Psychological Methods
2011, Vol. 16, No. 4, 391–405

1082-989X/11/$12.00 DOI: 10.1037/a0023352

Sample Size Planning for Longitudinal Models: Accuracy in Parameter Estimation for Polynomial Change Parameters

Ken Kelley
University of Notre Dame

Joseph R. Rausch
Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine

Longitudinal studies are necessary to examine individual change over time, with group status often being an important variable in explaining some individual differences in change. Although sample size planning for longitudinal studies has focused on statistical power, recent calls for effect sizes and their corresponding confidence intervals underscore the importance of obtaining sufficiently accurate estimates of group differences in change. We derived expressions that allow researchers to plan sample size to achieve the desired confidence interval width for group differences in change for orthogonal polynomial change parameters. The approaches developed provide the expected confidence interval width to be sufficiently narrow, with an extension that allows some specified degree of assurance (e.g., 99%) that the confidence interval will be sufficiently narrow. We make computer routines freely available, so that the methods developed can be used by researchers immediately.

Keywords: sample size planning, research design, accuracy in parameter estimation, longitudinal data analysis, group comparisons

Longitudinal studies have become a major source of knowledge generation in psychology and related disciplines. This is the case in part because of the rich information inherently provided by repeated measurement of the same set of individuals over time, as well as the sophisticated methods developed over the last three decades that allow a wide variety of questions about intraindividual change and interindividual differences in change to be addressed (see, for example, Collins & Horn, 1991; Collins & Sayer, 2001; Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2009; Singer & Willett, 2003, for reviews of longitudinal data analytic methods). While the analysis of longitudinal data gained widespread usage in psychology and related disciplines, comparisons of mean differences across groups continue to be widely used. Naturally, the idea of examining group differences over time itself became a widely used technique. Examining group-by-time interactions allows researchers to infer (a) whether groups are changing differently and (b) by how much groups are changing differently.

The question of "are groups changing differently" functionally is answered in a dichotomous manner via the results of a null hypothesis significance test. Namely, if the p value is less than the specified Type I error rate (e.g., .05), the null hypothesis of groups changing the same over time (i.e., the group-by-time interaction) is rejected, with the conclusion being that groups do indeed change differently. However, if the p value is greater than the specified Type I error rate, the null hypothesis is not rejected. Of course, the failure to reject a null hypothesis does not imply that the null hypothesis is in fact true. However, in such cases, the failure to find statistical significance at least does not show support for a difference. Obtaining a clear answer to the research question "are groups changing differently" is functionally answered when the null hypothesis of the group-by-time interaction is rejected.

The question of "by how much do groups change differently" is not answered with a null hypothesis significance test, but rather it is addressed continuously on the basis of a point estimate of the group-by-time interaction and the corresponding confidence interval for the population value. The magnitude of the group-by-time interaction, that is, how different the slopes of two groups are, is often an important outcome in longitudinal studies. Additionally, there is a one-to-one relationship between two-sided $(1 - \alpha)100\%$ confidence interval and a nondirectional null hypothesis significance test with a Type I error rate of α.[1] Namely, if the value of the specified null hypothesis (e.g., 0) is not contained within the $(1 - \alpha)100\%$ confidence interval limits, that same value would be rejected as the value of the null hypothesis using a Type I error rate of $\alpha 100\%$. Thus, it is known that a particular null hypothesis will be rejected if the corresponding confidence interval does not contain the specified null value. However, because the confidence interval contains those values that cannot be rejected as implausi-

This article was published Online First July 11, 2011.

Ken Kelley, Department of Management, Mendoza College of Business, University of Notre Dame; Joseph R. Rausch, Division of Behavioral Medicine and Clinical Psychology, Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine.

This research was supported in part by the University of Notre Dame Center for Research Computing through the use of the high performance computing cluster.

Correspondence concerning this article should be addressed to Ken Kelley, Department of Management, Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556. E-mail: kkelley@nd.edu

[1] There is also an analogous relationship between a one-sided confidence interval and a directional hypothesis tests. As noted, if a confidence interval contains the specified null value, then the corresponding null hypothesis, with the same value, will be rejected.

ble at the specified Type I error rate level, the confidence interval brackets those values that are plausible parameter values at the specified level of confidence (e.g., .95). Thus, a confidence interval by its very nature provides more information than does the result of a null hypothesis significance test.

Although long recognized, only relatively recently has the reporting of confidence intervals been essentially mandated. As the new edition of the *Publication Manual of the American Psychological Association* (American Psychological Association [APA], 2010) states, "Historically, researchers in psychology have relied heavily on null hypothesis significance testing (NHST) as a starting point for many (but not all) of its analytic approaches. APA stresses that NHST is but a starting point and that additional reporting elements such as effect sizes, confidence intervals, and extensive description are needed to convey the most complete meaning of the results" (2010, p. 33). Thus, in order to "convey the most complete meaning of the results" in the context of group longitudinal designs, the effect size, namely the regression coefficient for the group-by-time interaction, and the confidence interval for the population value of the regression coefficient for the group-by-time interaction should be reported. However, if the confidence interval for the population value of the group-by-time interaction is wide, it becomes clear that the population value has not been estimated with a high degree of statistical accuracy. Correspondingly, an "extensive description" of the value of the group-by-time interaction becomes more difficult because of the vagueness with which the group-by-time interaction has been estimated.

All other things being equal, from a statistical perspective a narrow confidence interval for an effect size of interest is preferred to a wider confidence interval. With regards to why researchers historically seldom reported confidence intervals, Cohen once speculated that it may be because their widths were often "embarrassingly large" (Cohen, 1994, p. 1102). In an effort to plan sample size so that "embarrassingly large" confidence intervals for the population values of interest are not obtained, the accuracy in parameter estimation (AIPE) approach to sample size planning was developed. AIPE has been developed for various widely used effect sizes (e.g., see Maxwell, Kelley, & Rausch, 2008, for a review of the AIPE rationale and how it compares to power analysis) but as of yet has not been developed in the context of multilevel models. We develop the AIPE approach to sample size planning here so that (a) sample size can be planned for the expected confidence interval width of the group-by-time interaction to be sufficiently narrow and (b) the confidence interval width of the group-by-time interaction will be sufficiently narrow with some specified degree of assurance (e.g., 99% assurance that the 95% confidence interval is sufficiently narrow). Our position is essentially that if a particular effect size, such as the regression coefficient for the group-by-time interaction, is the driving force behind the research, then a confidence interval should be calculated for the population value for the effect size. So as to not have a confidence interval that is "embarrassingly wide," one should plan sample size so that the confidence interval width is explicitly considered, ideally by having a large (e.g., 99%) assurance that the confidence interval will be sufficiently narrow.

Determining what sample size is necessary in order to achieve a particular goal with some specified probability is often a perplexing problem. Longitudinal studies are often designed in an effort to obtain a sample size that will lead to a false null hypothesis being rejected with some desired probability (i.e., have sufficient power). Planning sample size in such a way is termed a *power analysis*. Raudenbush and Xiao-Feng (2001) developed methods for planning sample size for a desired power and for calculating power given a particular sample size. Although the Raudenbush and Xiao-Feng's (2001) work is useful for planning sample size from a power analytic framework, under many circumstances, the research goal may be an accurate estimate of the parameter of interest, not merely the knowledge of whether the null hypothesis can be rejected.

There are (at least) two fundamentally different methods for planning sample size, namely (a) the power analytic approach and (b) the AIPE approach. Raudenbush and Xiao-Feng (2001) developed sample size planning methods for the power analytic approach for the group-by-time interaction. We developed sample size planning methods for the AIPE approach for the group-by-time interaction. What the Raudenbush and Xiao-Feng (2001) work did for the power analytic approach to sample size planning in the context of longitudinal polynomial change models, our work does for the AIPE approach to sample size planning. We do not claim that either approach is better than the other, as both methods are useful, but the method of choice is necessarily tied to the research question of interest. As we will explain, planning sample size from the AIPE approach is actually easier than planning from the power analytic approach. In particular, it is not necessary to specify the value of the group-by-time interaction when using the AIPE approach, but it is necessary in the power analytic approach.

Many times the ideal scenario is one where an estimate is accompanied with both a statistically significant hypothesis test as well as a narrow confidence interval. Depending on the exact situation, sometimes the power analytic approach demands a larger sample size, whereas at other times the AIPE approach demands a larger sample size. In general, if the effect size is close to the null value, all other things being equal, sample size for the power analytic approach will often exceed the necessary sample size for the AIPE approach. However, if the effect size is very large, all other things being equal, sample size for the AIPE approach will often exceed the necessary sample size for the power analytic approach. In general, the reason this holds is because the larger the effect size, the smaller the sample size needed to obtain the same degree of statistical power. However, obtaining an accurate estimate of the parameter does not change at all on the basis of the effect size or tends to do so rather slowly. The relationship between the necessary sample size for statistical power and AIPE relationship has been illustrated for regression coefficients (Kelley & Maxwell, 2003), mean differences (Kelley, Maxwell, & Rausch, 2003), and the standardized mean difference (Kelley & Rausch, 2006) among other effect sizes. Although the two sample size planning approaches can suggest largely discrepant sample sizes at times, such a situation is reasonable since they have very different goals. However, when a desire exists to have both a high degree of statistical power as well as a high degree of assurance that the confidence interval will be narrow, both sample size planning methods can be used, and the larger of the two sample sizes used.

An impediment to new methods being adopted by researchers is often the difficulty of implementation. As Revelle and Zinbarg (2009) noted when referring to the disconnect between new methodological developments and the use of such developments in

applied research, "it is likely that psychometric contributions would have greater impact if they were readily available in such open source programs" (p. 153). We have eliminated as much of the difficulty of implementation of our methods as possible by developing easy-to-use functions that have been incorporated into the R package (R Development Core Team, 2010) MBESS (Kelley, 2007a, 2007b; Kelley & Lai, 2010) that are both open source and freely available.[2] We demonstrate the ease of use of the functions developed in the appendix. We hope this article will benefit researchers who seek an accurate estimate of the group-by-time interaction in longitudinal studies.

