JSPR

Article

Power struggles: Estimating sample size for multilevel relationships research

Journal of Social and Personal Relationships
2018, Vol. 35(1) 7–31

DOI: 10.1177/0265407517710342
journals.sagepub.com/home/spr
SAGE

Sean P. Lane
Erin P. Hennes
Purdue University, USA

Abstract
Conducting research on human relationships entails special challenges of design and analysis. Many important questions benefit from the study of dyads and families, and studies of relationships in natural settings often involve longitudinal and/or clustered designs. In turn, power analyses for such studies require additional considerations, because multilevel statistical models (or structural equation modeling equivalents) are often used to analyze relationships data. Power calculations in multilevel models involve the difficult task of specifying hypothesized values for a large number of parameters. Planning studies can also involve power trade-offs, including whether to prioritize the number of dyads sampled or the number of repeated measurements per dyad. Unfortunately, the relationships literature provides limited guidance on how to deal with these issues. In this article, we present a data simulation method for estimating power for commonly used relationships research designs. We also illustrate the method using two worked examples from relationships research.

Keywords
Close relationships, dyadic data analysis, multilevel modeling, power analysis

Conducting research on close relationships brings with it many challenges that make data collection more difficult, time-consuming, and financially costly than many other topics within psychology. Studies frequently require substantial coordination between members of a large research team, as well as with the participants they wish to recruit

Corresponding author:
Sean P. Lane, Department of Psychological Sciences, Purdue University, 703 Third Street, Room 1242, West Lafayette, IN 47906, USA.
Email: seanlane@purdue.edu

and retain. Such coordination often involves repeated assessments of multiple relationship partners over periods of time that can stretch months and even years and may involve data collection during critical life stages such as marriage, childbirth, and death or dying. As a result, the stakes are high to successfully observe a hypothesized effect (if one actually exists in the population).

In the contemporary relationships literature, "successful" observation of a hypothesized effect generally means that one or more corresponding parameter estimates is statistically significant for some sample of data. However, in collecting samples from the population, we acknowledge the existence of sampling variability, which means that even if our effect exists, there is always some nonzero probability that we will fail to observe it in a single random sample (i.e., Type II error). The purpose of this article is to provide tools and suggestions for maximizing the chance that an existing hypothesized effect will be detected.[1] These tools are together referred to as *power analysis* (Cohen, 1962).

The concept of power is not new, but only recently have most relationships researchers begun to consider seriously the implications of statistically underpowered research (Bolger & Laurenceau, 2013; Finkel, Eastwick, & Reis, 2015). Moreover, even as a priori power analysis has risen in methodological prominence, sufficient tools and training to conduct such analyses for complicated relationships research have generally not followed. In the few cases where such tools do exist, virtually no previous work has provided guidance for the complex issue of determining reasonable parameter estimates for all but the most basic models or direct replications.

To address this critical gap, the current article offers a flexible procedure for conducting power analyses for relationships research. We focus on longitudinal and dyadic models, but the method can be extended to virtually any other empirical design. In order to contextualize this process, we introduce a substantive example from the relationships literature in which a researcher is interested in replicating and extending a previously reported finding. We consider four questions that are likely to arise when designing a study to test a hypothesized effect. Importantly, this includes being confident in the results regardless of whether the hypothesis is supported. These questions are as follows:

Q1: What information do I need to conduct a power analysis for a proposed study involving multiple individuals within dyads or families, at single or multiple time points?
Q2: How do I conduct a power analysis once I have collected the necessary information?
Q3: What factors are likely to have the strongest effect on power?
Q4: What if I do not have all of the information that I need?

After introducing our illustrative example, we provide a largely conceptual definition and discussion of power, with emphasis on the individual parameters that affect its calculation in between-subject (e.g., cross-sectional) designs. Importantly, we extend this description to illustrate how power calculations are complicated by designs that also vary within subjects (e.g., time intensive or dyadic). We then briefly review existing options available to researchers for conducting power analyses of multilevel models and note some advantages and disadvantages of each. We propose a flexible method for

power analysis using simulation and walk through a worked example in detail from the literature on support receipt in close relationships. We build upon this example in a second worked analysis that uses a more complex design often utilized by relationship researchers—that is, when relationship processes are simultaneously examined for partner dyads within relationships. We recommend sensitivity analyses and demonstrate how they can be an important tool for understanding the differential impact of various factors on power and how they can be used to inform efficient study design. Lastly, we discuss how to generate estimates of model parameters when the researcher has little previous information on which to base predictions. We conclude by providing a summary of our recommended guidelines.

Illustrative example: The effect of support receipt on stress

Throughout this article, we situate our discussion within an example from the close relationships literature on social support by Bolger, Zuckerman, and Kessler (2000). Decades of research has documented a robust positive relationship between individuals' perceptions of social support and various indicators of mental and physical well-being (House, Landis, & Umberson, 1988; LaRocco, House, & French, 1980; Taylor, 2007). The perceived availability of support is reliably associated with lower reports of psychological distress (Kessler & McLeod, 1985), improvements in individuals' abilities to cope (Leavy, 1983), better adjustment to chronic stressors, and accelerated recovery from acute stressors (Cobb, 1976). However, reviews of the literature find that the association between perceptions of support and partners' actual reports of enacted support is moderate (Barrera, 1986; Haber, Cohen, Lucas, & Baltes, 2007) and that often support provision attempts are associated with *increased* reports of stress and emotional distress on behalf of the recipient (Barrera, 1986).

In an early study of its type, Bolger et al. (2000) sought to examine if the inconsistent associations between support and (dis)stress could be accounted for by a support recipient's awareness of having received support. They conducted a longitudinal diary study of romantic couples who were followed as one of the partners approached an important examination. Surprisingly, they found that on days when examinees reported *receiving* support they reported more stress, but there was not a corresponding increase in examinee stress when partners reported *providing* support. In addition, the effect of examinee reported support receipt on stress was significantly increased during the week prior to the exam (i.e., stress phase), while the effect of partner reported support provision on stress the week before the exam was again nonsignificant. In other words, examinee stress seemed to be associated—positively—with examinees' report that they had received support from their partner but was independent of partners' reports that such support had actually been provided.

Based on Bolger et al.'s (2000) research, in this article, we will conduct two sets of a priori power analyses for subsequent hypothetical research testing four hypotheses. First, given that the negative effects of support are mixed in the literature, we will first conduct a direct replication testing the hypothesis that *on days in which examinees report receiving support, they report more stress* (H1). Next, we seek to examine the hypothesized effect of partner support provision that was nonsignificant in Bolger et al.'s

(2000) study. That is, *on days in which partners report providing support, examinees report less stress* (H2). Next, we will attempt to replicate Bolger et al.'s finding that *the negative effect of support receipt is heightened during the stress phase* (H3). Finally, we will test the hypothesis that *the positive effect of support provision is heightened during the stress phase* (H4).

In the second set of power analyses, we will expand Bolger et al.'s research to an explicitly dyadic design and test the reciprocal effects of support receipt and provision on both examinees and their partners. Our hypotheses for this study are the same as the previous (examinee-only) study; however, we expect the effects for partners to be somewhat weaker (since they are not under prolonged stress). In conducting power analyses, we will determine the number of individuals and time points needed to have a strong likelihood of observing statistically significant support for our hypotheses for both examinees and partners (assuming the effects exist in the population). In addition, we will assess other methodological factors besides individuals and time points that can increase our chance of success, particularly in cases in which initial sample-size-based power analyses indicate that we do not have sufficient resources to reliably observe an effect of interest.

Power for between-subject designs

We first review the basic concept of power and the factors that affect it using a basic unstandardized single-level regression model. Equation (1) depicts such a population model for a between-subjects version of the Bolger et al. (2000) social support model.

$$Y_i = b_0 + b_1 X_i + e_i \tag{1}$$

Here, Y_i is a continuous dependent variable (e.g., stress) that is modeled as a function of some continuous predictor variable, X_i, where i indexes individuals (e.g., support received from one's partner). Also in the model are parameters b_0, which indicates the average level of Y, and b_1, which is the association between X and Y. The error term, e_i, is the deviation of individual i in their actual Y value compared to what would be expected given their value of X. In most cases, our parameter of interest is b_1, the association between support receipt and stress.

