The sample size necessary to address a research question depends on the goals of the researcher. Do researchers want to know whether a particular effect exists or do researchers want to know the magnitude of the effect? We begin this chapter with two illustrations. In one, knowledge of the existence of an effect is of interest and, in the second, the size of an effect is of interest. We use these illustrations to show the types of research questions that can be addressed using the statistical frameworks commonly employed by psychologists in the 21st century. Ultimately, we argue that the particular research question is what drives how it should be addressed statistically, either to infer an effect exists or to make inferences about the magnitude of an effect.

First, consider a team sporting competition. In general, the most important outcome is which team wins the game or if the game ends in a tie, rather than the score. If the score is known, then the winner can be determined by deduction, but of primary interest is simply knowing who won. Second, consider a retirement portfolio. In general, the outcome of interest is the amount the portfolio’s value changed over some time period, either in raw dollars or as a percentage of the initial investments, not simply whether the portfolio increased or decreased in value over some time period. If the amount of change is known, it also is known whether the portfolio increased or decreased in value. Of primary interest, however, is how much the value of the portfolio changed.

We use the team sport and retirement portfolio examples as analogies for the types of questions that are generally of interest to psychologists and how statistical methods can help answer a particular question. In some situations, the dichotomous decision of a null hypothesis significance test (NHST; specifically to reject or fail to reject the null hypothesis) answers the question of interest and in some situations with an additional benefit of finding the direction of an effect. Examples of dichotomous research questions are (a) does the proposed model explain variance in the outcome variable, (b) is there a relationship between a predictor and the criterion variable after other predictors have been controlled, (c) do the treatment and control groups have different means, and (d) is there a correlation between the two variables? For the latter three questions, the direction of the effect is immediately known if the effect can be deemed statistically significant. The first question does not have a directional component. All of these and many other questions that fit into the reject or fail-to-reject NHST framework can be useful for evaluating theories and learning from data. Such questions, however, are not the only type of scientific questions of interest.

In some situations, rather than answering a research question with “the results were statistically significant,” the question concerns estimating the magnitude of an effect. For example, the four research questions can be reframed in the magnitude estimation context as follows: (a) How much of the variance in the outcome variable is explained by the proposed model? (b) How strong is the relationship between a predictor and the criterion variable after other predictors have been controlled? (c) How large is the mean difference between the treatment
group and the control group? (d) How strong is the correlation between the two variables? Each of these questions attempts to quantify the magnitude of an effect with an effect size and corresponding confidence interval, where a confidence interval is an interval that will bracket the population value with some specified degree of confidence. The confidence intervals may contain the null value but still be narrow, which from a magnitude-estimation framework is considered a success, even though it would not be considered a success if the null hypothesis is false but not rejected. Thus, the magnitude-estimation framework, like the NHST framework, can be used to evaluate theories and learn from data, but it does so in a different way.

The various types of questions that researchers may pose have different implications for the methodological techniques required to address the question of interest. Correspondingly, the way in which a study is designed with regard to sample size planning is very much dependent on the type of research question. In this chapter, we discuss both approaches and the implications that choosing one approach over the other have for sample size planning. We discuss sample size planning, including (a) when the existence of an effect or the direction of an effect is of primary interest (from the power analytic perspective) and (b) when the magnitude of an effect is of primary importance (from the accuracy in parameter estimation [AIPE] perspective).

The power analytic approach and the AIPE approach are two fundamentally different ways to plan sample size. For now, however, it is necessary only to understand that the power analytic approach helps to plan sample size when research questions concern whether an effect exists and in some cases the direction of the effect, whereas the AIPE approach helps to plan sample size when research questions concern the magnitude of an effect, which in some but not all cases can also be concerned with the direction of the effect. Because the two approaches to sample size planning are fundamentally different, the sample sizes necessary for the approaches may differ substantially. We do not claim that one approach is better than the other. Rather, we contend that the appropriate approach, and thus the appropriate sample size, is wedded to the research question(s) and the goal(s) of the study.

Kelly and Maxwell (2008) provided a scheme for sample size planning. Their scheme consists of a two-by-two table. One dimension of the table concerns whether the approach to planning sample size is statistical power or AIPE. The other dimension concerns whether the effect of interest is omnibus (relates to the overall model; e.g., the proportion of variance accounted for by group membership) or targeted (relates to a specific well-defined part of the model; e.g., the contrast between a treatment group and control group). Figure 11.1 provides a modified version of the conceptualization presented in Kelley and Maxwell.

It is important to realize that the same research project can pursue questions that fall into more than one of the cells in Figure 11.1. Thus, a researcher could have a goal of obtaining an accurate parameter estimate for a targeted effect, whereupon AIPE for the targeted effect should be used, and also an additional goal of achieving statistical significance for the omnibus effect, whereupon statistical power for the omnibus effect should be used. Even when multiple goals exist, the researcher has to choose a single sample size for the study. We recommend that the largest needed sample sizes be used so that the multiple goals will each individually have at least an appropriate sample size. Jiroutek, Muller, Kupper, and Stewart (2003) discussed methods where multiple goals can be satisfied simultaneously with a specified probability.

