Psychological Methods
2003, Vol. 8, No. 3, 305–321

1082-989X/03/$12.00 DOI: 10.1037/1082-989X.8.3.305

Sample Size for Multiple Regression: Obtaining Regression Coefficients That Are Accurate, Not Simply Significant

Ken Kelley and Scott E. Maxwell
University of Notre Dame

An approach to sample size planning for multiple regression is presented that emphasizes *accuracy in parameter estimation* (AIPE). The AIPE approach yields precise estimates of population parameters by providing necessary sample sizes in order for the likely widths of confidence intervals to be sufficiently narrow. One AIPE method yields a sample size such that the expected width of the confidence interval around the standardized population regression coefficient is equal to the width specified. An enhanced formulation ensures, with some stipulated probability, that the width of the confidence interval will be no larger than the width specified. Issues involving standardized regression coefficients and random predictors are discussed, as are the philosophical differences between AIPE and the power analytic approaches to sample size planning.

Sample size estimation from a power analytic perspective is often performed by mindful researchers in order to have a reasonable probability of obtaining parameter estimates that are statistically significant. In general, the social sciences have slowly become more aware of the problems associated with underpowered studies and their corresponding Type II errors, which can yield misleading results in a given domain of research (Cohen, 1994; Muller & Benignus, 1992; Rossi, 1990; Sedlmeier & Gigerenzer, 1989). The awareness of underpowered studies in the literature has led vigilant researchers attempting to curtail this problem in their investigations to perform a power analysis (PA) prior to data collection. Researchers who have used various power analytic procedures have undoubtedly strengthened their own research findings and added meaningful results to their respective research areas. However, even with PA becoming more common, it is known that null hypotheses of point estimates are rarely exactly true in nature (Cohen, 1994). Therefore, performing sample size planning solely for the purpose of obtaining statistically significant parameter estimates may often be improved by planning sample sizes that lead to accurate parameter estimates, not merely statistically significant ones.

The zeitgeist of null hypothesis significance testing seems to be losing ground in the behavioral sciences as the generally more informative confidence interval begins to gain widespread usage. Instead of simply testing whether a given parameter estimate is some exact and specified value, typically zero, forming a $100(1 - \alpha)$ percent confidence interval around the parameter of interest frequently provides more meaningful information. Although null hypothesis significance tests and confidence intervals can be thought of as complementary techniques, confidence intervals can provide researchers with a high degree of assurance that the true parameter value is within some confidence limits. Understanding the likely range of the parameter value typically provides researchers with a better understanding of the phenomenon in question than does simply inferring that the parameter is or is not statistically significant. With regard to *accuracy in parameter estimation* (AIPE), all other things being equal, the narrower the confidence interval, the more certain one can be that the observed parameter estimate closely approximates the corresponding population parameter. Accuracy in this

Editor's Note. Samuel B. Green served as action editor for this article.—SGW

Correspondence concerning this article should be addressed to Ken Kelley or Scott E. Maxwell, Department of Psychology, University of Notre Dame, 118 Haggar Hall, Notre Dame, Indiana 46556. E-mail: kkelley@nd.edu or smaxwell@nd.edu

sense is a measure of the discrepancy between an estimated value and the parameter it represents.[1]

One position that can be taken is that AIPE leads to a better understanding of the effect in question and is more important for a productive science than a dichotomous decision from a null hypothesis significance test. Many times obtaining a statistically significant parameter estimate provides a research community with little new knowledge of the behavior of a given system. However, obtaining confidence intervals that are sufficiently narrow can help lead to a knowledge base that is more valuable than a collection of null hypotheses that have been rejected or that failed to reach significance, given that the desire is to understand a particular phenomenon, process, or system.

If we assume that the correct model is fit, observations are randomly sampled, and the appropriate assumptions are met, $(1 - \alpha)$ is the probability that any given confidence interval from a collection of confidence intervals calculated under the same circumstances will contain the population parameter of interest. However, it is not true that a specific confidence interval is correct with $(1 - \alpha)$ probability, as a computed confidence interval either does or does not contain the parameter value. The meaning of a $100(1 - \alpha)$ percent confidence interval for some unknown parameter was summarized by Hahn and Meeker (1991) as follows: "If one repeatedly calculates such [confidence] intervals from many [technically an infinite number of] independent random samples, $100(1 - \alpha)\%$ of the intervals would, in the long run, correctly bracket the true value of [the parameter of interest]" (p. 31). It is important to realize that the probability level refers to the procedures for constructing a confidence interval, not to a specific confidence interval (Hahn & Meeker, 1991).[2]

Many of the arguments in the present article regarding the use and utility of confidence intervals echo a similar sentiment that has been long recommended, as well as the more recent discussions in Wilkinson and the American Psychological Association Task Force on Statistical Inference (1999), essentially an entire issue of *Educational and Psychological Measurement* (Thompson, 2001) devoted to confidence intervals and measures of effect size, Algina and Olejnik (2000), and Steiger and Fouladi (1997), as well as the still salient views offered by Cohen (1990, 1994). In fact, Cohen (1994) argued that the reason confidence intervals have previously seldom been reported in behavioral research is because the widths of the intervals are often "embarrassingly large" (p. 1002). The AIPE approach presented here attempts to curtail the problem of embarrassingly large confidence intervals and provides sample size estimates that lead to confidence intervals that are sufficiently precise and thereby produce results that are presumably more meaningful than simply being statistically significant.

In the context of multiple regression, sample size can be approached from at least four different perspectives: (a) power for the overall fit of the model, (b) power for a specific predictor, (c) precision of the estimate for the overall fit of the model, and (d) precision of the estimate for a specific predictor. The goal of the first perspective is to estimate the necessary sample size such that the null hypothesis of the population multiple correlation coefficient equaling zero can be correctly rejected with some specified probability (e.g., Cohen, 1988, chapter 13; Gatsonis & Sampson, 1989; S. B. Green, 1991; Mendoza &

[1] The formal definition of *accuracy* is given by the square root of the mean square error and can be expressed by the following formulation:

$$\text{RMSE} = \sqrt{\text{E}[\hat{\theta} - \theta)^2]} = \sqrt{\text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2] + (\text{E}[\hat{\theta} - \theta])^2},$$

where E is the expectation operator and $\hat{\theta}$ is an estimate of θ, the value of the parameter of interest (Hellmann & Fowler, 1999; Rozeboom, 1966, p. 500). The first component under the second radical sign represents precision, whereas the second component represents bias. Thus, when the expected value of a parameter is equal to the parameter value it represents (i.e., when it is unbiased), accuracy and precision are equivalent concepts and the terms can be used interchangeably.

[2] It should be noted that the interpretation of confidence intervals given in the present article follows a frequentist interpretation. The Bayesian interpretation of a confidence interval was well summarized by Carlin and Louis (1996), who stated that "the probability that [the parameter of interest] lies in [the computed interval] given the observed data **y** is at least $(1 - \alpha)$" (p. 42). Thus, the Bayesian framework allows for a probabilistic statement to be made about a specific interval. However, when a Bayesian confidence interval is computed with a noninformative prior distribution (which uses only information obtained from the observed data), the computed confidence interval will exactly match that of a frequentist confidence interval; the interpretation is what differs. Regardless of whether one approaches confidence intervals from a frequentist or a Bayesian perspective, the suggestions provided in this article are equally informative and useful.

Stafford, 2001). With the second perspective, sample size is computed on the basis of the desired power for the test of a specific predictor rather than the desired power for the test of the overall fit of the model (Cohen, 1988, chapter 13; Maxwell, 2000).

The precision of the overall fit of the model leads to another reason for planning sample size. One alternative within this perspective provides the necessary sample size such that the width of the one-sided (lower bound) confidence interval of the population multiple correlation coefficient is sufficiently precise (Darlington, 1990, section 15.3.4). Another alternative within this perspective provides the sample size such that the total width of the confidence interval around the population multiple correlation squared is specified by the researcher (Algina & Olejnik, 2000).

The final perspective for sample size estimation within the multiple regression framework provides the main purpose of the present article. Necessary sample size from this perspective is obtained such that the confidence interval around a regression coefficient is sufficiently narrow. Oftentimes confidence intervals are computed at the conclusion of a study, and only then is it realized the sample size used was not large enough to yield precise estimates. The AIPE approach to sample size planning allows researchers to plan necessary sample size, a priori, such that the computed confidence interval is likely to be as narrow as specified.