Power, formally defined, is the probability of correctly rejecting a false null hypothesis. That is, it is the likelihood of declaring that one has found a significant hypothesized effect (e.g., $\alpha \leq .05$) in a sample given that there is an effect in the population. A *power analysis* involves determining the sample size necessary to ensure some predetermined probability (e.g., 80%) of rejecting the null hypothesis given a hypothesized effect size. Power is a function of the Type I error rate (i.e., α, which is generally fixed a priori), the effect size (unstandardized or standardized), and the standard error (Cohen, 1988).

In conducting a power analysis, we explicitly acknowledge uncertainty due to variability in our estimate of the effect size if we were to run the same study repeatedly across multiple samples. The standard error of an effect is the statistical approximation of that uncertainty for an effect from a single study. We use the standard error in conjunction with the estimated effect size to construct test statistics that are then evaluated

Table 1. Single level regression model and factors that affect power.

Regression model	Slope standard error	
$Y_i = b_0 + b_1 X_i + e_i$	$\sigma_{b_1} = \sqrt{\frac{\sigma_e^2}{N\sigma_X^2}}$	
Variables		
Y_i	Dependent variable	
X_i	Independent variable	
e_i	Error (i.e., residual)	
i	Indexes each individual	
Parameters		Effect on power
b_0	Intercept estimate	↑
b_1	Slope estimate	↑
σ_{b_1}	Standard error of the slope	↓
σ_e^2	Error variance	↓
σ_X^2	Variance of independent variable	↑
N	Sample size (i.e., number of individuals, i)	↑

Note. ↑ = increasing the magnitude of the parameter increases power for the corresponding fixed effect; ↓ = increasing the magnitude of the parameter decreases power for the corresponding fixed effect.

for statistical significance. Our goal is to determine the minimum sample size necessary in order to have a high probability of concluding, in our single study, that we can be reasonably certain that our hypothesized effect size differs from zero in the population (i.e., that the association between support receipt and stress is statistically significant in our study).

The standard error is itself made up of three factors: the sample size, the variance in X (e.g., the degree to which some individuals receive more support than do others), and the error variance (e.g., the amount of variability in stress that is not due to differences in support receipt; see Table 1). Table 1 shows that increasing the sample size, N, will decrease the standard error (and increase power). Similarly, increasing the amount of variance in X (support receipt, σ_X^2) will increase power. However, increasing amounts of unexplained variance (i.e., error/residual, σ_e^2) will serve to decrease power.

Although we discuss at the end of this article strategies for increasing the effect size, or reducing the standard error by other means than increasing sample size, for the most part, the focus of a power analysis is to examine the likely impact on statistical significance of the number of observations at each level of analysis.[2] Next, we extend our discussion to the within-subject (multilevel) case, which is the primary focus of this article.

Power for within-subject (multilevel) designs

For demonstrating the factors that affect power in multilevel designs, we begin by extending the population model given in Equation (1) to that given in Equation (3) (cf., Raudenbush & Bryk, 2002).

$$Y_{it} = (b_0 + b_{0i}) + (b_1 + b_{1i})^* X_{it} + e_{it} \quad (2)$$

In this example, we can imagine that individuals (i) were measured in terms of their levels of support receipt (X) and stress (Y) not just once but t times. Thus, Y_{it} is the value of stress for individual i on measurement t, and X_{it} is the value of support receipt for individual i on measurement t. The error term, e_{it}, is the measurement-specific deviation from the model prediction for each individual observation. Most important are the coefficients b_{0i} and b_{1i}. These are random variables that represent individual-specific intercepts and slopes, respectively, in addition to the sample average intercept (b_0) and slope (b_1). In other words, because individuals were assessed multiple times, we can estimate and statistically test if individual i reported a higher average level of stress across the diary period than the sample average level of stress (b_0), as well as whether individual i had a lower association between support receipt and stress than the average association across the sample (b_1). In a multilevel framework, we assume that b_{0i} and b_{1i} are normally distributed with a population mean and variance represented by $\bar{b}_0$ and $\sigma^2_{b_0}$, and $\bar{b}_1$ and $\sigma^2_{b_1}$, respectively. That is, we expect individuals to have some sample average level of anxiety, and some sample average association between support receipt and anxiety, but we also anticipate variability between individuals on both average stress (the intercept) and the association between support receipt and stress (the slope).

As before, if we ran a study in which we sampled a group of individuals and measured support receipt (X) and stress (Y) a total of t times for each individual, we can show that the power to detect the average within-subject effect of $\bar{b}_1$ (the slope) can be characterized as a function of Type I error rate, effect size, and the standard error (Table 2).[3] The Type I error rate is defined as in the between-subjects case. The effect size, $\bar{b}_1$, is analogous to b_1 in Equation (1). However, measuring individuals repeatedly over time results in two new components to the standard error not found in the between-subject example. These are n, the number of repeated assessments, and $\sigma^2_{b_1}$, the variance of the random slope. Each is added to the equation for the sampling variance of the effect of X on Y from Equation (2) and is shown in Table 2 (cf., Snijders, 2005).

The multilevel sampling variance now consists of two components. The ratio to the left of the "+" sign (the within-subject component) is largely familiar, but now includes n in the denominator, which indicates that as the number of repeated measurements increases so will power. Our observation of repeated measurements now also allows us to identify individual differences, and thus there is an additional piece to the sampling variability to the right of the "+" sign (the between-subject component) that includes the variance of the individual slopes divided by N, the number of individuals.[4]

Available tools for power analysis

Having reviewed the components that influence power in a multilevel model, we now turn to available tools for estimating it. Until recently, there were relatively few resources available to conduct power analyses for multilevel models. Those that are available fall into two broad categories: formula based and simulation. Formula-based approaches can often be found in books specializing in multilevel modeling and longitudinal design (Ahn, Heo, & Zhang, 2015; Fitzmaurice, Laird, & Ware, 2012; Gelman & Hill, 2006; Hox, 2010; Liu & Liang, 1997; Moerbeek & Teerenstra, 2016; Moerbeek, Van Breukelen, & Berger, 2008; Snijders & Bosker, 1993, 1999). Such approaches are

able to provide exact estimates of power for various sets of models where the standard errors for certain parameters have been explicitly derived. However, given the considerable effort required to derive such formulae for the countless possible models researchers could be interested in, many available formulae are restricted to more basic multilevel models (e.g., Moerbeek & Teerenstra, 2016).

The approach that we typically prefer is the use of simulation methods (Bolger & Laurenceau, 2013; Gelman & Hill, 2006). This approach is supported by many of the statistical software packages psychologists use, including Mplus (Muthén & Muthén, 1998–2012), R (R Development Core Team, 2015), SAS (SAS Institute, 2013; Zhang & Wang, 2009), and SPSS (IBM, Inc., 2013) and has myriad uses in scientific research beyond power applications (cf., Paxton, Curran, Bollen, Kirby, & Chen, 2001). The rationale behind power estimation by simulation is to use the hypothesized population model and parameters to generate data for a hypothetical study. We can then analyze the data from this study using our hypothesized statistical model and record the significance of the effects of interest. Since the data are randomly generated using the population model, there will be sampling variability in the estimates of the individual effects and their standard errors, and we will not exactly recover the parameters of the population model. This can be done thousands of times, simulating thousands of hypothetical studies, and for each we record the significance of the hypothesized effects. The proportion of times each individual effect is significant across all of the simulations is the power of that effect—quite literally, the expected number of times you would observe a significant effect if an alternative hypothesis was true in the population.[5]

In an effort to make formulae or simulation-based procedures easier to apply and power analysis for such designs more widely accessible, a number of specialized software programs have been developed, including RMASS2 (Hedeker, Gibbons, & Waternaux, 1999), Optimal Design (Raudenbush et al., 2011), PinT (Bosker, Snijders, & Guldemond, 2007), and ML-DEs (Cools, Van den Noortgate, & Onghena, 2008). These programs generally provide the user with a limited set of models to choose from, and once chosen present a list of the various parameters the user must specify in order to conduct the analysis. Although these software vary in their degree of user-friendliness, they can be fast and convenient when researchers know the specific model they wish to estimate, that model is accommodated by the software, and they can generate reasonable estimates for the model parameters. However, as with the formula-based approach, the array of supported models across software programs can be limited, and there are few guidelines for generating parameter estimates when the model specifications are not well defined by previous research. Moreover, generating sample size estimates from software that cannot fully accommodate a predicted model comes with the risk of producing invalid power estimates via model misspecification.