We begin this chapter with a discussion of effect sizes and their role in research. We then discuss the interpretation of results from the NHST and the magnitude-estimation frameworks. Sample size planning to determine the existence of an effect and sample size planning to estimate the magnitude of the size of the effect are discussed. We then consider appropriate ways to specify the parameter(s)

1Of course, in research there are many other things to consider when designing a study besides sample size, such as the available resources. Nevertheless, because of the direct impact that sample size has on statistical power and accuracy in parameter estimation, sample size consideration is a necessary, but not sufficient, component of good study design.
necessary to begin a formal process of sample size
planning.

EFFECT SIZES AND THEIR ROLE IN RESEARCH

Effect sizes come in many forms with many purposes. Preacher and Kelley (2011) give a general definition of effect size as “any measure that reflects a quantity of interest, either in an absolute sense or compared with some specified value” (p. 95; for more information, see Volume 3, Chapter 6, this handbook). As they note, the quantity of interest might refer to such measures as variability, association, difference, odds, rate, duration, discrepancy, proportionality, superiority, and degree of fit or misfit, among other measures that reflect quantities of interest. Effect size estimation is the process of estimating the value of an effect size in the population (even if the population value is zero).

Omnibus and Focused Effect Sizes

An effect size of interest can be one that is omnibus or targeted. Consider multiple regression, a commonly used statistical method in psychology and related disciplines. A basic application of multiple regression considers both (a) the overall effectiveness of the model (i.e., an omnibus effect) via the squared multiple correlation coefficient and (b) the specific relationships linking each regressor (e.g., predictor, explanatory variable, independent variable) to the outcome or criterion variable (i.e., a targeted effect), while controlling for the other regressor(s) in the model, via the estimated regression coefficients. An
example of other omnibus and targeted effect sizes are the root-mean-square error of approximation (RMSEA) in structural equation modeling, an omnibus effect size that quantifies the overall goodness of fit of the model, whereas path coefficients are targeted effect sizes that quantify the strength of the relationship between two variables. As another example, in fixed-effect one-way analysis of variance, eta-squared is an omnibus effect size that quantifies the proportion of the total variance accounted for by group status, whereas contrasts are targeted effects that quantify the difference between some specified linear combination of means. Similarly, Cramér’s V is an omnibus effect size that quantifies the overall association among the rows and columns in a contingency table, whereas the odds ratio is a targeted effect size that quantifies the odds for a certain segment of the table to the odds for another segment of the table for the two-by-two subtable of interest. These examples of omnibus and targeted effect sizes are not exhaustive, but rather illustrate how some effect size measures quantify omnibus effects whereas other effect sizes quantify targeted effects and show how both types of effect sizes can be considered within the same study.2

Unstandardized and Standardized Effect Sizes

Effect sizes can be unstandardized or standardized. An unstandardized effect size is one in which a linear transformation of the scaling of the variable(s) changes the value of the effect size. For a standardized effect sizes, linear transformations of the scaling of the variable(s) do not affect the value of the calculated effect size. For example, suppose there are five regressors in a multiple regression model. The squared multiple correlation coefficient remains exactly the same regardless of whether the variables in the model are based on the original units or standardized variables (i.e., z-scores). The value of the regression coefficients, however, generally will differ depending on whether variables are standardized or unstandardized.

A two-by-two scheme for types of effect sizes can be constructed by considering whether effect sizes are unstandardized or standardized on one dimension and omnibus or targeted on another dimension. Often, multiple effect sizes are of interest in a single study. Nevertheless, what is important is explicitly linking the type of effect size(s) to the research question(s). Without a clear linkage of the type of effect size to the research question, planning sample size is an ill-defined task; appropriate sample size depends heavily on the particular effect size and the goal(s) of the study. To design a study appropriately, the primary effect size that will be used to answer the research question needs to be identified and clearly articulated.

INTERPRETATION OF RESULTS

As discussed, there are two overarching ways to make inferences about population quantities on the basis of sample data. The first way is NHST, which attempts to reject a specified null value. The second way is magnitude estimation, which attempts to learn the size of the effect in the population on the basis of an estimated effect size and its corresponding confidence interval. We review these two approaches to inference before discussing how sample size planning can be based on one of the two approaches to inference, or possibly a combination of the two.

Null Hypothesis Significance Testing

The rationale of null hypothesis significance testing is to specify a null hypothesis, often that the value of the population effect size of interest is zero, and then to test whether the data obtained are plausible given that the null hypothesis is true. If the results obtained are sufficiently unlikely under the null hypothesis, then the null hypothesis is rejected. “Sufficiently unlikely” is operationalized as the p value from the test of the null hypothesis being less than the specified Type I error rate (e.g., .05).3 The value of the null hypothesis depends on the situation and the test may concern an omnibus or

---

2To complicate matters somewhat, effect sizes can be partially omnibus or partially targeted. For example, Cramér’s V can be applied to a subtable larger than a two-by-two table from a larger contingency table. In such a situation, the effect size represents a semitargeted (or semionmnibus) effect size. We do not consider such semitargeted effect sizes here, as they are not used often and are rarely of primary interest for the outcome of a study.

3Recall that the technical meaning of a p-value calculated in the context of an NHST is the probability, given that the null hypothesis is true, of obtaining results as or more extreme than those obtained.
targeted effect size that is either unstandardized or standardized. Regardless of the specific circumstance, the logic of a NHST is exactly the same.