Multilevel Models for Change

Multilevel models, also referred to as hierarchical linear models, mixed effects models, and random coefficient models, which in some cases can be equivalent to latent growth curves, are widely used methods for modeling an outcome as it changes, or remains relatively constant, over time. In this section, we provide an overview of the multilevel model as it applies to longitudinal data and define much of the notation used in the article.

In longitudinal data, observations are nested within the individual (e.g., person), where the individual may also be nested within one or more organizational structures. We operationally define a longitudinal polynomial change model as

$$y_{it} = \sum_{m=0}^{M} \pi_{mi} a_{it}^{m} + \varepsilon_{it}, \tag{1}$$

where π_{mi} is the mth polynomial coefficient of change ($m = 0, \ldots, M$; $M < T$), and a_{it} is some basis of time (e.g., time itself, age, measurement occasion, and so forth) for the ith individual ($i = 1, \ldots, N$) at the tth measurement occasion ($t = 1, \ldots, T$), and ε_{it} is the error for the ith individual at the tth measurement occasion. Equation 1 is often termed the "Level 1" model. The degree of polynomial fit describes the type of individual change (e.g., $M = 1$ is a straight-line model, $M = 2$ is a quadratic model, $M = 3$ is a cubic model, and so on). The variance of ε_{it} across the individuals for the tth measurement occasion is denoted $\sigma^2_{\varepsilon_t}$.

Equation 1 models intraindividual change for each of the N individuals. This is the case because the coefficients of change (i.e., the π_{mi}) in Equation 1 have i subscripts. Each of the coefficients of change in Equation 1 can thus be specific to the individual, and thus individual differences in change are explicitly taken into consideration in the model.

For research design purposes when planning sample size, we restrict all measurement occasion to be the same across individuals, implying that there is a common vector of measurement occasions (i.e., time values) for each of the N individuals and that there are no missing data for the derivation of the model and development of the method. Later we will revisit the idea of missing data in the method that we develop. Thus, we remove the i subscript from a_{it} in Equation 1:

$$y_{it} = \sum_{m=0}^{M} \pi_{mi} a_{t}^{m} + \varepsilon_{it}. \tag{2}$$

Further, we assume that the errors at the different time points have a common variance, implying that $\sigma^2_{\varepsilon_1} = \sigma^2_{\varepsilon_2} = \ldots = \sigma^2_{\varepsilon_T} = \sigma^2_{\varepsilon}$, and errors are independent and identically distributed random variables.

After data have been obtained, assumptions made for statistical analysis purposes, such as (a) that there are no missing data, (b) that all individuals share a common vector of measurement occasions, (c) that errors have homogeneous variances across time, and (d) that the errors are independent and identically distributed random variables, are not always tenable. Fortunately, these assumptions can generally be relaxed easily in the specification and estimation of the change model of interest so that certain types of missing data, interindividual differences in measurement occasions, and more complicated error structures can be incorporated explicitly. However, for purposes of designing a longitudinal study, it is generally unnecessary to incorporate these potential model modifications unless there are strong a priori reasons to do so.

Equation 2 is a general polynomial model used to map a mathematical model onto observed scores over time. A polynomial change model implies that the set of exponents of the time basis is a complete set of whole numbers (i.e., nonnegative integers) from 0 to M.[3] Special cases of Equation 2 are

$$y_{it} = \pi_{0i} + \pi_{1i} a_t + \varepsilon_{it} \tag{3}$$

for a straight-line change model,

$$y_{it} = \pi_{0i} + \pi_{1i} a_t + \pi_{2ij} a_t^2 + \varepsilon_{it} \tag{4}$$

for a quadratic change model, and

$$y_{it} = \pi_{0i} + \pi_{1i} a_t + \pi_{2i} a_t^2 + \pi_{3i} a_t^3 + \varepsilon_{it}. \tag{5}$$

for a cubic change model, and so forth. Extensions beyond the cubic change model are certainly possible, provided that enough measurement occasions have been collected. Although time-varying covariates can be included in Equation 1, and thus in Equations 2–5, we do not consider such models in the present article.

The multilevel nature of the model is evidenced when the coefficients of change in Equation 2 are themselves modeled by what is termed a Level 2 model. Each coefficient of change from Equation 2 can itself be modeled as a dependent variable from interindividual difference variables (e.g., sex, age, education level, etc.). In particular, as dependent variables, the M π_{mi}s from Equation 2 can be explained by an overall fixed effect (i.e., the β_{mk}s, K regressor variables (denoted x_{ki}, and potentially a unique effect (denoted υ_{mi} associated with the ith individual for the mth change coefficient. Such a conceptualization can be written as

[2] Originally, MBESS stood for "Methods for the Behavioral, Educational, and Social Sciences." However, MBESS is now an orphaned acronym, meaning that what was an acronym is now literally its name.

[3] The general polynomial change model can easily be relaxed so that time values can be exponentiated to nonwhole numbers. Doing so does not change our presentation in any way. For example, it is not problematic to exponentiate a number to 1/2, implying that the square root of time is used.

$$\pi_{mi} = \beta_{m0} + \sum_{k=1}^{K} \beta_{mk} x_{ki} + \upsilon_{mi}, \quad (6)$$

where β_{mk} represents the effect of the kth ($k = 1, \ldots, K$) x variable (e.g., group status) for the ith individual on the mth individual specific polynomial change coefficient (e.g., $m = 1$ for a straight-line model, $m = 2$ for a quadratic model, and so on).

The υ_{mi} values are assumed to be independent across individuals and independent of the ε_{it} values. The K x variables are regressors (e.g., predictor or explanatory variables) that can be used to account for the variance of the individual uniqueness on the mth change coefficient. The variance of $\sigma^2_{\pi_m}$ (the last component of Equation 6) is denoted. Although only a model with two levels has been illustrated, the β_{mk}s can themselves be modeled as dependent variables if one or more additional organizational structures exist, and so forth.

In the present article, we are concerned primarily with the test of the group effect for polynomial change coefficients or, more precisely, the coefficient of the group-by-trend interaction. Without loss of generality, we will use the term *group-by-time interaction* to refer to what is formally the *group-by-trend interaction*. The term *group-by-trend interaction* is more general in that it applies to any particular trend. However, the linear trend is so often the only trend included in a model (Kelley & Maxwell, 2008; Mehta & West, 2000), and correspondingly it is the way that the effect of group status on individual change parameters is often discussed in the literature. We are focused explicitly on multilevel models with two levels and with two groups due to their prevalence in the applied literature. We use the values 0 and 1 to represent the two groups. Of course, other coding schemes can be used to identify group membership (e.g., [−.5, .5], [−1, 1]), but we regarded the 0/1 scheme as generally the most intuitive for our purposes. In such a coding scheme, the group effect is the effect of Group 1 as compared with Group 0. In this notational framework, Equation 6 can be written as

$$\pi_{mi} = \beta_{m0} + \beta_{m1} Group_i + \upsilon_{mi} \quad (7)$$

for a straight-line change model, which itself can be written as

$$\pi_{mi} = \beta_{m0} + \upsilon_{mi} \quad (8)$$

for individuals in Group 0 and

$$\pi_{mi} = \beta_{m0} + \beta_{m1} + \upsilon_{mi} \quad (9)$$

for individuals in Group 1. Thus, if there is a difference in the mth change coefficient due to group, that difference will be manifested in β_{m1}. In our notational scheme, β_{m1} is the group-by-trend interaction, with β_{11} being the group-by-time interaction that quantifies the differences in group change over time for straight-line change models.

Using Orthogonal Polynomials

We use orthogonal polynomials for the time values rather than the observed time values themselves in the remainder of the article. For the orthogonal polynomial approach, Equation 2 can be rewritten as

$$y_{it} = \sum_{m=0}^{M} \pi_{mi}\, c_{mt} + \varepsilon_{it}, \quad (10)$$

where c_{mt} is the mth orthogonal polynomial coefficient for the tth measurement occasion, and the value of c_{mt} depends only on T and M. Orthogonal polynomials can facilitate the interpretation of fitted model parameters. In contrast to an approach based on observed time values, orthogonal polynomials make it "straightforward to derive simple expressions for estimators and exact standard errors that apply in studies of arbitrary length and for polynomials of any degree" (Raudenbush & Xiao-Feng, 2001, p. 389). The ability to obtain exact analytic standard errors for longitudinal studies of arbitrary length and polynomial degree facilitates the sample size planning procedures in the work of Raudenbush and Xiao-Feng (2001) as well as in our work. Regardless of whether orthogonal polynomials or the observed time values are used, the value of the highest order polynomial change coefficient will remain the same. Correspondingly, for a straight-line change model, the value of the slope is equivalent in models where orthogonal polynomials are used or where the observed values of time are used. Similarly, for a quadratic change model, the value of the quadratic term is equivalent in models where orthogonal polynomials are used or where the observed values of time were used, and so forth. Finally, the overall fit of the model is left unaltered, implying that the proportion of variation accounted for by the model is the same for both scalings of time values, regardless of whether orthogonal polynomials coefficients or raw time values are used.[4]

