Bolger, N., Stadler, G., & Laurenceau, J-P. (2012). Power analysis for intensive longitudinal measurement designs. In M. Mehl & T. Conner (Eds.), *Handbook of Research Methods for Studying Daily Life*. New York: Guilford.

CHAPTER 16

Power Analysis for Intensive Longitudinal Studies

NIALL BOLGER
GERTRAUD STADLER
JEAN-PHILIPPE LAURENCEAU

Intensive longitudinal studies are not for the faint of heart: Each one typically requires a lot of time, effort, and financial resources. Before commencing such a study, therefore, it seems worthwhile to assess its statistical power: Namely, given the proposed design and sample size, what is the probability of detecting a hypothesized effect if one actually exists? If that probability is, for example, 0.5, then it may not make sense to proceed. Why embark on an arduous project if it has only a 50:50 chance of success? Until recently, however, researchers wishing to conduct power analyses for intensive longitudinal studies had few resources on which to draw. The situation has recently improved to the extent that it is now possible to give basic advice on how to carry out a power analysis using widely available software packages. That is the subject of our chapter.

First, the good news: The exercise of conducting a power analysis for an intensive longitudinal study greatly increases understanding of those designs and of how that can be used to capture the phenomenon of interest. The not so good news: Conducting a power analysis for intensive longitudinal studies is considerably more challenging than for simpler designs. Because intensive longitudinal studies involve multiple sources of random variation, one needs to make assumptions about each source in order to do the required calculations. It is especially helpful, therefore, to have some prior data available upon which to base the assumptions.

Researchers who want to conduct a power analysis for an intensive longitudinal study can choose from several options: (1) working with the power formulae available in books covering multilevel modeling, repeated measures designs, and longitudinal designs (Fitzmaurice, Laird, & Ware, 2004; Gelman & Hill, 2007; Hox, 2010; Moerbeek, Van Breukelen, & Berger, 2008; Snijders & Bosker, 1993, 1999); (2) using specialized software designed for power analyses for multilevel and longitudinal models, for example, the freely available PinT (Bosker, Snijders, & Guldemond, 2007), Optimal Design (Rauden-

bush, Spybrook, Liu, & Congdon, 2006), and RMASS2 (Hedeker, Gibbons, & Waternaux, 1999); and (3) using simulation methods in general purpose programming software, such as Mplus 6.2 (Muthén & Muthén, 1998–2007), Statistical Analysis Software (SAS; Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006; SAS Institute, Inc., 2010), R (R Development Core Team, 2011), or MATLAB (MathWorks, Inc., 2011).

Each of these approaches has advantages and drawbacks. The approach we advocate and demonstrate in this chapter is to use simulation methods in Mplus, as proposed by Muthén and Muthén (2002). A free demonstration version of Mplus is available at *www.statmodel.com*. Unfortunately, to run the model described in this chapter with the demonstration version of Mplus, one needs to drop one predictor variable. We return to the justification of the model and the risks of using a simpler version later in the chapter.

The approach we advocate requires some initial effort to translate the research question into Mplus syntax. This effort could be substantial for those not already familiar with Mplus and the logic of simulation. Nevertheless, the payoff will be large because this approach is flexible enough to accommodate a wide variety of intensive longitudinal designs, and it can be used for simple multilevel models.

The Basics

The problem: Researchers want to draw a conclusion about a phenomenon they hypothesize to exist in a population, but usually they rely only on data from a sample. Now statistical theory tells us how we can make the inferential leap from sample to population, but it forces us to accept that there will be considerable uncertainty in our inferences. At its core, a *power analysis* is an assessment of whether the population effect size or signal strength the researcher wants to be able to detect is large enough to be detected given the uncertainty or noise due to the use of a sample.

The population *effect size* or *signal strength* is often a mean difference or regression slope. For example, a researcher might hypothesize that each additional hour of exercise by the average subject in a population reduces end-of-day depression by 0.15 units. The effect can also be a variance, such as how much people in the population differ from one another in their regression slopes. Some readers may find our use of the term *effect size* off-putting given that we use it in cases where the effect is based on an experimental manipulation, and also where it is simply a summary of a relationship between two nonmanipulated variables. This usage, however, is extremely common, and rather than risk confusing readers with an unfamiliar but more accurate term, we opted to follow common practice.

Uncertainty or *noise*, in statistical terms, is how much the sample estimate of the population effect varies from sample to sample. For example, if the previous exercise slope predicting depression were –0.15 in one sample and –1.10 in a replication sample, then researchers should not trust either as being an accurate reflection of the true population value. When the effect size of interest involves a single degree of freedom (e.g., a mean difference or regression slope) rather than multiple degrees of freedom (e.g., differences among three conditions), the standard error of the effect under investigation is used to assess the amount of statistical noise in the estimate (Rosenthal, Rosnow, & Rubin, 2000).

As noted, hypothesis tests can be thought of as assessment of signal-to-noise ratios. It is not surprising, then, that many common statistical tests involve computing the ratio of the sample effect size (signal strength) to its standard error (noise strength), resulting in a t or a z value. To conduct a power analysis, a researcher is fundamentally asking, "If I want to have a good chance of detecting the hypothesized effect size, and if I carry out the study as planned, how big a standard error is the effect size of interest likely to have?" Designs that result in a smaller standard error are better able to detect a given effect size and therefore have greater power.