The null hypothesis significance testing framework has been criticized on many occasions (e.g., Bakan, 1966; Cohen, 1994; Meehl, 1967; RozÃ©boom, 1960; for reviews, see also Harlow, Mulaik, & Steiger, 1997; Morrison & Henkel, 1970; Nickerson, 2000). It is not our intent to criticize or defend this framework. Nevertheless, by requiring statistical significance to infer that there is an effect and in certain cases to be able to infer the direction of an effect requires that a null hypothesis be specified and tested, generally in the formal NHST framework. Such a framework is a useful guide for evaluating research results so that there is assurance that what was observed in the sample is consistent with what is true in the population.

In some cases, the basic research question involves directionality. For example, the questions might be as follows: (a) Is the population correlation coefficient positive or negative? (b) Does the treatment group have a larger population mean than the control group? (c) Is the population effect of a particular variable positive or negative after controlling for a set of variables? Inference for directionality usually is meaningful only for clearly defined targeted research questions. For example, consider a fixed-effects one-way analysis of variance with more than two groups. Knowledge that there is a statistically significant F-test for the null hypothesis that all group means are equal does not signify how the groups are different from one another. This phenomenon is true for many types of situations in which an NHST is used for an omnibus effect size: Rejection of an omnibus null hypothesis does not convey which one or which combination of the multiple target comparisons is responsible for the null hypothesis rejection. One group might have a different mean from the remaining groups with equal means, or, all of the groups might have group means that differ from one another. For a targeted test, however, such as the difference between two group means, rejection of the null hypothesis clearly indicates the direction of the difference.

### Effect Sizes and Confidence Intervals

The rationale of confidence interval formation for population parameters comes from the realization that a point estimate almost certainly differs from its corresponding population value. Assuming that the correct model is used, 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. The meaning of a $(1 - \alpha)$100% confidence interval for some unknown parameter was summarized by Hahn and Meeker (1991) as follows: “if one repeatedly calculates such intervals from many independent random samples, $(1 - \alpha)$100% of the intervals would, in the long run, correctly bracket the true value of [the parameter of interest]” (p. 31). Because the values contained within the $(1 - \alpha)$100% confidence interval limits are those that cannot be rejected with the corresponding NHST at a Type I error rate of $\alpha$, they often are regarded as being plausible values of the parameter. The values outside of the $(1 - \alpha)$100% confidence limits can be rejected with the corresponding significance test at a Type I error rate of $\alpha$ and often are regarded as implausible values of the parameter.

When a confidence interval is wide, which is a relative statement made on the basis of a particular context, there is much uncertainty about the size of the population value. As Krantz (1999) stated, “a wide confidence interval reveals ignorance [about the population effect size] and thus helps dampen optimism about replicability” (p. 1374). All other things being equal, a narrower confidence interval is preferred to a wider confidence interval when interest concerns the magnitude of the population effect size. As the confidence interval width decreases, such as when sample size increases or when sampling variability decreases, more values are excluded from the confidence interval.

In some situations the confidence interval need not be exceedingly narrow for the confidence interval to be useful. A confidence interval width can be

---

*Confidence intervals can be used to evaluate NHST, too, in the sense that a reject or fail-to-reject conclusion is drawn on the basis of whether a specified null hypothesis is contained within the confidence interval. We believe the real strength of using confidence intervals is when estimating the magnitude of an effect size. We momentarily discuss confidence intervals but want to point out that confidence intervals generally can be used as a substitute for NHST.*
relatively wide but still exclude parameter values that would support an alternative theory or demonstrate practical significance. For example, if the goal is only to reject a false null hypothesis but not necessarily to bracket the population value with a narrow confidence interval, the confidence interval can be wide but still exclude the specified null value.

The Relationship Between Hypothesis Testing and Confidence Intervals

There is a well-defined relationship between the rejection of a particular NHST and the corresponding upper and lower limits of the confidence interval. If the value corresponding to the null hypothesis falls within the $(1 - \alpha)100\%$ confidence interval limits, that particular null hypothesis value would not be rejected by the corresponding NHST at a Type I error rate of $\alpha$. The converse is also true: All values outside of the $(1 - \alpha)100\%$ confidence interval limits would be rejected using a Type I error rate of $\alpha$ if any of those values were used as the value of the null hypothesis. Thus, there is a one-to-one relationship between NHST framework using a Type I error rate of $\alpha$ and inference based on the corresponding $(1 - \alpha)100\%$ confidence interval. For example, a confidence interval for an estimated standardized mean difference of 0.25 with equal sample sizes across two groups of 100 participants each has a 95% confidence interval of $[-.03, .53]$. Zero, a common value of the null hypothesis for the difference between two groups, is contained within the 95% confidence interval, which implies that the null hypothesis of a population standardized mean difference of zero cannot be rejected.

The existence of this relationship between hypothesis testing and confidence intervals has led some authors to suggest that confidence intervals can supplant NHSTs because the confidence interval immediately reveals whether any specific hypothesized value can be rejected. Other authors have maintained that the NHST provides information not available in a confidence interval, such as a $p$-value.