Means and Variances of Estimates

Like the procedure developed by Raudenbush and Xiao-Feng (2001), our sample size planning procedures depend on properties of least squares estimators of the individual change coefficients. In practice, longitudinal models will be fitted with maximum likelihood methods. However, under certain conditions, an approach based on least square estimates used in a two-stage analysis procedure and one based on maximum likelihood estimates are conceptually equivalent and numerically similar. More specifically, when π_{mi} is modeled with an intercept and group status (i.e., as in Equation 7) in situations where the vector of time values is the same across individuals and there are no missing data, the two estimation methods will generally yield essentially the same results (e.g., see Rogosa and Saner, 1995, for illustrative examples of the similarity between use of the two-stage least squares estimates and maximum likelihood estimates for longitudinal data analysis). Thus, although maximum likelihood estimation methods are generally more advantageous (e.g., when there are missing data, an unequal number of measurement occasions, more complex measurement occasions, more complex models, and so on) than least squares estimation, in the situation described here, the least

[4] Although the overall model fit is the same regardless of the legitimate scaling used, interpretation of the lower ordered effects is altered.

squares and the maximum likelihood estimates are essentially the same. However, expressions for the least squares estimates are much more straightforward to use than are maximum likelihood estimation, which are generally based on difficult or analytically intractable expressions.[5]

The estimated individual specific least squares estimates of the change coefficients can be written as

$$\hat{\pi}_{mi} = \beta_{m0} + \beta_{m1}Group_i + \upsilon_{mi} + (\hat{\pi}_{mi} - \pi_{mi}), \tag{11}$$

where $\hat{\pi}_{mi} - \pi_{mi}$ is the difference (i.e., the error) between the least squares estimate of π_{mi}, denoted $\hat{\pi}_{mi}$, and π_{mi} itself (e.g., Raudenbush & Xiao-Feng, 2001).[6] Notice that in Equation 11 $\hat{\pi}_{mi}$ is on both the left-hand side and the right-hand side of the equals sign. Thus, Equation 11 can thus be rewritten as

$$0 = \beta_{m0} + \beta_{m1}Group_i + \upsilon_{mi} - \pi_{mi}. \tag{12}$$

which in turn can be rewritten as

$$\pi_{mi} = \beta_{m0} + \beta_{m1}Group_i + \upsilon_{mi} \tag{13}$$

for the true individual specific least squares estimate of the *m*th change coefficient.

Of course, the parameters of Equation 13 are not known in practice and must be estimated, but the grouping variable will be known for any individual. Returning to Equation 11 and recalling the assumption of independence between υ_{mi} and ε_{it} (which implies that υ_{mi} and $\hat{\pi}_{mi} - \pi_{mi}$ are independent), covariance algebra shows that

$$\begin{aligned}\mathrm{Var}(\hat{\pi}_{mi}) = \mathrm{Var}(\beta_{m0}) + \mathrm{Var}(\beta_{m1}Group_i) + \mathrm{Var}(\upsilon_{mi}) \\ + \mathrm{Var}(\hat{\pi}_{mi} - \pi_{mi}),\end{aligned} \tag{14}$$

where all the variances are conditional on group. The interpretation of Equation 14 is the variance of the estimated change coefficients within a group for the *m*th change coefficient. The interpretation of the components of Equation 14 is that $\mathrm{Var}(\beta_{m0})$ is the variance of the fixed effect of the intercept for the *m*th change coefficient, $\mathrm{Var}(\beta_{m1})$ is the variance of the fixed effect for the slope for the *m*th change coefficient, $\mathrm{Var}(\upsilon_{mi})$ is the variance of the unique effects for the *m*th change coefficient within the groups, and $\mathrm{Var}(\hat{\pi}_{mi} - \pi_{mi})$ is the variance of the error in estimating the *m*th change coefficient. As we will show shortly, the last two terms are of primary importance (since the first two terms will drop from the equation). The $\mathrm{Var}(\upsilon_{mi})$ component is the true variance within a group for the *m*th change coefficient, whereas $\mathrm{Var}(\hat{\pi}_{mi} - \pi_{mi})$ is the variance across individuals within a group for the difference between the estimated and true *m*th change coefficient.

Because β_{m0} is a constant, Equation 14 can be reduced to

$$\mathrm{Var}(\hat{\pi}_{mi}) = \beta_{m1}^2\mathrm{Var}(Group_i) + \mathrm{Var}(\upsilon_{mi}) + \mathrm{Var}(\hat{\pi}_{mi} - \pi_{mi}). \tag{15}$$

Of interest, however, is the within-group variance. Because $Group_i$ is a constant within either group (either 0 or 1) Equation 15 can be rewritten as

$$\mathrm{Var}(\hat{\pi}_{mi})_j = \mathrm{Var}(\upsilon_{mi})_j + \mathrm{Var}(\hat{\pi}_{mi} - \pi_{mi})_j. \tag{16}$$

However, because we assume homogeneity of variance across groups in the planning stage, Equation 16 need not have a grouping subscript and thus can be simplified to

$$\mathrm{Var}(\hat{\pi}_{mi}) = \mathrm{Var}(\upsilon_{mi}) + \mathrm{Var}(\hat{\pi}_{mi} - \pi_{mi}). \tag{17}$$

By letting $\mathrm{Var}(\hat{\pi}_{mi}) = \sigma^2_{\hat{\pi}_m}$ (i.e., the variance of the estimated *m*th change coefficients), $\mathrm{Var}(\upsilon_{mi}) = \sigma^2_{\upsilon_m}$ (i.e., the variance of the unique effects for the *m*th change coefficient across the individuals), and $\mathrm{Var}(\hat{\pi}_{mi} - \pi_{mi}) = \sigma^2_{\hat{\pi}_m - \pi_m}$ (i.e., the variance of the difference between the *m*th estimated and true change coefficients for each individual across the individuals), Equation 17 can be rewritten as

$$\sigma^2_{\hat{\pi}_m} = \sigma^2_{\upsilon_m} + \sigma^2_{\hat{\pi}_m - \pi_m}, \tag{18}$$

where

$$\sigma^2_{\hat{\pi}_m - \pi_m} = \frac{\sigma^2_\varepsilon}{\sum_{t=1}^{T} c^2_{mt}} \tag{19}$$

due to the orthogonal approach we use.

Recalling the standard error of the two independent groups *t* test, there is an exact analog with the standard error of β_{m1}. The variance for the *m*th change coefficient for the *j*th group ($j = 0, 1$) is $\hat{\sigma}^2_{\upsilon_{mj}} + \hat{\sigma}^2_{(\hat{\pi}_m - \pi_m)_j}$, with a pooled estimate when $n_0 = n_1$ simply being the mean of the two group estimates of variance, denoted as before without a "*j*" subscript. Because the value of β_{m1} represents a difference in the slope for the two groups, the variance of the difference is needed. The variance of a difference is the sum of the variances when the groups are independent; thus, twice the pooled variance is the variance of the difference. To obtain the standard error in this context, we multiply the pooled variance by the quantity $(1/n_0 + 1/n_1)$ in an analogous way as the pooled standard deviation is used in the two independent group *t* tests:

$$\widehat{\mathrm{SE}}(\hat{\beta}_{m1}) = \sqrt{2(\hat{\sigma}^2_{\upsilon_m} + \hat{\sigma}^2_{\hat{\pi}_m - \pi_m})(1/n_0 + 1/n_1)}. \tag{20}$$

[5] Technically, the framework used for design purposes is not the same as maximum likelihood estimation. One clear difference between these estimation approaches becomes evident when an estimate of the random effect variance for one of the polynomial change parameters is negative using the least squares two-stage estimation approach. In this case, the maximum likelihood estimate of the parameter cannot be negative because from the definition of a maximum likelihood estimate, an estimate must be within the parameter space (e.g., Stuart, Ord, & Arnold, 1999, chapter 18). Clearly, a population variance cannot be negative and thus is not in the parameter space. Consequently, the maximum likelihood estimation of the random effect variance is typically set to 0 in such a scenario, which yields different values for these parameter estimates across these two estimation approaches, sometimes yielding different standard errors for other parameter estimators in the model as well. This is an important issue and leads to the distributional properties of the maximum likelihood estimators being substantially more complex than those of the two-stage parameter estimators (e.g., McCulloch, Searle, & Neuhaus, 2008). However, in many "well-behaved" cases, the results of the two approaches are virtually identical.

[6] A circumflex above a parameter always denotes an estimate of that parameter.

However, because the group sample sizes are equal, Equation 20 simplifies to

$$\widehat{\mathrm{SE}}(\hat{\beta}_{m1}) = \sqrt{\frac{2}{n}(\hat{\sigma}^2_{\upsilon_m} + \hat{\sigma}^2_{\hat{\pi}_m - \pi_m})}. \quad (21)$$

In order to evaluate the null hypothesis that β_{m1} equals some specified value, 0 in most cases and that which is assumed here, a t statistic can be formed:

$$t = \frac{\hat{\beta}_{m1}}{\widehat{\mathrm{SE}}(\hat{\beta}_{m1})}. \quad (22)$$

If the null hypothesis of no group effect on the mth polynomial change coefficient is true, the t statistic will follow a central t distribution with $2n - 2$ degrees of freedom. Of course, ultimate interest is in β_{m1}, not $\hat{\beta}_{m1}$. Thus, it is generally desirable to know the confidence interval limits for β_{m1}. Computing the confidence interval for β_{m1} is a simple matter. The margin of error for the mth polynomial change coefficient is the standard error multiplied by the appropriate critical t value. For a symmetric two-sided confidence interval, the critical value is $t_{(2n-2,\ 1-\alpha/2)}$, where there are $2n - 2$ degrees of freedom because of the two-stage estimation procedure used for the derivation, which is literally a two independent group t test. Thus, the confidence interval for the mth polynomial change coefficient is

$$probability[\hat{\beta}_{m1} - \widehat{\mathrm{SE}}(\hat{\beta}_{m1})t_{(2n-2,1-\alpha/2)} \leq \beta_{m1} \leq \hat{\beta}_{m1} + \widehat{\mathrm{SE}}(\hat{\beta}_{m1})t_{(2n-2,\ 1-\alpha/2)}) = 1 - \alpha, \quad (23)$$

where $1 - \alpha$ is the desired confidence level (e.g., .95).