Which factors influence power? Almost 80 years ago, Neyman and Pearson (1933) developed the appropriate statistical framework to answer this question. Whereas the existing approach to hypothesis testing developed by Fisher (1925) was concerned only with accepting or rejecting a null hypothesis (usually the null hypothesis of no effect in the population), Neyman and Pearson introduced the idea of an alternative hypothesis, and through this, the idea of a population effect size. In their framework, it was possible to hypothesize a particular population effect size and ask about the probability that samples from such a population were likely to detect that effect.

So, for a power analysis, we work from the assumption that the population effect size is a particular nonzero value. Jacob Cohen, beginning in the 1960s, simplified this process by suggesting particular nonzero values that could be used for particular research questions. For tests of mean differences, for example, he suggested that population differences of 0.2, 0.5, and 0.8 standard deviation units be regarded as "small," "medium," and "large," respectively (Cohen, 1969, 1988). Not surprisingly, power depends on the size of the effect that one wishes to detect. A particular study could have sufficient power to detect, in Cohen's terminology, a medium or large effect size, but not a small one. Also, a study could have sufficient power to test one hypothesis (e.g., a hypothesis regarding group differences) but insufficient power to test a more complicated hypothesis (e.g., a hypothesis that group differences depend on the age of the participants).

At this point we provide some statistical notation that will be helpful in building a concrete example of a power analysis for intensive longitudinal data. Equation 1 represents a population regression model, where Y_i, a continuous dependent variable, is regressed on one predictor, X_i, where i indexes subjects. Assume that X_i can be continuous or a binary categorical variable. For the purposes of exposition, assume that Y_i is subject i's score on a depression scale, and that X_i is a binary treatment versus control variable, where subjects in the control group are coded 0 and those in the treatment group are coded 1. In this case, β_0 is the level of depression of people in the control group, $\beta_0 + \beta_1$ is the level of depression in the treatment group, and β_1 is the difference between the two groups, namely, the treatment effect on depression. The error term, ε_i, represents each subject i's deviation from the true mean score of their respective groups.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1}$$

Assume now that we draw a sample from a population where this statistical model holds. For this model, the power of any given sample to detect the effect β_1 of the treatment X_i is determined by five main factors. These are summarized in Equation 2:

$$Pwr = f(\beta_1, N, \sigma_X^2, \sigma_E^2, \alpha) \tag{2}$$

The first factor is β_1, the effect size for the treatment. If, for example, the treatment lowered depression by 2.5 units, the effect size would be $\beta_1 = -2.5$. Not surprisingly, the bigger the absolute value of the effect size, the easier it is for any given sample to detect it. Although it is often useful to express effect sizes in standardized units, such as those developed by Cohen, unstandardized effect sizes are perfectly legitimate. Note that the effect size in Equation 2 is an unstandardized regression coefficient.

The second factor is the sample size, *N*. Larger sample sizes make it easier to detect a population effect. The third factor is σ_X^2, the extent to which *X* varies. For example, a treatment study with most of the participants in the control condition would have less power than a design with the same sample size but equal numbers of participants in each condition. The fourth factor is σ_E^2, the extent to which there is a lot of or a little unexplained variance in *Y*. If, in our example, we were sampling from a population where depression scores had high between-subject variability, our study would have less power than it would have were we sampling from a population with lower between-subject variability. Finally, power is a function of what α, or Type I error probability, we choose to use. The more stringent the α (e.g., using .01 instead of .05), the lower the power.

In experimental or intervention studies, the investigators have control over some but not all of these factors. Thus, a stronger manipulation in an experiment will increase the effect size β_1 and therefore the probability of concluding that an effect actually exists. Similarly, it is common in experimental studies to have some control over the sample size, as well as ensuring equal sample sizes across the conditions of *X*. The unexplained variability can be reduced by standard experimental techniques of blocking or matching. The Type I error rate is under the investigator's control in theory, but accumulated norms have become so strong that in practice most investigators use .05, and it is rare to see investigators use alphas more lenient than .10.

Power in Intensive Longitudinal Studies

All things being equal, within-subject designs tend to have more power than between-subject designs (Maxwell & Delaney, 2004). In within-subject experimental designs, for example, we compare subjects to themselves under different conditions; in between-subject designs we compare subjects in one condition to subjects in other conditions. Thus, in within-subject experimental designs, stable, systematic differences between subjects (e.g., neighborhood quality, social desirability, extraversion, or liking of physical activity) cannot contribute to the within-subject error variance because these individual differences affect all the repeated measures of a subject in the same way. By contrast, stable features of the individual and his or her environment do contribute to error variance in between-subject experimental designs.

This benefit does not necessarily apply to nonexperimental within-subject designs, of which intensive longitudinal studies are a prominent example. In these nonexperimental designs we are often interested in assessing possible causal effects of within-subject changes in *X*, but we fail to realize that between-subject differences in mean *X* can bias estimates of such effects. This problem has been recognized for decades in econometrics (e.g., Judge, Griffiths, Hill, & Lee, 1980) but was poorly understood in other branches of social science until it was highlighted in recent years by Allison (2005, 2009; see also

Curran & Bauer, 2011; Hoffman & Stawski, 2009). There are various ways of dealing with this problem; the one we use requires (1) that all within-subject X's are decomposed into their between- and within-subject components, (2) that both components are included as predictors in analyses, and (3) that only the coefficients for the within-subject components are taken as evidence of within-subject effects.