Regardless of how one views this debate, we will show that both the NHST and AIPE approaches are necessary for sample size planning. The AIPE approach focuses solely on ensuring that an interval will be sufficiently narrow and gives secondary attention to where the center of the interval is likely to be. This is perfectly appropriate when the goal is to obtain an accurate estimate of the parameter. Conversely, power analysis for NHST requires specifying a hypothesized parameter value corresponding to the alternative hypothesis. This generally can be thought of as specifying the expected value of the center of the confidence interval. Sample size planning from the NHST perspective ensures that the confidence interval will not contain the value specified by the null hypothesis with the specified probability. This is fundamentally different from AIPE for estimating the magnitude of an effect with a narrow confidence intervals.

Although a relationship exists between confidence intervals and NHST, each offers its own advantages and each approach dictates its own sample size-planning procedures. Effect size estimates and their accompanying confidence intervals are good at placing boundaries on the magnitude of the corresponding population effect sizes. NHSTs are good at detecting whether effects exist and in certain cases the direction of such effects. Depending on the question of interest, one framework may be more useful than the other, but both should often be used in the same study.

METHODS OF SAMPLE SIZE PLANNING

We have discussed NHST, effect sizes, types of effect sizes, confidence interval formation, and the relationship between NHST and confidence interval

---

5 We say NHST is useful in some cases for detecting direction because some statistical tests allow directionality to be determined, whereas other types of statistical tests only test if an effect exists. For example, when $t$ tests are used to test whether a difference between two group means exists, because the $t$-distribution has only a single degree of freedom parameter, it is known in the sample which group has the larger mean and thus the directionality in the population (e.g., $\mu_1 > \mu_2$) is probabilistically established. If the $F$-distribution is used to assess whether three or more means are all equal in the population, a statistically significant $F$-statistic identifies only that a difference among the means exists, but says nothing about where the difference(s) exists. The $F$-distribution has two degrees of freedom parameters and tests an omnibus hypothesis (except in the special case in which the numerator degrees of freedom equal one, such as when only two groups are involved). A twist on this can be found in Jones and Tukey (2000), in which they discussed a three-outcome framework for an NHST: (a) reject the null hypothesis in favor of a negative effect, (b) reject the null hypothesis in favor of a positive effect, or (c) fail to reject the null hypothesis. The real twist is that the direction of an alternative hypothesis is not prespecified but is dictated from the sample data. This approach makes the presumption that the null hypothesis cannot literally be true, which in and of itself can be considered a controversial statement.
formation. We have discussed these topics extensively because designing a study with an appropriate sample size depends on these considerations, in addition to other issues. In particular, the approach taken to sample size planning and the planned sample size depends on how these factors come together to answer the research question(s). Without careful consideration of these issues, answering one of the most important questions in planning research, namely, “What sample size do I need?” cannot adequately be answered.

This brings us to two methods of sample size planning. Simply put, when interest concerns the existence or the direction of an effect, a power analysis should be performed to plan an appropriate sample size. If interest is in estimating the magnitude of an effect size, the AIPE approach should be used to plan an appropriate sample size. Recall Figure 11.1, which makes clear the distinction between the goal of establishing the existence of an effect and establishing the magnitude of an effect.

**Determining Existence or Direction: Statistical Power and Power Analysis**

Statistical power is the probability of correctly rejecting the null hypothesis. Statistical power is based on four quantities: (a) the size of the effect, (b) the model error variance, (c) the Type I error rate (i.e., $\alpha$), and (d) the sample size. Often, the size of the effect and the model error variance are combined into a standardized effect size. The Type I error rate is a design factor known a priori. In many cases $\alpha = .05$, which is essentially the standard value used in psychology and related disciplines, unless there are justified reasons for some other value to be used. After the Type I error rate is specified and a particular value is chosen for the hypothesized value of the standardized effect size in the population (or effect size and the error variance), statistical power depends only on sample size. Sample size can be planned to obtain a desired level of statistical power on the basis of the specified conditions.

When testing a null hypothesis, the sampling distribution of the effect size of interest is transformed to a test statistic, which is a random variable that, provided the null hypothesis is true, follows a particular statistical distribution (e.g., a central $t$, $\chi^2$, $F$).

When the null hypothesis is false, however, the test statistic follows a different distribution, namely, the noncentral version of the statistical distribution (e.g., a noncentral $t$, $\chi^2$, $F$). The noncentral version of a statistical distribution has a different mean, skewness, and variance, among other properties, as compared with its central distribution analog. Whereas $100\%$ of the sampling distribution under the null hypothesis is beyond the critical value(s), the noncentral distribution has a larger proportion of its distribution beyond the critical value(s) from the central distribution. If the effect size actually came from a distribution in which case the null hypothesis is false, then the probability will be higher of rejecting the null hypothesis than the specified value of $\alpha$.

It is advantageous to have a sufficiently large area, which translates into a high probability, of the noncentral distribution beyond a critical value under the null distribution. The area of the noncentral distribution beyond a critical value of the central distribution can be quantified and is termed statistical power. Holding everything else constant, increases in sample size will lead to a larger area (i.e., higher probability) of the noncentral distribution being beyond the critical value from the central distribution. This happens for two reasons, specifically because (a) a larger sample size decreases variability of the null and alternative distributions and (b) larger sample sizes lead to a larger noncentrality parameter, both of which magnify the difference between the null and alternative distributions.

