Effects of Study Duration, Frequency of Observation, and Sample Size on Power in Studies of Group Differences in Polynomial Change

Stephen W. Raudenbush
University of Michigan

Liu Xiao-Feng
University of South Carolina

Consider a study in which 2 groups are followed over time to assess group differences in the average rate of change, rate of acceleration, or higher degree polynomial effect. In designing such a study, one must decide on the duration of the study, frequency of observation, and number of participants. The authors consider how these choices affect statistical power and show that power depends on a standardized effect size, the sample size, and a person-specific reliability coefficient. This reliability, in turn, depends on study duration and frequency. These relations enable researchers to weigh alternative designs with respect to feasibility and power. The authors illustrate the approach using data from published studies of antisocial thinking during adolescence and vocabulary growth during infancy.

In assessing treatment effects, evaluating prevention programs, and describing the correlates of human growth or change, a common analytic goal is to assess group effects on individual trajectories (cf. Bryk & Raudenbush, 1987; Francis, Fletcher, Stuebing, Davidson, & Thompson, 1991; Huttenlocher, Haight, Bryk, & Seltzer, 1991; Laird & Ware, 1982; Muthen & Curran, 1997; Rogosa & Willett, 1985; Willett & Sayer, 1994). In one simple but important case, participants are randomly assigned to treatments and are then monitored over time. Each participant's trajectory is a linear function of age or time, characterized by an intercept and a rate of change. The primary aim of the study is to assess the average treatment effect on the rate of change. In other studies, random assignment does not occur, but the aim remains similar: to compare groups on linear rates of change. Some-

times, however, growth or change is curvilinear, and the aim is to assess treatment effects on the rate of acceleration. For example, Huttenlocher et al. (1991) studied gender differences in the acceleration of vocabulary during the second year of life. These cases have two features in common.

First, an individual trajectory of growth or change is characterized by a polynomial function. A first-degree polynomial characterizes individual trajectories in terms of an intercept and a rate of change, whereas a second-degree polynomial adds an acceleration parameter. Higher degree polynomials, although less prominent in the literature, are possible. For example, a third-degree polynomial allows study of changes in acceleration. Second, the key hypothesis involves group differences in one or more polynomial change parameters. These include group differences in intercepts, linear rates of change, rates of acceleration, and so on.

We refer to such examples as studies of group differences in polynomial change. In planning these studies, one must consider trade-offs between the duration of the study, frequency of observation, and number of participants. We propose and illustrate a framework for managing such trade-offs. We develop a model for treatment effects on polynomial change coefficients and represent the variance of an estimated group difference as an explicit function of duration, frequency, and sample size. We introduce a standardized effect size for group differences in polynomial

change components and show that power depends on a noncentrality parameter, λ, for the F distribution, defined as $\lambda = n\delta^2\alpha/4$, where n is the sample size, δ is the standardized effect size, and α is the person-specific reliability coefficient commonly printed out by software for hierarchical models.[1] The reliability coefficient α, in turn, is shown to be a simple function of study duration, frequency of observation, between-person variance in change, and within-person variance. This approach yields simple power computations that enable researchers to study these trade-offs as they plan longitudinal studies. We illustrate the approach using data on antisocial thinking during adolescence, which focused on a first-order polynomial, and a second study of vocabulary growth that focused on a second-order polynomial.

Background

A growing literature provides guidance on sample size and statistical power for group comparisons using repeated measures designs. For example, Bloch (1986) considered sample size and power for repeated measures designs assuming an underlying compound symmetry model for the variances and covariances of the repeated measures. That article also considered how variation between and within participants, together with cost considerations, affects power. Rochon (1991) extended this logic to include repeated measures having an autoregressive covariance structure. Muthen and Curran (1997) considered power for detecting group differences under the assumption of randomly varying linear growth rates. They varied the study duration and frequency of time points as well as number of participants per treatment in assessing power and also considered power for testing interaction effects among age, treatments, and participants' initial status. This work provided large-sample power approximations for a range of linear growth models for sample sizes larger than 100. Brown (1998) extended Muthen and Curran's (1997) work by developing a web-based program that computes power for a variety of designs using straight-line models for individual change.

Liu and Liang (1997) studied sample size and power for the case of linear change based on continuous outcomes, but they also considered repeatedly obtained binary outcome data. Their general approach, based on the method of generalized estimating equations, can readily be specialized to a wide variety of outcome types and to nonlinear link functions (e.g.,

logit or log linear models). Hedeker, Gibbons, and Waternaux (1999) allowed for sample attrition in considering sample size determination for repeated measures studies. Their approach assumed continuous outcomes and linear link function. They illustrated their approach in the case of compound symmetry, first-order autoregressive structure, and random effects structure. Related contributions to sample size determination and power for longitudinal designs may be found in Muller, LaVange, Ramey, and Ramey (1992), Overall and Doyle (1994), and Kirby, Galai, and Munoz (1994). Maxwell (1998), starting from a pre–post design, considered how adding intermediate observations affects power.

Schlesselman (1973) considered the consequences of study duration and frequency of time points for the standard error of an estimated group difference in linear growth rates. He found that increasing duration had a greater effect than increasing frequency on reducing the standard error. We extend this approach to include higher order polynomials and to consider implications for statistical power.

Approach

The current article focuses on the trade-offs among study duration, frequency of observation, and sample size in planning studies to have adequate power. Unlike previous authors, we represent statistical power for an estimated group difference as an explicit function of study duration, frequency of observation, and sample size. This approach facilitates the comparison of alternative designs with respect to power and feasibility. Estimated variances and noncentrality parameters can be computed simply with a hand calculator, and power calculations are exact rather than approximate, making them useful even for small-sample research. To achieve this simplicity, we restrict our attention to orthogonal designs. However, the approach applies generally to higher order polynomials in contrast to much previous work on straight-line growth models. To bring our key issues into focus, we limit our attention to the case of continuous outcomes, linear link functions, a random-effects covariance structure, homogeneous covariance structures within treatments, and complete data. The consequences of

---

[1] Throughout this article, α refers to a reliability coefficient rather than to a significance level. We use the .05 significance level throughout for simplicity.

increasing any aspect of data collection (duration, frequency, and sample size) depend on which polynomial effect (intercept, linear, quadratic, or cubic) is of interest, on the magnitude of variation between participants on the polynomial coefficient of interest, and on the magnitude of the variation within participants. We illustrate how to use extant data to optimize power over feasible designs using data from two studies in developmental psychology.