As we did earlier for standard single-level power analysis, we begin by specifying a population statistical model that we regard as adequate to handle basic analyses of intensive longitudinal data. We use a continuous outcome, but in this case it varies over subjects and over time. Thus in Equation 3 below, Y_{it} is the score on Y for subject i on measurement occasion t. For the sample dataset presented later, we analyze end-of-day depression scores for $N = 66$ subjects assessed for $T = 9$ consecutive days. The predictor variable X_{it} also varies over subjects and time, and in our empirical example this will be the amount of physical activity a subject i engaged in on day t. However, as noted, to avoid potential bias in assessing within-subject effects, in the analysis we do not use X_{it} in its original form; rather, we split it into two components, XB_i, the component that varies between subjects only, and XW_{it}, the component that varies within subjects only. XB_i is each subject's mean value on X_{it} averaging over all occasions T; XW_{it} is how much on occasion t a subject i's score is higher or lower than his or her mean level XB_i. Given that the original X_{it} is simply the sum of XB_i and XW_i, no information is lost in creating these new variables.

Equation 3 represents a population multilevel regression with three predictors. Please note: We use a single overall equation in Equation 3 to facilitate comparison with the single-level regression model presented in Equation 1. This contrasts with many expositions of multilevel statistical models, where at least one separate equation is presented for each level of analysis. Also, we use the letter β where it is more common to use γ, again to facilitate comparison with Equation 1.

$$Y_{it} = \beta_{0i} + \beta_{1i}XW_{it} + \beta_2 XB_i + \beta_3 T_t + \upsilon_{it} \tag{3}$$

The coefficients β_{0i} and β_{1i} are random variables representing subject-specific intercepts and XW slopes, respectively. We assume that the coefficients β_{0i} and β_{1i} are normally distributed with a population mean and variance that we represent as $\bar{\beta}_0$ and $\sigma^2_{\beta_0}$, and $\bar{\beta}_i$ and $\sigma^2_{\beta_1}$, respectively. The third coefficient, β_2, shows how strongly between-subject differences in X relate to Y. It is not unusual for it to be very different in size from $\bar{\beta}_1$, the coefficient for within-subject X for the average subject (see, e.g., Bolger & Schilling, 1991). The fourth coefficient, β_3, is the time slope, the average change in outcome Y per time unit T. We assume in this case that the time effect does not vary across participants. It is important to include time in the model because time can be an influence on both X and Y and can lead to a spurious relationship between the two. We also include an error term, υ_{it}, specific to each subject and occasion. As is common in longitudinal and time series analysis, we allow adjacent error terms to be correlated. This permits us to decompose the error term into an autocorrelation component $\rho\upsilon_{it-1}$, and a pure random error component ε_{it}. We assume that ε_{it} is a normally distributed random variable with a mean of 0 and a variance of σ^2_ε that is constant over subjects and occasions. Equation 4 is as follows:

$$\upsilon_{it} = \rho\upsilon_{it-1} + \varepsilon_{it} \tag{4}$$

Now that we have specified the complete model, we can, as before, identify the factors that influence the power of any given sample to detect an effect. In this case, the effect of interest is $\overline{\beta}_i$, the effect of the within-subject predictor *XW*. There are eight factors in all, three more than we discussed for the between-subject regression model, as Equation 5 shows.

$$Pwr = f(\overline{\beta}_i, N, T, \sigma^2_{XW}, \sigma^2_{\beta_1}, \rho, \sigma^2_\varepsilon, \alpha) \quad (5)$$

The first is $\overline{\beta}_i$, the effect size for within-subject *X* for the average subject in the population. The bigger this is, the more power a given sample has to detect it. Next is *N*, the number of subjects in the sample; more subjects means more power. The third factor is *T*, the total number of measurement occasions per subject. Again, the more occasions *T* over which *XW* is measured, the more power to detect its effect. The fourth factor is the variance in *XW*, σ^2_{XW}. The more *X* varies within each subject, the more power the study has to detect within-subject effects. The fifth factor is $\sigma^2_{\beta_1}$, how much people differ in the within-subject effect of *X* on *Y*. The more people in the population differ from one another in this effect, the harder it is to be confident that the average person has a nonzero coefficient. The sixth is ρ, the autocorrelation coefficient. The larger the autocorrelation, the less new information is gained when an additional measurement occasion is added to a study and, therefore, the lower the power of the study to detect effects of *XW*. The seventh is the true within-subject random noise, σ^2_ε; the greater this is, the lower the power. Finally, as before, as the significance level α is made more stringent (i.e., made smaller), power is decreased.

Physical Activity and Depression Example

The best way to learn about a power analysis for intensive longitudinal data is by doing one, and a goal of this chapter is to provide the reader with the necessary example data and software commands. In this section we illustrate the process using a simulated dataset on daily physical activity and depressive symptoms. The data were simulated to resemble an actual dataset of 66 participants who wore accelerometers and completed evening diaries for a period of 9 consecutive days.

The power analysis involves two main steps. Step 1 involves using the sample dataset to get descriptive statistics for the predictor variables and to estimate the equivalent of Equations 3 and 4 using standard multilevel or mixed model software that allow for autocorrelated within-subject errors. Below we provide the syntax for doing so in SAS and IBM SPSS (IBM Inc., 2010). We believe that comparable analyses can be accomplished using other software such as R, HLM (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011), and MLwin (Rasbash, Charlton, Browne, Healy, & Cameron, 2011).

Step 2 involves using the parameter estimates obtained in Step 1 together with predictor descriptive statistics (means and variances) as population values for Monte Carlo simulations using Mplus (as described in Muthén & Muthén, 2002). One uses Mplus to (1) specify a population multilevel model with specific parameter values (e.g., for $\overline{\beta}_1$, σ^2_{XW}, $\sigma^2_{\beta_1}$, from Equation 3), (2) generate many (e.g., 1,000) samples with one specific combination of *N* and *T* (e.g., $N = 66$, $T = 9$) from such a population model, (3) analyze each simulated sample as if it were a real one in order to obtain estimates of those same specific

population parameters, and (4) report what proportion of the samples resulted in statistically significant results (using $p < .05$ as the α level) for any given parameter.