Figure 11.2 shows the null (central) and alternative (noncentral) distributions in the situation in which there are four groups in a fixed effects analysis of variance design, where the null hypothesis is that the four groups have equal population means. In this situation the supposed population means

---

6 In cases in which directionality is sensible, such as for a $t$ test of the difference between means, in addition to specifying only the Type I error rate, the type of alternative hypothesis (e.g., directional or nondirectional) also must be specified. In other situations for tests that are inherently one tailed, such as for an analysis of variance, such a distinction is unnecessary.

7 One place in which a Type I error rate other than .05 might be used is in the multiple comparison context, where, for example, a Bonferroni correction is used (e.g., .05/4 = .0125 if there were four contrasts to be performed).
are \([-0.5, 0, 0, 0.5]\) for a common within group standard deviation of 1. The effect size, \(f\), is then .35 (see Cohen, 1988, section 8.2, for information on \(f\)). To have power of .80, it is necessary to have a sample size of 23 per group (92 total). Figure 11.2 displays this situation, where the alternative distribution has a noncentrality parameter of 11.5. As can be seen in Figure 11.2, the mean of the noncentral \(F\) distribution (i.e., \(H_a\)) is much greater than it is for the central \(F\) distribution (i.e., \(H_0\)). The larger the noncentral parameter, the larger the mean of the alternative distribution, implying that, holding everything else constant, increases in the noncentral parameter will increase statistical power. The effect of this is that as the noncentral parameter becomes larger and larger (e.g., from increasing sample size), the bulk of the alternative distribution is shifted farther and farther to the right. At the same time, the null distribution is unchanged, which implies that a smaller area of the alternative distribution is less than the critical value from the null distribution (i.e., the probability of a Type II error is decreased, where a Type II error is the failure to reject a null hypothesis when the null hypothesis is false).

It is often recommended that statistical power should be .80, that is, there should be an 80% chance that a false null hypothesis will be rejected. Although we do not think anything is wrong with a power of .80, we want to make it clear that when \(\alpha\) is set to the de facto standard value of .05 and power is set at .80, the probability of a Type II error (1 − .80 = .20) is four times larger than the probability of a Type I error. Whether this is good is determined on the basis of a trade-off between false positives (Type I errors) and false negatives (Type II errors). The most appropriate way to proceed in such a situation is to consider the consequences of a Type I versus a Type II error. If an effect exists and it is not found (i.e., a Type II error has been committed), what are the implications? In some research settings, there may be little if any consequences. If
finding the effect would have been used to advance a major initiative that would greatly benefit society or a specific population, a Type II error could have major implications. In such situations, the Type II error can be regarded as more important than a Type I error, and consistent with the importance, setting the probability of a Type II error to be more like, or even less than, the probability of a Type I error would be desirable.

**Determining the Magnitude of an Effect: Accuracy in Parameter Estimation**

In general, when an effect size is of interest, the corresponding confidence interval for that effect should be of interest as well. The estimated effect size from a sample almost certainly differs from its corresponding population value. The ultimate interest in research usually is the population value of an effect, rather than an estimate from some particular sample that necessarily contains sampling error. Thus, in addition to reporting the estimated effect size (i.e., the best single point estimate of the population parameter), researchers should report the confidence interval limits for the population value.

Confidence intervals are important in part because of a belief in the “law of small numbers,” which states that people often employ the flawed intuition that there is a lack of variability of estimates from samples (Tversky & Kahneman, 1971). Tversky and Kahneman suggested that confidence intervals be reported because they provide “a useful index of sampling variability, and it is precisely this variability that we tend to underestimate” (1971, p. 110). Kelley (2005) has noted that reporting any point estimate “in the absence of a confidence interval arguably does a disservice to those otherwise interested in the phenomenon under study” (p. 52). Similarly, Bonett (2008) has argued that “the current practice of reporting only a point estimate of an effect size is actually a step backwards from the goal of improving the quality of psychological research” (p. 99; see also Thompson, 2002). The American Psychological Association (APA; 2010) *Publication Manual* clearly states the importance of effect sizes and confidence intervals: “It is almost always necessary to include some measure of effect size in the Results section,” and “whenever possible, provide a confidence interval for each effect size reported to indicate the precision of estimation of the effect size” (p. 34). Psychology and related disciplines have emphasized including effect sizes and their confidence intervals as an alternative to or in addition to NHSTs (e.g., see Grissom & Kim, 2005; Harlow, Mulaik, & Steiger, 1997; Hunter & Schmidt, 2004; Schmidt, 1996; Thompson, 2002; Wilkinson & the APA Task Force on Statistical Inference, 1999).

A wide confidence interval reflects the uncertainty with which a parameter has been estimated. Historically, confidence intervals were seldom reported in psychology and related disciplines. Although not routinely reported now, their use seems to have increased in at least some journals. Cumming et al. (2007) have shown that for 10 leading international psychology journals, the use of confidence intervals increased from 3.7% in 1989 to 10.6% in 2005–2006. This is not ideal, but it is a start. Cohen (1994) once suggested that the reason more researchers did not report confidence intervals was that their widths were often “embarrassingly large” (p. 1002). Wide confidence intervals often can be avoided with proper sample size planning from an AIPE approach, which differs from traditional methods of sample size planning whose goal is to obtain adequate statistical power. It may seem that because there is a one-to-one relationship with regards to rejecting the null hypothesis and the null value being outside of the confidence interval, there might be a one-to-one relationship between sample size planning for statistical power and sample size planning for AIPE. This is not the case, however. In particular, the goals of AIPE are satisfied even if the null value is contained within the confidence interval as long as the interval width is sufficiently narrow. The goals of power analysis can be satisfied even with a wide confidence interval as long as the hypothesized value is not contained within the interval. Although for any confidence interval it is known whether the null hypothesis value can be rejected, for sample size–planning purposes, the goals are fundamentally different and approached in an entirely different way. When magnitude is of interest, the computed confidence interval should be narrow, as that signifies less uncertainty about the
value of the population parameter. So, the question confronting a researcher interested in magnitude estimation is how to choose a sample size that will provide a sufficiently narrow confidence interval.