Model

Following Bryk and Raudenbush (1987, 1992), we represent the model in its hierarchical form. It is a two-level model wherein the first level units are occasions within persons and the second units are the persons.

Level 1 Model for Individual Change

The Level 1 model describes the trajectory of change for person i as a polynomial function of degree $P - 1$ defined at equally spaced occasions of observation. Thus, we have an outcome Y_{mi} for person i ($i = 1, \ldots, n$) at occasion m ($m = 1, \ldots, M$):

$$Y_{mi} = \sum_{p=0}^{P-1} \pi_{pi} c_{pm} + e_{mi}. \qquad (1)$$

Here c_{pm} is the orthogonal polynomial contrast coefficient of degree p at occasion m ($P < M$), $c_{0m} = 1$ for all m, π_{0i} is the mean outcome for participant i, c_{1m} is a linear contrast coefficient, π_{1i} is the linear rate of change for participant i, and π_{pi} ($p > 1$) are higher order polynomial effects. The within-participant random effects, e_{mi}, are assumed independently and identically distributed as $N(0, \sigma^2)$. One must take care in defining the individual change coefficients, the πs. Thus, for example, in a quadratic model, π_{1i} is the average rate of increase during the course of the study, whereas π_{2i} represents acceleration. In a cubic model, π_{2i} represents the average rate of acceleration, whereas π_{3i} is the rate at which acceleration changes. In general, adding higher order terms will change the definitions of the lower order change parameters other than the mean.

Orthogonal polynomial contrast coefficients appear in the tables of many experimental design texts (cf. Kirk, 1982). With equally spaced time points, the first four coefficients may be computed as

$$c_{0m} = 1$$

$$c_{1m} = m - \sum_{m=1}^{M} m/M$$

$$c_{2m} = \frac{1}{2}\left(c_{1m}^2 - \sum_{m=1}^{M} c_{1m}^2/M \right) \qquad (2)$$

$$c_{3m} = \frac{1}{6}\left(c_{1m}^3 - \frac{\sum_{m=1}^{M} c_{1m}^4}{\sum_{m=1}^{M} c_{1m}^2} c_{1m} \right).$$

For example, with $M = 5$, we have

$$\begin{aligned}
c_0 &= (1, 1, 1, 1, 1) \\
c_1 &= (-2, -1, 0, 1, 2) \qquad (3) \\
c_2 &= (1, -0.5, -1, -0.5, 1) \\
c_3 &= (-0.2, 0.4, 0, -0.4, 0.2).
\end{aligned}$$

There are several benefits to choosing an orthogonal polynomial model. First, the Level 1 parameters have clear substantive definitions, as described previously. Second, it is straightforward to derive simple expressions for estimators and exact standard errors that apply in studies of arbitrary length and for polynomials of any degree. These lead to simple computations of power that apply even in small-sample research.

Least squares estimates of each participant's change parameters along with their conditional variances[2] are simply computed as

$$\hat{\pi}_{pi} = \frac{\sum_{m=1}^{M} c_{pm} Y_{mi}}{\sum_{m=1}^{M} c_{pm}^2}$$

$$Var(\hat{\pi}_{pi}|\pi_{pi}) \equiv V_p = \frac{\sigma^2}{\sum_{m=1}^{M} c_{pm}^2}. \qquad (4)$$

---

[2] These least squares estimates and their variances facilitate our presentation without loss of generality. However, we do not generally recommend them for person-specific inference. Bryk and Raudenbush (1992, chap. 6) illustrated how and why empirical Bayes estimators will typically provide better predictions of status than will the least squares estimators.

In general, the denominator of the variance expression is

$$\sum_{m=1}^{M} c_{pm}^2 = K_p \frac{(M+p)!}{(M-p-1)!},$$ (5)

where K_p is a constant for each polynomial parameter, p. For the first four polynomial contrasts, we have $K_0 = 1$, $K_1 = 1/12$, $K_2 = 1/720$, $K_3 = 1/100{,}800$ with

$$\sum_{m=1}^{M} c_{0m}^2 = M$$

$$\sum_{m=1}^{M} c_{1m}^2 = \frac{(M+1)M(M-1)}{12}$$ (6)

$$\sum_{m=1}^{M} c_{2m}^2 = \frac{(M+2)(M+1)M(M-1)(M-2)}{720}$$

$$\sum_{m=1}^{M} c_{3m}^2 = \frac{(M+3)(M+2)(M+1)M(M-1)(M-2)(M-3)}{100{,}800}.$$

Level 2 Model for Variation Between Persons

Whereas the Level 1 model defines the change parameters for each participant, the Level 2 model describes how those parameters vary over a population of persons. In a two-group study comparing an experimental group with a control group, we have, for each change parameter, π_{pi}, $p = 0, \ldots, P - 1$,

$$\pi_{pi} = \beta_{p0} + \beta_{p1} X_i + u_{pi},$$ (7)

where X_i takes on a value of 0.5 for members of the experimental group and -0.5 for the members of the control group. Thus, β_{p0} is the population mean of polynomial effect π_{pi}, and β_{p1} is the mean difference between experimentals and controls with respect to that polynomial effect. The random effects u_{pi}, $p = 1, \ldots, P - 1$ are assumed multivariate normal in distribution with means of zero, variances τ_{pp}, with the covariance between u_{pi} and $u_{p'i}$ denoted $\tau_{pp'}$. These random effects are assumed independent across persons and independent of the Level 1 random effect, e_{mi}.