Accuracy in Parameter Estimation for Polynomial Change Models

Although planning sample size for sufficient statistical power is an important research design consideration, statistical power can be high while the corresponding confidence interval for the parameter of interest is wide. Conversely, the expected confidence interval width can be sufficiently narrow, but the statistical power low, albeit this tends to happen only when the population parameter is relatively close to the null value. Kelley and Rausch (2006) discussed the potential discrepancy in the sample size from the power and AIPE approaches to sample size planning in the context of the standardized mean difference, but conceptually the issue is the same here in the context of longitudinal polynomial change models for the group-by-time interaction. When interest concerns the population value of the parameter, a reject or fail-to-reject outcome of a null hypothesis significance test does not provide much information. Sample size planning with the goal of obtaining a narrow confidence interval dates back to at least Guenther (1965) and Mace (1964) and has been discussed in the recent literature (e.g., Algina & Olejnik, 2000; Bonett, 2008; Bonett & Wright, 2000, in press; Kelley, 2007, 2008; Kelley & Lai, 2011; Lai & Kelley, in press; Liu, 2010) as an alternative or supplement to power analysis as effect sizes and confidence intervals have become a more widely used approach to making scientifically based inferences. Of course, in some situations, there may be a desire for sufficient statistical power as well as a narrow confidence interval (e.g., Jiroutek, Muller, Kupper, & Stewart, 2003).

To begin, from the confidence interval procedure discussed earlier, the observed confidence interval width for any realization of a confidence interval is

$$w = 2\widehat{\mathrm{SE}}(\hat{\beta}_{m1})t_{(2n-2,1-\alpha/2)}. \quad (24)$$

Let ω be the desired confidence interval width. The basic AIPE approach seeks to find the minimum sample size so that the expected confidence interval is sufficiently narrow. That is, the AIPE approach seeks the minimal sample size so that

$$\mathrm{E}[w] \leq \omega. \quad (25)$$

In many cases, the theoretical sample size where $\mathrm{E}[w] \leq \omega$ will be fractional. The necessary sample size is always rounded to the next largest integer, so in reality $\mathrm{E}[w]$ is generally slightly less than ω.

Operationalizing a "sufficiently narrow" confidence interval for the group-by-time interaction is very context specific and will vary greatly from one situation to another. For example, educational psychologists studying vocabulary development over time for third-graders randomly assigned to treatment group (e.g., an advanced work immersion program) or control group (e.g., traditional third-grade vocabulary instruction) may set the desired confidence interval width (i.e., ω) for the confidence interval for the group-by-time interaction to a width of 20 vocabulary words. However, industrial–organizational psychologists studying the productivity of new investment generation over time for account representatives randomly assigned to treatment group (e.g., representatives who use a comprehensive customer relations database for maintaining client information) or control group (representatives who choose their own approach for maintaining client information) may set the desired confidence interval width (i.e., ω) for the confidence interval for the group-by-time interaction to a width of $1,000. The point is, the desired confidence interval width will be based on a minimum of the particular measurement scale and the specific goals of the researcher with regards to how accurately estimated the coefficient for the group-by-time interaction is. Using another strategy, rather than choosing the desired confidence interval width on the basis of a specified number of raw-score units, the researcher could plan the desired confidence interval width on the basis of (a) a certain percentage (e.g., 20%) of the estimated standard deviation of the group-by-time standard deviation (e.g., set $\omega = .20 \times \sqrt{\hat{\sigma}^2_{\hat{\pi}_m}}$), (b) a certain percentage of the estimated standard deviation of the errors (e.g., set $\omega = .20 \times \sqrt{\hat{\sigma}^2_{\varepsilon}}$), or (c) a certain percentage of the estimated value of the group-by-time coefficient itself (e.g., set $\omega = .20 \times \beta_{m1}$).

Recalling the standard error of a polynomial change coefficient from Equation 21, suppose that $\sigma^2_{\upsilon_m}$ and σ^2_{ε} could be known or well approximated. Given this, the standard error would be a function only of n. Furthermore, the sample size could be planned such that the confidence interval width had an approximate expected value of ω. In other words, the research question addressed is the following: Given $\sigma^2_{\upsilon_m}$ and σ^2_{ε}, at what sample size is $\mathrm{E}[w] \leq \omega$ for a particular design? Recalling that T and $\sum_{t=1}^{T} c^2_{mt}$ are fixed design factors, $\sigma^2_{\hat{\pi}_m - \pi_m}$ is implicitly known if σ^2_{ε} is known, and conse-

quently could be determined if σ_ε^2 was known or could be well approximated.

The expectation of the confidence interval width is not a precisely known quantity but can be approximated. Using the confidence interval formulation for β_{m1} from Equation 23, we first write

$$E[w] = E[2\widehat{SE}(\hat{\beta}_{m1})t_{(2n-2,1-\alpha/2)}]. \quad (26)$$

Realizing that $2t_{(2n-2,\ 1-\alpha/2)}$ is a constant, we can rewrite Equation 26 as

$$E[w] = 2t_{(2n-2,1-\alpha/2}E[\widehat{SE}(\hat{\beta}_{m1})]. \quad (27)$$

We substitute for the sample variances used to compute $\widehat{SE}(\hat{\beta}_{m1})$ in Equation 21 the population variances when computing $E[\widehat{SE}(\hat{\beta}_{m1})]$ in order to obtain an approximate expectation. That is, we set

$$E[\widehat{SE}(\hat{\beta}_{m1})] \approx \sqrt{\frac{2}{n}(\sigma_{\upsilon_m}^2 + \sigma_{\hat{\pi}_m-\pi_m}^2)}. \quad (28)$$

In general, the expectation of a function is not the same as the function of the expectation, that is, $E[\varphi(z)]$ does not generally equal $\varphi(E[z])$, where $\varphi(\cdot)$ is some function and z is some random variable. Nevertheless, the delta method (i.e., a commonly used way of obtaining the standard error for a parameter estimate) says that under some fairly general conditions, this approximation is reasonable, especially as sample size increases (e.g., Oehlert, 1992). Also, using $\widehat{SE}(\hat{\beta}_{m1})$ with population values substituted for their sample analogs to approximate $E[\widehat{SE}(\hat{\beta}_{m1})]$ is justified because the variance estimates are unbiased and consistent, implying that as sample size grows sufficiently large the estimates of $\hat{\sigma}_{\upsilon_m}^2$ and $\hat{\sigma}_\varepsilon^2$ converge to $\sigma_{\upsilon_m}^2$ and σ_ε^2, their respective population values.[7]

With this approximation (i.e., Equation 28), an approximate expression for the expected value of the standard error is available that allows the approximate confidence interval width to be determined:

$$E[w] \approx 2t_{(2n-2,\ 1-\alpha/2)}\sqrt{\frac{2}{n}(\sigma_{\upsilon_m}^2 + \sigma_{\hat{\pi}_m-\pi_m}^2)}. \quad (29)$$

Given our approximation for the expected confidence interval width, we set E[w] from Equation 29 to the desired value of the confidence interval width,

$$\omega \approx 2t_{(2n-2,\,1-\alpha/2)}\sqrt{2\left(\frac{\sigma_{\upsilon_m}^2 + \sigma_{\hat{\pi}_m-\pi_m}^2}{n}\right)}. \quad (30)$$

Solving Equation 30 for sample size yields

$$n \approx \text{ceiling}\left\{\frac{(\sigma_{\upsilon_m}^2 + \sigma_{\hat{\pi}_m-\pi_m}^2)8t_{(2n-2,\ 1-\alpha/2)}^2}{\omega^2}\right\} \quad (31)$$

or, equivalently, because of Equation 18

$$n \approx \text{ceiling}\left\{\frac{\sigma_{\hat{\pi}_m}^2 8t_{(2n-2,\ 1-\alpha/2)}^2}{\omega^2}\right\}. \quad (32)$$

Thus, solving Equation 31 for the desired confidence interval width, which requires knowledge of $\sigma_{\upsilon_m}^2$ and $\sigma_{\hat{\pi}_m-\pi_m}^2$, or solving Equation 32 which requires only $\sigma_{\hat{\pi}_m}^2$, yields the necessary sample size so that the expected confidence interval width for the mth polynomial change coefficient will be sufficiently narrow.

Notice that the sample size procedure of Equation 31 (or Equation 32) is the per group sample size. The total sample size necessary is thus $2n = N$. In the power analysis framework, not only do $\sigma_{\upsilon_m}^2$ and $\sigma_{\hat{\pi}_m-\pi_m}^2$, or just $\sigma_{\hat{\pi}_m}^2$, need to be specified, so too does β_{m1}.[8] This is the case because statistical power depends not only on the sampling variability of the estimate, but also on the location of the parameter. Thus, an advantage of the AIPE approach over the power analysis approach in terms of implementation is that the AIPE approach does not require specification of an extra parameter, namely the population value of β_{m1} or a value of β_{m1} that is of minimal interest.[9] Although difficult to discern from Equation 32 initially, halving the desired confidence interval width will lead to approximately a four-fold increase in the planned sample size. This is the case because if $w/2$ is substituted for w in Equation 32, the net effect is that 4 goes to the numerator because the fraction (i.e., $w/2$) in the denominator is squared, which functionally moves 4 (i.e., 2 squared) into the numerator of the equation.