The approach to sample size planning when the goal is to obtain narrow confidence intervals has been termed AIPE. The goal of the AIPE approach is to plan sample size so that the confidence interval for the parameter of interest will be sufficiently narrow, where sufficiently narrow is necessarily context specific. The AIPE approach to sample size planning has taken on an important role in the research design literature, as it is known that researchers (a) tend to overestimate how precise an estimate is, (b) prefer not to have wide confidence intervals, and (c) are expected to report effect sizes and their corresponding confidence intervals. This push in psychology (e.g., Journal Article Reporting Standards [JARS]; APA, 2010) and education (e.g., American Educational Research Association, 2006) to report effect sizes and confidence intervals is consistent with medical research, where the Consolidated Standard of Reporting Trials (CONSORT; Moher et al., 2010) and the Transparent Reporting of Evaluations with Nonrandomized Designs (TREND; Des Jarlais, Lyles, Crepaz, & the TREND Group, 2004) both state that effect sizes and confidence intervals should be reported for primary and secondary outcomes (see item 17 in both checklists). Because of the need for confidence intervals that are not “embarrassingly wide” (Cohen, 1994, p. 1002), we believe that the AIPE approach to sample size planning will play an integral role in research in the coming years.

Confidence interval width is in part a function of sample size. To understand why sample size planning for a narrow confidence interval is termed accuracy in parameter estimation, it is helpful to consider the definition of accuracy. In statistics, accuracy is defined as the square root of the mean square error (RMSE) for estimating some parameter $\theta$ of interest, which is formally defined as follows:

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

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

\[
= \sqrt{\sigma^2 + B^2}, 
\]

where $\sigma^2$ is the variance of the estimated parameter, which is inversely proportional to the precision of the estimator and $B^2$ is the squared bias of the estimator. From Equation 1c, it can readily be seen for a fixed $\sigma^2$, an increase in $B^2$ yields a less accurate estimate (i.e., larger RMSE), with the converse also being true. Because we seek to obtain an accurate estimate, we must consider precision and bias simultaneously. In general, the AIPE approach uses estimates that are unbiased, or at least consistent (i.e., that converge to their population value as sample size increases). It would be entirely possible to have a precise estimate that was biased. For example, regardless of the data, a researcher could estimate a parameter on the basis of a theory-implied value. Doing so would not be statistically optimal, in general, because the theory-implied value is unlikely to be correct, but the estimate would be highly precise (e.g., its variance could be zero). Because AIPE simultaneously considers precision and bias when estimating sample size, the approach is most appropriately referred to as accuracy in parameter estimation.

Just as the Type I error rate usually is fixed at .05, as previously discussed for statistical power, the confidence level is essentially a fixed design factor, which almost always is set to .95. With the level of confidence fixed and with estimates for the model error variance and in some situations the size of the effect considered fixed, sample size is a factor that can be planned so that the expected confidence interval width is sufficiently narrow. Because the confidence interval width is itself a random variable, even if the expected confidence interval width is sufficiently narrow, any particular realization almost certainly will be either narrower or wider than desired. An optional specification allows a researcher to have some specified degree of assurance (e.g., 99%) that the obtained confidence interval will be sufficiently narrow. That is, a modification to the standard procedure, where only the expected width is sufficiently narrow, answers the question, “What size sample is necessary so that there is 99% assurance that the 95% confidence interval is sufficiently narrow?” What is “sufficiently narrow” necessarily depends on the particular context. Most important, the confidence interval will bracket the population value with the specified level
of confidence, which implies that the estimated
effect size on which the confidence interval was
based, or its unbiased or more unbiased version, will
be contained within a smaller range of plausible
parameter values. Because of the narrow range of
plausible parameter values and the estimate of
choice being within that narrow range, the estimate
is one that is more accurate than one with a confi-
dence interval wider than that observed, all other
things being equal.

Parameter Specification When
Planning Sample Size
One or more nondesign factors must be specified in
all of the sample size planning procedures com-
monly used. The values that must be specified to
plan the sample size depend on the statistics of
interest and the type of sample size planning pro-
cedure. For example, planning sample size for statisti-
cal power for an unstandardized mean difference
between two groups requires that a mean difference
and the standard deviation must be separately speci-
fied. Alternatively, a standardized mean difference
can be specified. For an unstandardized mean dif-
fERENCE, from the AIPE approach, only the error
variance (or standard deviation) must be specified,
as the confidence interval width is independent of
the size of the mean difference. Choosing relevant
values to use for the approaches is often difficult
because the many times the values are unknown.
The values used, however, can have a sizable impact
on the resulting sample size calculation. For this
reason, the difficulty in estimating an effect size to
use for purposes of sample size planning has been
labeled the “problematic parameter” (Lipsey, 1990,
p. 47).