We present here a syntax-based procedure for conducting power analysis using simulation. While the learning curve can be steep, investing in the process will pay considerable dividends in fully understanding the model one wants to estimate, identifying all of the factors that can affect hypothesized statistical tests, and providing an exact template of the desired analysis once the data have been collected. Furthermore, the simulation approach can be extended to a virtually unlimited variety of models in order to accommodate researchers' idiosyncratic needs.

Longitudinal random effects model example (support receipt and anxiety)

In this section, we provide an example of how to conduct a power analysis using simulation in Mplus (Muthén & Muthén, 1998–2012). We chose Mplus because it requires the least amount of formal programming knowledge and has the ability to accommodate a wide variety of statistical models without changing the analysis framework. In both examples, we will attempt to fully describe each element of the syntax and how it relates back to the information we gathered to conduct the power analysis. We present the analogous syntax for R (R Development Core Team, 2015), SAS (SAS Institute, 2013; Zhang & Wang, 2009), and SPSS (IBM, Inc., 2013) in the Online Supplement to this article.

We describe a power analysis in which we gather information from that provided by Bolger et al. (2000) and link it to syntax generated in Mplus. Our first example represents a particularly complete reporting of the information necessary to conduct a power analysis. We use it partially as an exemplar of how to report such analyses in the interest of facilitating power analyses aimed at promoting future research and replication. Later we describe methods for estimating parameters when the model is less well defined.

The information needed broadly consists of four components:

1. The statistical model
2. The measurement scale for each variable
3. Expected patterns of missingness
4. All means, variances, and parameter estimates (i.e., covariances) for variables defined by the model

Bolger et al. (2000, p. 956) explicitly define their statistical model, which we restate in Equation (3). They operationalized stress using reports of anxiety and depression in separate analyses; for the current example, we focus their analysis using anxiety as the dependent variable. We recommend writing out the hypothesized statistical model as a first step in any power analysis, both to help identify the information needed and to prepare for syntax generation (which will require this model to be explicitly stated)

$$\begin{aligned} \text{Anx}_{it} = {} & (b_0 + b_{0i}) + (b_1 + b_{1i})^{*}\text{LagAnx}_{it} + b_2{}^{*}\text{Phase}_{it} \\ & + (b_3 + b_{3i})^{*}\text{Provision}_{it} + (b_4 + b_{4i})^{*}\text{Receipt}_{it} \\ & + b_5{}^{*}\text{Provision}_{it}{}^{*}\text{Phase}_{it} + b_6{}^{*}\text{Receipt}_{it}{}^{*}\text{Phase}_{it} + e_{it} \end{aligned} \tag{3}$$

Several components of this model are consistent with Equation (2). Specifically, the outcome, examinees' anxiety (Anx_{it}) is a function of a fixed and random intercept (b_0 and b_{0i}) and support receipt (b_4 and b_{4i}). This basic model is expanded by Bolger et al. (2000) to also account for the previous measurement's anxiety (($b_1 + b_{1i}$)*LagAnx_{it}; cf., Castro-Schilo & Grimm, this issue), whether or not it was the week before the exam (b_2*Phase_{it}), and if the examinees' partner reported providing support (($b_3 + b_{3i}$)*Provision_{it}). The authors also estimated two two-way interactions, between provision and phase and between receipt and phase (b_5*Provision_{it}*Phase_{it}, b_6*Receipt_{it}*Phase_{it}).

Now that we have defined our model (Step 1), we must determine the scale on which all of the variables are measured (Step 2), which is in this case easily determined from the Measures section of the original manuscript (Bolger, Zuckerman, & Kessler, 2000, p. 955): Examinee anxiety is a continuous variable and support and phase are binary variables. The authors indicate that their final sample was composed of 68 couples, each of whom provided 32 days of data (Bolger et al., 2000, p. 955). However, they also note that approximately 2% of the data were missing (Step 3).

Finally, we must acquire estimates for all means, variances, and parameters (Step 4). We begin by determining the parameter estimates for all of the fixed effects. Fortunately, Bolger et al. (2000, p. 957) provide the unstandardized fixed effects parameter estimates, which we reproduce in Table 3. In the original study, we see that both support receipt and its interaction with phase were statistically significant, while both support provision and its interaction with phase were not. Next, we determine the parameter estimates for the random effects. The random effects are not reported in the main body of Bolger et al. (2000) but rather in their Footnote 6 (p. 956). They report standard deviations for the random effects of the intercept, lagged anxiety, and support receipt to be .306, .147, and .229, respectively. For the power analysis, we square these values to estimate the variances for those random effects (Table 2). The authors were unable to estimate a random effect for support provision, suggesting that it was very small in actuality, and therefore constrained it to 0.0 in the final model. We estimated this random effect but assigned it a very small variance ($\sigma^2_{b_3} = .001$).

Lastly, we estimate the means and variances for each of the variables in the model. Means and variances of the phase and support variables are not directly provided by the authors, but they can be easily derived. The authors note that of the 32 days that individuals completed diaries, 7 belonged to the week prior to the exam.[6] A dichotomized variable was created in which 22% (7/32) of the entries belonged to the stressed phase and the other 78% belonged to the nonstressed phase. We can then multiply these values to create an estimate of the variance for the phase variable (i.e., $p(1-p)$). Therefore, phase has a mean of .22 and a variance of .17. The authors also present the proportions of support provision and support receipt days across the study (p. 956). From this, we can estimate that partners reported support provision on 58% of days and examinees reported support receipt on 56% of days. As with the phase variable, we can use these values to estimate each support variable's variance to be .24 and .25, respectively. The authors do not report a mean or variance for the lagged anxiety predictor; however, this is not a major concern as its primary use was as a covariate, for which the other parameter estimates are already adjusted. We use an estimated mean of 0.0 and a variance of .5 in the simulation. Lastly, we require an estimate of the error variance. Unfortunately, the authors do not report this nor do they provide a direct means for estimating it. However, another published article using the same data does provide us with a useful estimate of the error variance. Shrout, Herman, and Bolger (2006) reanalyzed the data from this study using different statistical models and examining different moods with the same hypotheses in mind. They report error variances for their models, which, although not exact due to the different modeled parameters, provide reasonable estimates for the error variance used here.

Table 2. Multilevel regression model and factors that affect power.

Regression model	Slope standard error	
$Y_{it} = (b_0 + b_{0i}) + (b_1 + b_{1i})^* X_{it} + e_{it}$	$\sigma_{\bar{b}_1} = \sqrt{\frac{\sigma_e^2}{Nn\sigma_X^2} + \frac{\sigma_{b_1}^2}{N}}$	
Variables		
Y_{it}	Dependent variable	
X_{it}	Independent variable	
e_{it}	Error (i.e., residual)	
i	Indexes each individual	
t	Indexes each repeated assessment	
Parameters		Effect on power
b_0	Intercept estimate	↑
b_{0i}	Individual-specific intercept	↑
b_1	Slope estimate	↑
b_{1i}	Individual-specific slope	↑
$\sigma_{\bar{b}_1}$	Standard error of the average (between-person) slope	↓
σ_e^2	Error variance	↓
σ_X^2	Variance of independent variable	↑
$\sigma_{b_1}^2$	Variance of the individual slopes (multilevel)	↓
N	Sample size (i.e., number of individuals, i)	↑
n	Cluster size (i.e., number of repeated assessments, t)	↑

Note. ↑ = increasing the magnitude of the parameter increases power for the corresponding fixed effect; ↓ = increasing the magnitude of the parameter decreases power for the corresponding fixed effect.

