10.1037/1082-989X.9.2.147

McNeish, D., & Stapleton, L. M. (2016a). Modeling clustered data with very few clusters. *Multivariate Behavioral Research, 51,* 495–518. http://dx.doi.org/10.1080/00273171.2016.1167008

McNeish, D. M., & Stapleton, L. M. (2016b). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28,* 295–314. http://dx.doi.org/10.1007/s10648-014-9287-x

Meyer, B., & Glenz, A. (2013). Team faultline measures: A computational comparison and new approach to multiple subgroups. *Organizational Research Methods, 16,* 393–424. http://dx.doi.org/10.1177/1094428113484970

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass, 2,* 842–860. http://dx.doi.org/10.1111/j.1751-9004.2007.00059.x

O'Keefe, D. J. (2007). Brief report: Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Communication Methods and Measures, 1,* 291–299. http://dx.doi.org/10.1080/19312450701641375

Onwuegbuzie, A. J., & Leech, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics, 3,* 201–230. http://dx.doi.org/10.1207/s15328031us0304_1

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716. http://dx.doi.org/10.1126/science.aac4716

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2017). *A user's guide to MLwiN* (Version 3.00). Bristol, England: Centre for Multilevel Modelling, University of Bristol. Retrieved from http://www.bristol.ac.uk/cmm/software/mlwin/download/manuals.html

Rast, P., & Hofer, S. M. (2014). Longitudinal design considerations to optimize power to detect variances and covariances among rates of change: Simulation results based on actual longitudinal studies. *Psychological Methods, 19,* 133–154. http://dx.doi.org/10.1037/a0034524

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2,* 173–185. http://dx.doi.org/10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: SAGE.

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

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

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,* e35564. http://dx.doi.org/10.1371/journal.pone.0035564

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2,* 110–114. http://dx.doi.org/10.2307/3002019

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12,* 347–367. http://dx.doi.org/10.1177/1094428107308906

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 473–485. http://dx.doi.org/10.1037/0021-9010.61.4.473

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316. http://dx.doi.org/10.1037/0033-2909.105.2.309

Shieh, Y.-Y., & Fouladi, R. T. (2003). The effect of multicollinearity on multilevel modeling parameter estimates and standard errors. *Educational and Psychological Measurement, 63,* 951–985. http://dx.doi.org/10.1177/0013164403258402

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational and Behavioral Statistics, 18,* 237–259. http://dx.doi.org/10.3102/10769986018003237

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: SAGE.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles, CA: SAGE.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). *Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software*. Retrieved from http://www.rmcs.buu.ac.th/statcenter/HLM.pdf

Stroup, W. W. (2002). Power analysis based on spatial effects mixed models: A tool for comparing design and analysis strategies in the presence of spatial variability. *Journal of Agricultural Biological & Environmental Statistics, 7,* 491–511. http://dx.doi.org/10.1198/108571102780

van der Leeden, R., Busing, F. M. T. A., & Meijer, E. (1997). *Bootstrap methods for two-level models*. Retrieved from www2.soziologie.uni-halle.de/langer/buecher/mehrebenen/literatur/busing1997.pdf

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54,* 426–482. http://dx.doi.org/10.1090/S0002-9947-1943-0012401-3

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement, 4,* 587–602. http://dx.doi.org/10.1177/001316448904900309

Received June 1, 2017
Revision received May 24, 2018
Accepted June 28, 2018 ■