It is useful to reexpress the Level 2 model in terms of the person-specific least squares estimate:

$$\hat{\pi}_{pi} = \beta_{p0} + \beta_{p1} X_i + u_{pi} + (\hat{\pi}_{pi} - \pi_{pi}),$$ (8)

from which it follows that the marginal within-treatment variance of the least squares estimator is

$$Var(\hat{\pi}_{pi}) = Var[u_i + (\hat{\pi}_{pi} - \pi_i)]$$
$$= \tau_{pp} + V_p.$$ (9)

The minimum variance, unbiased estimate of the treatment effect for polynomial p and its variance is

$$\hat{\beta}_{p1} = \overline{\pi}_{pE} - \overline{\pi}_{pC}$$
$$Var(\hat{\beta}_{pl}) = 4 \, (\tau_{pp} + V_p)/n,$$ (10)

where $\overline{\pi}_{pE}$ is the mean of the least squares estimators $\hat{\pi}_{pi}$ among those in the experimental group and $\overline{\pi}_{pC}$ is the mean of the least squares estimators $\hat{\pi}_{pi}$ among those in the control group.[3]

Hypothesis Testing and Statistical Power

We test the null hypothesis

$$H_0\colon \beta_{p1} = 0$$

against the alternative hypothesis

$$H_0\colon \beta_{p1} \neq 0.$$

When the null hypothesis is true, the test statistic

$$F = \hat{\beta}_{p1}^2 / Var(\hat{\beta}_{p1})$$ (11)

follows a central F distribution with degrees of freedom of 1 and $n - 2$. However, when the alternative hypothesis is true, the test statistic F follows a noncentral F distribution with the same degrees of freedom and noncentrality parameter

$$\lambda_p = \beta_{p1}^2 / Var(\hat{\beta}_{p1}) = \frac{n\beta_{p1}^2}{4\left(\tau_{pp} + \sigma^2 \Big/ \sum_{m=1}^{M} c_{pm}^2\right)}.$$ (12)

Power is, of course, influenced primarily by the magnitude of the noncentrality parameter and only secondarily by the critical value of F. Inspection of the noncentrality parameter reveals five important facts.

First, increasing the number of time points, M, will increase power. This will occur because large values of M will increase $\sum_{m=1}^{M} c_{pm}^2$, which reduces the denominator of the noncentrality parameter. Inspection of Equation 6 shows that this tendency of power to depend on M is more pronounced for higher order polynomials than for lower order polynomials. For example, in testing treatment effects on the mean, the noncentrality parameter diminishes as a function of

---

[3] Recall that n here is the total sample size with $n/2$ participants in each group.

M; in testing treatment effects on the linear contrast, it is a function of $(M + 1)M(M - 1)$; and so on.

Second, increasing the sample size, n, will also increase power. In the limit, increasing n will generally have a greater effect on power than will increasing M. This can be seen by noting that, as M increases without bound, $\sigma^2/\Sigma_{m=1}^{M} c_{pm}^2$ will vanish, leaving $\lambda = n\beta_{p1}^2/(4\tau_{pp})$. Thus, power will achieve an upper bound less than 1.0 as M increases if n is held constant unless τ_{pp} is null. In contrast, as n increases without bound, λ will increase without bound regardless of the value of τ_{pp}, driving power inevitably toward 1.0.

Third, power depends on the effect size, β_{p1}. The larger the mean difference between treatments on polynomial p, the greater is the power.

Fourth, power depends on τ_{pp}, the magnitude of variation across persons (within treatments) on the polynomial contrast of interest. When persons within treatments are quite heterogeneous with respect to π_{pi}, power is less than when those persons are more nearly homogeneous.

Fifth, and finally, power depends on σ^2, the within-person variance. Assuming a correctly specified model for person-specific change, σ^2 represents measurement error. The larger the measurement error, then, the weaker is the power.

These five points help clarify the logic of planning research. Adding repeated observations within persons helps most when the degree of polynomial is high and there is considerable within-person variance. Adding participants is most helpful when between-person heterogeneity on the effect of interest is large. Given that the number of time points is the minimum necessary to estimate the model (i.e., $M > P$), adding participants generally increases power without bound.

Although these general principles are useful, the general guidance they provide is not adequate. First, there are various ways of increasing the precision with which we can estimate a given participant's parameters of change. One may increase the duration of the study (holding constant the frequency of observation), one may increase the frequency of observation (holding constant the duration of the study), or one may increase both the study's duration and frequency. Second, one must be quite specific about the magnitude of variation at each level to get serious about planning. Finally, an effect size or a range of effect sizes must be specified to plan for adequate power. We first consider how to represent standardized effect sizes within our modeling framework conveniently and

then elaborate the model to allow for changes in study duration and frequency. We then consider power.

Standardized Effect Sizes

The power of a study to detect a group difference depends on the effect size, that is, the magnitude of the true group difference. Standardized effect sizes are scale-invariant measures of effect magnitude that are often useful in planning new research. The popularity and utility of Cohen's (1988) text on power result in part from its use of standardized effect sizes, which put a variety of otherwise disparate problems on a common footing. In a simple two-group cross-sectional design with no clustering, the standardized effect size is simply the mean difference between groups divided by the within-group standard deviation (or, in some cases, the standard deviation of the control group). This concept has been extended to clustered designs (Raudenbush, 1997; Raudenbush & Liu, 2000).

We define a standardized effect size for polynomial trend p as

$$\delta_p = \beta_{p1}/\sqrt{\tau_{pp}};\qquad(13)$$

that is, the group difference on polynomial trend p divided by the population standard deviation of the polynomial trend of interest.[4] For example, if a study of school learning represents the growth in cognitive skill as a linear function of age, $\sqrt{\tau_{11}}$ is the population variation in annual growth rates and δ_1 is the standardized mean difference between two groups on this annual growth rate. In our second example, based on Huttenlocher et al. (1991), interest focuses on acceleration in expressive vocabulary during the second year of life; τ_{22} is the population variation in accel-

---

[4] Our definition of a standardized effect size departs from the definition often used in cross-sectional research, that is, the mean difference on an outcome divided by the standard deviation of that outcome. The standard deviation used in the denominator will typically include measurement error. The analogy in our repeated measures case would define the standardized effect size as $\beta_{p1}\sqrt{\tau_{pp} + V_p}$. The denominator would depend on the sampling variance V_p, which, in turn, depends not only on the amount of measurement error of the outcome but also on the design of the study. Given our aim of evaluating alternative designs, such a definition would not be useful. We, therefore, have opted to define the standardized effect size as the ratio of the group mean difference to the standard deviation of the true change component, that is, $\sqrt{\tau_{pp}}$.

eration (i.e., the change in vocabulary per month squared), and δ_2 is the standardized mean difference between two groups in acceleration.