Figure 1 provides the input syntax to conduct a Monte Carlo simulation study in Mplus using the population values just described and the statistical model from Equation (3). Individual lines are commented using exclamation points (!) and include heading identifiers that correspond to the following explanations:

a. A list of variables in the model: anxiety (anx), lagged anxiety (lanx), phase (phase), support provision (prov), support receipt (rec), phase by provision interaction (pprov), phase by receipt interaction (prec)
b. Total number of observations: N multiplied by n (2,176)
c. Number of individuals (68) and repeated measurements (32)
d. Arbitrary value for the random number generator (20160215) so that one can rerun the simulation and get the same results
e. Number of simulated studies[7] (1,000)
f. Generate patterns of missingness (missing completely at random): In this case, 1 pattern ($p = 1.0$) where anx is missing on 2% of cases
g. Predictors that only vary at the within-subjects level (lanx, phase, prov, rec, pprov, prec)
h. Predictors that only vary on the between-subjects level (none in current example)
i. Type of model: a two-level random effects model

```
! Exclamation point indicates a comment
MONTECARLO:
    NAMES ARE anx lanx phase prov rec    ! a)
                phprov phrec;            ! a)
    NOBSERVATIONS = 2176;                ! b)
    NCSIZES = 1;
    CSIZES = 68 (32);                    ! c)
    SEED = 20160215;                     ! d)
    NREPS = 1000;                        ! e)
    PATMISS = anx(.02); PATPROBS = 1; ! f)
    WITHIN = lanx phase prov rec         ! g)
             phprov phrec;               ! g)
    !BETWEEN = ;                        ! h)
ANALYSIS:
    TYPE = TWOLEVEL RANDOM;              ! i)
MODEL POPULATION:                        ! j)
    %WITHIN%                             ! k)
      slope1 | anx ON lanx;              ! define the random slope of lanx main effect
      slope2 | anx ON prov;              ! define the random slope of prov main effect
      slope3 | anx ON rec;               ! define the random slope of rec main effect
      anx ON phase*.40;                  ! phase main effect
      anx ON phprov*-.03;                ! phase*prov interaction effect
      anx ON phrec*.17;                  ! phase*rec interaction effect
      [lanx*0]; lanx*.50;                ! lanx has a mean of 0.0 and variance of 0.5
      [phase*.22]; phase*.17;            ! phase has mean of 0.22 and a variance of 0.17
      [prov*.58]; prov*.24;              ! prov has a mean of 0.58 and a variance of 0.24
      [rec*.56]; rec*.24;                ! rec has a mean of 0.56 and a variance of 0.24
      [phprov*0]; phprov*.041;           ! pprov has a mean of 0.0 and a variance of 0.041
      [phrec*0]; phrec*.033;             ! prec has a mean of 0.0 and a variance of 0.033
      anx*.42;                           ! residual variance of anx
    %BETWEEN%                            ! l)
      [anx*.14]; anx*.094;               ! intercept estimate (0.14) and variance (0.094)
      [slope1*-.50]; slope1*.022;        ! lanx slope (-0.50) and variance (0.022)
      [slope2*-.04]; slope2*.001;        ! prov slope (-0.04) and variance (0.001)
      [slope3*.12]; slope3*.052;         ! rec slope (0.12) and variance (0.052)
MODEL:                                   ! m)
    !copy MODEL POPULATION: syntax here
OUTPUT: TECH9;
```

Figure 1. Mplus syntax for power analysis of Bolger et al. (2000).

j. Specify the generating population model
k. Specification for the within-subjects effects (see Figure 1 for details)
l. Specification for the between-subjects effects (see Figure 1 for details)
m. Specify the model to be estimated: It is the same and the population model syntax should be copied into this section

The results of the power analysis using the specifications from the original model reported by Bolger et al. (2000) are shown in Table 3 and Supplementary Material (for additional examples of annotated Mplus output see Bolger & Laurenceau, 2013; Bolger, Stadler, & Laurenceau, 2012). From these results, we can see that a replication would be acceptably powered to test H1 (the main effect of support receipt) but underpowered to test the other hypotheses (e.g., support provision, both support by phase interactions).

To ensure that a future study is well powered to detect all hypothesized effects, we can vary the number of individuals (N) and/or repeated assessments (n) we wish to collect. In

Table 3. Parameter estimates and power results for Bolger et al. (2000).

				Power		
Parameter	Mplus parameter	Estimate	*p*	$N = 68$	$N = 110$	$N = 310$
Main effects						
Intercept	BTW – Means ANX	.140	.002	.863	.972	1.000
Lag anxiety	BTW – Means SLOPE1	−.500	.000	1.000	1.000	1.000
Phase	WTH – ANX ON PHASE	.400	.000	1.000	1.000	1.000
Supp. provision	BTW – Means SLOPE2	−.040	.350	.234	.343	.798
Supp. receipt	BTW – Means SLOPE3	.120	.025	.844	.963	1.000
Provision × Phase	WTH – ANX ON PHPROV	−.030	.748	.065	.083	.141
Receipt × Phase	WTH – ANX ON PHREC	.170	.047	.555	.792	.995
Random effects						
Intercept	BTW – Variances ANX	.094	<.050	1.000	1.000	1.000
Lag anxiety	BTW – Variances SLOPE1	.022	<.050	.794	.959	1.000
Supp. provision	BTW – Variances SLOPE2	.001	>.050	.010	.010	.022
Supp. receipt	BTW – Variances SLOPE3	.052	<.050	.943	.997	1.000
Error/residual variance	WTH – Residual Variances ANX	.420[a]	<.010[a]	1.000	1.000	1.000
Means						
Lag anxiety	WTH – Means LANX	.000				
Phase	WTH – Means PHASE	.220				
Supp. provision	WTH – Means PROV	.580				
Supp. Receipt	WTH – Means REC	.560				
Provision × Phase	WTH – Means PHPROV	.000				
Receipt × Phase	WTH – Means PHREC	.000				
Variances						
Lag anxiety	WTH – Variances LANX	.500				
Phase	WTH – Variances PHASE	.170				
Supp. provision	WTH – Variances PROV	.240				
Supp. receipt	WTH – Variances REC	.250				
Provision × Phase	WTH – Variances PHPROV	.041				
Receipt × Phase	WTH – Variances PHREC	.033				
Covariances						
Cov(Provision, Receipt)	WTH – PROV WITH REC	.051				

Note. *N* corresponds to the number of subjects in the power analysis. The number of assessments (*n*) is fixed at 32. BTW—between-level results. WTH—within-level results.
[a]Residual variance and *p*-value estimated from Shrout et al. (2006).

this case, holding the number of assessments fixed at 32, we wish to see the number of individuals required so that the effect of the support receipt by phase interaction is powered to 80% (H3). Table 1 shows that this will be accomplished by recruiting a sample of 110 individuals. We note that this is approximately double the original sample, despite the original sample producing a statistically significant result. This is because the initial study obtained a *p*-value for the effect of .047, which is close to the critical value

($\alpha = .05$) where power is 50% (Cohen, 1988). As such, considerably more individuals would be necessary to observe the same effect at greater than chance probability. Given that Bolger et al. (2000) also hypothesized (but did not find) that the provision of support would be associated with a decrease in examinees' anxiety, we may wish to see the sample required to power the support provision effect ($b = -.04$, $p = .350$) to 80% (H2). Rerunning the power analysis, we see that this would be accomplished by recruiting a sample of 310 individuals. Consistent with this power analysis, a replication of Bolger et al. (2000) conducted by Shrout et al. (2010) using a sample of 312 examinees and 310 of their partners over the course of 35 days did observe significant effects for both support receipt and support provision. However, we note that even in a sample of 310 individuals, we are still heavily underpowered to detect a significant provision by phase interaction. Given the current model and parameter estimates, we would require 3,000 individuals to have 80% power to detect such an effect. This highlights that researchers may find that the sample size required to test their hypothesis may go far beyond their available resources. In such situations, they might reconsider whether the effect, even if it does exist in the population, is sufficiently large to be of substantive importance.

Dyadic longitudinal random effects/actor–partner interdependence model example (coupled support receipt and anxiety)

Our second example departs from the previous model in two important ways. First, although Bolger et al. (2000) collected couples data, only the examinees' outcome was estimated. In our second example, we demonstrate a model in which we simultaneously predict stress for both the examinees and their partners. We model the data together in a dyadic data analysis using the actor–partner interdependence model (Kashy & Kenny, 1999; Kenny, 1996; see Iida, Seidman, & Shrout, current issue; Rogers, Wood, & Furr, current issue; and Stern & West, current issue, for further details and potential alternatives). In doing so, we can fit a multivariate version of Equation (3) in which each partner is modeled simultaneously (Laurenceau & Bolger, 2005; Raudenbush, Brennan, & Barnett, 1995; Raudenbush & Bryk, 2002). This model is presented as follows:

$$\begin{aligned}\text{Anx}_{itd} = {} & E_d{}^*[(b_{0E} + b_{0Ei}) + (b_{1E} + b_{1Ei})^*\text{LagAnx}_{itd} + b_{2E}{}^*\text{Phase}_{itd} \\ & +(b_{3E} + b_{3Ei})^*\text{Provision}_{itd} + (b_{4E} + b_{4Ei})^*\text{Receipt}_{itd} \\ & + b_{5E}{}^*\text{Provision}_{itd}{}^*\text{Phase}_{itd} + b_{6E}{}^*\text{Receipt}_{itd}{}^*\text{Phase}_{itd}] \\ & + \\ & P_d{}^*[(b_{0P} + b_{0Pi}) + (b_{1P} + b_{1Pi})^*\text{LagAnx}_{itd} + b_{2P}{}^*\text{Phase}_{itd} \\ & +(b_{3P} + b_{3Pi})^*\text{Provision}_{itd} + (b_{4P} + b_{4Pi})^*\text{Receipt}_{itd} \\ & + b_{5P}{}^*\text{Provision}_{itd}{}^*\text{Phase}_{itd} + b_{6P}{}^*\text{Receipt}_{itd}{}^*\text{Phase}_{itd}] \\ & + e_{itd}\end{aligned} \tag{4}$$