Figure 1 illustrates the relation between confidence intervals and null hypothesis significance testing as they relate to the issue of sample size for AIPE and PA. Specifically, the figure shows the limits of a confidence interval for a standardized regression coefficient in each of four hypothetical studies with a different predictor variable in each instance. In all four studies the null hypothesis that the regression coefficient equals zero is false.

From a purely power analytic perspective, Study 1 is considered a "success." The confidence interval in this study shows that the parameter is not likely to be zero and is thus judged to be statistically significant. However, the confidence interval is wide, and thus the parameter is not accurately estimated. In this study little information about the population parameter is learned other than it is likely to be some positive value, a "failure" according to the goals of AIPE. This study had an adequate sample size from the perspective of power, but a larger sample is needed in order to obtain a more precise estimate.

Study 2, on the other hand, not only indicates that the null hypothesis should be rejected but also provides precise information about the size of the population parameter. Here the confidence interval is narrow, and thus the population parameter is precisely estimated. Study 2 is a success according to both the PA and AIPE frameworks.

Study 3 shows a nonsignificant effect that is accompanied by a wide confidence interval, illustrating a failure by both methods. Had a larger sample size

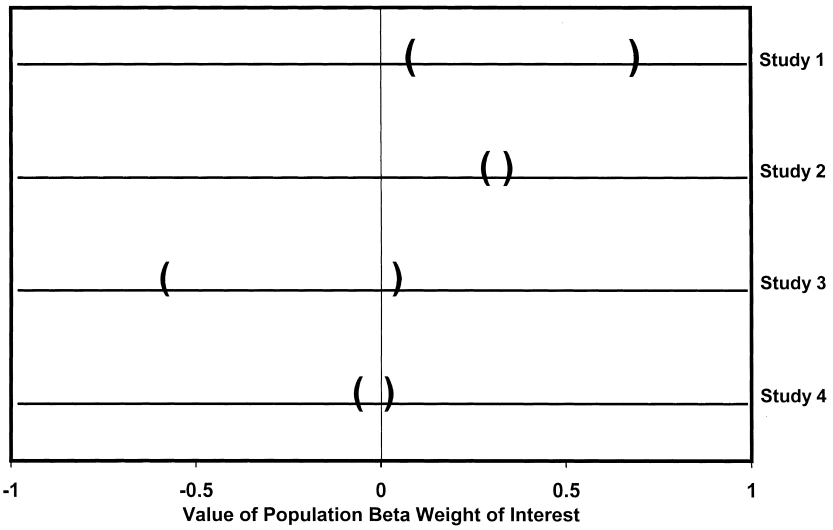

Figure 1. Illustration of possible scenarios in which planned sample size was considered a "success" or "failure" according to the accuracy in parameter estimation and the power analysis frameworks. Parentheses are used to indicate the width of the confidence interval.

been used and had the effect been of approximately the same magnitude, the width of the confidence interval would have likely been smaller, leading to a potential rejection of the null hypothesis. Thus, the sample size of Study 3 was inadequate from both perspectives.

Study 4 illustrates a case in which the confidence interval contains zero, yet the parameter is estimated precisely. Study 4 exemplifies a failed PA but a successful application of AIPE, as the population parameter is bounded by a narrow confidence interval. Of course, one could argue that this study is not literally a failure from a PA perspective, because as a conditional probability, power depends on the population effect size. In this study the population effect size may be smaller than the minimal effect size of theoretical or practical importance.

The goals for PA and AIPE are fundamentally different. The goal of PA is to obtain a confidence interval that correctly excludes the null value, thus making the direction of the effect unambiguous. The necessary sample size from this perspective clearly depends on the value of the effect itself. On the other hand, the goal of AIPE is to obtain an accurate estimate of the parameter, regardless of whether the interval happens to contain the null value. Thus, sample size from the AIPE perspective does not depend on the value of the effect itself. However, these two methods of sample size planning are not rivals; rather they can be viewed as complementary. In general, the most desirable study design is one in which there is enough power to detect some minimally important effect while also being able to accurately estimate the size of the effect. In this sense, designing a study can entail selecting a sample size based on whichever perspective implies the need for the largest sample size for the desired power and precision. We revisit this possibility in the Power Analysis Versus Accuracy in Parameter Estimation section, in which AIPE and PA are formally compared in a multiple regression framework.

For the moment let us suppose that a researcher has decided to adopt the AIPE perspective. Provided the input population parameters are correct, the techniques that are presented in this article allow researchers to plan sample size in a multiple regression framework such that the confidence interval around the regression coefficient of interest is sufficiently narrow.[3] One approach provides the necessary sample size such that the expected width of the confidence interval will be the value specified. However, achieving an interval no larger than the specified width will be realized only (approximately) 50% of the time. A reformulation provides the necessary sample size such that there is a specified degree of assurance that the computed confidence interval will be no larger than the specified width. The precision of the confidence interval and the degree of assurance of this precision depend on the goals of the researcher. Not surprisingly, all other things being equal, greater precision and greater assurance of the precision necessitate a larger sample size. It is believed that if AIPE were widely applied, it would facilitate the accumulation of a more meaningful knowledge base than does a collection of studies reporting only parameters that are statistically significant but which do not precisely bound the value of the parameter of interest.

Sample Size Estimation for Regression Coefficients

In order to develop a general set of procedures for determining the sample size needed to obtain a desired degree of precision for confidence intervals in multiple regression analysis, we use standardized regression coefficients.[4] Standardized regression coefficients are used for two reasons in developing procedures for determining sample size using an AIPE approach. First, due to the arbitrary nature of the many measurement scales used in the behavioral sciences, standardized coefficients are more directly interpretable. Second, standardized coefficients provide a more general framework in that variances and covariances need not be estimated when planning an appropriate sample size.[5]

---

[3] Although the present article illustrates AIPE in a multiple regression framework, the extension to other applications of the general linear model is not difficult, many of which can be thought of as special cases of multiple regression.

[4] The use of standardized regression coefficients may give rise to technical issues that are addressed in a later section of this article. Standardizing regression coefficients in the presence of random predictors has many appealing characteristics with regard to interpretability, but under certain circumstances problems can develop when using this popular technique.

[5] If the desire is to form confidence intervals around unstandardized regression coefficients, the techniques presented here are equally useful. The desired width of the computed confidence interval is measured in terms of the

The formula for a $100(1 - \alpha)$ percent symmetric confidence interval for a single population standardized regression coefficient, β_j, can be written as follows:

$$\hat{\beta}_j \pm t_{(1-\alpha/2;N-p-1)} \sqrt{\frac{1 - R^2}{(1 - R^2_{XX_j})(N - p - 1)}}, \quad (1)$$

where $\hat{\beta}_j$ is the observed standardized regression coefficient, j represents a specific predictor ($j = 1, \ldots, p$), p is the number of predictors (independent or concomitant variables, covariates, or regressors), R^2 is the observed multiple correlation coefficient of the model, $R^2_{XX_j}$ represents the observed multiple correlation coefficient predicting the jth predictor (X_j) from the remaining $p - 1$ predictors, and N is the sample size (Cohen & Cohen, 1983; Harris, 1985).[6] The value that is added to and subtracted from $\hat{\beta}_j$ to define the upper and lower bounds of a symmetric confidence interval is defined as w, which is the half-width of the entire confidence interval. Thus, the total width of a confidence interval is $2w$. The value of w is of great importance for accuracy in estimation, because the width of the interval determines the precision of the estimated parameter.

In the procedure for planning sample size, the critical value for $t_{(1-\alpha/2;N-p-1)}$ is replaced by the critical $z_{(1-\alpha/2)}$ value. Justification for this can be made because precise estimates generally require a relatively large sample size, and replacing the critical $t_{(1-\alpha/2;N-p-1)}$ value with the critical $z_{(1-\alpha/2)}$ value has virtually no impact on the outcome for the sample size in most cases.[7] The formula used to determine the planned sample size, such that confidence intervals around a particular population regression coefficient, β_j, will have an expected value of the width specified, is obtained by solving for N in Equation 1 and by making use of the presumed knowledge of the population multiple correlation coefficients:

$$N = \left(\frac{z_{(1-\alpha/2)}}{w}\right)^2 \left(\frac{1 - \boldsymbol{R}^2}{1 - \boldsymbol{R}^2_{XX_j}}\right) + p + 1, \quad (2)$$

where $\boldsymbol{R}^2$ represents the population multiple correlation coefficient predicting the criterion (dependent) variable Y from the p predictor variables and $\boldsymbol{R}^2_{XX_j}$ represents the population multiple correlation coefficient predicting the jth predictor from the remaining $p - 1$ predictors. The calculated N should be rounded to the next larger integer for sample size. The w in the above equation is the desired half-width of the confidence interval. It should be kept in mind that this procedure yields a planned sample size that leads to a confidence interval width for a specific predictor. In practice, both $\boldsymbol{R}^2$ and $\boldsymbol{R}^2_{XX_j}$ must be estimated prior to data collection, a complication we address momentarily. Although not frequently acknowledged in the behavioral literature on regression analysis, Equation 1 is derived assuming predictors are fixed and unstandardized. Equation 2 is a reformulation of Equation 1 and thus is based on the same assumptions. Results from a Monte Carlo study are provided later in the article indicating that sample size estimates based on Equation 2 are reasonably accurate when predictors are random and have been standardized.