Although Equation 31 (or Equation 32) yields the desired sample size, implementation of Equation 31 (or Equation 32) is not as simple as it might first appear. This is the case because sample size is on the left-hand side of the equation and implicitly on the right-hand side of the equation via the critical t value by way of its degrees of freedom, $2n - 2$. Thus, Equation 31 (or Equation 32) requires an iterative solution in order for it to be solved. Solving Equation 31 (or Equation 32) can be implemented by first selecting an initial sample size that has an expected width that is wider than desired. From that sample size, which is too small, sample size is increased until successive iterations of the sample size do not change, and thus an equilibrium is reached on the left-hand and right-hand sides of the equation. The sample size where successive iterations of the sample size remain the same is the minimum sample size necessary so that the inequality is satisfied.

Achieving a Narrow Confidence Interval With a Desired Degree of Assurance

Although sample size from Equation 31 (or Equation 32) provides necessary sample size so that the *expected* confidence interval width for the group-by-time interaction on the mth polynomial change coefficient will be sufficiently narrow, any given realization of a confidence interval will almost certainly be larger or smaller than the desired width. That is, the expectation only

[7] Although we have based the presentation thus far on a two-stage approach, as McCulloch et al. (2008, chapter 2, section b) pointed out, even in the simplest case of a multilevel model, the expectation of the standard error of a fixed effect can be difficult to determine. Thus, the value of $E[\widehat{SE}(\hat{\beta}_{m1})]$ is not known exactly for the general case for maximum likelihood estimation.

[8] Rather than specifying β_{m1} itself for the power analysis, the minimum value of β_{m1} that would be practically important could be specified.

[9] A standardized version of β_{m1} could be supplied (see footnote 4 of Raudenbush & Xiao-Feng, 2001, p. 391), which technically requires only a single parameter, but implicitly is based on β_{m1} and $\sigma_{\upsilon_m}^2$.

satisfies approximately the mean confidence interval width but does not guarantee a sufficiently narrow confidence interval in any particular situation. This issue is similar to when a mean is estimated from a normal distribution. Even though the sample mean is an unbiased estimator of the population mean, the sample mean will either be smaller or larger than the population value, almost certainly. This is the case because the sample mean is a continuous random variable, as is the confidence interval width, due to the fact that both are based on random data. Thus, approximately half of the time, the computed confidence interval will be narrower than the desired width specified, and the other half of the time the computed confidence interval will be wider than desired.

Because the confidence interval width actually depends only on $\sigma^2_{\hat{\pi}_m}$, the sum of the two random quantities $\hat{\sigma}^2_{\upsilon_m}$ and $\hat{\sigma}^2_{\hat{\pi}_m - \pi_m}$, understanding the distributional properties of $\sigma^2_{\hat{\pi}_m}$ will be beneficial. As can be seen from Equation 29, if a realization of $\hat{\sigma}^2_{\hat{\pi}_m}$ (i.e., $\hat{\sigma}^2_{\upsilon_m} + \hat{\sigma}^2_{\hat{\pi}_m - \pi_m}$) is larger than $\sigma^2_{\hat{\pi}_m}$, the confidence interval will tend to be too wide. (Recall, however, that w may be less than ω in this case because of the requirement that n be a whole number, and thus the expected width is slightly narrower than ω.) However, when a realization of $\hat{\sigma}^2_{\hat{\pi}_m}$ is obtained that is smaller than the corresponding population value on which the sample size was based, the confidence interval width will be narrower than desired (i.e., $w < \omega$). If sample size could be planned so that $\hat{\sigma}^2_{\hat{\pi}_m}$ would be smaller than $\sigma^2_{\hat{\pi}_m}$ with some specified degree of probabilistic assurance (e.g., 99%), then the observed confidence interval would be sufficiently narrow more than the roughly half of the time that it is sufficiently narrow with the expected width approach.

Let γ be the desired degree of assurance, a probabilistic statement, that the confidence interval will be sufficiently narrow. That is,

$$probability(w \leq \omega) \geq \gamma.$$

Further, let ${}_{\gamma}\sigma^2_{\hat{\pi}_m}$ be the value of $\hat{\sigma}^2_{\hat{\pi}_m}$ that will not be exceeded with $\gamma 100\%$ assurance for the particular condition. That is,

$$probability(\hat{\sigma}^2_{\hat{\pi}_m} \leq {}_{\gamma}\sigma^2_{\hat{\pi}_m}) \geq \gamma.$$

Substituting ${}_{\gamma}\sigma^2_{\hat{\pi}_m}$ for $\sigma^2_{\hat{\pi}_m}$ from the standard procedure (i.e., Equation 32) will then ensure that the confidence interval will be sufficiently narrow with assurance. A method of finding ${}_{\gamma}\sigma^2_{\hat{\pi}_m}$ would thus be desirable.

Obtaining ${}_{\gamma}\sigma^2_{\hat{\pi}_m}$ is possible because $\hat{\sigma}^2_{\hat{\pi}_m}$ is an estimated variance (i.e., the pooled variance of the unique effects for the mth change coefficient) of normally distributed values (i.e., the unique effects of the mth change coefficient), which is assumed for the unique effects in multilevel models fitted via maximum likelihood estimation methods. Confidence limits can be found for the variance (or standard deviation via transformation) of normally distributed values using a chi-square distribution (e.g., Hays,1994). In particular,

$$\frac{\hat{\sigma}^2_{\hat{\pi}_m}(2n-2)}{\sigma^2_{\hat{\pi}_m}} \sim \chi^2_{(2n-2)}, \tag{33}$$

where $\sim$ means "is distributed as" and the $2n - 2$ in the numerator is the value of the degrees of freedom. Note that Equation 33 is analogous to the general properties of variances as they relate to chi-square distributions. In order to find the value of $\hat{\sigma}^2_{\hat{\pi}_m}$ that will not be exceeded $\gamma 100\%$ of the time, one needs to find the γ quantile from a (central) chi-square distribution with $2n - 2$ degrees of freedom and convert that to the scale of $\hat{\sigma}^2_{\hat{\pi}_m}$. More formally,

$${}_{\gamma}\sigma^2_{\hat{\pi}_m} = \frac{\sigma^2_{\hat{\pi}_m}\chi^2_{(2n-2,\,\gamma)}}{2n-2}, \tag{34}$$

where $\chi^2_{(2n-2,\,\gamma)}$ is the γth quantile from a chi-square distribution with $2n - 2$ degrees of freedom. The degrees of freedom in this situation, as before, are $2n - 2$ because an estimate of the variance is obtained by pooling from across the two groups.

Now, given ${}_{\gamma}\sigma^2_{\hat{\pi}_m}$ (i.e., the value not to be exceeded by $\hat{\sigma}^2_{\hat{\pi}_m}$ with assurance γ) has been obtained, substituting ${}_{\gamma}\sigma^2_{\hat{\pi}_m}$ for $\sigma^2_{\hat{\pi}_m}$ in Equation 32 and solving, which involves iteration due to n being on the left-hand side and implicitly on the right-hand side of the equation, will provide the sample size such that the confidence interval will be no wider than desired with assurance no less than γ. That is, solving

$$n \approx \text{ceiling}\left\{\frac{{}_{\gamma}\sigma^2_{\hat{\pi}_m} 8 t^2_{(2n-2,\,1-\alpha/2)}}{\omega^2}\right\}. \tag{35}$$

will provide the necessary sample size such that the obtained confidence interval will not be wider than desired with assurance γ. Like with the standard procedure, the requirement that fractional sample size be a whole number implies that sample size is always slightly larger than desired, and thus the empirical will tend to be larger than the nominal. This poses no practical problem provided that the sample size based on is the minimal sample size where

$$probability(w \leq \omega) \geq \gamma$$

holds.

Monte Carlo Simulation Study of the Effectiveness of the Procedures

O'Brien and Mueller (1993), in discussing the exactness of power computations that are based on population parameters, recognized that "strict numerical accuracy of the power computations is usually not critical," as "the population parameters [used in the power analysis] are conjectures or estimates" anyway (p. 23). Nevertheless, ideally if the population parameters were in fact known, the sample size planning procedures would produce the desired result, at the very least to a reasonable approximation. Because we used properties of least squares when developing sample size planning methods in which maximum likelihood is used for estimation, in an analogous fashion as Raudenbush and Xiao-Feng (2001) did, our methods are approximate to a certain degree. Further, relying upon the consistency property of maximum likelihood for finite sample sizes is potentially problematic, in that the particular finite sample size suggested by the procedure might not be large enough for the desirable properties of maximum likelihood estimates to be realized. In order to assess the effectiveness of the proposed sample size planning procedures, we conducted a Monte Carlo simulation study. The Monte Carlo simulation study was based on two different situations taken from the literature, where we examine the expected confidence interval width as well as when a desired degree of assurance is incorporated.

It should be noted that any sample size planning procedures that require a population parameter to be specified will necessarily be

limited by the appropriateness of the value(s) used for the population parameter(s). The appropriateness of the sample size planning procedure output to the misspecification of the population parameter input speaks to the robustness of the procedure. Although the robustness of a statistical procedure is important, it is beyond the scope of this article. What is important here is that the proposed sample size planning methods perform as they should, given that the appropriate values for the population parameters have been specified. That is, under the ideal circumstances, we seek to answer the question "Does the method we propose perform in an optimal way?" We used our Monte Carlo simulation study to evaluate the appropriateness of the procedure given that the correct values are supplied. We now outline the two studies we used for the bases of the parameter values.