With this simple and intuitively appealing definition in mind, the noncentrality parameter of Equation 12 reduces to a simple and illuminating expression:

$$\lambda_p = n\delta_p^2\alpha_p/4, \tag{14}$$

where α_p is the reliability of the least squares estimator $\hat{\pi}_p$, that is

$$\alpha_p = \frac{Var(\pi_p)}{Var(\hat{\pi}_p)} = \frac{\tau_{pp}}{\tau_{pp} + V_p}, \tag{15}$$

with V_p defined as before (Equation 4). This reliability statistic is commonly cited in literature on hierarchical linear models (cf. Bryk & Raudenbush, 1992, Equation 3.35). It denotes the reliability with which one can discriminate between participants on their growth or change parameters using least squares estimates.[5] Thus, past experience with repeated measures designs and data creates some basis for estimating reliabilities, which can then be easily used in power calculations using Equation 14. However, reliability will depend strongly on study duration and frequency of observation, a topic to which we now turn.

The Consequences of Study Duration and Observation Frequency for Statistical Precision and Power

Our model for individual change defines equally spaced occasions of measurement $m = 1, \ldots, M$. Implicitly, we have defined a time metric $t_m = m - 1$, where t_m is the elapsed time since the onset of the study at observation occasion m, $D = M - 1$ is duration of the study, and the frequency of observation is $f = 1.0$ observations per unit time. If we require data to be collected at equal intervals but allow f to vary, where f is a positive rational number, observation m will occur at Time $t_m = (m - 1)/f$, and the study will terminate at time $D = (M - 1)/f$. Clearly, a study with duration D and frequency f will require $M = fD + 1$ time points.

Our interest focuses on the consequences for statistical power of varying D and f as well as the sample size, n. To examine these consequences, we assume that the model parameters remain invariant under alternative designs. We evaluate the tenability of this assumption in each example. For any n, changing D or f affects precision and power through their effect on the reliability $\alpha_p = \tau_{pp}/(\tau_{pp} + V_p)$ and then only through the denominator of V_p.

Effect of Modifying Duration

Changing duration (holding frequency constant at $f = 1$) changes V_p simply by changing the number of time points. Thus, if we change duration from D to D' without changing f, the new number of time points is $M' = fD' + 1$. We simply select the set of orthogonal coefficients appropriate for M' time points (Equation 2) and compute V_p by applying Equations 4 and 6.

Effect of Modifying Frequency

For $f \neq 1$, observations occur at times $t_m = (m - 1)/f$. This change suggests a new set of the orthogonal coefficients c'_{pm}. Given $M = fD + 1$ time points, we have

$$c'_{1m} = t_m - \frac{\sum\limits_{m=1}^{M} t_m}{M}$$

$$= (m - 1)/f - \frac{\sum\limits_{m=1}^{M}(m - 1)/f}{M}$$

$$= m/f - \frac{\sum\limits_{m=1}^{M} m/f}{M}$$

$$= c_{1m}/f. \tag{16}$$

Similarly, $c'_{2m} = c_{2m}/f^2$ and $c'_{3m} = c_{3m}/f^3$. This recoding leads to a new definition of the sampling variance, V_p:

$$V_p = \frac{\sigma^2 f^{2p}(M - p - 1)!}{K_p(M + p)!}, \tag{17}$$

where $M = fD + 1$. This general formula translates to the specific formulas for $p = 0, 1, 2,$ and 3 given by Equation 6 except for the factor f^{2p}.

Changing Duration While Holding Constant the Number of Time Points

If we increase duration by a factor of $\omega = D'/D$ without changing M, we must necessarily reduce f by a factor of $1/\omega$. Thus, $f'/f = 1/\omega$. Setting $M' = M$ and $f' = f/\omega$ in Equation 17, we have

$$\frac{V'_p}{V_p} = \frac{1}{\omega^{2p}}. \tag{18}$$

---

[5] The reliability statistic also plays a central role in empirical Bayes estimates of individual growth (Bryk & Raudenbush, 1992, chap. 3).

Changing Frequency While Holding Constant the Number of Time Points

By the reasoning of the previous paragraph, to change the frequency without changing the number of time points is equivalent to changing the duration without changing the number of time points.

In sum, a choice of f and D, along with the within-person variation, σ^2 determines V_p, which, along with the between-person variation, determines the reliability α_p. This reliability, together with the sample size n and standardized effect size δ_p, then determines the noncentrality parameter and thence statistical power. Power computations are simple and exact, as illustrated in the next section.

Illustrative Examples

We now illustrate how the prior framework may be applied in practical settings to assess consequences of study duration, frequency of observation, and sample size on power. We begin with a study for which a straight-line model for individual change is a reasonable assumption and then turn to a case in which a quadratic function is needed to model individual change. All powers are computed at the .05 significance level.

Antisocial Thinking During Adolescence: A Straight-Line Model for Individual Change

Our first example is based on data from the National Youth Survey (NYS; Elliot, Huizinga & Menard, 1989), a nationally representative survey of young people with annual interviews conducted from 1976 to 1980. We focus on Cohort 1 ($n = 239$), who were 11 years old at the study's outset and 15 years old by 1980. If we measure time in years, this design entails $f = 1$ (one observation per year), $D = 4$ (the study lasts 4 years [1976–1980]), and the number of time points is $M = fD + 1 = 5$.

The repeatedly observed outcome is "tolerance of antisocial behavior," a nine-item scale indicating the extent to which a participant viewed as "wrong" acts such as lying, cheating on school tests, stealing, deliberately destroying property, attacking a person with intent to do harm, and using illegal drugs. A natural logarithmic transformation reduced skewness. The data have been reported extensively elsewhere (Miyazaki & Raudenbush, 2000; Raudenbush & Chan, 1992, 1993; Willett & Sayer, 1994) and so are not described in detail here.

Our aim is to use these data to plan a future inter-vention study designed to reduce tolerance to antisocial thinking during early adolescence. On the basis of the finding that change during this age range tends to be linear (Raudenbush & Chan, 1993), we adopt a straight-line change model at Level 1:

$$Y_{mi} = \pi_{0i}c_{0m} + \pi_{1i}c_{1m} + e_{mi}, \quad e_{mi} \; N(0, \sigma^2), \quad (19)$$

where $c_{0m} = (1, 1, 1, 1, 1)$ and $c_{1m} = (-2, -1, 0, 1, 2)$ at $m = 1, 2, 3, 4, 5$ in accord with the orthogonal polynomials in Equations 2 and 3. This model defines π_{0i} as the expected outcome for participant i at age 13, and π_{1i} as the expected annual rate of increase during the age interval for participant i.