The existing literature generally should be
the guiding sources of information for choosing
parameter values when planning sample size. Some
areas have an ample body of literature that exists
and perhaps meta-analyses that estimate the effect
size of interest. When such information is available,
it should be used. When such information is not
readily available, planning sample size often is diffi-
cult. Guidelines that seem to make the process eas-
ier generally have been shown to be inappropriate
(e.g., Green, 1991; MacCallum, Widaman, Zhang, &
Hong, 1999; Maxwell, 2004), largely because they
tend to ignore one or more of the following: (a) the
question of interest, (b) the effect size, (c) character-
istics of the model to be used, (d) characteristics of
the population of interest, (e) the research design,
(f) the measurement procedure(s), (g) the failure to
distinguish between targeted and omnibus effects,
and (h) the desired level of statistical power or the
width of the desired confidence interval. In the
remainder of this section, we discuss the specifica-
tion of parameters for power analysis and accuracy
in parameter estimation.

For power analysis, there are two common
approaches for choosing parameter values. The first
approach uses the best estimates available for the
necessary population parameter(s), which usually is
made on the basis of a review of the literature (e.g.,
via meta-analytic methods) or data from a pilot
study. The second approach uses the minimum
parameter value of interest (MPVI) by linking the
size of an effect to its meaningfulness for a theory or
application. In using the MPVI, there will be suffi-
cient power if the magnitude of the effect size is as
large as specified and even more power if the effect
size is larger in the population.

Both the literature review and MPVI approaches
are potentially useful when planning sample size,
but the appropriate choice will depend on the infor-
mation available and the research goals. If the goal is
to plan sample size for what is believed true in the
population, then the literature review approach gen-
erally should be used. If the goal is to plan sample
size on the basis of the minimum value of the effect

---

To better understand how sample size influences the width of a confidence interval, consider the case of normally distributed data when the
population standard deviation, \( \sigma \), is known and a 95% two-sided confidence interval is of interest. Such a confidence interval can be expressed as
\( \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} \), where \( \bar{x} \) is the sample mean and \( n \) is the sample size. The estimated mean, \( \bar{x} \), is necessarily contained within the confidence interval
limits. As the sample size increases, the margin of error (i.e., \( 1.96 \frac{\sigma}{\sqrt{n}} \)) decreases. For example, suppose \( \sigma = 15 \). For a sample size of \( n = 10 \), the mar-
gin of error is 9.30, whereas the margin of error is 2.94 when \( n = 100 \). Because the range of plausible parameter values is 5.88 for the later case, it is
preferred, holding everything else constant, to the former case with a confidence interval width of 18.60. A similar relationship holds for confidence
intervals for other quantities.

We use standardized effect size here so that both the unstandardized effect size and the model error variance are considered simultaneously.
size that is of scientific or practical importance and interest, then the MPVI approach generally should be used.

As an example of using the MPVI, suppose that only a standardized mean difference of 0.20 in magnitude or larger would be of interest in a particular setting. Choosing this value for the MPVI essentially implies that any value of the population standardized mean difference less than 0.20 in magnitude is of little to no value in the particular research context. If this is the case, then the parameter value used for the sample size planning procedure need not be literally the true but unknown population effect size. Instead, it can be the minimum parameter value that is of theoretical or practical interest. If the population value of the effect size is larger than the chosen MPVI, then statistical power will be greater than the nominal level specified. Senn (2002) addressed this idea by saying,

> The difference you are seeking is not the same as the difference you expect to find, and again you do not have to know what the treatment will do to find a figure. This is common to all science. An astronomer does not know the magnitude of new stars until he [or she] has found them, but the magnitude of star he [or she] is looking for determines how much he [or she] has to spend on a telescope. (p. 1304)

Although using the MPVI to plan sample size can be useful in some situations, O’Brien and Castelloe (2007) illustrated that such a method potentially can be misguided, specifically in the power analytic approach. But, the issue is also relevant to some applications of the AIPE approach. O’Brien and Castelloe highlighted this potential problem with the MPVI approach using a hypothetical example in the context of lowering the mortality rate of children with malaria, which is estimated to be 15% for “usual care.” They noted that reasonable people would tend to agree that a reduction in the rate of mortality by 5% of the usual care (which would lower the mortality rate of 15% by 5% to 14.25%) would be “clinically relevant” (2007, pp. 244–245). In their scenario, however, the necessary sample size increases from 2,700 patients to detect a 33% reduction in mortality rate (which would lower the mortality rate of 15% by 33% to 10%) to 104,700 patients to detect a 5% reduction. This is too large of a study to conduct in their specific context. With such a large number of patients required to detect the MPVI value, the study might not be conducted because resources may not be available for 104,700 patients. So, although a 5% reduction in mortality rate may be the MPVI, the sample size is such that the study might never be conducted. If a 33% reduction is a reasonable value for the size of the effect in the population, it almost certainly would be preferable to carry out the study with only 2,700 patients. In so doing, the study actually may be conducted because resources may be available for only 2,700 patients. O’Brien and Castelloe (2007) argued that their hypothetical scenario “exemplifies why confirmatory trials are usually designed to detect plausible outcome differences that are considerably larger than ‘clinically relevant’” (p. 248). Thus, although the MPVI approach can be useful in certain situations, it is not necessarily universally preferred as a way to plan sample size from the power analytic perspective.