Study 1: Tolerance of Antisocial Thinking During Adolescence

We used the Elliot, Huizinga, and Menard (1989) study of the tolerance of antisocial thinking during adolescence as the basis for part of our Monte Carlo simulation study. The Elliot et al. (1989) used data from the National Youth Survey, where a dependent variable of interest was "tolerance of antisocial behavior" and five measurement occasions for Cohort 1 (age 11 years at the beginning and 15 years at the end of the study) with a sample size of 239. This study was also used by Raudenbush and Xiao-Feng (2001) in the context of power analysis, as well as in other works to illustrate various methodological issues (e.g., Miyazaki & Raudenbush, 2000; Raudenbush & Chan, 1992, 1993; Willett & Sayer, 1994). Like Raudenbush and Xiao-Feng (2001), we used sex as a grouping variable so as to estimate the group-by-time (i.e., sex-by-time) interaction (i.e., β_{11}). The estimates used are those reported in Raudenbush and Xiao-Feng (2001): the error variance (i.e., $\hat{\sigma}^2$) is 0.0262, the true variance of the intercept is 0.0333, and the true variance of the slope is .0030. Note that it is not necessary to specify a population value or a parameter of minimal interest for the slope, as is the case in power analysis, as the width of the confidence interval is independent of the value of the slope.[10]

For the tolerance of antisocial thinking during adolescence data, a 3 (number of measurement occasions = 3, 5, and 10) by 2 (widths = .25 and .05) by 2 (sample size procedure for the expected confidence interval width will be sufficiently narrow and there will be 85% assurance that the confidence interval will be sufficiently narrow) Monte Carlo simulation was used. Our reading of the literature suggested that there often tend to be fewer rather than many measurement occasions in psychology and related disciplines. The number of measurement occasions of three, five, and 10 seemed quite reasonable, given what has been found in the literature for typical longitudinal designs (e.g., Kwok, West, & Green, 2007).

Each of the 12 conditions was based on 10,000 replications using the PROC MIXED procedure in SAS (Version 9.2). Such a large number of replications were used so that we could very accurately estimate the mean and median confidence interval widths for the expected width case and the percentile and percentile rank of the desired width for the assurance case. In the Monte Carlo simulation, all assumptions were satisfied, illustrating the ideal conditions in order to evaluate the effectiveness of our sample size planning methods. Our combination of conditions led to planned sample sizes that ranged from small (e.g., 43) to relatively large (e.g., 793); demonstrating the relative variety of situations of the conditions used in the Monte Carlo simulation study to examine the effectiveness of the sample size planning procedures.

Table 1 shows the results of the Monte Carlo simulation based on the tolerance of antisocial thinking during adolescence data for the expected confidence interval width. As Table 1 shows, the mean and median confidence interval widths were nearly identical to the desired width in most cases. The biggest discrepancy was for the widest confidence interval condition, where the desired width (ω) was 0.05 and necessary sample size was only 43 per group. In this most discrepant condition, the mean of the confidence interval widths was 0.0487, illustrating the mean confidence interval widths that were 0.0013 units smaller than specified. As the sample sizes became larger, the desired width and the empirical widths converged and became nearly identical. Thus, in this situation, the procedure developed for planning sample size to ensure that the expected width would be sufficiently narrow worked very well.

Table 2 shows the results of the Monte Carlo simulation based on the tolerance of antisocial thinking during adolescence data when an assurance parameter is incorporated produced the desired proportion of confidence intervals that were sufficiently narrow no less than the specified assurance of .85. The biggest discrepancy was again for the 0.05 condition, where the procedure implied sample size was 49. Analogous to the expected width situation, as the sample size becomes larger, the empirical assurance approaches the specified value. Thus, in this situation, the procedure developed for planning sample size to provide a desired degree of assurance worked very well.

Study 2: Quality of Marriage

Karney and Bradbury (1995) provided a tutorial on how change models can be used to better understand the way in which the quality of marriage changes over time that is based on repeatedly measuring the same set of married individuals. Karney and Bradbury (1995) provided illustrative data from a study of newlywed couples. In particular, the data were from 25 newlywed wives from five measurement occasions over the first 30 months of marriage (measured approximately every 6 months), where the participants self-reported marital quality using the Marital Adjustment Test (MAT; Locke & Wallace,1959). In general, a sample size of 25 is inordinately small for an application of a multilevel change model. However, we used their data simply for illustrative purposes, where the estimate of the error variance (i.e., $\hat{\sigma}^2_\varepsilon$) is 134.487, the estimate of the true variance of the intercept is 447.393, and the

[10] For power analysis, a value for the group-by-time interaction, or a standardized version which implicitly includes the slope as well as the variance, must be specified, as the noncentral parameter depends on it. However, because the confidence interval width is independent of the variability, as is the case for a normal distribution, the slope is not specified in the AIPE approach to sample size planning. This is true for AIPE whenever the effect size is independent of its variance, which is not the case for all effect sizes (e.g., standardized mean difference, coefficient of variation, squared multiple correlation coefficient, and so on). Thus, because one less parameter value needs to be specified in the AIPE approach, it is easier to plan sample size from an AIPE perspective.

Table 1
Monte Carlo Simulation Results for the Tolerance of Antisocial Thinking Example: Proposed AIPE Sample Size Planning Method For Expected Width of the Confidence Interval (Equation 31)

Per-group sample size	ω	M_w	Mdn_w	SD_w
		$T = 3$		
793	.0250	0.0249932	0.0249965	0.000442782
200	.0500	0.0498566	0.0498429	0.0017698
		$T = 5$		
278	.0250	0.0249251	0.0249327	0.000743270
71	.0500	0.0493865	0.0494098	0.0029557
		$T = 10$		
165	.0250	0.0248532	0.0248318	0.000978308
43	.0500	0.0486610	0.0485635	0.0037387

Note. M_w, Mdn_w, and SD_w denote the mean, median, and standard deviation, respectively, of the observed ws. AIPE = accuracy in parameter estimation.

estimate of the true variance of the slope is 27.928. Although the Karney and Bradbury (1995) study was for only a single group, assuming homogeneity of variance across groups implies that only a single group is necessary for the estimation of the necessary parameters to plan sample size. Imagine, for example, that interest concerned newlywed first-time wives and newlywed second-time wives. In such a situation, the group-by-time interaction may be of interest and estimating the difference between the two groups with a sufficiently narrow confidence interval may provide a valuable research outcome. Thus, data from single-group designs can potentially be helpful for planning sample size when multiple groups are involved.

Again, as before, the number of measurement occasions chosen was three, five, and 10, which we believe is representative of the literature (e.g., Kwok et al., 2007), not considering two measurement occasions, which is not sufficient to estimate the fixed and random effects of both an intercept and a slope in a multilevel model (although certain restrictions, such as a fixed slope with no random effect, would allow a multilevel model to be fitted to two measurement occasions). Further, rather than having specified absolute values of the expected width directly, we set the specified width to be a proportion of the estimated true variance of the slope. In particular, we specified the desired confidence interval widths to be 5% (ω = 1.396), 10% (ω = 2.793), and 20% (ω = 5.586) of the true variance of the slope. These proportions are arbitrary, but we believe they are reasonable and provide what many might consider small, medium, and large sample sizes. Although the confidence interval widths in terms of an absolute value as well as a proportion are arbitrary in general, for purposes of a Monte Carlo simulation, we have covered what we regard as adequate coverage of the range of sample sizes used in applied psychological research. Our combination of conditions led to a wide range of planned sample sizes that ranged from small (e.g., 42) to relatively large (e.g., 1,566), demonstrating the relative variety of situations of the conditions used in the Monte Carlo simulation study to examine the effectiveness of the sample size planning procedures.

Table 3 shows the results of the Monte Carlo simulation based on the quality of marriage data for the expected confidence interval width. As Table 3 shows, the mean and median confidence interval widths were very close for the narrow and medium sample size for

Table 2
Monte Carlo Simulation Results for the Tolerance of Antisocial Thinking Example: Proposed AIPE Sample Size Planning Method for Obtaining an 85% Assurance of Sufficiently Narrow Confidence Interval (Equation 35)

Per-group sample size	ω	PR(ω)	85%ile
	$T = 3$		
822	.0250	0.8547	0.024991
214	.0500	0.8595	0.049930
	$T = 5$		
295	.0250	0.8733	0.024926
79	.0500	0.8884	0.049542
	$T = 10$		
178	.0250	0.8814	0.024856
49	.0500	0.9104	0.048978

Note. PR(ω) is percentile rank of ω, and 85%ile is the 85th percentile. AIPE = accuracy in parameter estimation.

Table 3
Monte Carlo Simulation Results for The Quality of Marriage Example: Proposed AIPE Sample Size Planning Method for Expected Width of the Confidence Interval Skip (Equation 31)

Per-group sample size	Desired width	M_w	Mdn_w	SD_w
	$T = 3$			
1,503	5% of slope variance ($\omega = 1.396$)	1.3955940	1.3955972	0.0177483
377	10% of slope variance ($\omega = 2.793$)	2.7888950	2.7884883	0.0716498
95	20% of slope variance ($\omega = 5.586$)	5.5718544	5.5697285	0.2865597
	$T = 4$			
866	5% of slope variance ($\omega = 1.396$)	1.3948369	1.3949514	0.0236478
218	10% of slope variance ($\omega = 2.793$)	2.7817167	2.7810272	0.0932619
56	20% of slope variance ($\omega = 5.586$)	5.5022490	5.4976906	0.3708220
	$T = 5$			
654	5% of slope variance ($\omega = 1.396$)	1.3945582	1.3943652	0.0275262
165	10% of slope variance ($\omega = 2.793$)	2.7765749	2.7746552	0.1063374
42	20% of slope variance ($\omega = 5.586$)	5.5079918	5.5065839	0.4338801

Note. M_w, Mdn_w, and SD_w denote the mean, median, and standard deviation, respectively, of the observed ws. AIPE = accuracy in parameter estimation.

each of the three measurement occasions (six situations). The biggest discrepancy was for the widest confidence interval conditions for four and five measurement occasions (four situations). In these two situations, the procedure implied sample sizes were 56 and 42, respectively. The mean confidence interval width in these conditions was approximately 5.50 and 5.51, respectively, whereas the desired confidence interval widths were both 5.86. Thus, there were discrepancies of −.36 and −.35 units for the widest confidence interval conditions for four and five measurement occasions, or a −6.14% and −5.97% disparity, respectively. Although not ideal, the discrepancy is so slight that the implications for applied research seem to be, at worst, negligible. However, because the confidence interval widths are narrower than desired, this is actually an advantageous problem. Overall, the method to plan sample size, so that the expected width is sufficiently narrow, works well.