To be more concrete: Imagine that one wishes to determine the power to detect a $\bar{\beta}_1$ effect size of 0.20 units (e.g., the within-subject association between time-varying physical activity and depression for the average subject), given that one has specified values for the other parameters in the model (e.g., a $\sigma^2_{\beta_1}$ of 0.50) and one has chosen a particular N and T study design. If the simulations revealed that estimates of $\bar{\beta}_1$ were significant in 60% of the samples, this would mean that this particular study design had a power of .60 to detect the specified $\bar{\beta}_1$ effect size.

This procedure is complex, but in our opinion it is the best that is currently available for assessing power in intensive longitudinal studies. If you found the overview paragraph above hard to follow, then try it again after reading through the worked example below. You can download the full sample dataset (exampledata.sas7bdat for SAS and exampledata.sav for SPSS) and syntaxes from our website at *www.columbia.edu/~nb2229/publications.html*. As is typical for such datasets, it is in a vertical or long format: Each subject has multiple data lines, one for each time point.

Table 16.1 shows the structure of the simulated dataset, including data lines for particular subjects. To facilitate interpretation of the analysis results, all independent

TABLE 16.1. Data Example: Daily Physical Activity and Evening Depression

id	day	dayc5	depr	steps	stepsB	stepsW	stepsB_GM	stepsBc
1	1	−4	3.09	1.61	0.91	0.71	0.98	−0.08
1	2	−3	2.63	0.79	0.91	−0.12	0.98	−0.08
1	3	−2	3.17	1.08	0.91	0.17	0.98	−0.08
1	4	−1	1.98	1.67	0.91	0.76	0.98	−0.08
1	5	0	3.00	0.75	0.91	−0.15	0.98	−0.08
1	6	1	3.32	0.80	0.91	−0.11	0.98	−0.08
1	7	2	2.06	0.26	0.91	−0.64	0.98	−0.08
1	8	3	2.02	0.81	0.91	−0.10	0.98	−0.08
1	9	4	1.31	0.38	0.91	−0.53	0.98	−0.08
2	1	−4	2.29	0.93	0.62	0.31	0.98	−0.36
2	2	−3	2.95	0.42	0.62	−0.19	0.98	−0.36
2	3	−2	2.86	0.68	0.62	0.06	0.98	−0.36
2	4	−1	3.33	0.50	0.62	−0.12	0.98	−0.36
2	5	0	2.56	0.39	0.62	−0.22	0.98	−0.36
2	6	1	3.18	0.56	0.62	−0.06	0.98	−0.36
2	7	2	2.96	0.80	0.62	0.18	0.98	−0.36
2	8	3	4.18	0.65	0.62	0.04	0.98	−0.36
2	9	4	2.26	0.62	0.62	0.00	0.98	−0.36
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
66	1	−4	1.45	1.37	1.07	0.30	0.98	0.09
66	2	−3	3.08	0.45	1.07	−0.62	0.98	0.09
66	3	−2	2.64	1.52	1.07	0.45	0.98	0.09
66	4	−1	3.04	1.67	1.07	0.60	0.98	0.09
66	5	0	1.75	0.97	1.07	−0.10	0.98	0.09
66	6	1	2.91	1.01	1.07	−0.06	0.98	0.09
66	7	2	3.27	0.58	1.07	−0.49	0.98	0.09
66	8	3	2.67	1.13	1.07	0.06	0.98	0.09
66	9	4	1.68	0.92	1.07	−0.15	0.98	0.09

variables were centered either on the grand mean or the subject mean (see Cohen, Cohen, West, & Aiken, 2003, pp. 564–565). Table 16.1, column 4 contains scores on the dependent variable, depr. The depression variable was simulated to resemble a mean of several depression items, each scaled from 0 to 5, as might be reported at the end of each day in an online diary. Variables indexing time are in columns 2 and 3: Day is simply the study day coded 1–9. Dayc5 is a version of day centered on the middle day of the study (Day 5).

Column 5 shows the simulated physical activity variable, steps; this is the number of steps taken each day (where 1 unit equals 10,000 steps), as might be obtained from a portable accelerometer. Physical activity was simulated to vary both within subject and between subject. Therefore, as with the between- and within-subject X variable in Equation 3 earlier, steps was split into two predictor variables: (1) stepsB, the between-subjects means of steps across all diary days and (2) stepsW, the within-subject deviations from these between-subject means. StepsBc, a centered version of stepsB, was created by subtracting the grand mean (stepsB_GM) from stepsB.

Because the later simulations require means and variances for all predictor variables, these descriptive statistics need to be calculated (e.g., using PROC UNIVARIATE in SAS or using DESCRIPTIVES in SPSS). StepsBc, the grand mean–centered subject means across all available diary days, has a mean of 0.00 (due to centering) and a variance of 0.096. StepsW, the within-subject deviations from each subject's mean, has a mean of 0.00 and variance of 0.157.

We simulated the dataset to have missing observations on 15% of days; this level of missingness corresponds to what has been observed in a real dataset on physical activity and depression. Therefore, instead of a complete dataset with 9 time points for all 66 participants (i.e., 594 total observations), the current dataset contains a total of 507 observations, which corresponds to having an average of 7.7 out of 9 observations.