This model allows for the estimation of parallel effects for each of the dyad members, d, while also allowing for each dyad member to have unique random variation that correlates with that of their partner. Indicators E_d and P_d signify which partner each line of data comes from so that separate examinee and partner components of the model can

be estimated. Determining parameter estimates in our first example was largely straightforward, as we were conducting a hypothetical direct replication. In many real-world research scenarios, however, precise estimates of our hypothesized parameters are not available. In our second example, we demonstrate a power analysis in which we adjust findings from a similar model to offer best guesses of our proposed effects. Later, we discuss strategies for making such guesses when the models are even more weakly predefined.

Bolger et al. (2000) hypothesized that when examinees acknowledged support receipt from their partners they would feel more distressed and that this effect would be particularly exaggerated the week prior to their stressful examination (i.e., phase). The authors also collected, but did not analyze, similar data on the partners of examinees. By their logic, we might hypothesize that partners also feel more distressed after acknowledging support provision, but since they are not preparing for a difficult examination and not under systematic prolonged stress, the effects may be weaker.

Table 3 presents parameter estimates for a simulation study in which both examinees and partners report on their anxiety, if they provided support, and if they received support as the examinees' exam approached. Many of the estimates are the same as those presented in Table 2, just repeated twice, once for each dyad member. We may expect many of the individual-level parameter estimates to be the same across dyad members, and so we retain the examinee estimates in many cases. This includes assuming 2% missing data for partners.

Because partners are not preparing for a difficult examination and are not under systematic prolonged stress, overall we expect them to report less anxiety. Therefore, we estimated the intercept to be .07 instead of .14. We also expect a smaller main effect of phase (.20 instead of .40), a smaller effect of support receipt on anxiety (.06 instead of .12), and a smaller interaction between support receipt and phase (.04 instead of .17). Lastly, since partners' reports are likely to be correlated, we specified that partners' random intercepts, slopes, and residual variances correlated at a moderate level ($r \approx .30$; Kenny, Kashy, & Cook, 2006). This was accomplished by using the formula for calculating the correlation between two variables, in which we know the approximate correlation (Kenny et al., 2006) and the variances of partners' parameters (Bolger et al., 2000), and so can estimate the shared covariance between each pair (see Table 4).[8] Figure 2 presents the Mplus syntax for this model.

Table 4 presents results from the power simulation of the hypothetical dyadic model using the original sample size and number of repeated assessments reported by Bolger et al. (2000). As expected, the hypothesized smaller support receipt effect for partners would be underpowered (34.7%; H1). However, using the sample size from Shrout et al. (2010), replication of this study offers good power for the main effect of partner support receipt (89.3%), as well as partner support provision (83.8%; H2). Alternatively, with the exception of the examinee support receipt by phase interaction, the interaction effects are still heavily underpowered in the event that the researchers expect them to differ from zero in the population (H3 and H4).

Table 4. Power results for dyadic example.

Parameter	Mplus parameter	Estimate	Power	
			$N = 68$	$N = 310$
Main effects				
Intercept (E)	BTW – Means EANX	.140	.872	1.000
Intercept (P)	BTW – Means PANX	.070	.326	.921
Lag anxiety (E)	BTW – Means ESLOPE1	−.500	1.000	1.000
Lag anxiety (P)	BTW – Means PSLOPE1	−.500	1.000	1.000
Phase (E)	WTH – EANX ON EPHASE	.400	1.000	1.000
Phase (P)	WTH – PANX ON PPHASE	.200	1.000	1.000
Supp. provision (E)	BTW – Means ESLOPE2	−.040	.276	.831
Supp. provision (P)	BTW – Means PSLOPE2	−.040	.259	.838
Supp. receipt (E)	BTW – Means ESLOPE3	.120	.854	1.000
Supp. receipt (P)	BTW – Means PSLOPE3	.060	.347	.893
Prov. × Phase (E)	WTH – EANX ON EPHPROV	−.030	.086	.148
Prov. × Phase (P)	WTH – PANX ON PPHPROV	−.030	.076	.163
Receipt × Phase (E)	WTH – EANX ON EPHREC	.170	.565	.998
Receipt × Phase (P)	WTH – PANX ON PPHREC	.040	.097	.167
Random effects				
Intercept (E)	BTW – Variances EANX	.094	.999	1.000
Intercept (P)	BTW – Variances PANX	.094	1.000	1.000
Lag anxiety (E)	BTW – Variances ESLOPE1	.022	.826	1.000
Lag anxiety (P)	BTW – Variances PSLOPE1	.022	.811	1.000
Supp. provision (E)	BTW – Variances ESLOPE2	.001	.012	.026
Supp. provision (P)	BTW – Variances PSLOPE2	.001	.011	.031
Supp. receipt (E)	BTW – Variances ESLOPE3	.052	.946	1.000
Supp. receipt (P)	BTW – Variances PSLOPE3	.052	.944	1.000
Random effect covariances				
Intercept (E-P)	BTW – EANX WITH PANX	.030	.526	.993
Lag anxiety (E-P)	BTW – ESLOPE1 WITH PSLOPE1	.007	.214	.695
Supp. prov. (E-P)	BTW – ESLOPE2 WITH PSLOPE2	.000	.000	.002
Supp. receipt (E-P)	BTW – ESLOPE3 WITH PSLOPE3	.017	.253	.844
Residual (E-P)	WTH – EANX WITH PANX	.100	1.000	1.000

Note. *N* corresponds to the number of subjects in the power analysis. The number of assessments (*n*) is fixed at 32. Means and variances are the same for both partners as in Table 3 and so are excluded. (E)—Examinee. (P)—Partner. BTW—between-level results. WTH—within-level results.

Sensitivity analysis and factors that affect power

In Example 1, we demonstrated a power analysis in a case in which the model parameters were generally well defined, and the main source of uncertainty was merely due to sampling variability. Such methods are particularly useful in instances where the researcher is seeking to conduct a direct replication and has a large amount of resources. However, in most cases, there is also uncertainty surrounding the predicted values of model parameters themselves. This may be due to novel hypotheses regarding effect sizes that have not been previously established in the literature, as well as limited information regarding the expected magnitudes or structure of the variance or residual

```
MONTECARLO:
    NAMES ARE eanx elanx ephase eprov erec ephprov ephrec   ! examinee
              panx planx pphase pprov prec pphprov pphrec; ! partner
    NOBSERVATIONS = 2176;
    NCSIZES = 1;
    CSIZES = 68 (32);
    SEED = 20160215;
    NREPS = 1000;
    PATMISS = eanx(.02) panx(.02); PATPROBS = 1;
    WITHIN = elanx ephase eprov erec ephprov ephrec              ! examinee
             planx pphase pprov prec pphprov pphrec;             ! partner
ANALYSIS:
    TYPE = TWOLEVEL RANDOM;
    MODEL POPULATION:
    %WITHIN%
    eslope1 | eanx ON elanx; pslope1 | panx ON planx;
    eslope2 | eanx ON eprov; pslope2 | panx ON pprov;
    eslope3 | eanx ON erec; pslope3 | panx ON prec;
    eanx ON ephase*.40 ephprov*-.03 ephrec*.17;
    panx ON pphase*.20 pphprov*-.03 pphrec*.04;
    [elanx*0]; [ephase*.22]; [eprov*.58]; [erec*.56]; [ephprov*0]; [ephrec*0];
    eanx*.42; ephase*.17; elanx*.50; eprov*.24; erec*.25; ephprov*.041; ephrec*.033;
    [planx*0]; [pphase*.22]; [pprov*.58]; [prec*.56]; [pphprov*0]; [pphrec*0];
    panx*.42; pphase*.17; planx*.50; pprov*.24; prec*.25; pphprov*.041; pphrec*.033;
    eprov WITH erec*.051; pprov WITH prec*.051;
    eanx WITH panx*.1;                                           ! residual covariance
    %BETWEEN%
    [eanx*.14]; [eslope1*-.50]; [eslope2*-.04]; [eslope3*.12];
    eanx*.094; eslope1*.022; eslope2*.001; eslope3*.052;
    [panx*.07]; [pslope1*-.50]; [pslope2*-.04]; [pslope3*.06];
    panx*.094; pslope1*.022; pslope2*.001; pslope3*.052;
    eanx WITH panx*.03;                                          ! random intercept covriance
    eslope1 WITH pslope1*.007;                                   ! lanx random slope covariance
    eslope2 WITH pslope2*.0003;                                  ! prov random slope covariance
    eslope3 WITH pslope3*.017;                                   ! rec random slope covariance
MODEL:
    !copy MODEL POPULATION: syntax here
OUTPUT: TECH9;
```

Figure 2. Mplus syntax for dyadic power analysis.

covariance components. Researchers may also conduct a power analysis and discover that they do not have the resources available to have a high likelihood of observing particular hypothesized effects.