Equation 2 is intended to determine N such that the expected half-width of an interval is under the researcher's control. However, there is approximately only a 50% chance that the interval will be no larger than specified. The reason for this can be seen from Equation 1. Notice that the width of an interval will depend in part on R^2 and $R^2_{XX_j}$, both of which will vary from sample to sample. Thus, for a fixed sample size, the interval width will also vary over replications. However, it is possible to modify Equation 2 in order to increase the likelihood that the obtained interval will be no wider than desired.

ratio of the standard deviation of Y to the standard deviation of X_j. Thus, following the methods presented for standardized regression coefficients, application to unstandardized coefficients is straightforward.

[6] We introduce the notational system used throughout the article. A boldface italicized $\boldsymbol{R}$ denotes the population multiple correlation coefficient, while a standard-print italicized R is used for its corresponding sample value. A population correlation matrix is denoted by a nonitalicized, boldface, nonserif-font $\mathbf{R}$. A population zero-order correlation coefficient is denoted as a lowercase rho (ρ), whereas a vector of population zero-order correlation coefficients is denoted as a boldface lowercase rho ($\boldsymbol{\rho}$).

[7] The z approximation is poor if the correlations between the predictors and the criterion are large and the correlations among the predictors are small. In this case, the standard error of $\hat{\beta}_j$ is small, producing a relatively small estimated sample size. Under these conditions, the degrees of freedom of the critical t value are small, and thus the critical t value will not closely match the critical z value. We do not believe that this occurs frequently in behavioral research. The al-

If γ is the desired degree of uncertainty of the computed confidence interval being the specified width, Equation 2 can be modified with a multiplicative factor that will provide a modified N such that a researcher can have approximately $100(1 - \gamma)$ percent assurance that a computed confidence interval will be of the specified width or less. For example, if there were a desire to be 80% confident that the obtained w would be no larger than the desired half-width, γ would be defined as 0.20 and there would be only a 20% chance that the half-width of the confidence interval around β_j would be larger than the specified w.

Hahn and Meeker (1991, section 8.3) showed how to plan sample size for confidence intervals when a specified width around the mean of a normal distribution is desired, as well as modifying that formula to obtain $100(1 - \gamma)$ percent confidence that the interval will be of the desired width or less. Taking similar logic and applying it to multiple regression leads to the creation of a formula for a modified N, N_M. This modified formulation provides the necessary sample size in order for researchers to be $100(1 - \gamma)$ percent confident that the β_j of interest will have a corresponding confidence interval width that is no larger than specified. The formula for N_M is given as follows:

$$N_M = \left(\frac{z_{(1-\alpha/2)}}{w}\right)^2 \left(\frac{1 - \boldsymbol{R}^2}{1 - \boldsymbol{R}^2_{XX_j}}\right) \left(\frac{\chi^2_{(1-\gamma;N-1)}}{N - p - 1}\right) + p + 1, \tag{3}$$

where N is the value obtained in Equation 2 and $\chi^2_{(1-\gamma;N-1)}$ is the critical value from a chi-square distribution at the $1 - \gamma$ quantile having $N - 1$ degrees of freedom. Like N, N_M should also be rounded to the next larger integer.

Rather than using the parameter value of the variance for $\hat{\beta}_j$ as was done in the calculation of N, to compute N_M, Equation 3 uses the upper bound of the $100(1 - \gamma)$ percent confidence interval for the variance of $\hat{\beta}_j$. Recall that in any given sample the obtained variance of $\hat{\beta}_j$ will be either larger or smaller than the parameter value specified in Equation 2. Equation 3 uses the maximum value expected for the variance of $\hat{\beta}_j$ at the $100(1 - \gamma)$ percent confidence level. This value is substituted into Equation 2 for the variance of $\hat{\beta}_j$ and thus leads to Equation 3. Because the only random variable in Equation 2 is the variance of $\hat{\beta}_j$, use of Equation 3 provides probabilistic assurance that the obtained confidence interval of interest around β_j will have a half-width no larger than the specified w with $100(1 - \gamma)$ percent confidence.

With regard to choosing a $100(1 - \gamma)$ percent confidence interval for estimation, when compared with a $100(1 - \alpha)$ percent confidence interval for hypothesis testing, important distinctions arise. The most obvious difference in the present context is that γ represents the probability of obtaining a confidence interval with an observed w that is larger than the specified w, whereas alpha is the probability of rejecting a null hypothesis that is true. When making use of Equation 3, a researcher is expected to obtain a w that is larger than the value specified only 100γ percent of the time, regardless of whether or not the null hypothesis is true. Whereas alpha is typically thought of as one of two essentially constant values, .05 or .01, γ is chosen by the researcher in order to achieve some desired degree of assurance that the precision of the estimated parameter will be realized. Thus, confidence intervals formed in the realm of hypothesis testing represent an attempt to accomplish a different goal than those formed when a researcher's interest is in obtaining a precise estimate of the parameter of interest.

Specifying Population Parameters as Input Values

As illustrated in the last section, determining sample size through an AIPE approach requires one to know, or anticipate, $\boldsymbol{R}^2$ and $\boldsymbol{R}^2_{XX_j}$. This is by no means an easy task, but with some careful planning and sound theoretical judgment, it is possible to develop appropriate estimates of the two parameters. In the remainder of this section we suggest different methods for anticipating the values of $\boldsymbol{R}^2$ and $\boldsymbol{R}^2_{XX_j}$, such that sample size planning can be accomplished.

Given that estimates are available for the $p(p + 1)/2$ zero-order population correlation coefficients, the squared multiple correlation coefficient predicting Y from the p predictors can be calculated using the following equation:

$$\boldsymbol{R}^2 = \boldsymbol{\rho}'_{YX}\mathbf{R}^{-1}_{XX}\boldsymbol{\rho}_{YX}, \tag{4}$$

where $\boldsymbol{\rho}_{YX}$ is the population $p \times 1$ column vector of correlations of each X_j regressor with Y (and $\boldsymbol{\rho}'_{YX}$, its transpose), and $\mathbf{R}_{XX}$ is the $p \times p$ population intercor-

ternative method is to solve for the appropriate sample size iteratively, which generally adds unnecessary complications.

relation matrix of all of the predictor variables with one another.[8]

Finding the squared multiple correlation coefficient of variable j from the other $p - 1$ predictors can be readily computed from $\mathbf{R}_{XX}$ in two steps. The first step is to calculate r_{jj}, which for the jth predictor variable is defined as the jth principal diagonal element of $\mathbf{R}_{XX}^{-1}$ (Harris, 1985). In the second step, $\boldsymbol{R}^2_{XX_j}$ for the jth predictor variable is found from the following expression:

$$\boldsymbol{R}^2_{XX_j} = 1 - \frac{1}{r_{jj}}. \quad (5)$$

The inverse of r_{jj} is known as the tolerance of variable j with the other $p - 1$ predictors. The tolerance ($1 - \boldsymbol{R}^2_{XX_j}$) is the proportion of variance of a predictor that cannot be explained by the remaining $p - 1$ predictor variables included in the model. As the tolerance of X_j approaches zero, X_j becomes highly correlated with the remaining predictor variables and $\boldsymbol{R}^2_{XX_j}$ becomes larger, which means there is more predictability, or collinearity, of predictor X_j from the other $p - 1$ predictors (Darlington, 1990, p. 128).