At Level 2 (between participants), the Level 1 coefficients depend on treatment ($X_i = 0.5$ if experimental, $X_i = -0.5$ if control):

$$\pi_{pi} = \beta_{p0} + \beta_{p1}X_i + u_{pi}$$

for $p = 0, 1$. We assume the pair of random effects (u_{0i}, u_{1i}) to be distributed independently as bivariate normal with zero means, variances τ_{00}, τ_{11} and covariance τ_{01}.

Our interest focuses on β_{11}, the treatment effect on the rate of increase in the outcome. Of course, there was no prevention treatment in the NYS. Rather, we are using NYS data to obtain estimates for purposes of planning. To simulate the analysis, we code gender as X_i (0.5 for female, -0.5 for male). Gender has at most a modest association with the intercept and slope. We obtain estimates

$$\hat{\sigma}^2 = 0.0262$$
$$\hat{\tau}_{00} = 0.0333, \quad\quad (20)$$
$$\hat{\tau}_{11} = 0.0030.$$

To assess power, we assume a standardized effect size of $\delta_1 = -0.40$. This represents an unstandardized effect size of $\beta_{11} = \delta_1\sqrt{\tau_{11}} = -0.40 \cdot 0.0548 = -0.0219$. The practical significance of such an effect size can be assessed by reasoning that assignment to the experimental rather than the control group would produce a mean difference between groups of $4 \cdot 0.0219 = 0.0876$ after a treatment duration of 4 years, nearly 30% of the standard deviation of the outcome.[6] We assume a sample size of $n = 238$ with $n/2 = 119$ per treatment group.

We can now apply the results of the previous sec-

---

[6] The standard deviation at age 15 was 0.30 (Raudenbush & Chan, 1993, Table 1).

Table 1

Effect of Study Duration (D) and Frequency of Observation on Power (f), Holding Sample Size Constant at 238

	f						
D	0.5	1	2	3	4	5	6
2	.26	.26	.31	.35	.39	.43	.46
3	.26	.46	.54	.59	.64	.67	.69
4	.57	.61	.69	.73	.76	.77	.79
5	.57	.71	.76	.79	.81	.82	.83
6	.73	.77	.80	.82	.83	.84	.84
7	.73	.80	.83	.84	.84	.85	.85
8	.80	.82	.84	.85	.85	.86	.86

Note. Outcome is average change rate in antisocial thinking. Effect size (δ_1) is −0.40 with a significance level of .05.

tions to compute the variance of the expected treatment contrast using simple calculations. First, $p = 1$, $f = 1$, and $\sigma^2 = 0.0262$ and $K_1 = 12$ (see Equation 6). Equation 17 then translates into

$$V_1 = \frac{12 \cdot f^2 \sigma^2}{(M+1) \cdot M \cdot (M-1)} \tag{21}$$

$$\frac{12 \cdot (1) \cdot (0.0262)}{6 \cdot 5 \cdot 4} = 0.00262 . \tag{22}$$

Knowing the sampling variance V_1 leads to a simple computation of the reliability

$$\alpha_1 = \tau_{11}/(\tau_{11} + V_1)$$
$$= 0.0030/(0.0030 + 0.00262)$$
$$= .53,$$

which leads to the noncentrality parameter

$$\lambda_1 = n\delta_1^2\alpha_1/4 = (238) \cdot (.40)^2(.53)/4 = 5.046.$$

Table 2

Effect of Study Duration (D) and Sample Size on Power, Holding Constant Frequency of Observation at 1.0

	n						
D	100	200	300	400	500	600	700	800
2	.14	.23	.32	.41	.49	.56	.63	.68
3	.22	.40	.55	.67	.77	.84	.89	.93
4	.30	.54	.71	.83	.90	.95	.97	.98
5	.37	.63	.81	.90	.95	.98	.99	1.0
6	.41	.69	.85	.94	.97	.99	1.0	1.0
7	.44	.73	.88	.95	.98	.99	1.0	1.0
8	.46	.75	.90	.96	.99	1.0	1.0	1.0

Note. Outcome is average change rate in antisocial thinking. Effect size (δ_1) is −0.40, with a significance level of .05.

Table 3

Effect of Frequency of Observation (f) and Sample Size on Power, Holding Duration Constant at 4

	n						
f	100	200	300	400	500	600	700	800
0.5	.28	.49	.67	.79	.87	.92	.95	.97
1	.30	.54	.71	.83	.90	.95	.97	.98
2	.35	.61	.78	.89	.94	.97	.99	.99
3	.38	.65	.82	.92	.96	.98	.99	1.0
4	.40	.68	.85	.93	.97	.99	1.0	1.0
5	.42	.70	.86	.94	.98	.99	1.0	1.0
6	.43	.71	.87	.95	.98	.99	1.0	1.0

Note. Outcome is average change rate. Effect size (δ_1) is −0.40, with a significance level of .05.

Inspection of a table for the noncentral F distribution with $\lambda_1 = 5.046$ with $df = 238 - 2 = 236$ yields a power of $1 - \text{Prob } [F(1, 236, 5.046) < F_0] = .61$, where $F_0 = 3.88$ is the critical value of $F(1, 236)$ at the .05 significance level.[7] This value appears in Table 1, where $D = 4$ and $F = 1$.

Tables 1–3 show how variations in this design would affect power to detect the hypothetical treatment effect of interest. Table 1 provides powers for alternative designs that vary in terms of the duration, D, and the frequency, f, holding the sample size constant at $n = 238$. As we have computed, the NYS design is estimated to yield a power of .61 (see Table 1, row 3, column 2). As f increases (moving across the third row of the table), power increases. Similarly, as duration increases (moving down the second column), power also increases. Note that if duration was doubled to $D = 8$ without increasing the number of time points (i.e., by reducing frequency to $f = 0.5$), power would become .80. In contrast, doubling f while holding constant the number of time points (thereby halving D) would reduce power to .31.

Table 2 holds $f = 1$ and allows power to vary as a function of D and n. Note that increasing the sample size to $n = 400$ while holding duration constant at $D = 4$ would boost power to .83. Table 3 holds duration constant at $D = 4$ and allows power to vary as a function of f and n. Increasing n is more efficient than increasing f for boosting power.

---

[7] Alternatively, this power can be computed using a statistical package. Using the SAS syntax in the Appendix, we computed power = .61.

An important caveat arises in using a past study to predict power in a future study. Our estimates of the effect of duration on power assume that the straight-line model for individual change would continue to hold. Thus, for example, we might increase duration from $D = 4$ to $D = 8$ by expanding the age range of interest from 11 to 15 years to 9 to 17 years. However, the data at hand provide no evidence that change is linear over the 9 to 17-year age range. Thus, the model estimates at hand may not represent the parameters at work in the new design.