We now turn to accomplishing Step 1 using PROC MIXED in SAS. To review, Step 1 involves analyzing the sample dataset to obtain estimates of the parameters that we will use together with predictor means and variances in the power analysis in Step 2. Below is the exact SAS PROC MIXED syntax:

```
*SAS Mixed Model for Physical Activity and Depression Dataset;
PROC MIXED DATA=power.exampledata METHOD=ml COVTEST;
   CLASS id day ;
   MODEL depr = stepsBc stepsW dayc5 /S DDF = 64, 65, 65;
   RANDOM int stepsW /SUBJECT = id TYPE = un;
   REPEATED day /SUBJECT = id TYPE = ar(1);
RUN;
```

Note that the SAS syntax requires two variables for time: a version for use in the MODEL statement and a version for use in the REPEATED statement. For the first, we used dayc5, the variable day centered at Day 5; for the second, we used the uncentered time variable, called day. The variable dayc5 is used in the model statement to get an estimate of the fixed effect of time (β_3 in Equation 3); this variable could be included in the random statement if we wanted to include a random effect for time. The time variable, day, in the repeated statement is used in obtaining an estimate of ρ, the first-order autocorrelation parameter. This variable also has to appear in the CLASS statement.

First-order autocorrelation is specified using TYPE = AR(1) in the REPEATED statement. An alternative choice for the error covariance structure would be a Toeplitz structure; for example, to capture a lag – 1 correlation with zero correlation at longer lags, we substitute `type = toep(1)` for `type = ar(1)`.

Note also that the estimation method we used was full information maximum likelihood (METHOD=ML) rather than the more standard restricted maximum likelihood (METHOD=REML). We did so in order to be consistent with the simulation software to be used in Step 2; that software, Mplus, does not allow REML estimation. For intensive longitudinal datasets, the difference in estimator will usually matter very little. Finally, rather than use the PROC MIXED defaults, we used the DDF option on the MODEL statement to specify degrees of freedom for tests of fixed effects. For the between-subject predictor stepsBc, we used *N* minus 2 degrees of freedom, that is, 66 – 2 = 64, because there were two between-subject predictors, the intercept and stepsBc. For the within-subject predictors stepsW and dayc5, we used *N* minus 1 degrees of freedom, that is, 66 – 1 = 65, because there were no cross-level interactions. For more details on relevant options in SAS PROC MIXED, see Bolger and Laurenceau (in press).

The simulated example data used in this chapter do not show much remaining autocorrelation ($\hat{\rho} = -0.048$), but many intensive longitudinal datasets do show appreciable autocorrelation, and we believe it is important to include this parameter in the model. Failure to adjust for autocorrelation results in overly small estimates of within-subject error variance, and causes an upward bias in test statistics, such at *t* values and *F* values (Fitzmaurice et al., 2004).

Running this model in SAS gives the following output.

Dimensions

Covariance Parameters	5
Columns in X	4
Columns in Z Per Subject	2
Subjects	66
Max Obs Per Subject	9

Number of Observations

Number of Observations Read	507
Number of Observations Used	507
Number of Observations Not Used	0

Solution for Fixed Effects

	Effect	Estimate	Standard Error	DF	t value	Pr > |t|
	Intercept	2.6624	0.0575	64	46.30	<.0001
	stepsBc	-0.6748	0.1899	64	-3.55	0.0007
⇨	stepsW	-0.1857	0.0978	65	-1.90	0.0621
	dayc5	-0.0138	0.0110	65	-1.25	0.2166

Covariance Parameter Estimates					
Cov Parm	**Subject**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
UN(1,1)	**id**	0.1688	0.0386	4.38	<.0001
UN(2,1)	**id**	-0.0152	0.0464	-0.33	0.7437
UN(2,2)	**id**	0.2154	0.0947	2.27	0.0115
AR(1)	**id**	-0.0475	0.0605	-0.78	0.4330
Residual		0.3946	0.0285	13.85	<.0001

The following SPSS syntax gives the same results:

```
*SPSS Mixed Model for Physical Activity and Depression Dataset.
MIXED depr WITH stepsBc stepsW dayc5
 /FIXED=stepsBc stepsW dayc5 | SSTYPE(3)
 /METHOD=ML
 /PRINT=SOLUTION
 /RANDOM=INTERCEPT stepsW | SUBJECT(id) COVTYPE(un)
 /REPEATED=day | SUBJECT(id) COVTYPE(AR1).
```

Table 16.2 organizes these mixed model results in a form recommended for journal articles (see American Psychological Association, 2009, pp. 147–148; Bolger & Laurenceau, in press). Given that diary and intensive longitudinal studies are usually carried out to assess within-subject associations, the single most important parameter estimate is $\hat{\beta}_1$, the within-subject relationship between daily physical activity and depression for the average subject: On days when the typical subject's physical activity was one unit above his or her typical level, daily depression was lower by −0.19 units. An effect of this size is small but not trivial: It is equivalent to a Cohen's *d* of 0.28 within-person *SD* units (see notes at the bottom of Table 16.2). The sampling error for the unstandardized estimate is sufficiently large, however, that using a 95% confidence interval we cannot rule out zero as a possible population value (95% CI = −0.38, 0.01).

We simulated these data to produce inconclusive results because this is a common situation in the early stages of a research program; for example, when researchers are in the process of developing a grant proposal they often conduct a small and underpowered pilot study. It is in this situation that one needs guidance on how much to increase the number of persons and time points to ensure sufficiently powered future studies. Fortunately, the ability to conduct such power simulations for diary and intensive longitudinal studies has recently become available.