Table 4 shows the results of the Monte Carlo simulation based on the quality of marriage data with an assurance parameter incorporated that produced the desired proportion of confidence intervals that were sufficiently narrow (no less than the specified assurance of 95%). An assurance of 95% was used here to show the generality of the procedure. The 85% used previously is what we regard as about the smallest assurance that seems to be meaningfully justified. We chose 95%, rather than an assurance of 99%, so that we could determine how much over the specified assurance the empirical assurance would be. For example, had we specified an assurance of 99% and 100% of the confidence intervals been sufficiently narrow, we would not have been able to tell if the sample size was in fact appropriate or if the planned sample size was much too large. However, setting the assurance to 95% allowed us to gauge the appropriateness of the sample size by ensuring that the empirical assurance was not too much larger than the nominal assurance value.[11]

Out of all of the empirical assurances, the biggest discrepancy was again for the widest desired confidence interval condition, where the procedure implied sample size was 53. In this situation, 97.28% of the confidence intervals were sufficiently narrow, whereas the assurance parameter was such that 95% of the intervals should have been sufficiently narrow. Thus, the worst case was off by only 2.28 raw percentage points in a situation with a reasonably small sample size (i.e., 53 per group). Analogous to Table 2, as the sample size becomes larger, the empirical assurance approaches the specified value. Thus, in this situation, the procedure developed for planning sample size so that there was a desired degree of assurance worked very well.

Discussion

Longitudinal research involving two groups is a common way to assess group differences in change. In their work on the power analytic approach to sample size planning for the group-by-time interaction, Raudenbush and Xiao-Feng (2001) provided compelling reasons why the group-by-time or, more generally, the group-by-trend (e.g., linear, quadratic, cubic) effect is often important in applied research contexts. Realizing the importance of the group-by-time interaction and the fact that the estimate will almost certainly not equal the population value illustrates the value of providing a confidence interval for the population value so that the population value can be bracketed with the specified level of confidence. However, a wide confidence interval illustrates the uncertainty with which the population value has been estimated. Planning sample size so that the confidence interval width for the group-by-time interaction will (a) have a sufficiently narrow expectation or (b) have a desired degree of assurance that the confidence interval will be sufficiently narrow is a natural way to plan sample size in longitudinal studies involving two groups.

Although our work provides a way to plan sample size to have a narrow confidence interval for the group-by-time interaction in a two-group design, there are limitations. An important limitation is that we do not allow for missing data in the design phase of the research study, but we presume that all available data will be used in the

[11] Note that it is purely coincidental that the necessary sample sizes from the $T = 10$ case for the tolerance of antisocial thinking example and the $T = 5$ case from the quality of marriage example are so similar.

Table 4
Monte Carlo Simulation Results for the Quality of Marriage Example: Proposed AIPE Sample Size Planning Method for 95% Assurance of Sufficiently Narrow Confidence Interval (Equation 31)

Per-group sample size	Desired width	PR (ω)	95%ile
	$T = 3$		
1,566	5% of slope variance (ω = 1.396)	0.9535	1.39535
408	10% of slope variance (ω = 2.793)	0.9533	2.79021
111	20% of slope variance (ω = 5.586)	0.9561	5.57026
	$T = 4$		
914	5% of slope variance (ω = 1.396)	0.9488	1.39631
241	10% of slope variance (ω = 2.793)	0.9570	2.78601
67	20% of slope variance (ω = 5.586)	0.9627	5.54140
	$T = 5$		
696	5% of slope variance (ω = 1.396)	0.9592	1.39380
185	10% of slope variance (ω = 2.793)	0.9609	2.78202
53	20% of slope variance (ω = 5.586)	0.9728	5.47623

Note. PR(ω) is percentile rank of ω, and 95%ile is the 95th percentile. AIPE = accuracy in parameter estimation.

analysis phase. Under fairly general assumptions—such as data missing at random or missing completely at random (e.g., Rubin, 1976), sample size not too small, and an approximately correct model—multilevel models tend to provide quality results when there is some degree of missing data. However, when the sample size is less than that suggested by the planning procedure, the goal set forth in the procedure will likely not be met with the same probabilistic properties.

Hedeker, Gibbons, and Waternaux (1999) developed sample size planning methods for statistical power in the context of longitudinal designs that are anticipated to have a known missing data structure. Hedeker et al. (1999) noted that the assumptions in methodological works on longitudinal study design (e.g., power analysis) are (a) that sample size is constant across time and (b) that a minimum expected sample size is used at any measurement occasion when sample size is planned. For example, if the procedure implied sample size was 1,000 per group, with the greatest rate of attrition at any of the measurement occasions being 10%, the recommendation would be to use a sample size of 1,112 per group: ceiling(1000/.9) = 1,112. That is, if 1,112 is the sample size used per group and if the data for 10% of the individuals are missing at one particular measurement occasion, the sample size at that particular measurement occasions will still be 1,000 per group, which is the procedure-implied value (if no missing data existed). This is a generally conservative approach (Hedeker et al., 1999) because the "extra" 112 individuals will contribute some information to the analysis based on their other observed data. Correspondingly, the actual statistical power will be greater than desired. Having a greater than desired power is not in and of itself problematic, but because each additional individual in the study takes additional resources, it is undesirable to have more participants than necessary to accomplish the stated goal (e.g., 80% statistical power). Nevertheless, consideration of missing data in this manner leads to boundaries upon which the theoretically ideal sample size will be contained within (here between 1,000 per group and 1,112 per group). A difficulty in implementing the Hedeker et al. (1999) approach is that one must anticipate the percentage of missing data (i.e., retention rates) for each of the two groups at each of the measurement occasions. Nevertheless, if such information could be estimated, using the Hedeker et al. (1999) approach would provide the theoretically ideal sample size, rather than the theoretically ideal sample size being bounded, as exemplified above. Although we think missing data needs to be carefully considered when designing a longitudinal study, we deal only with the idealized case of no missing data in this article.

Like all mathematical models, if the initial assumption(s) on which the model is based are not correct, then the output value(s) may not be correct. Correspondingly, another limitation to our method is the requirement of specifying the input value of $\sigma^2_{\hat{\pi}_m}$. An obvious question is, "How could one ever know $\sigma^2_{\hat{\pi}_m}$?" Unfortunately, it is very likely that $\sigma^2_{\hat{\pi}_m}$ will not be known exactly in any particular situation. However, reasonable estimates of $\sigma^2_{\hat{\pi}_m}$ can often be obtained from a previous study, meta-analysis, or pilot study. Like many sample size planning procedures, where estimation of one or more population values is required, the degree to which the value specified conforms to the population values in reality is usually unknown.

With regards to the requirement of specifying the input value of $\sigma^2_{\hat{\pi}_m}$, we believe that using values available from the literature, in particular those based on several studies in a meta-analysis context, is the ideal way to estimate the necessary parameters, provided that appropriate studies exist. If such studies are not available, then estimates obtained from a pilot study could be used to implement the procedure, with the realization that the estimates from the pilot study may differ considerably from the population values. That is to say, in cases of pilot studies, the sampling variability of $\hat{\sigma}^2_{\hat{\pi}_m}$ may be large (e.g., Kraemer, Mintz, Noda, Tinklenberg, & Yesavage, 2006). One possible solution is to use the upper limit of a plausible value for the variance of the slope, as the upper limit of the set of plausible values for $\sigma^2_{\hat{\pi}_m}$ could be obtained and used as if it was the population value. Such a procedure would help to ensure with a high degree of confidence that the sample size will be no smaller than necessary. Of course, caution is also warranted if a previous study or meta-analysis was based on a population that differs from the one from which the sample will be taken.

There is no question that designing a research study is difficult, especially when parameter values that are unknown to the researcher are required. The difficulty in determining the population parameter(s) has led many researchers to not plan sample size. However, we believe that the alternative of not formally planning sample size is sufficiently worse than planning sample size on the basis of a reasonable estimate, in this case, $\sigma^2_{\hat{\pi}_m}$ (see Equation 32). To assist in the implementation of the sample size planning procedure, we have developed freely available software (see the Appendix). We hope the methods developed assist researchers who are interested in an accurate estimate of the group-by-time interaction in longitudinal models.

References

Algina, J., & Olejnik, S. (2000). Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research, 35,* 119–136.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Bonett, D. G. (2008). Confidence intervals for standardized linear contrasts of means. *Psychological Methods, 13,* 99–109.

Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall, and Spearman correlations. *Psychometrika, 65,* 23–28.

Bonett, D. G., & Wright, T. A. (in press). Sample size requirements for multiple regression interval estimation. *Journal of Organizational Behavior.*

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Collins, L. M., & Horn, J. L. (Eds.). (1991). *Best methods for the analysis of change.* Washington, DC: American Psychological Association.

Collins, L. M., & Sayer, A. G. (Eds.). (2001). *New methods for the analysis of change.* Washington, DC: American Psychological Association.

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

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds.). (2009). *Longitudinal data analysis.* New York, NY: Chapman & Hall/CRC.

Guenther, W. C. (1965). *Concepts of statistical inference.* New York, NY: McGraw–Hill.

Hays, W. L. (1994). *Statistics* (5th ed.). New York, NY: Harcourt Brace.

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.

Jiroutek, M. R., Muller, K. E., Kupper, L. L., & Stewart, P. W. (2003). A new method for choosing sample size for confidence interval based inferences. *Biometrics, 59,* 580–590.