The Appendix provides SAS code for computing Tables 1 to 3 or similar tables in the case of studies in which β_{p1} is of interest (i.e., studies that focus on linear growth or change rates). A program that produces power graphs is available from Stephen W. Raudenbush.

Vocabulary Growth During the 2nd Year of Life: A Quadratic Growth Model

Our second example is based on data from Hutten-locher et al.'s (1991) study of vocabulary growth from 12 to 26 months of age. That study found a marginally significant gender gap in acceleration rates of child vocabulary during the second year of life. We suppose that our interest lies in replicating this effect on a new sample.

The original study used a design having $n = 22$ with a maximum of $M = 8$ observations per child. Observations occurred every 2 months. If we measure time in months, the study used a frequency of $f = 0.5$ with $D = 14$ and $M = fD + 1 = 8$ time points per participant. We vary n and f to assess power of alternative designs. We hold D fixed at 14 because substantive interest focuses on a fixed age interval, the second year of life.

We estimated a model that was quadratic in age 12. We re-parameterize the model in terms of orthogonal polynomials. We first show how to assess power under the original design.

To calculate power for the original design, we need to use orthogonal polynomials for the case in which $f = 0.5$, $D = 14$, and $M = Df + 1 = 8$. To see how this can be done, let us first find the orthogonal polynomial in the case of $f = 1$, $D = 7$, $M = 8$. We then show how these polynomials change when $f = 0.5$. When $f = 1$, our Level 1 model for individual growth is

$$Y_{mi} = \pi_{0i}c'_{0m} + \pi_{1i}c'_{1m} + \pi_{2i}c'_{2m} + e_{mi}, \quad e_{mi} \sim N(0, \sigma^2),$$
(23)

with $c_{0m} = (1, 1, 1, 1, 1, 1, 1, 1)$ $c_{1m} = (-3.5, -2.5, -1.5, -0.5, 0.5, 1.5, 2.5, 3.5)$, and $c_{2m} = (3.5, .5, -1.5, -2.5, -2.5, -1.5, .5, 3.5)$ at $m = 1, 2, 3, 4, 5, 6, 7, 8$ in accord with the orthogonal polynomials in Equation 2. However, with $f = 0.5$, we must modify these coefficients as described by Equation 16. Thus, we have

$$Y_{mi} = \pi_{0i}c'_{0m} + \pi_{1i}c'_{1m} + \pi_{2i}c'_{2m} + e_{mi}, \quad e_{mi} \sim N(0, \sigma^2),$$
(24)

with

$$c'_{0m} = c_{0m}, c'_{1m} = c_{1m}/f = (-7, -5, -3, -1, 1, 3, 5, 7),$$
and

$$c'_{2m} = c_{2m}/f^2 = (14, 2, -6, -10, -10, -6, 2, 14) \text{ at}$$
$$m = 1, 2, 3, 4, 5, 6, 7, 8.$$

This model defines π_{0i} as the expected vocabulary for participant i at age 19 months, π_{1i} is the "average velocity" (i.e., the average monthly increase in vocabulary between ages 12 months and 26 months for participant i), whereas π_{2i} is the "acceleration" (i.e., the rate of increase in velocity in words per month for participant i).

At Level 2 (between participants), we model the child-specific growth parameters as a function of gender:

$$\pi_{pi} = \beta_{p0} + \beta_{p1}X_i + u_{pi}$$

for $p = 0, 1, 2$. We assume the triplet of random effects (u_{0i}, u_{1i}, u_{2i}) to be distributed independently as trivariate normal with zero means, variances $\tau_{00}, \tau_{11}, \tau_{22}$, and covariances $\tau_{01}, \tau_{02}, \tau_{12}$.

The estimates relevant to the evaluation of power in this case are

$$\hat{\beta}_{21} = 1.4545$$
$$\hat{\sigma}^2 = 677.506$$
(25)
$$\hat{\tau}_{22} = 1.48575.$$

To assess power, we assume an effect size of $\beta_{21} = 1.4545$, the effect of gender estimated in these data. This is equivalent to a standardized effect size of 1.19, a large effect size.

Huttenlocher et al. (1991) invested heavily in frequency ($f = 0.5$, once every 2 months) and less heavily in sample size ($n = 22$). Assessments of child vocabulary are intensive. We wonder in particular whether recruiting more participants would improve power, even if frequency were suitably reduced. Our sense is that duration is constrained by the substantive focus of vocabulary growth during the second year of

life. We provide the results before showing how they are computed.

Table 4 gives the results, providing powers for varied frequency and sample size, holding duration constant at $D = 14$. We see that doubling the sample size (from 22 to 44) while reducing the frequency by about half would increase power from .54 to .79.

To illustrate how these powers are computed, consider the computation of power for the design as implemented, with $f = 0.5$ and $D = 14$ so that $M = fD + 1 = 8$. We are interested in the contrast for the quadratic coefficient, thus $p = 2$. Therefore, Equation 17 translates into

$$V_2 = \frac{720 \cdot f^{2p}\sigma^2}{(M+2)(M+1) \cdot M \cdot (M-1)(M-2)}$$

$$= \frac{(720) \cdot (.5)^4 \cdot (677.506)}{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6} = 1.00819. \tag{26}$$

Substituting $V_2 = 1.00819$ and $\tau_{22} = 1.48575$ into the formula for the reliability yields $\alpha_2 = \tau_{22}/(\tau_{22} + V_2) = 1.48575/(1.48575 + 1.00819) = .60$, thus yielding noncentrality parameter $\lambda_2 = n\delta_2^2\alpha_2/4 = (22) \cdot (1.19)^2 \cdot (.60)/4 = 4.67$. Inspection of a table for the noncentral F distribution with $\lambda_2 = 4.67$ with $df = 22 - 2 = 20$ yields a power of $1 - \text{Prob} [F(1, 20, 4.67) < F_0] = .54$, where $F_0 = 4.35$ is the critical value of $F(1, 20)$ at the .05 level.[8] This value appears in Table 4, where $n = 22$ and $f = 0.5$.