Step 2 begins with the researcher conducting a power simulation using the *N* and *T* for the initial dataset, the predictor means and variances, and the parameter estimates from the multilevel model in Step 1. Note that although the simulated pilot study (and most real diary studies) had missing data, we ran the power simulations under the assumption of no missing data. Making such an assumption did not alter the power estimates appreciably and, as will be seen, it simplifies the task of creating "what-if" scenarios where the number of persons and time points are systematically varied.

The model used in the power simulations differed from the model estimated on the pilot data in two other ways: it did not include an autocorrelation parameter and it did

TABLE 16.2. Parameter Estimates for Multilevel Model of Daily Physical Activity Predicting Evening Depression

		Parameter	Estimate	*SE*
Fixed effect				
Intercept	Intercept	$\overline{\beta}_0$	2.66 **	0.058
Activity	stepsW	$\overline{\beta}_1$	−0.19 †	0.098
	stepsBc	β_2	−0.68 **	0.190
Time	dayc5	β_3	−0.01	0.011
Random effects				
Within-subject:	Error	σ^2_ε	0.39 **	0.029
	Autocorrelation	ρ	−0.05	0.061
Between-subject:	Intercept	$\sigma^2_{\beta_0}$	0.17 **	0.039
	stepsW	$\sigma^2_{\beta_1}$	0.22 *	0.095
	Covariance	$\sigma_{(\beta_0, \beta_1)}$	−0.02	0.046

Note. The number of subjects was 66; the number of days averaged 7.7 out of a possible 9; and the total number of observations was 507 out of a possible 594 (85%). The variance decomposition for evening depression was 0.65, 0.20, and 0.45, for total, between-subject, and within-subject variances, respectively. The corresponding standard deviations were 0.81, 0.45, and 0.67. The estimate for $\overline{\beta}_1$ in within-person *SD* units is 0.19/0.67 = 0.28. This corresponds to a small effect size in Cohen's (1988) terminology (see Bolger & Laurenceau, in press, for discussion of effect sizes in diary and intensive longitudinal designs).
$^{\dagger}p < .10$; $^{*}p < .05$; $^{**}p < .01$.

not include a time trend. The first omission is because Mplus cannot specify autocorrelation when data are analyzed in the long form (as they almost invariably are for diary and intensive longitudinal data). The second omission is because the free demonstration version of Mplus allows no more than two independent variables. Given the choice of dropping stepsW, stepsB, or dayc5, we dropped dayc5, the nonsignificant time trend variable.

The reader may wonder about the value of a power simulation that omits parameters and variables that would be included in typical analysis models for diary and intensive longitudinal data. The reason it is possible to omit these in the simulations is that the omitted variables and parameters are controls only. Their inclusion in the PROC MIXED analyses was to ensure that the other, central parameters such as the fixed and random effects of stepsW would be free of bias. By conducting simulation using these presumably debiased parameter estimates, we are in effect obtaining the results that would be obtained had we included these controls. We urge readers who possess the full version of Mplus to run simulations with the fully parametrized model to convince themselves that this is the case (more on this point later).

The initial power simulation below is one that represents a baseline against which subsequent simulations will be compared. Specifically, it uses the parameter estimates from the pilot data as population values and also uses the number of persons and time points that would have been observed had there been no missing data. With these initial settings, the results of the simulation should tell us what we already know, that the pilot study was underpowered for estimating the within-subject effect of physical activity on depression. Having conducted this baseline simulation, we then proceed to specifying

various what-if scenarios of the kind that might be used in developing a grant application.

The Mplus syntax is given next. Syntax lines with specific comments are explained later in the text (e.g., the line with the comment ! (a) is explained under the heading (a) in the later text).

```
!Exclamation point indicates a comment
MONTECARLO:NAMES ARE depr stepsW stepsBc; !{a}
      NOBSERVATIONS = 594; !{b}
      NCSIZES = 1;
      CSIZES = 66 (9); !{c}
      SEED = 5859; !{d}
      NREPS = 1000; !{e}
      WITHIN = stepsW ; !{f}
      BETWEEN = stepsBc; !{g}
ANALYSIS: TYPE = TWOLEVEL RANDOM; !{h}

MODEL POPULATION:
      %WITHIN% !{i}
      slope | depr ON stepsW;
      [ stepsW*0.00 ];
      depr*0.395; stepsW*0.157;
      %BETWEEN% !{j}
      depr ON stepsBc*-0.675;
      depr WITH slope*-0.015;
      [ depr*2.662 ]; [ slope*-0.186 ]; [ stepsBc*0 ];
      depr*0.169; slope*0.215; stepsBc*0.096;
MODEL: !{k}
      %WITHIN%
      slope | depr ON stepsW;
      [ stepsW*0.00 ];
      depr*0.395; stepsW*0.157;
      %BETWEEN%
      depr ON stepsBc*-0.675;
      depr WITH slope*-0.015;
      [ depr*2.662 ]; [ slope*-0.186 ]; [ stepsBc*0 ];
   depr*0.169; slope*0.215; stepsBc*0.096;
```