The second method of finding $\boldsymbol{R}^2$ is a variation of the first method and depends on the notion of exchangeability. An *exchangeable structure* (Maxwell, 2000) is one in which the intercorrelations of the predictors are all the same and the correlations of the predictors with the criterion variable are all the same (but ρ_{XX} and ρ_{YX} need not be equal to one another, where ρ represents a population zero-order correlation coefficient). Thus, instead of estimating the $p(p + 1)/2$ zero-order correlations, it is necessary to estimate only two correlations, one for the correlation of each of the predictors with one another and another correlation for each of the predictors with the criterion variable. The two zero-order correlations used in exchangeable structures should be of the general magnitude as the set of correlations they represent. Since B. F. Green (1977) showed that "many linear composites [that is, predicted scores] are barely different from using equal weights" (p. 274), the exchangeable structure offers a potentially useful tool when planning necessary sample size (see Maxwell, 2000, for a thorough treatment and rationale of the exchangeable structure, as well as a similar correlational structure that is somewhat relaxed). Many times an exchangeable structure may be a sensible place to start when planning sample size for a multiple regression analysis, unless there are obvious theoretical reasons not to do so (B. F. Green, 1977; Raju, Bilgic, Edwards, & Fleer, 1999; Wainer, 1976).

If a researcher does not have a good idea of the relationship of the zero-order correlations, conventions such as Cohen's (1988, section 3.2) small ($\rho = .10$), medium ($\rho = .30$), and large ($\rho = .50$) effect sizes for correlations can be used. These correlations can be used directly in Equation 4 or used in an exchangeable structure. For example, if exchangeability seems reasonable and the predictor variables are moderately or highly correlated with one another, a researcher could fill the off-diagonal elements of the $\mathbf{R}_{XX}$ intercorrelation matrix with values of .30, .40, or .50. Further, suppose that it is reasonable to expect that the correlations of the predictors with the criterion are, in general, small or medium. In this case the vector $\boldsymbol{\rho}_{YX}$ can be filled with correlations of .10, .20, or .30. Once acceptable estimates for the two types of correlations have been determined, the multiple correlations can be obtained from Equations 4 and 5.

The third way to determine values for $\boldsymbol{R}^2$ and $\boldsymbol{R}^2_{XX_j}$ is to consult previous literature in order to determine likely values for these two parameters or for likely values of the zero-order correlation coefficients (whether the data follow an exchangeable structure or not). Meta-analytic studies may be of help when estimating the required population parameters; however, in many domains of research, meta-analytic studies have not yet been conducted or the construct of interest may differ from those previously examined.

The final method is presented here more as a warning than a recommendation. This method is based on the commonly recommended approach of sample size planning based on parameter estimates obtained from pilot studies. Pilot studies are sometimes undertaken when literature reviews provide little or no information about the population parameter(s) necessary for sample size planning. However, a potential problem with pilot studies is that these small-scale investigations may yield parameter estimates that do not closely correspond with the parameter values they represent. Thus, basing Equations 2 and 3 on param-

[8] A caution is warranted when estimating the $p(p + 1)/2$ zero-order correlation coefficients, as it is feasible to estimate an impossible set of correlations. If an impossible set is estimated, the multiple correlation coefficient can be greater than one. If this were to occur, adjustments to $\mathbf{R}_{XX}$ and/or $\boldsymbol{\rho}_{YX}$ must be made, such that a realistic set of parameter values can be used for estimating N and N_M.

eter estimates obtained from pilot studies may yield inappropriate estimates of the required sample size if the obtained estimates do not closely approximate their corresponding parameter values.

When planning an appropriate sample size, regardless of whether it is for an application of PA or AIPE, it is typically unrealistic to proceed as if the values of the necessary population parameters are known exactly. Given that, a researcher who uses methods of sample size planning should conduct a sensitivity analysis. A sensitivity analysis involves calculating appropriate sample sizes using a range of realistic values of the necessary population parameters. In the context of the present article, a researcher would specify likely values of $\boldsymbol{R}^2$ and $\boldsymbol{R}^2_{\mathrm{XX}_j}$ in order to determine their effects on N and N_{M}. For the values of N and N_{M} computed with the various parameter values in the sensitivity analysis, the most appropriate estimate of sample size is chosen given what is deemed to be the most appropriate input parameter values. It is also advantageous to triangulate planned sample sizes from multiple methods, rather than focusing only on a single technique. The suggestion of a sensitivity analysis and multiple methods of obtaining estimates of sample size are provided in order for the researcher to have a firm grasp on the nonlinear relationship between the required sample size and the unknown parameter values.

Although the particular value of w is arbitrary and depends only on the desired width for the confidence interval, researchers should keep in mind the likely range of β_j when choosing w, even though the value of β_j itself need not be known. Although there have been conventions established regarding the magnitude of particular effect sizes (e.g., Cohen's, 1988, conventions for the standardized mean difference and the zero-order correlation coefficient), no such conventions have been established for standardized regression coefficients. For example, a medium standardized regression coefficient might be viewed as resulting from medium zero-order correlations. In reality, however, the population β_j will depend greatly on the number of predictors, even when all zero-order correlations are medium. In such multiparameter situations, it becomes very difficult to develop a meaningful scale for small, medium, and large effect sizes.[9]

Even though effect size conventions do not exist for the relative size of the standardized regression coefficient, the likely value of β_j is in the interval $[-1, 1]$. In the special case in which there is only one predictor, β_j is literally the population correlation coefficient between the predictor and the criterion variable. However, if there is more than one predictor variable, the β_js are not confined to the interval $[-1, 1]$, as they do not represent correlations. Thus, the choice of w is not necessarily obvious, in large part because of the interpretation of the standardized regression coefficient and its interrelatedness with the other predictors in the model. Not surprisingly, all other things being equal, the smaller the specified w, the larger the required sample size.

Example and Application of the Procedures

Suppose that a researcher is interested in performing an analysis using multiple regression. Further suppose that the researcher is interested in obtaining a precise estimate of a particular population standardized regression coefficient. In particular, rather than having an embarrassingly large confidence interval around the estimated β_j of interest, the researcher decides that a confidence interval with an expected width of 0.20 will provide a sufficiently precise estimate of β_j; thus, w is defined as 0.10. The researcher is also interested in calculating N_{M}, such that there will be an 80% chance that the β_j of interest will have a corresponding confidence interval that has a half-width no larger than the specified w of 0.10.

Suppose that after consulting past research and in line with theory, the researcher determines that an exchangeable correlational structure seems reasonable, and the five predictor variables that are to be used in the analysis are hypothesized to correlate with one another at .40. Further, suppose there is reason to believe that there is likely to be a medium effect, a correlation of .30, between each of the predictor variables and the criterion.

Following Equation 4, the $\boldsymbol{R}^2$ can be shown to equal .17, and from Equation 5, the $\boldsymbol{R}^2_{\mathrm{XX}_j}$ predicting the jth regressor from the remaining $p - 1$ predictors equals .29. The researcher then solves for the estimated N by use of Equation 2, which yields a value of 453.98. When rounded to the next largest integer, the estimated N from Equation 2 provides the researcher with an estimated sample size of 454. Accordingly, if the

[9] Cohen (1988) even acknowledged the difficulties and inconsistencies in conventions for effect size measures in the context of multiple regression. These inconsistencies are due to the interrelatedness of p, the multiple correlation coefficients, and the zero-order correlation coefficients (Cohen, 1988, p. 413; see also Maxwell, 2000, p. 438).

input parameter values were correct, using a sample size of 454 will yield a confidence interval around β_j that has an expected half-width of 0.10.

To compute N_M, such that there is an 80% chance of obtaining a confidence interval for β_j with a half-width no larger than 0.10, the researcher uses Equation 3. Implicit in Equation 3 for this example is the fact that the sample variance of $\hat{\beta}_j$ is expected to be less than the parameter value 80% of the time. Because the obtained w will be less than the w specified if the variance of $\hat{\beta}_j$ is smaller in the sample than the parameter value used to estimate sample size, the obtained w will be no greater than the specified w with a probability of .80.

The .80 quantile of the chi-square distribution with $N - 1$ degrees of freedom is 478.12. This critical chi-square value is then divided by $N - p - 1$, yielding a variance correction factor of 1.07. Following Equation 3, N_M is estimated at 484.10 and after being rounded up to the next largest integer yields a value of 485. If the parameter values estimated by the researcher were correct, using an N_M of 485 will provide the researcher with approximately an 80% chance of obtaining a w of 0.10 or less for the confidence interval around the beta weight of interest. Notice that sample size increases by only 31 (or 6.83%) when specifying 80% confidence that the obtained w would be less than the specified width. Typically N_M is not considerably greater than N and should be considered for the added assurance it provides for a precise estimate with what generally amounts to a relatively small cost.