Final Remarks

We have considered power to detect treatment effects on polynomial change. In particular, we have shown how one may evaluate the consequences of study duration, frequency of observation, and sample

size on power. In general, how these three factors affect power depends on (a) which polynomial is of interest, (b) the magnitude of variation between participants on the polynomial effect of interest, and (c) the magnitude of variation within participants. Increasing frequency or duration increases power to an upper limit that depends on the variation between participants in the polynomial coefficient of interest, whereas increasing sample size raises power inexorably toward 1.0.

We noticed that changing frequency or duration changes statistical power only by changing the reliability α_p of the least squares estimator $\hat{\pi}_p$ of the individual change coefficient π_p. The effect of frequency and duration are easily computed using a hand calculator or simple computer program (see Appendix). Once the reliability is known, the noncentrality parameter is computed as $\lambda_p = n\delta_p^2\alpha_p/4$ where n is the sample size and δ_p is the standardized effect size $\beta_{p1}/\sqrt{\tau_{pp}}$. This logic leads to a six-step procedure for evaluating alternative designs:

1. Enumerate the frequencies and durations under consideration.
2. Select the standardized effect size, δ_p, of interest.
3. Given past estimates of variances τ_{pp} and σ^2, compute the reliability of $\hat{\pi}_p$.
4. For each possible n, compute the noncentrality parameter, λ_p, and power for each combination of frequency and duration.
5. Compare the statistical power yielded by alternative designs defined by combinations of frequency, duration, and sample size.
6. Consider the cost and feasibility of designs that yield adequate power and select a design.

In our second example, duration was fixed by substantive considerations. In this setting, we modify the procedure described previously to allow frequency and sample size to vary. In any given setting, one or two of the three factors (sample size, duration, frequency) may be fixed and the procedure modified accordingly.

A natural extension of the work presented here is to specify cost functions for duration, frequency, and sample size. The cost of increased duration depends in part on the cost of tracking participants. Increasing

Table 4
Effect of Frequency of Observation (f) and Sample Size on Power, Holding Constant Duration at 14

f	n							
	11	22	33	44	55	66	77	88
0.214[a]	.24	.48	.66	.79	.87	.93	.96	.98
0.5	.28	.54	.73	.85	.92	.96	.98	.99
1	.31	.60	.79	.90	.95	.98	.99	1.0
2	.35	.66	.84	.93	.97	.99	1.0	1.0
3	.37	.69	.86	.94	.98	.99	1.0	1.0

Note. Outcome is acceleration rate in vocabulary. Effect size (δ_2) is 1.19, with a significance level of .05.
[a] Frequency is reduced from .250 to .214 to guarantee an integer number of $fD + 1 = (.214)(14) + 1 = 4$ time points.

[8] Alternatively, this power can be computed using a statistical package. Using the SAS syntax provided in the Appendix, we computed power = .54.

frequency will be expensive when assessments are expensive. Increasing sample size increases cost of recruitment. Increasing duration increases the risk of attrition, whereas increasing frequency increases participant burden and may induce unfortunate training effects. The costs of these increased risks may be difficult to quantify. Once a cost function is specified, one may choose an optimal frequency and duration and then sample enough participants to achieve adequate power.

Other extensions include evaluating power (a) under alternative assumptions about attrition (Hedeker et al., 1999); (b) under more complex assumptions about the variation within persons; (c) under alternative assumptions about the effect of treatments on within treatment variation (cf. Muthen & Curran, 1997); (d) for continuously measured explanatory variables; (e) for nonnormal error models and nonlinear link functions (Liu & Liang, 1997); and (f) for more complex designs, for example, designs involving repeated measures on persons nested within clusters that have been randomly assigned to treatments (Feldman & Mc-Kinlay, 1994). These extensions will generally require approximations rather than exact computation of power.

References

Bloch, D. A. (1986). Samples size requirements and the cost of a randomized clinical trial with repeated measurements. *Statistics in Medicine, 5,* 663–667.

Brown, H. (1998). Power calculations for latent growth modeling. Retrieved April 4, 2001, from http://www.biostat.coph.usf.edu/research/psmgold/Power/LGM.v2.html

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

Bryk, A., & Raudenbush, S. W. (1992). *Hierarchical linear models for social and behavioral research: Applications and data analysis methods.* Newbury Park, CA: Sage.

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

Elliot, D., Huizinga, D., & Menard, S. (1989). *Multiple problem youth: Delinquency, substance use, and mental health problems.* New York: Springer-Verlag.

Feldman, H., & McKinlay, S. (1994). Cohort versus crosssectional design in large field trials: Precision, sample size, and a unifying model. *Statistics in Medicine, 13,* 61–78.

Francis, D. J., Fletcher, J. M., Stuebing, K. K., Davidson, K. C., & Thomspon, N. M. (1991). Analysis of change: Modeling individual growth. *Journal of Consulting and Clinical Psychology, 39,* 27–37.

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.

Huttenlocher, J., Haight, W., Bryk, A., & Seltzer, M. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27,* 236–249.

Kirby, A., Galai, N., & Munoz, A. (1994). Sample size estimation using repeated measurements on biomarkers as outcomes. *Controlled Clinical Trials, 15,* 165–172.

Kirk, R. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

Laird, N., & Ware, J. (1982). Random-effects models for longitudinal data. *Biometrika, 65,* 581–590.

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

Maxwell, S. E. (1998). Longitudinal designs in randomized group comparisons: When will intermediate observations increase statistical power? *Psychological Methods, 3,* 275–290.

Miyazaki, Y., & Raudenbush, S. W. (2000). Tests for linkage of multiple cohorts in an accelerated longitudinal design. *Psychological Methods, 5,* 44–63.

Muller, K., LaVange, L., Ramey, S., & Ramey, C. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association, 87,* 1209–1226.

Muthen, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2,* 371–402.

Overall, J., & Doyle, S. (1994). Estimating sample sizes for repeated measurement designs. *Controlled Clinical Trials, 15,* 100–123.

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

Raudenbush, S., & Chan, W. (1992). Growth curve analysis in accelerated longitudinal designs with application to the National Youth Survey. *Journal of Research on Crime and Delinquency, 29,* 387–411.

Raudenbush, S., & Chan, W. (1993). Application of hierarchical linear model to the study appendix of adolescent

deviance in an overlapping cohort design. *Journal of Clinical and Consulting Psychology, 61,* 941–951.

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

Rochon, J. (1991). Sample size calculations for two-group repeated-measures experiments. *Biometrics, 47,* 1383–1398.