The syntax contains the following pieces of information:

a. a list of the variables in the model (in our example, depr, stepsW, stepsBc);
b. the total number of observations calculated as the product of the number of subjects N and the number of time points T (i.e., 594 = 66 subjects * 9 time points);
c. the number of subjects N (66) and time points T (9);
d. an arbitrary numerical seed (5859) that enables one to rerun the simulation at a later time and get the same results;
e. the number of simulations (1000);
f. predictors that vary within subjects only (stepsW);
g. predictors that vary between subjects only (stepsBc);

h. the type of model to be run, a two-level random effects model (TYPE = TWOLEVEL RANDOM);
i. specifications for the within-subject part of the model, including
   - a random slope for the within-subject predictor (slope | depr ON stepsW), the within-subject residual variance (depr*0.395) obtained from the SAS/SPSS multilevel model;
   - the variance (stepsW*0.157) and the mean of the within-subject predictor ([stepsW*0.00]) obtained from the descriptive statistics of stepsW;
j. specifications for the between-subject part of the model, including
   - the fixed effect of the between-subject predictor (depr ON stepsBc*−0.675), the covariance of intercept and the within-subject predictor's slope (depr WITH slope*−0.015), and the means of the intercept and the within-subject predictor's slope ([depr*2.662], [slope*−0.186]), all obtained from the SAS/SPSS multilevel model;
   - the mean of the between-subject predictor ([stepsBc*0]) obtained from the descriptive statistics of stepsBc;
   - the variance of the intercept and the variance of the within-subject predictor's slope obtained from the SAS/SPSS multilevel model (depr*0.169; slope*0.215);
   - the variance of the between-subject predictor (stepsBc*0.096) obtained from the descriptive statistics of stepsBc;
k. a specification of the analysis model that is (in our case) identical to the simulation model.

The following are the results obtained using the demonstration version of Mplus version 6.1.

```
MODEL RESULTS
                ESTIMATES              S. E.    M. S. E.  95%  %Sig
          Population  Average   Std. Dev.  Average          Cover Coeff
Within Level

 Means
   STEPSW       0.000   0.0007     0.0166   0.0161   0.0003   0.938   0.062

 Variances
   STEPSW       0.157   0.1567     0.0091   0.0090   0.0001   0.936   1.000

 Residual Variances
   DEPR         0.395   0.3939     0.0259   0.0253   0.0007   0.934   1.000

Between Level

 DEPR ON
   STEPSBC     -0.675  -0.6728     0.1937   0.1800   0.0375   0.919   0.933

 DEPR WITH
   SLOPE       -0.015  -0.0151     0.0436   0.0409   0.0019   0.941   0.074

 Means
   STEPSBC      0.000   0.0017     0.0380   0.0378   0.0014   0.948   0.052
   SLOPE       -0.186  -0.1795     0.0892   0.0907   0.0080   0.950   0.502
```

Intercepts							
DEPR	2.662	2.6642	0.0577	0.0565	0.0033	0.942	1.000
Variances							
STEPSBC	0.096	0.0950	0.0164	0.0161	0.0003	0.918	1.000
SLOPE	0.215	0.2117	0.0915	0.0907	0.0084	0.911	0.689
Residual Variances							
DEPR	0.169	0.1627	0.0374	0.0352	0.0014	0.897	1.000

The partial output above summarizes the results of the 1,000 simulated samples. The last column of the output (with the header `% Sig Coeff`) gives power estimates for each parameter, that is, the percentage of samples in which the parameter estimate was statistically significant at $\alpha = .05$. The key power estimate is the one for the fixed effect of stepsW, which is the within-subject relationship between daily physical activity and depression for the average subject ($\bar{\beta}_1$ in Equation 3). The estimate is 0.502, which means that in only 50% of the 1,000 simulated samples did the slope for the within-subject predictor reach statistical significance. This, of course, is what one would expect to find given that simulation is based on the original underpowered pilot study.

Power Curves for the Within-Subject Effect for Varying Numbers of Subjects and Time Points

We have now reached the point where we can conduct power simulations that would result in grant writing and planning follow-up studies. We focus on the following pragmatic question that is common in diary and intensive longitudinal studies: Other things being equal, in order to detect a within-subject effect, is it better to increase the number of time points per person or the number of persons per time point? With the earlier Mplus syntax in hand, it is relatively easy to answer this question. To do so, we rerun the syntax, keeping all the simulation parameters the same except for number of persons and time points.

Figure 16.1 displays the results of 10 Mplus power simulations, involving combinations of increased participants and time points. To understand the figure, recall that the baseline power simulation used a combination of 66 subjects and 9 time points, resulting in a total of 594 observations. Now consider a total number of observations of 660. As shown below the horizontal axis in Figure 16.1, this can be accomplished in two ways: by increasing the number of subjects from 66 to 73 (approximately) while keeping the time points at 9, or by increasing the number of time points from 9 to 10 while keeping the subjects at 66. The net result is approximately the same increase in power, from 50% to 59%.

As we move, however, from 660 observations to 990, the power increases begin to diverge. To reach 990 observations, we can keep the time points at 9 and increase the subjects from 73 to 110, or we can keep the subjects at 66 and increase time points from 10 to 15. Figure 16.1 shows that power is improved more by increasing subjects than time points, and the differential increases as we examine further equivalent combinations of persons and time points. A power of 80%, the magic number for grant applications, can