When the assumption of exchangeability does not hold, generally a different sample size will be estimated for each of the p predictors. In the following example, suppose a researcher hypothesizes the following population parameters for the $\mathbf{R}_{XX}$ intercorrelation matrix and the $\boldsymbol{\rho}_{YX}$ vector, respectively:

$$\mathbf{R}_{XX} = \begin{bmatrix} 1 & & \\ .40 & 1 & \\ .60 & .05 & 1 \end{bmatrix} \qquad \boldsymbol{\rho}_{YX} = \begin{bmatrix} .50 \\ .30 \\ .10 \end{bmatrix}.$$

Further suppose the desired half-width and alpha were set to 0.15 and .05, respectively. In this scenario, the planned sample sizes would be estimated as 237, 154, and 201 for Predictors 1, 2, and 3, respectively. Furthermore, if the researcher wanted to have 90% confidence that the obtained w would be less than or equal to 0.15, N_M would be 268, 180, and 229 for Predictors 1, 2, and 3, respectively. Thus, when exchangeability does not hold, planning sample size for a specific predictor may provide expected ws narrower or wider than the specified value for the remaining $p - 1$ predictors, depending on the tolerance of the predictor for which sample size was calculated.

When interest lies in the w for a specific predictor, no problems arise regardless of whether the correlational structure is or is not exchangeable. Sample size is calculated for the specific predictor regardless of whether the tolerance for the predictor of interest is smaller or larger than any of the remaining $p - 1$ predictors. Under this strategy, researchers are concerned foremost with the width of the confidence interval for the beta of interest and less so for the remaining $p - 1$ predictors. For example, in the scenario in the previous paragraph, a researcher whose question pertains specifically to estimating the relationship between X_3 and Y controlling for X_1 and X_2 should choose an N of 201 or an N_M of 229.

Another strategy in situations in which exchangeability does not hold leads to the expected value of all of the confidence intervals being as narrow as or narrower than the specified w. In this approach the sample size used for the study is the largest of the p different sample sizes. Thus, the expected half-width for the predictor with the lowest tolerance is w, whereas the expected half-widths for the remaining $p - 1$ confidence intervals will be less than w; to what degree depends on the tolerance of the other predictors. For example, given N_M values of 268, 180, and 229 for the three predictors, respectively, a researcher interested in a narrow confidence interval for each and every predictor should choose an N_M of 268.

Power Analysis Versus Accuracy in Parameter Estimation

Estimating sample size from a PA perspective is conceptually different than estimating sample size to achieve AIPE. This conceptual difference can potentially translate into very different practical implications. This section considers the relative sample sizes required by the two approaches. Maxwell (2000) showed that sample size could be estimated for a given predictor to obtain a specified power using the following formula:

$$N = \left(\frac{\lambda}{\beta_j^2}\right)\left(\frac{1 - \boldsymbol{R}^2}{1 - \boldsymbol{R}^2_{XX_j}}\right) + p - 1, \qquad (6)$$

where λ is a noncentrality parameter from an F distribution with 1 numerator and $N - p - 1$ denominator degrees of freedom. The λ value in Equation 6 is a

tabled critical value that determines the power of a given statistical test for a predictor of interest. The required value of λ for a specified degree of power can be obtained from Cohen's (1988, pp. 448–455) tables or from the appropriate noncentral *F* distribution.

The relative sample size required for AIPE versus PA can be compared by the following two multiplicative ratios found in Equations 2 and 6, respectively:

$$\left(\frac{z_{(1-\alpha/2)}}{w}\right)^2$$

versus

$$\left(\frac{\lambda}{\beta_j^2}\right).$$

Unless *p* is very large, the ratio of required sample size for AIPE compared with PA is approximately $(z_{(1-\alpha/2)}\beta_j)^2/(\lambda w^2)$ to 1. Note that the population standardized regression coefficient is the only one of the four values beyond the researcher's control. Whereas α, λ, and *w* are chosen to coincide with the goals of the research project, the PA approach requires that the parameter value or the minimally important value of the standardized regression coefficient be specified. Note that a value for the standardized regression coefficient is not necessary when planning sample size for precision. For this reason, planning sample size from the AIPE perspective is actually easier than approaching sample size planning from the PA perspective.

Unless *p* is very large, sample size for PA is approximately

$$N = \mathrm{M_{PA}}\left(\frac{1-\boldsymbol{R}^2}{1-\boldsymbol{R}^2_{\mathrm{XX}_j}}\right), \tag{7}$$

where $\mathrm{M_{PA}} = \lambda/\beta_j^2$, which is the multiplier used for the PA approach. Similarly, sample size for AIPE is approximately

$$N = \mathrm{M_{AIPE}}\left(\frac{1-\boldsymbol{R}^2}{1-\boldsymbol{R}^2_{\mathrm{XX}_j}}\right), \tag{8}$$

where $\mathrm{M_{AIPE}} = (z_{(1-\alpha/2)}/w)^2$, which is the multiplier used in the AIPE approach. Figure 2 depicts the re-

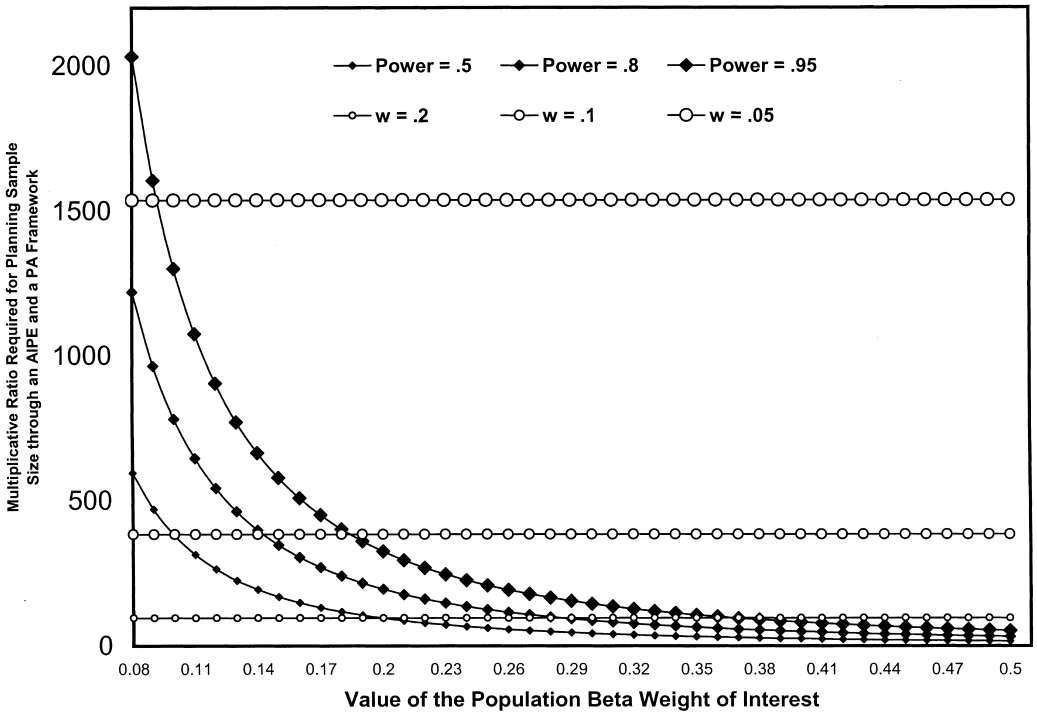

Figure 2. Relationship of the relative planned sample size for the accuracy in parameter estimation (AIPE) and the power analytic (PA) approaches to sample size planning as a function of the population beta weight (approximate sample size in the special case when $\boldsymbol{R}^2 = \boldsymbol{R}^2_{\mathrm{XX}_j}$).

lationship of the multipliers for PA and AIPE for population betas for various values of power and precision ($\alpha = .05$). As Equations 7 and 8 show, multiplying the corresponding value on the ordinate for either power or precision in Figure 2 by the ratio $(1 - \boldsymbol{R}^2)/(1 - \boldsymbol{R}^2_{\mathrm{XX}_j})$ yields an approximate sample size. More generally, the relative elevation of a curve or line represents the relative sample size required to achieve a desired level of power or precision.