To address these issues, in this section, we briefly discuss sensitivity analysis. Sensitivity analyses are iterative analyses that systematically assess the impact on power of various model parameters. Sensitively analyses can fulfill two related objectives: First, they can identify lower limits of expected power across a range of parameter values and combinations when there is limited preexisting information about the likely magnitude of the effects. For instance, a researcher may be unsure of the size of a dyad-level random slope effect but can use sensitivity analyses to ensure that a study is sufficiently powered even in the event that the random effect is quite large. Second, sensitivity analyses can examine trade-offs when resources are limited. For instance, the researcher can assess the impact of increasing number of participants versus number of measurements when funding for participant remuneration is fixed or can help researchers identify which

moderators or mediators of interest are most likely to be fruitful to assess given a fixed amount of survey time. In all cases, we recommend that researchers create tables or plots of multiple specifications to examine the sensitivity of their model design to various exogenous or endogenous factors and optimize their design to maximize study success. Below we provide some examples for doing so using the support provision examples described above and offer guidance for anticipating the impact of some common design decisions on power.

Assuming that we hold the Type I error rate constant, the remaining factors that influence power are the effect size and standard error. So far, we have restricted our conversation to the influence of the number of participants, but the researcher also has direct control over the number of repeated assessments. From a practical standpoint, researchers conducting studies that involve repeated assessments are often faced with the issue of balancing the number of repeated assessments and the number of individuals recruited. Increasing participants can be time-consuming and costly. Increasing the number of assessments can be associated with participant fatigue, leading to increased dropout and decreased quality of data. Relatively, it may ultimately be easier to increase the number of assessments compared to recruiting more participants, but from the viewpoint of statistical power, is it worth it (see also Rast & Hofer, 2014)?

To provide perspective in attempting to answer this question, we conducted a series of power analyses using the model from Table 3. We focus on the interaction effect between support receipt and phase, which in the original example with 68 couples and 32 assessments had 55.5% power. Assuming that the total number of measurements that a researcher was able to collect was fixed, in Figure 3 we show the relative increase in power for this effect as either the number of individuals or the number of assessments is increased. As is suggested by Table 2, increasing the number of participants increases power more than increasing the number of assessments. This is because increasing the number of participants decreases both the within-subject and between-subject components of the standard error, whereas increasing the number of assessments only decreases the within-subject component (Snijders & Bosker, 1999). In this example, it is critical to note that increasing the number of assessments while holding the number of participants constant will never reach 80% power. This is an important point more broadly, as in certain designs, such as diary studies, it is often less expensive to have participants complete more assessments than it is to recruit more participants. Figure 3 demonstrates that even a few extra participants may be more impactful for testing a hypothesis than hundreds of additional measurements. However, a researcher might also require a minimum number of assessments to assure variability in variables of interest. For example, interpersonal conflict might be rare compared to support receipt. Thus, while increased subjects may gain a researcher more power than increased assessments holding all other parameters equal, this may come with a trade-off of realistically also reducing predictor variance, which will undermine power.

Beyond the number of participants and assessments, Table 2 shows that other study elements can affect the probability that a hypothesized effect will be significant. These are less directly under the researcher's control, but the researcher can make efforts to mitigate them. For example, researchers might also measure constructs they believe are associated with their dependent variable and include them in their analyses as covariates in order to decrease error variance (σ_e^2). This will have the effect of increasing power.

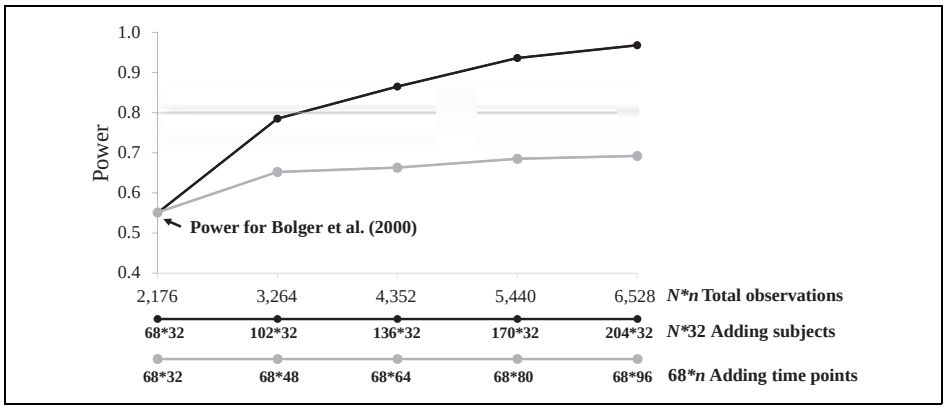

Figure 3. Sensitivity analysis of the effect of adding participants versus adding assessments on the estimated power of the phase by support receipt interaction in Bolger et al. (2000).

Experimental techniques like matching or blocking participants or engaging methodological strategies for reducing measurement noise can also decrease error variance and increase power. Researchers may also attempt to maximize the variance in their predictor by assigning participants equally to groups if the predictor is categorical or sampling subpopulations with more variability on X or from the extremes of the distribution if the predictor is continuous (σ_X^2), which also would increase power.

One factor that affects power that is probably the least under researchers' control is the random variation of an effect across subjects ($\sigma_{b_1}^2$). These are the random effects in a multilevel model, and as their variance increases, power for the associated fixed effect decreases. Given that researchers generally have less control over how much a hypothesized effect will vary across subjects, and they may not even have an expectation as to how much an effect varies across subjects to begin with, it is useful to investigate how much different levels of variation in a random effect can impact the power of the associated fixed effect. To do this, we again use the Bolger et al. (2000) example and, in this case, focus on the main effect of support receipt. They reported that this effect had an associated random effect with variance .052. That is, the fixed effect of support receipt on stress of .12 could vary across participants, with some participants being quite negatively impacted by support receipt while others showed a very small or negligible negative effect. However, such a random effect might be very difficult to estimate in the absence of prior research. It may also be difficult to estimate empirically and requires a researcher to estimate such an effect as fixed (Barr, Levy, Scheepers, & Tily, 2013). Therefore, we can conduct a sensitivity analysis to assess the impact of this random effect on our overall power. We have plotted this effect in reference to three other factors that affect power in order to illustrate its influence.

First we show how increasing the number of participants (Figure 4(a)) can mitigate the impact of an increasingly large random effect. In the first power analysis, we saw that the power of the support receipt main effect in Bolger et al. (2000) was 84.4%.

Assuming the same number of participants, we can see that as the random effect of support receipt increases power decreases steadily. However, increasing the number of participants mitigates this effect, importantly in a nonlinear way. As the number of participants increases, the slope of the power curve decreases, initially sharply, but less so with each increment. This indicates that increasing the number of participants can protect against large random effects, but after a certain number of additional participants, it is unnecessary. We next illustrate the same effect for increasing the number of assessments (Figure 4(b)). In contrast to Figure 4(a), adding more assessments is not as protective as adding participants. These two examples represent direct ways in which researchers can hedge themselves against uncertainty with regard to the size of random effects. Lastly, we illustrate the impact of increasingly large random effects on the power of the fixed effect as the size of the fixed effect changes (Figure 4(c)). A researcher might be interested in how robust their hypothesized effect is to the random between-subject variation observed within their sample. Importantly, Figure 4(c) shows that smaller effects are disproportionately affected by random slope variation. Larger effects are more resilient, and random slope variation is not as much of a concern.