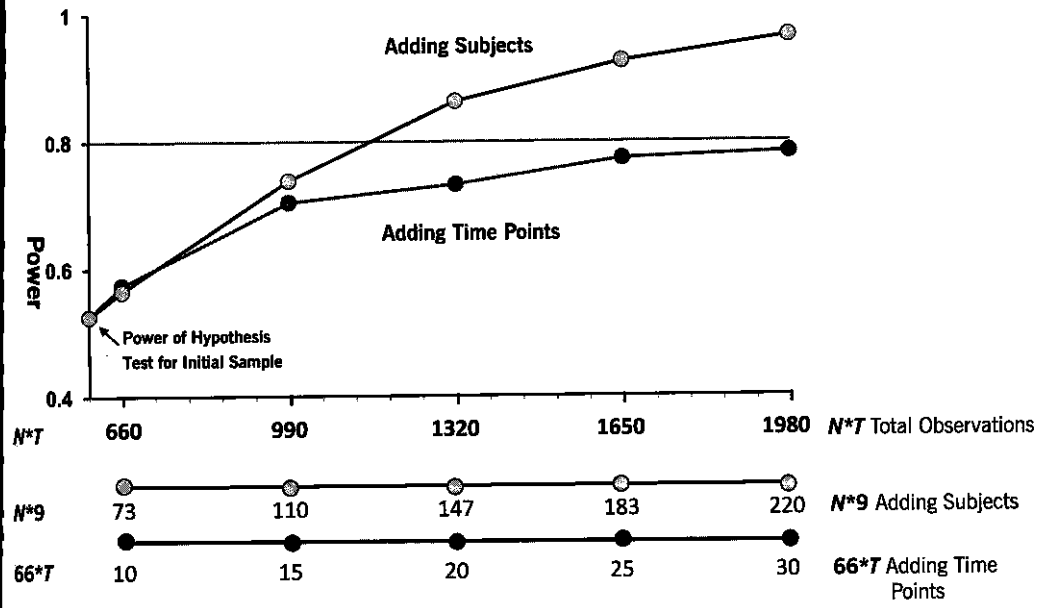

FIGURE 16.1. Power curves for the fixed effect of physical activity: What is the benefit of adding subject versus time points to the sample?

be achieved by increasing the number of subjects to just under 120 while keeping the time points to 9; by contrast, even were we to increase the number of time points to 30, while keeping subjects at 66, we would still only reach a power of 78%.

Figure 16.1 shows nicely the phenomenon that has been highlighted in the past (e.g., Snijders & Bosker, 1999), that increasing upper-level units can often result in more power than increasing the number of lower-level units. The tradeoff, however, can be different when the cost of increasing subjects is substantially greater than the cost of increasing time points. In diary studies, for example, it can often be more expensive to add subjects than time points.

The simulation approach described in this chapter has some limitations. First, it does not take account of missing data and the accompanying loss of power. One solution is to draw on previous research to make assumptions about how many participants and time points will be lost, and adjust the sample size used in the Mplus simulation. Another is to use Mplus's ability to simulate datasets with particular patterns of missing data (see Muthén & Muthén, 1998–2010, chap. 12). Yet another solution is to use Zhang and Wang's (2009) SAS power analysis macro that allows one to specify missing data.

The second limitation is that in order to use the demonstration version of Mplus we illustrated the power simulation with a scaled-back model of the data generating process. Although we believe that this approach does not bias the power estimates, it is likely that they are more imprecise than estimates derived from simulating the full model that provided us with the parameter values in the first place. On our chapter website, *www.columbia.edu/~nb2229/publications.html*, we provide the Mplus syntax for a power simulation based on the full model.

Summary

We hope that this chapter illustrates how to conduct a simulation-based power analysis for within-subject effects in models of diary and intensive longitudinal data. To conduct a power analysis based on Monte Carlo simulation, one needs a lot of information about the model to be tested. In most cases researchers will need to have already conducted at least some pilot work in order to specify hypothesized parameter values. It may be possible in some cases, however, to make informed guesses about the parameters based on similar published studies. For a more general and in-depth treatment of power simulation using Mplus, see Muthén and Muthén (2002). For a more extended introduction to power analysis for diary and intensive longitudinal studies, with a variety of datasets and statistical models, see Bolger and Laurenceau (in press).

References

Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS.* Cary, NC: SAS Institute.

Allison, P. D. (2009). *Fixed effects regression models.* Thousand Oaks, CA: Sage.

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

Bolger, N., & Laurenceau, J.-P. (in press). *Analyzing intensive longitudinal data methods.* New York: Guilford Press.

Bolger, N., & Schilling, E. A. (1991). Personality and the problems of everyday life: The role of neuroticism in exposure and reactivity to daily stressors. *Journal of Personality, 59,* 355–386.

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.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* San Diego, CA: Academic Press.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

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.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh, UK: Oliver & Boyd.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis.* Hoboken, NJ: Wiley.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

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 Behavioral Statistics, 24,* 70–93.

Hoffman, L., & Stawski, R. S. (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in Human Development, 6,* 97–120.

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

IBM Inc. (2010). IBM SPSS Statistics (Version 19). Armonk, NY: Author.

Judge, C. G., Griffiths, W. E., Hill, R. C., & Lee, T. C. (1980). *The theory and practice of econometrics.* New York: Wiley.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, NC: SAS Institute.

MathWorks, Inc. (2008). *Matlab* (Version 2008b). Sherborn, MA: Author.

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

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

Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide* (5th ed.). Los Angeles: Author.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4*, 599–620.

Neyman, J., & Pearson, E. S. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society, 29*, 492–510.

R Development Core Team. (2008). *R: A language and environment for statistical computing* (Version 2.7.2). Vienna: R Foundation for Statistical Computing.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., Jr., & du Toit, M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W., Spybrook, J., Liu, X.-F., & Congdon, R. (2006). *Optimal Design* (Version 1.77). Ann Arbor, MI: HLM Software. Retrieved from *sitemaker.umich.edu/group-based/files/od-manual-20080312-v176.pdf.*

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* Cambridge, UK: Cambridge University Press.

SAS Institute, Inc. (2008). *SAS 9.1.3.* Cary, NC: Author.

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: Sage.

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