Several practical implications emerge from Figure 2. First, as the curves and lines show, as the population β_j becomes larger, sample size for power can be much smaller than it is for precision. Conversely, when the β_j is small, sample size for power can be much larger than is required for precision. For example, when the β_j equals 0.30, the sample size required to obtain a confidence interval with an expected half-width of 0.10 is just over 4 times as large as the sample size needed to obtain a power of .80. However, when β_j is 0.08, the sample size needed for a power of .80 is more than 3 times larger than that needed to obtain a confidence interval with an expected half-width of 0.10. Note that these relationships hold true regardless of the values of $\boldsymbol{R}^2$ and $\boldsymbol{R}^2_{\mathrm{XX}_j}$, as both of these values play the same role in Equations 7 and 8. Second, for constant values of $\boldsymbol{R}^2$ and $\boldsymbol{R}^2_{\mathrm{XX}_j}$, sample size for precision is independent of the value of β_j, whereas smaller samples can provide adequate power for larger values of β_j. Third, implicit in Equation 8 and as depicted in Figure 2, halving the width of a confidence interval for β_j requires approximately a fourfold increase in sample size. Fourth, in the special case in which $\boldsymbol{R}^2$ is equal to $\boldsymbol{R}^2_{\mathrm{XX}_j}$—that is, $(1 - \boldsymbol{R}^2)/(1 - \boldsymbol{R}^2_{\mathrm{XX}_j}) = 1.00$—the values on the ordinate based on the curve for power and the line for precision are approximately the required sample sizes.

Thus, it is clear that the two methods are different from the outset and can yield very different estimates of sample size in the same study. Each is designed to answer a different question, and as can be seen, they do just that. The two approaches differ on a philosophical level, one designed to achieve a narrow interval and one designed to obtain an interval that does not contain the specified null value. The point is that depending on what the researcher's question is and the desired outcome, a different approach to sample size estimation will be needed. Neither approach is necessarily "right" or "wrong" for a given problem; these approaches are merely different in the questions that they attempt to answer. It is recommended that the two approaches be used in conjunction with one another in order to achieve reasonable statistical power while obtaining confidence intervals that are sufficiently narrow.

Random Versus Fixed Predictors and the Issue of Standardization

In the present article it was assumed that the predictor variables were random and that all variables were standardized. The reason that standardized values were discussed exclusively is because correlations tend to be easier to hypothesize and work with than variances and covariances, which would be necessary to carry out AIPE in the unstandardized case. Another reason why standardized regression coefficients are beneficial is because of the arbitrariness of most scales of measurement used in the behavioral sciences. Furthermore, a widely used convention of the magnitude of effect is available for correlations in psychology (Cohen, 1988, section 3.2). It should be clear, however, that if the hypothesized values are correct when finding N and N_{M} for standardized values, they will provide the same relative degree of precision around the unstandardized regression coefficients. The relative degree of precision regarding w is scaled in terms of the ratio of the standard deviation of the criterion to the standard deviation of the jth predictor (s_Y/s_{X_j}).

With regard to random and fixed predictor values in the unstandardized case, Sampson (1974) showed that regardless of the predictors being fixed or random, "we obtain the same estimates for the regression coefficients and the variance of the error" (p. 684 from Theorem 1). There is, however, a difference between the two cases. Note that if $\boldsymbol{R}^2 = 0$, then the distribution of R^2 is identical in both cases and follows a central F distribution. However, the distribution of R^2 is different for the two cases when $\boldsymbol{R}^2 \neq 0$ (Stuart, Ord, & Arnold, 1999, section 28.29). In fact, the distribution of R^2 is a noncentral F distribution in the case of fixed predictors, whereas it is not in the case of random predictors (Rencher, 2000, pp. 240–241). Accordingly, the distribution of the test statistic under the null hypothesis is the same for the fixed as well as the random X case, but the power functions for the test statistic are different for the two cases (Rencher, 2000, chapter 10). Gatsonis and Sampson (1989) showed that Cohen's (1988) power tables for determining sample size are approximations, because Cohen treated random predictors as though they were fixed. However, Gatsonis and Sampson concluded that "Cohen's approximation works quite well in

many situations" (p. 519). Thus, practically speaking, random versus fixed X values have little effect on applied research because the consequences, in most cases, are trivial. The issue of standardization, however, is quite different, especially when standardization is performed on random predictor variables.

Even though multiple regression using standardized random predictors is common practice in behavioral research, as well as in many other fields, there are nuances associated with this strategy that are not widely known and are potentially problematic. As previously stated, the formula (see Equation 1) for the standard error of a regression coefficient that is random and standardized is approximate. The formula, as given explicitly in sources such as Cohen and Cohen (1983) and Harris (1985) and implicitly in many others, treats the standard deviation of each predictor as a constant value. This is obviously not the case when the predictors are random, as the standard deviation of the predictor is itself a random variable. This is contrasted with the situation in which the values of the predictor variables are preset in advance and thus the standard deviation of those predictors would not vary across replications of the study.

In order to transform an unstandardized regression coefficient to a standardized regression coefficient, one can multiply the raw score regression coefficient by s_{X_j}/s_Y, so as to remove the (generally arbitrary) scaling of Y and X_j. Likewise, this same procedure is commonly done in order to obtain the standard error of the standardized regression coefficient.[10] However, "standard errors of standardized parameters, in general, are not a simple rescaling of the standard errors of the original parameter estimates" (Jamshidian & Bentler, 2000, p. 74). The problem with scaling the standard error of a standardized regression coefficient in the random predictor case can be seen by a well-known property of variances. If C is a constant and V is a random variable, $\mathrm{Var}(CV) = C^2\mathrm{Var}(V)$, where $\mathrm{Var}(\cdot)$ represents the variance of the quantity in parentheses. However, if $\tilde{C}$ is itself a random variable, then $\mathrm{Var}(\tilde{C}V) \neq \tilde{C}^2\mathrm{Var}(V)$. Common formulas for the standard error of standardized regression coefficients (e.g., Equation 1) assume that the standard deviation of the predictor is fixed. In the case of random predictor variables, such an assumption implies that $\mathrm{Var}(\tilde{C}V) = \tilde{C}^2\mathrm{Var}(V)$. Because this assumption is false, the variability of X_j is not taken into consideration when calculating the standard error of standardized regression coefficients from the random X case, which generally leads to incorrect standard errors.

In structural equation modeling (SEM), which can be viewed as a generalization of multiple regression, several authors have illustrated the potential problems of analyzing a correlation matrix as if it were a covariance matrix (e.g., Babakus, Ferguson, & Jöreskog, 1987; Browne, 1982; Cudeck, 1989; Jamshidian & Bentler, 2000). Steiger (2001) concluded that SEM parameter estimates based on a correlation matrix (analogous to standardized coefficients in multiple regression) may be correct, whereas their standard errors are incorrect (see also Lawley & Maxwell, 1971, chapter 7, for technical details). MacCallum and Austin (2000) stated that when a correlation matrix is analyzed as if it were a covariance matrix in SEM, "in all cases, standard errors of parameter estimates as well as confidence intervals and test statistics for parameter estimates will be incorrect," and they further emphasized that the "correct standard errors will generally be smaller than the incorrect values which results in narrower confidence intervals and larger test statistics" (p. 217). For the reasons outlined in this section regarding the approximate nature of Equation 1, a simulation study was conducted to verify the integrity of the procedures suggested throughout the article.

Results of Monte Carlo Simulations

If Equation 1 was exact, the assumptions were met, and the multiple correlation coefficients were correctly specified, the sample size estimation procedures presented here yield correct estimates of required sample size. However, whenever the values of the predictors are random and standardized, rather than being fixed, Equation 1 is an approximation. In applications of multiple regression to observational studies in the behavioral sciences, predictors are typically random, not fixed. Further, standardization often occurs in the behavioral sciences because of the in-

[10] The reason that multiplying the standard error of the unstandardized regression coefficient by s_{X_j}/s_Y removes the scaling of the jth predictor can be seen by the formula for the standard error of the unstandardized regression coefficient: $(s_Y/s_{X_j})\sqrt{(1-R^2)/[(1-R^2_{XX_j})(N-p-1)]}$. Multiplying this formula by s_{X_j}/s_Y removes the scaling of Y and X_j from the standard error and is commonly, yet inappropriately, assumed to be the correct standard error for the jth standardized regression coefficient when the predictor is random.

terpretational problems associated with arbitrary scales of measurement. Under these circumstances, it was unclear whether basing planned sample size on Equation 2 would produce an interval with the desired width. In addition to ensuring that Equation 2 consistently yields accurate estimates of sample size, a Monte Carlo study was necessary because Equation 3 implicitly assumes Equation 2 is correct.