Determining parameter estimates

Although we recommend sensitivity analyses regardless of the certainty a researcher may have in the parameters of their proposed model, it can still be hard to know where to start. Certainly, it is unfeasible (and unhelpful) to conduct power analyses on an infinite combination of parameter values. In this section, therefore, we provide some initial (and incomplete) pointers for obtaining starting values for your power sensitivity analyses.

Echoing decades of other scholars who have written on methodological and design issues in relationship research, our strongest recommendation is to pilot, pilot, pilot! Having concrete data on which to base multilevel power analyses, however small or unrepresentative the sample, can result in considerably less effort in estimating the various parameters for the hypothesized population model and less uncertainty regarding the parameter specifications in the model. Pilot data are particularly useful for the estimation of random effects in power analyses, as these are infrequently reported in published research. While pilot study estimates may themselves be imprecise and subject to sampling variability, they at least provide a starting point from which to conduct sensitivity analyses.

Drawing on previously conducted studies using similar measures or employing similar designs can also be extremely useful, although researchers may have to contact the authors of the original study to obtain the necessary parameter estimates. For instance, an existing study of support receipt on stress may be a useful starting point for estimating parameters modeling the effect of support receipt on intimacy; or a study of relationship satisfaction approaching and following marriage may provide some clues for a new study examining weight fluctuation approaching and following marriage. Researchers may also turn (cautiously) to between-person estimates as proxies for proposed within-person processes. Most importantly, the

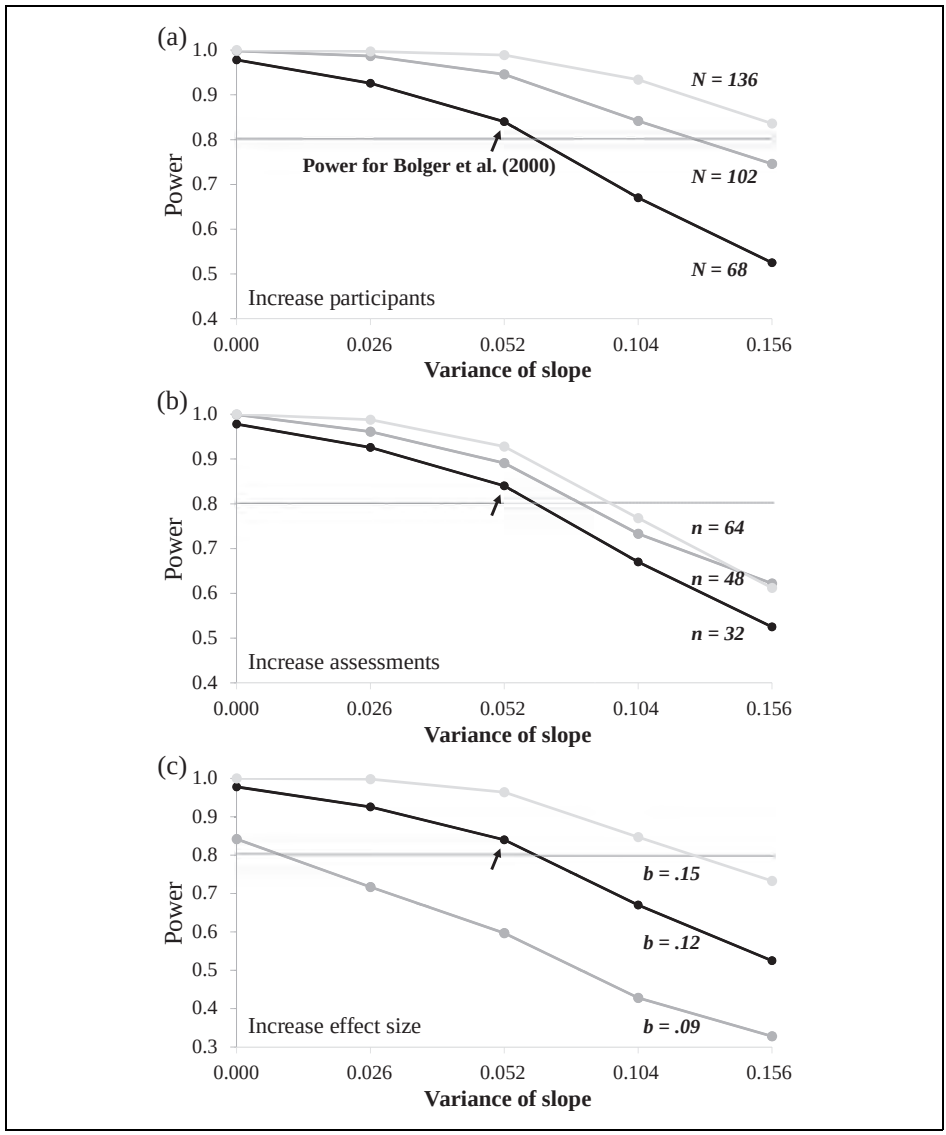

Figure 4. Sensitivity analyses for Bolger et al. (2000) of the effect of random slope variance on power at varying levels of (a) participants, (b) assessments, and (c) effect sizes.

researcher should become familiar with their literature. Many patterns, such as residual autocorrelation structures or self- versus other effect sizes are largely consistent even across quite varied studies. In sum, although the values of many parameter estimates may not be intuitive, the existing literature may provide useful initial approximations for a wide variety of parameters, which, combined with

sensitivity analyses, may offer considerable insight into the resources needed to maximize successful studies and minimize failed ones.

Conclusion

We detailed above a method for conducting power analyses for designs commonly used in relationships research, including all of those reviewed in this special issue (Castro-Schilo & Grimm, this issue; Iida, Seidman, & Shrout, this issue; Rogers et al., this issue; Stern & West, this issue). Although it requires effort, we believe that this method can be used by relationships researchers in planning future studies and in writing grant applications. To recapitulate, conducting a power analysis involves answering the following questions:

Q1: What information do I need to conduct a power analysis for a proposed study involving multiple individuals in dyads or families, at single or multiple time points?

A1: Researchers will need to know in advance their hypothesized statistical model, including estimates for all parameters. This includes means, variances, and coefficients (i.e., covariances/correlations) for the effects at each level of analysis.

Q2: How do I conduct a power analysis once I have collected the necessary information?

A2: Several formulae and software programs exist for conducting power analyses for a subset of common models. We offer tools for conducting power analyses using simulation, an approach that can accommodate virtually any model.

Q3: What factors are likely to have the strongest effect on power?

A3: The factors that influence power are the Type I error rate (usually set a priori to $\alpha = .05$), effect size, and standard error (which comprises the sample size at each level of analysis and variance components). Of primary interest to most researchers are the within-subject (n) and between-subject (N) sample sizes. Increasing both increases power, but as Figure 3 illustrates, increasing the between-subjects sample size boosts power more than increasing the number of within-subjects assessments. We draw special attention to the variance of the random within-subject slope ($\sigma^2_{b_1}$): The greater its variance, the lower the power to detect within-subject effects.

Q4: What if I do not have all of the information that I need?

A4: It is unlikely that researchers will have precise estimates of all parameters unless they are conducting a direct replication. We strongly recommend collecting pilot data or drawing on prior research using similar measures and designs. However, when even imprecise estimates are difficult to locate, we encourage anchoring on the conservative end of the distribution.

In closing, we hope that the procedures detailed here will help relationships researchers plan informative studies that make the best use of financial resources. For additional examples of power simulations using models that are relevant to relationships researchers, see Bolger & Laurenceau (2013) and Bolger, Stadler and Laurenceau (2012).

Acknowledgments

The authors would like to especially thank Niall Bolger and Patrick E. Shrout for their advisory roles in developing and guiding the work that formed the basis of this manuscript. The authors would also like to thank them, the Columbia University and New York University Couples Labs, and three anonymous reviewers for their helpful feedback on earlier versions of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the National Institutes of Health grant T32 AA013526.

Supplemental material

Supplementary material is available for this article online.