Karney, B. R., & Bradbury, T. N. (1995). Assessing longitudinal change in marriage: An introduction to the analysis of growth curves. *Journal of Marriage and the Family, 57,* 1091–1108.

Kelley, K. (2007a). Confidence intervals for standardized effect sizes: Theory, aplication, and implementation. *Journal of Statistically Software, 20,* 1–20.

Kelley, K. (2007b). Methods for the Behavioral, Educational, and Social Sciences: An R package. *Behavior Research Methods, 39,* 979–984.

Kelley, K. (2007c). Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods, 39,* 755–766.

Kelley, K. (2008). Sample size planning for the squared multiple correlation coefficient: Accuracy in parameter estimation via narrow confidence intervals. *Multivariate Behavioral Research, 43,* 524–555.

Kelley, K., & Lai, K. (2010). MBESS 3.0.4 (or greater) [Computer software and manual]. Retrieved from http://www.cran.r-project.org/

Kelley, K., & Lai, K. (2011). Accuracy in parameter estimation for the root mean square error of approximation: Sample size planning for narrow confidence intervals. *Multivariate Behavioral Research, 46,* 1–32.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8,* 305–321.

Kelley, K., & Maxwell, S. E. (2008). Delineating the average rate of change in longitudinal models. *Journal of Behavioral and Educational Statistics, 33,* 307–332.

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample size planning. *Evaluation and the Health Professions, 26,* 258–287.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11,* 363–385.

Kraemer, H. C., Mintz, J., Noda, A., Tinklenberg, J., & Yesavage, J. A. (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry, 63,* 484–489.

Kwok, O., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research, 42,* 557–592.

Lai, K., & Kelley, K. (2011). Accuracy in parameter estimation for targeted effects in structural equation modeling: Sample size planning for narrow confidence intervals. *Psychological Methods, 16,* 127–148.

Liu, X. S. (2010). Sample size for confidence interval of covariate-adjusted mean difference. *Journal of Educational and Behavioral Statistics, 35,* 714–725.

Locke, H. J., & Wallace, K. M. (1959). Short marital adjustment prediction tests: Their reliability and validity. *Marriage and Family Living, 21,* 251–255.

Mace, A. E. (1964). *Sample size determination.* New York, NY: Reinhold.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59,* 537–563.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models* (2nd ed.). Hoboken, NJ: Wiley.

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods, 5,* 23–43.

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

O'Brien, R., & Muller, K. (1993). A unified approach to statistical power for *t* tests to multivariate models. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral sciences* (pp. 297–344). New York, NY: Dekker.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician, 46,* 27–29.

Raudenbush, S. W., & Chan, W. (1992). Grown 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. W., & Chan, W. (1993). Application of hierarchical linear model to the study of appendix of adolescent deviance in an overlapping cohort design. *Journal of Clinical and Consulting Psychology, 61,* 941–951.

Raudenbush, S. W., & Xiao-Feng, L. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods, 6,* 387–401.

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika, 74,* 145–154.

Rogosa, D. R., & Saner, H. M. (1995). Longitudinal data analysis examples with random coefficient models. *Journal of Educational and Behavioral Statistics, 20,* 149–170.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Press.

Stuart, A., Ord, J. K., & Arnold, S. (1999). *Kendall's advanced theory of statistics: Vol. 2A. Classical inference and the linear model* (6th ed.). New York, NY: Oxford University Press.

Venables, W. N., Smith, D. M., & the R Development Core Team. (2010). *An introduction to R.* Vienna, Austria: The R Development Core Team.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116,* 363–381.

Appendix

Using R and MBESS to Implement the Methods Discussed

A function for sample size planning from the AIPE perspective for polynomial change models was written and incorporated into the MBESS (Kelley, 2007a, 2007b, 2007c; Kelley & Lai, 2010) R package (R Development Core Team, 2010).[12]

Throughout the appendix, sans serif font denotes R functions, options of functions, or output. Sans serif font followed by an open parenthesis and immediately by a closed parenthesis denotes a particular R function. When specifications are given within the parentheses of a function, that function is directly executable in R, after the MBESS packages have been installed and loaded. The easiest way to install MBESS is with the `install.packages()` function in the following manner

```
install.packages(pkgs="MBESS")
```

assuming that the machine has an active Internet connection, which may require the user to select one of many download (i.e., mirror) sites. Alternatively, MBESS can be installed via the Package Manager drop-down menu (in the Windows and Macintosh versions) from the R toolbar, where the user selects from the many packages available to install onto his or her system. After MBESS is installed, it is loaded into the current session with the `require()` function, which is implemented as follows:

```
require(MBESS).
```

A set of help files also accompanies MBESS. For any function in MBESS (or R more generally), the help file can be displayed with the help function, `help()`. For example, the associated help files for the `ss.aipe.pcm()` function, the function that implements the sample size planning methods developed in the article,

```
help(ss.aipe.pcm).
```

Additionally, when the exact name of a function is not known, one can search for functions and help files by using the `help.search()` function. For example, if one were interested in computing a covariance matrix on a data set, searching for "covariance matrix" via the `help.search()` function as follows

```
help.search("covariance matrix")
```

returns information on functions that pertain to covariance matrices. More details on the way in which R is installed and used is available for download via the freely available book *An Introduction to R* (Venables, Smith, & the R Development Core Team, 2010).

For the `ss.aipe.pcm()` function, which is the function that implements the methods developed in this article, the parameters of the function are

```
ss.aipe.pcm(true.variance.trend, error.variance, variance.true.minus.estimated.trend=NULL, duration, frequency, width, conf.level=.95, trend="linear", assurance=NULL),
```

some of which need to be specified on any implementation of the function, where true.variance.trend is the variance of the individuals' true change coefficients (i.e., $\sigma^2_{v_m}$, the first component on the right-hand side of Equation 18), error.variance is the true error variance (i.e., σ^2_ε from the numerator of the right-hand side of Equation 19), and variance.true.minus.estimated.trend is the variance of the difference between the *m*th true change coefficient minus the *m*th estimated change coefficient (i.e., $\sigma^2_{\hat{\pi}_m - \pi_m}$ from Equation 19). Because of the one-to-one relationship between σ^2_ε and $\sigma^2_{\hat{\pi}_m - \pi_m}$, only one of the two values needs to be specified. Further, the parameters of duration, frequency, width, confidence level (e.g., .90, .95, .99, and so forth), trend (either linear, quadratic, or cubic), and assurance (e.g., NULL for only an expected width, .85, .95, .99, and so forth) each need to be specified.

To illustrate how the `ss.aipe.pcm()` MBESS function is used, we will use the previously discussed tolerance of antisocial behavior example from Elliot et al. (1989), which was used as an exemplar by Raudenbush and Xiao-Feng (2001) for their contribution on sample size planning in the context of polynomial change model for the power analytic approach.

[12] R and MBESS are both open source and thus freely available. R is available for download via the Comprehensive R Archival Network (CRAN; http://www.r-project.org/) for computers running Microsoft Windows, Linux/Unix, and Apple Macintosh operating systems. The direct link to the MBESS page on CRAN, where the most up-to-date version of MBESS and its corresponding manual are available, is http://cran.r-project.org/web/packages/MBESS/index.html (note that these Internet addresses are case sensitive).

(*Appendix continues*)

Suppose that a researcher would like to plan sample size so that the straight-line change coefficient has an expected 95% confidence interval width of 0.025 units, which the researcher believes is sufficiently narrow for the purposes of establishing an accurate difference between a treatment group and a control group. The study will have a duration of 4 years with one measurement occasions per year, for a total of five measurement occasions. The supposed variance of the linear trend (i.e., $\sigma^2_{v_m}$) of 0.003 and the supposed error variance (σ^2_e) of 0.0262, both of which are obtained from literature (i.e., in Raudenbush & Xiao-Feng, 2001 based on the data of Elliot et al., 1989).

In this situation, the way in which the `ss.aipe.pcm()` MBESS function is implemented, after MBESS has been installed and loaded via the `require()` function, is as follows

```
ss.aipe.pcm(true.variance.trend=0.003,
error.variance=0.0262, duration=4,
frequency=1, width=0.025, conf.level=.95),
```

which returns the following output

```
"Results for expected width to be sufficiently
narrow"
278.
```

Thus, a sample size of 278 is required when the duration of the study will be 4 units and the frequency of measurement occasions is 1 year in order for the expected confidence interval width to be 0.025 units.

Suppose that the researcher was not happy with having *only* an expected confidence interval width for the group-by-time interaction of 0.025 units. Rather, suppose that the researcher wanted to have 99% assurance that the 95% confidence interval would be sufficiently narrow. The way in which sample size can be planned in this situation with the `ss.aipe.pcm()` MBESS function is as follows,

```
ss.aipe.pcm(true.variance.trend=.003,
error.variance=.0262,
duration=4, frequency=1, width=.025,
conf.level=.95, assurance=.99),
```

which returns the following output

```
"Results for Assurance"
316.
```

Thus, a sample size of 316 will be required to ensure that the 95% confidence interval will be sufficiently narrow (i.e., have a width less than .025 units) at least 99% of the time.

As can be seen, the functions are easy to use and require only minimal knowledge of R. Even if R will not be used for the analysis of the results, R can easily be used for sample size planning purposes. An additional function in the MBESS R package is the `ss.power.pcm()` function, which implements sample size planning for statistical power in this context. That is, the `ss.power.pcm()` function implements the methods developed by Raudenbush and Xiao-Feng (2001) for planning sample size in order to have a desired statistical power. Detailed information on the `ss.power.pcm()` function is available in the MBESS manual or from R via the command

```
help(ss.power.pcm)
```

after MBESS has been installed and loaded.

Received February 5, 2010
Revision received January 19, 2011
Accepted January 26, 2011 ■