One scenario studied in the Monte Carlo simulation was the aforementioned exchangeable structure with five predictors and where $\rho_{XX} = .40$ and $\rho_{YX} = .30$. The simulation revealed that Equations 2 and 3 produced very accurate results in this situation. Recall that when w is specified as 0.10 for this scenario, Equation 2 dictates a necessary sample size of 454. The mean w for the five betas, each based on 10,000 replications, using a sample size of 454, was 0.101, with a standard deviation of 0.003; the median w was also 0.101. Recall that having an 80% chance of obtaining a w no larger than the specified value of 0.10 requires a necessary sample size of 485 based on Equation 3. The mean and the median confidence interval half-width using a sample size of 485 was 0.098, with a standard deviation of 0.003. Most important, 81.64% of the obtained ws were no larger than the specified value of 0.10. Further, the 80th percentile for the empirical distribution of the obtained ws was 0.10. In summarizing the results for this scenario, the suggested procedures yielded an original sample size such that the mean of the ws was 0.101 and a modified sample size that led to just over 80% of the confidence intervals being no larger than specified.

This example was selected because we thought it was reasonably typical of a behavioral research scenario. However, this single scenario cannot address the extent to which the approximation is accurate for other situations. To investigate the general accuracy of the procedures, we undertook a large Monte Carlo simulation study to address the appropriateness of Equation 2. In the simulation study 166 different conditions were examined. In the different conditions a variety of correlational structures were used. The ws were specified to be 0.025, 0.05, 0.10, 0.15, 0.20, 0.15, and 0.35, using ps of 2, 5, and 10. Presumably the simulations encompass the likely ranges of w and p that is commonly of interest to behavioral researchers, combined with a variety of correlation structures to show generality. Each condition in the simulation study was based on 10,000 replications. The results showed that the suggested procedures generally performed very well. Because of the large number of conditions that were studied, the tabled results could not be presented; however, detailed descriptions of the results follow.[11]

The mean, median, and standard deviation of the percentage of error were determined for each of the 166 conditions that were examined. The percentage of error was determined by subtracting the specified w from the mean of the obtained ws, dividing this difference by the specified w, and then multiplying by 100. For example, if the mean of the obtained ws was 0.204 when the specified w was 0.20, the percentage of error would be computed as follows: $100(0.204 - 0.20)/0.20 = 2.00$. Thus, in this condition the mean of the obtained ws was 2.00% larger than the specified w.

In the simulation conditions in which p was 2, all combinations of small, medium, and large correlations among the predictors as well as the criterion (27 total) were completely crossed with ws of 0.05, 0.10, and 0.20. Thus, a total of 81 different conditions were examined for $p = 2$. The mean and median of the percentage of error were 0.33 and 0.17, respectively, with a standard deviation of 0.34. The minimum percentage of error was 0.01 for a case in which w was 0.05, and the maximum percentage of error was 1.85 for a case in which w was 0.20. Thus, in the worst case out of the 81 different conditions for $p = 2$, the mean of the obtained w was less than 0.01 units larger than expected.

In the case in which p was 5, the results are reported separately for two different types of correlational structures. In the first type of correlational structure, 25 different exchangeable structures were examined. In any single one of the 25 combinations, all predictors correlated equally among themselves and each correlated equally with the criterion variable. Correlations among predictors consisted of ρ_{XX} values of .10, .20, .30, .40, and .50. Correlations of the predictors with the criterion consisted of ρ_{YX} values of .10, .20, .30, .40, and .50. Thus, ρ_{XX} and ρ_{YX} each varied from small to large by .10 and yielded a 5×5 factorial design.

Two combinations of correlations are excluded

[11] The complete set of simulation results is available in tabular format from Ken Kelley or Scott E. Maxwell. The code, which was written in R/S-PLUS, is also available on request. Note that the anonymous reviewers were provided with the simulation results as part of their assessment of our procedures.

from the following descriptive statistics because their multiple correlations between the predictors and criterion are greater than .80 and not representative of most psychological research.[12] The mean and median percentage of error for the remaining 23 *w*s were 1.87 and 1.03, respectively, with a standard deviation of 2.22. The minimum percentage of error was 0.22, and the maximum was 10.00. This worst case occurred when the correlations among the predictors were .10 and the correlations between the predictors and criterion were .40. This correlational structure is unlikely in most behavioral research because $\boldsymbol{R} = .76$. However, even this condition had a mean *w* that was only 0.01 units larger than expected.

The other simulations that were conducted for $p = 5$ were based on two published correlational structures. The first was a subset of a correlation matrix obtained from the developmental literature (Smari, Petursdottir, & Porsteinsdottir, 2001), and the other was obtained from an example given in an SEM text (Table 7.1 in Loehlin, 1998). The mean and median of the absolute percentage of error for the 30 conditions (15 from each example) were 0.55 and 0.23, respectively, with a standard deviation of 0.76. The minimum of the absolute percentage of error was 0.01 in a condition in which *w* was 0.025, and the maximum was 2.75 in a condition in which *w* was 0.35. Thus, the worst condition in this situation produced a mean *w* of 0.36 when the specified *w* was 0.35.

For $p = 10$, the correlation matrix used was a subset of one obtained from the clinical–counseling literature that had previously been cited in an SEM text (Worland, Weeks, Janes, & Strock, 1984, as cited in Kline, 1998, p. 254). The mean and median of the percentage of error for the 30 conditions that were examined were 0.18 and 0.09, respectively, with a standard deviation of 0.19. The smallest absolute percentage of error was less than 0.01 for a case in which *w* was 0.05, and the largest percentage of error was 0.67 for a condition in which *w* was 0.20. Thus, the condition with the largest discrepancy had a percentage of error less than 1%.

Recall the cited SEM literature in which it has been shown that the standard errors of parameter estimates are generally inflated when a correlation matrix is treated as a covariance matrix. Because ordinary least squares (OLS) multiple regression is a special case of SEM, it follows that the standard errors of OLS multiple regression are often inflated when predictor variables are random and standardized. In 130 of the 166 conditions investigated (78.31%), the confidence interval coverage was greater than 95% (the nominal alpha was set to .05). The mean and median percentage of coverage were 95.53 and 95.24, respectively, with a standard deviation of 0.78. Whereas the smallest percentage of coverage was 94.34, the largest percentage of coverage was 97.89. Thus, the results of the simulations have shown empirically the approximate nature of Equation 1 and the fact that OLS multiple regression tends to have inflated standard errors when predictor variables are random and have been standardized.[13]

The fact that Equation 1 is approximate and generally provides confidence intervals wider than necessary raises some questions regarding its use as well as the use of Equations 2 and 3 in the context of sample size planning for precise estimates of standardized regression coefficients. For example, in the case in which the largest confidence discrepancy occurred, 97.89% of the computed confidence intervals bracketed the population parameter. Applying Equation 1 to this condition ($w = 0.10$, $\rho_{YX_1} = .50$, $\rho_{YX_2} = .10$, $\rho_{X_1X_2} = .10$, $p = 2$), we found that the population correlations would suggest that the standard error was 0.051. A simulation based on 1,000,000 replications showed that, consistent with the SEM literature, the standard deviation of the regression coef-

[12] The two excluded cases consisted of unlikely scenarios for much behavioral research. The first excluded scenario consisted of correlations among the predictors of .10 and correlations between the predictor and the criterion of .50. Such a combination of correlations leads to an $\boldsymbol{R}$ of .95 and where the requirement of a positive definite correlation matrix is nearly violated. In this case the mean *w* was 0.151 when it was specified to be 0.10. Poor performance of the technique in this particular scenario is not surprising, given that many statistical procedures fail when parameters approach their theoretical bounds. The second excluded case is similar to the first and consisted of correlations among the predictors of .20 and correlations between the predictor and the criterion of .50. This combination of correlations leads to an $\boldsymbol{R}$ of .83. In this second excluded scenario, the mean *w* was 0.112 when it was specified to be 0.10.

[13] Many behavioral scientists would see no problem with an empirical alpha smaller than the nominal alpha level and thus with being more conservative. However, a toxicologist or bioscientist working with chemical agents or medicine would likely argue that a Type II error may be more costly than a Type I error, as concluding that there is "no effect" of a noxious substance could be a harmful mistake. Further, power and precision will be sacrificed if the actual Type I error rate is smaller than the nominal alpha level.

ficients was 0.044, a value smaller than implied by Equation 1. This result suggests that the sample size calculated from Equation 2, which assumes the standard error from Equation 1 is correct, is approximate and in this particular case somewhat negatively biased. Unfortunately, no exact formula for the standard error is known to exist when predictors are random and standardized. Thus, given the current state of knowledge, researchers need to continue to use Equation 1 for forming confidence intervals around regression coefficients for predictors that are random and standardized. Equations 2 and 3 can then be used in the research design phase in order to determine approximate sample sizes for precise estimates of the regression coefficients of interest.