Notes

1. Although not elaborated in detail in the current article, the method described here is flexible and complementary to alternative hypothesis testing strategies (e.g., confidence intervals (Kelley & Rausch, 2011) or Bayesian analysis).
2. We note that here we are referring to a single hypothesized effect. If multiple hypotheses (corresponding to multiple parameters) are tested within the same study, each will have its own standard error and corresponding power. Hypotheses involving multiple parameters (e.g., mediation) will also have their own standard errors and associated power.
3. A parallel description exists for the power to detect an average intercept ($\bar{\beta}_0$). However, since researchers are more often interested in associations between variables than mean levels, and rarely focus on significance tests of the intercept, we focus on the slope here.
4. The factors affecting power for a given hypothesis depend on the statistical model. In addition to the factors we note, researchers could also choose to model factors such as the correlations between residuals. Or if individuals were recruited as dyads, a three-level or multivariate multilevel model that incorporated dyad and individual-specific intercepts and slopes may be estimated. If included, these additional model parameters would also affect power—greater residual correlation reduces power as would greater between-dyad random variance. Extensions for the standard error for many such expanded models can be found in Moerbeek and Teerenstra (2016). Moreover, the choice of statistical test can affect the calculation of the standard error and corresponding estimates of power (Berkhof & Snijders, 2001).
5. Notably, this approach differs slightly from common practices for a priori power analysis, in which an effect size and desired power are input and a sample size is output. Here, the effect size and sample size are input to the simulation and power is estimated. This requires minimal calibration to adjust the input sample size to obtain the desired power.
6. The authors later note that they ultimately excluded the last day of reports from their analyses to not confound support effects with those having to do with the exam being over, resulting in 31 total days, 6 of which belonged to the stress phase. We retain 32 as the number of repeated assessments in our example as it allows for convenient factorizations in the sensitivity analyses described later but note that using 6/31 for calculating the phase mean and variance trivially affects power estimates.

7. As with any simulation, more replications are better. However, for power analysis, usually 1,000 replications is sufficient, as the tiny difference in power that might be obtained when increasing the replications by an order of magnitude is almost always of little substantive interest. Additionally, this order of magnitude increment is expounded when conducting sensitivity analyses (i.e., iterative power analyses) and can often exceed the physical memory limits of many computers.
8. We note that in addition to dyad members' intercepts, slopes, and residuals correlating with that of their partners, the parameters may also correlate with one another within an individual. Bolger et al. only estimated the variances of the random effects and implicitly assumed that they did not correlate with one another. In both of our power analysis examples, we make the same assumption, though empirically, this is likely not accurate. However, even less information is usually available for estimating these associations. As with other parameters, we recommend conducting sensitivity analyses for assessing the degree to which misspecification/exclusion of parameters such as random effects correlations from one's final statistical model can impact the power of primary hypotheses.

References

Ahn, C., Heo, M., & Zhang, S. (2015). *Sample size calculations for clustered and longitudinal outcomes in clinical research*. Boca Raton, FL: CRC Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Barrera, M. (1986). Distinctions between social support concepts, measures, and models. *American Journal of Community Psychology*, *14*, 413–445.

Berkhof, J., & Snijders, T. A. B. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics*, *26*, 133–152.

Bolger, N., & Laurenceau, J-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York, NY: Guilford Press.

Bolger, N., Stadler, G., & Laurenceau, J. P. (2012). Power analysis for intensive longitudinal studies. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 285–301). New York, NY: Guilford Press.

Bolger, N., Zuckerman, A., & Kessler, R. C. (2000). Invisible support and adjustment to stress. *Journal of Personality and Social Psychology*, *79*, 953–961.

Bosker, R. J., Snijders, T. A. B., & Guldemond, H. (2007). *PinT (Power in two-level designs): Estimating standard errors of regression coefficients in hierarchical linear models for power calculations (Version 2.12)*. Groningen, The Netherlands: Author.

Castro-Schilo, L., & Grimm, K. J. (this issue). Using residualized change versus difference scores for longitudinal research. *Journal of Social and Personal Relationships*.

Cobb, S. (1976). Social support as a moderator of life stress. *Psychosomatic Medicine*, *38*, 300–314.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *3*, 145–153.

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

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

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, *62*, 583–619.

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology*, *108*, 275–297.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis* (Vol. *998*). Hoboken, NJ: John Wiley & Sons.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, England: Cambridge University Press.

Haber, M. G., Cohen, J. L., Lucas, T., & Baltes, B. B. (2007). The relationship between self-reported received and perceived social support: A meta-analytic review. *American Journal of Community Psychology*, *39*, 133–144.

Hedeker, D., Gibbons, R. D., & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. *Journal of Educational and Behavior Statistics*, *24*, 70–93.

House, J. S., Landis, K. R., & Umberson, D. (1988). Social relationships and health. *Science*, *241*, 540–545.

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

IBM, Inc. (2013). *IBM SPSS Statistics (Version 22)*. Armonk, NY: Author.

Iida, M., Seidman, & Shrout, P. E. (this issue). Models of interdependent individuals and dyadic process in relationship research. *Journal of Social and Personal Relationships*.

Kashy, D. A., & Kenny, D. A. (1999). The analysis of data from dyads and groups. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social psychology*. New York, NY: Cambridge University Press.

Kelley, K., & Rausch, J. R. (2011). Sample size planning for longitudinal models: Accuracy in parameter estimation for polynomial change parameters. *Psychological Methods*, *16*, 391–405.

Kenny, D. A. (1996). Models of nonindependence in dyadic research. *Journal of Social and Personal Relationships*, *13*, 279–294.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York, NY: Guilford Press.

Kessler, R. C., & McLeod, J. D. (1985). Social support and mental health in community samples. In S. Cohen & S. L. Syme (Eds.), *Social support and health* (pp. 219–240). San Diego, CA: Academic Press.

LaRocco, J. M., House, J. S., & French, J. R. P., Jr. (1980). Social support, occupational stress, and health. *Journal of Health and Social Behavior*, *21*, 202–218.

Laurenceau, J. P., & Bolger, N. (2005). Using diary methods to study marital and family processes. *Journal of Family Psychology*, *19*, 86–97.

Leavy, R. L. (1983). Social support and psychological disorder: A review. *Journal of Community Psychology*, *11*, 3–21.

Liu, G., & Liang, K.-Y. (1997). Sample size calculations for studies with correlated observations. *Biometrics*, *53*, 937–947.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

Moerbeek, M., & Teerenstra, S. (2016). *Power analysis of trials with multilevel data*. Boca Raton, FL: CRC Press.

Moerbeek, M., Van Breukelen, G. J., & Berger, M. P. (2008). Optimal designs for multilevel studies. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 177–205). New York, NY: Springer.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed). Los Angeles, CA: Author.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*, 287–312.

R Development Core Team. (2015). *R: A language and environment for statistical computing (Version 3.2.3)*. Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W., Brennan, R. T., & Barnett, R. C. (1995). A multivariate hierarchical model for studying psychological change within married couples. *Journal of Family Psychology*, *9*, 161–174.

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

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X.-F., Martinez, A., & Bloom, H. (2011). *Optimal design* (Version 3.01). Ann Arbor, MI: HLM Software. Retrieved from http://hlmsoft.net/od/

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.

Rogers, K., Wood, D., & Furr, R. M. (this issue). Assessment of similarity and self-other agreement in dyadic relationships: A guide to best practices. *Journal of Social and Personal Relationships*.

SAS Institute, Inc. (2013). *SAS 9.4*. Cary, NC: Author.

Shrout, P.E., Bolger, N., Iida, M., Burke, C., Gleason, M.E.J., & Lane, S.P. (2010). The mixed effects of daily support transactions during acute stress: Results from a diary study of bar exam preparation. In K. Sullivan & J. Davila (Eds.), *Support processes in intimate relationships* (pp. 175–199). New York, NY: Oxford University Press.

Shrout, P. E., Herman, C. M., & Bolger, N. (2006). The costs and benefits of practical and emotional support on adjustment: A daily diary study of couples experiencing acute stress. *Personal Relationships*, *13*, 115–134.

Snijders, T. A. (2005). Power and sample size in multilevel modeling. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. *3*, pp. 1570–1573). Chichester, UK: Wiley.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*, 237–259.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, England: Sage.

Stern, C., & West, T. V. (this issue). Assessing accuracy in close relationships research: A truth and bias approach. *Journal of Social and Personal Relationships*.

Taylor, S. E. (2007). Social support. In H. S. Friedman & R. C. Silver (Eds.), *Foundations of health psychology* (pp. 145–171). New York, NY: Oxford University Press.

Zhang, Z., & Wang, L. (2009). Statistical power analysis for growth curve models using SAS. *Behavior Research Methods*, *41*, 1083–1094.