Rogosa, D., & Willett, J. (1985). Understanding correlates

of change by modeling individual differences in growth. *Psychometrika, 90,* 726–748.

Schlesselman, J. (1973). Planning a longitudinal study: II. Frequency of measurement and study duration. *Journal of Chronic Disease, 26,* 561–570.

Willett, J., & Sayer, A. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116,* 363–380.

Appendix

SAS Program That Produces Tables 1 to 4

Effects of Study Duration, Frequency of Observation, and Sample Size on Power in Studies of Treatment Effects on Polynomial Change

Example 1: National Youth Survey (Produces Tables 1 to 3)

```
original design was based on
n=238        number of subjects,
D=4 years    duration of the study,
f=1          frequency of observation,
M=5          total number of observation.

data dsn1 (keep=D f power)
     dsn2 (keep=D n power)
     dsn3 (keep=f n power)
;
 format power 3.2;
 siglevel=0.05;
 tau11=0.003     ; * interpersonal variance;
 sigma2=0.0262   ; * within person variance;
 beta11=-0.4*sqrt(tau11) ; * coefficient=standardized es x sqrt(tau11);

%let func= 1 - probf(f0, ndf, ddf, lambda) ;
%let sum_p=(M+1)*M*(M-1)/12 ;
%let V_1=(f**2)*sigma2/sum_p;          * Equation 21;
%let delta=beta11/sqrt(tau11);         *standardized effect size;
%let alpha= tau11/(tau11 + V_1);       *reliability estimate;
%let lambda= 0.25*n*(delta**2)*alpha;

 do D=2 to 8;
  do f=0.5,1,2,3,4,5,6;
    n=238;
    ndf=1;     * df for the numerator;
    ddf=n-2;   * df for the denominator;
    f0=finv(1-siglevel, ndf, ddf);   * f0 is the critical value

    M=floor(D*f+1);

    sum_p=&sum_p;
    V_1=&V_1;
    delta=&delta;
    alpha=&alpha;
```

```
    lambda=&lambda;

    power=&func;
    output dsn1;
  end;
end;

do D=2 to 8;
 do n=100 to 800 by 100;

    ndf=1;     * df for the numerator;
    ddf=n-2;  * df for the denominator;
    f0=finv(1-siglevel, ndf, ddf);   * f0 is the critical value;

    M=floor(D*1+1);         * f=1 ;
    f=1;       * f ratio between new and old frequency f=1;

    sum_p=&sum_p;
    V_1=&V_1;
    delta=&delta;
    alpha=&alpha;

    lambda=&lambda;

    power=&func;
    output dsn2;
  end;
end;

do f=0.5,1,2,3,4,5,6;
 do n=100 to 800 by 100;

    ndf=1;     * df for the numerator;
    ddf=n-2;  * df for the denominator;
    f0=finv(1-siglevel, ndf, ddf);   * f0 is the critical value;

    D=4;
    M=floor(D*f+1);

    sum_p=&sum_p;
    V_1=&V_1;
    delta=&delta;
    alpha=&alpha;

    lambda=&lambda;

    power=&func;
    output dsn3;
   end;
  end;
run;

proc transpose data=dsn1 out=matrix1(drop=_NAME_);
 by D; id f;
run;

proc transpose data=dsn2 out=matrix2(drop=_NAME_);
 by D; id n;
run;

proc transpose data=dsn3 out=matrix3(drop=_NAME_);
 by f; id n;
run;
```

```
data table1;
set matrix1;
 label _0D5='f=0.5 /yr' _1='f=1 /yr' _2='f=2 /yr' _3='f=3 /yr'
       _4='f=4 /yr' _5='f=5 /yr' _6='f=6 /yr'
;
run;

data table2;
set matrix2;
 label _100='n=100' _200='n=200' _300='n=300' _400='n=400'
       _500='n=500' _600='n=600' _700='n=700' _800='n=800'
;

data table3;
set matrix3;
 label _100='n=100' _200='n=200' _300='n=300' _400='n=400'
       _500='n=500' _600='n=600' _700='n=700' _800='n=800'
;
 label f='f per year''
run;

title "Table 1: Effect of D and f on power, holding n=238";
proc print data=table1 label noobs; run;

title "Table 2: Effect of D and n on power, holding f=1";
proc print data=table2 label noobs; run;

title "Table 3: Effect of f and n on power, holding D=4";
proc print data=table3 label noobs; run;
```

Example 2: Vocabulary Growth During the Second Year of Life: A Quadratic Growth Model (Produces Table 4)

```
original design was based on
n=22                   number of subjects,
D=14 months            duration of the study,
f=.5 per month         frequency of observation,
M=8                    total number of observations.

the program generates
table 4 : fxn

data dsn4 (keep=f n power)
 format power 3.2;
 siglevel=0.05;
 tau22=1.48575      ; * interpersonal variance;
 sigma2=677.506     ; * within person variance;
 beta21=1.4545      ; * coefficient of treatment effect for quadratic;

%let func= 1 - probf(f0, ndf, ddf, lambda) ;
%let sum_p=(M+2)*(M+1)*M*(M-1)*(M-2)/720;
%let V_2=(f**4)*sigma2/sum_p;            * Equation 26;

%let delta=beta21/sqrt(tau22);           *standardized effect size;
%let alpha= tau22/(tau22 + V_2);         *reliability estimate;
%let lambda= 0.25*n*(delta**2)*alpha;

 do f=(4-1)/14, 0.5, 1, 2, 3;  * f frequency of observation per month;
  do n=11 to 88 by 11;

    ndf=1;    * df for the numerator;
    ddf=n-2;  * df for the denominator;
```

```
    f0=finv(1-siglevel, ndf, ddf);   * f0 is the critical value

    D=14;
    M=floor(D*f+1);
    f=(M-1)/D;

    sum_p=&sum_p;
    V_2=&V_2;
    delta=&delta;
    alpha=&alpha;
    lambda=&lambda;

    power=&func;
    output dsn4;
   end;
  end;

run;

proc transpose data=dsn4 out=matrix4(drop=_NAME_);
 by f; id n;
run;

data table4;
set matrix4;
 label _11='n=11' _22='n=22' _33='n=33' _44='n=44'
       _55='n=55' _66='n=66' _77='n=77' _88='n=88'
;
 label f='f per month';
run;
Title "Table 4: Effect of f and n on power, holding D=14 months";
proc print data=table4 label noobs; run;
```