Limitations of the Procedure

Although the distribution of R^2 is asymptotically normal throughout most of its domain (Stuart et al., 1999, section 28.33), this is not the case as $\boldsymbol{R}^2$ approaches its limits. When $\boldsymbol{R}^2$ begins to approach zero, the distribution of the observed R^2 values becomes positively skewed because of the lower bound at zero. The converse is true as $\boldsymbol{R}^2$ begins to approach one, and thus the distribution of the observed R^2 values will be negatively skewed.

The fact that the distribution of R^2 becomes negatively or positively skewed affects sample size estimation in two ways. Recall from Equations 2 and 3 that there are two multiple correlations in the equations for determining sample size, the model $\boldsymbol{R}^2$ in the numerator and $\boldsymbol{R}^2_{\mathrm{XX}_j}$ in the denominator. As $\boldsymbol{R}^2$ approaches zero in the population, the estimated sample size for a planned study based on Equation 2 or Equation 3 will, with everything else held constant, tend to be larger than necessary. One way to understand why overestimation occurs is to inspect Equation 1. On the basis of this equation, a confidence interval becomes narrower as $1 - R^2$ becomes smaller. As $\boldsymbol{R}^2$ approaches zero and thus the distribution of R^2 becomes more positively skewed, the mean R^2 tends to be greater than $\boldsymbol{R}^2$, implying that the mean $1 - R^2$ tends to be less than $1 - \boldsymbol{R}^2$. Accordingly, the observed confidence intervals will tend to be narrower than expected based on the value of $\boldsymbol{R}^2$. The estimated sample size from Equation 2 or Equation 3 is a function of $\boldsymbol{R}^2$; thus, confidence intervals based on sample size estimates from these equations will tend to be narrower than specified when the model $\boldsymbol{R}^2$ approaches zero. In other words, for a desired degree of precision, sample size estimates become inflated as $\boldsymbol{R}^2$ approaches zero. The opposite pattern of results occurs when $\boldsymbol{R}^2$ begins to approach one. In this case the proportion of variance unaccounted for is, on average, larger in the sample than is implied by $\boldsymbol{R}^2$. Consequently, the use of Equation 2 or Equation 3 will tend to underestimate sample size.

The same phenomenon happens in the denominator with $\boldsymbol{R}^2_{\mathrm{XX}_j}$ as it does in the numerator with $\boldsymbol{R}^2$; the only difference is that the relationship is the exact opposite. Because $\boldsymbol{R}^2_{\mathrm{XX}_j}$ is in the denominator of Equation 2, the sample size is over- or underestimated in a reverse fashion as was illustrated for $\boldsymbol{R}^2$.

For simplicity, the discussion has been limited to regression models that include only main effects and no interaction or other higher order (polynomial) terms, as there are certain nuances associated with multiplicative terms that have been scaled in multiple regression models (see chapter 3 of Aiken & West, 1991, for details regarding multiplicative effects in multiple regression). Furthermore, the procedures given here assume that all predictors are included in the regression model and that no selection of predictors occurs (as would be the case in, e.g., a stepwise regression analysis).

Discussion

Approaching sample size estimation from a perspective of AIPE rather than one exclusively emphasizing power is beneficial for a productive science. Although planning sample size through PA studies is important and undeniably improves research findings, the accuracy in those parameter estimates should be at least as much of a concern as their probability value, perhaps even more so. An optimal experimental design consists of an adequate sample size from an AIPE perspective as well as an adequate sample size from the PA perspective. Ensuring that sample size is adequate from both perspectives leads to parameter estimates that will likely be accurate as well as statistically significant.

A special case in which precision is especially important occurs when the goal is to provide evidence in support of the null hypothesis. If a confidence interval is sufficiently narrow and power is of sufficient strength (say, power > .90), at times it may be appropriate to show support for the null hypothesis, in the sense that the value of the parameter is not meaningfully different from the null value. Note that this is not "accepting the null hypothesis" but is merely showing support for it (Greenwald, 1975).

The simulation study showed that the procedures presented here were effective in accomplishing their respective goals. The mean and median of the observed *w*s were very close to their specified values when the estimated *N* (Equation 2) was used to select sample size. When using *N*, researchers are reminded that this provides the necessary sample size such that the expected half-width of the confidence interval is, on average, the specified width. However, this does not ensure that the particular observed *w* will be the specified width in any given sample. The modified sample size (Equation 3) takes into consideration the variability of the standard error of $\hat{\beta}$ and adjusts the sample size accordingly, such that one can be approximately $100(1 - \gamma)$ percent confident that the width around a particular β_j will have a corresponding *w* that is no larger than the specified *w*.

A caution is given because of the problems that can arise when using standardized variables from random *X* values in the context of multiple regression. Although there are numerous reasons to use standardized values as input into multiple regression models, and thus make use of their corresponding estimates for interpretational reasons, the standard errors of such estimates are generally not exact. Even though the simulations show that the common method of standardizing random predictors produces confidence intervals for standardized regression coefficients that are generally wider than they should be, the sample size procedures we present typically produce the desired degree of precision.

In conclusion, the AIPE procedures presented here are applicable to researchers working within the framework of OLS multiple regression who want to determine sample size a priori in order to obtain accurate parameter estimates. Given reasonably accurate input parameters, use of these procedures provides researchers with confidence intervals around regression coefficients whose expected widths are the values specified or, alternatively, with some degree of probabilistic assurance. As with all sample size planning, the AIPE procedures will be less accurate to the extent that the input parameters deviate from their true values. However, the problem with the choice of input parameters should not be used as a reason to avoid sample size planning. In addition, we have shown that planning sample size for precise estimates of standardized regression coefficients requires less a priori knowledge (i.e., fewer input parameters) than the corresponding planning necessary to obtain sufficient statistical power. We believe that obtaining accurate parameter estimates, not merely statistically significant ones, leads to a more productive science and yields research findings that are more beneficial to a given area of inquiry.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Algina, J., & Olejnik, S. (2000). Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research, 35,* 119–137.

Babakus, E., Ferguson, C. E., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distribution assumptions. *Journal of Marketing Research, 24,* 222–228.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkings (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). New York: Cambridge University Press.

Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis.* New York: Chapman & Hall.

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

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

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105,* 317–327.

Darlington, R. B. (1990). *Regression and linear models.* New York: McGraw-Hill.

Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin, 106,* 516–524.

Green, B. F. (1977). Parameter sensitivity in multivariate methods. *Multivariate Behavioral Research, 12,* 263–288.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26,* 499–510.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

Hahn, G. J., & Meeker, W. Q. (1991). *Statistical intervals: A guide for practitioners.* New York: Wiley.

Harris, R. J. (1985). *A primer of multivariate statistics* (2nd ed.). New York: Academic Press.

Hellmann, J. J., & Fowler, G. W. (1999). Bias, precision, and accuracy of four measures of species richness. *Ecological Applications, 9,* 824–834.

Jamshidian, M., & Bentler, P. M. (2000). Improved standard errors of standardized parameters in covariance structure models: Implications for construct explication. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment* (pp. 73–94). Dordrecht, the Netherlands: Kluwer Academic.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford Press.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). London: Butterworth.

Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis* (3rd ed.). Mahwah, NJ: Erlbaum.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51,* 201–226.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5,* 434–458.

Mendoza, J. L., & Stafford, K. L. (2001). Confidence intervals, power calculation, and sample size estimation for the squared multiple correlation coefficient under the fixed and random regression models: A computer program and useful standard tables. *Educational and Psychological Measurement, 61,* 650–667.

Muller, K. E., & Benignus, V. A. (1992). Increasing scientific power with statistical power. *Neurotoxicology and Teratology, 14,* 211–219.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement, 23,* 99–115.

Rencher, A. C. (2000). *Linear models in statistics.* New York: Wiley.

Rossi, J. C. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58,* 646–656.

Rozeboom, W. W. (1966). *Foundations of the theory of prediction.* Homewood, IL: Dorsey Press.

Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association, 69,* 682–689.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Smari, J., Petursdottir, G., & Porsteinsdottir, V. (2001). Social anxiety and depression in adolescents in relation to perceived competence and situational appraisal. *Journal of Adolescence, 24,* 199–207.

Steiger, J. H. (2001). Driving fast in reverse: The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association, 96,* 331–338.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical methods. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, NJ: Erlbaum.

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

Thompson, B. (Ed.). (2001). Confidence intervals around effect sizes [Special issue]. *Educational and Psychological Measurement, 61* (4).

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 213–217.

Wilkinson, L., & the American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Received December 11, 2001
Revision received March 18, 2003
Accepted April 23, 2003 ■