For power analysis, holding everything else constant, the larger the magnitude of the effect size, the more statistical power a study has to detect the effect. This is not necessarily the case with the AIPE approach, however. For example, Kelley and Rausch (2006) have showed that, holding everything else constant, confidence intervals for a sufficiently narrow expected width will require a larger sample size for larger population standardized mean differences than for smaller differences. That is, for a given expected confidence interval width (e.g., say 0.25 units), the sample size required for a population standardized mean of 0.05 requires a smaller sample size (347 per group) than does a population standardized mean difference of 1.00 (390 per group).10

---

10For fixed sample sizes, the larger the population standardized mean difference, the larger the noncentrality parameter, which implies a larger variance for the sampling distribution of observed standardized mean differences when the population standardized mean difference is larger (see Hedges & Olkin, 1985, for more details). Thus, the larger the population standardized mean difference, the larger the sample size will need to be for the same confidence interval width.
The relationship between necessary sample size for a particular expected confidence interval width and the size of the population standardized mean difference is monotonic in the AIPE approach—the larger the population standardized mean difference the larger the necessary sample size, holding everything else constant. For narrow confidence intervals, however, for the population squared multiple correlation coefficient, the relationship between the size of the population multiple correlation coefficient and the necessary sample size for a specified width is nonmonotonic—necessary sample size for an expected confidence interval width is maximized around a population squared multiple correlation value of around .333 (Kelley, 2008). In other situations, the size of the effect is unrelated to the confidence interval width. Kelley, Maxwell, and Rausch (2003) showed that for an unstandardized mean difference, desired confidence interval width is independent of the population difference between means. Correspondingly, sample size from the AIPE approach does not consider the size of the unstandardized mean difference, and thus it is not necessary to specify the mean difference when planning sample size from the accuracy in parameter estimation approach, unlike the power analytic approach where it is necessary. The point is that the relationship between the size of the effect and the necessary sample size from the AIPE approach, holding everything else constant, depends on the particular effect size measure.

To summarize, for power analysis, the larger the magnitude of an effect size, the smaller the necessary sample size for a particular level of statistical power, holding all other factors constant. For AIPE, however, such a universal statement cannot be made. This is the case because the way in which the value of an effect size is used in the confidence interval procedure depends on the particular effect size measure. There are two broad categories of effect sizes as they relate to their corresponding confidence intervals: (a) those whose confidence intervals depend on the value of the effect size and (b) those whose confidence intervals are independent of the value of the effect size. The literature on confidence intervals is a guide for the way in which confidence intervals are formed for the population value of an effect size (e.g., Grissom & Kim, 2005; Smithson, 2003). In the next section, we implement some of the ideas discussed in the context of multiple regression. In particular, we provide an example in which sample size is planned using the MBESS (Kelley, 2007b; Kelley & Lai, 2011) package in the program R (R Development Core Team, 2011) for a targeted (regression coefficient) and an omnibus (the squared multiple correlation coefficient) effect size from both the power analytic and AIPE perspectives.

Sample Size Planning for Multiple Regression: An Organizational Behavior Example

Core self-evaluations can be described as fundamental, subconscious conclusions that individuals reach about themselves, their relationships to others, and the world around them (e.g., Judge & Bono, 2001; Judge, Locke, Durham, & Kluger, 1998). Judge et al. (1997) argued from a theoretical perspective that individuals’ appraisals of objects are affected by their assumptions about themselves in addition to the attributes of the objects and their desires with respect to the objects. Judge et al. (1998) found empirical support for core self-evaluations having a consistent effect on job satisfaction, with core self-evaluations largely being unrelated to the attributes of the job.

Grant and Wrzesniewski (2010) noted that “little research has considered how core self-evaluations may interact with other kinds of individual differences to moderate the relationship between core self-evaluations and job performance” (p. 117). Grant and Wrzesniewski tested a model in which a measure of Performance (financial productivity over a 3-week period) was modeled as a linear function of Core Self-Evaluations, Duty, Anticipated Guilt, Anticipated Gratitude, and the three two-way interactions of Core Self-Evaluations with the remaining three regressors.

Interestingly, in the power analytic approach, the relationship between necessary sample size and the size of the population standardized mean difference is also monotonic, but in the other direction—the larger the population standardized mean difference the smaller the necessary sample size for a particular level of statistical power (e.g., .90), holding everything else constant.
A primary focus of the research was an “other-orientation,” which Grant and Wrzesniewski (2010) described as the “extent to which employees value and experience concern for the well-being of other people” (p. 109; see also De Dreu & Nauta, 2009). The Duty regressor, which quantifies the “tendency toward dependability and feelings of responsibility for others” (p. 111), is an other-oriented attribute and thus a driving force of the research, especially on how it is moderated by Core Self-Evaluation.

To test the model, data were collected from 86 call center employees engaged in an outbound calling campaign raising money for job creation at a university. Descriptive statistics obtained (means, standard deviations, and correlations) or derived (variances and covariances) from Grant and Wrzesniewski (2010) are contained in Table 11.1, with the results from a standardized regression model (i.e., a regression model applied to standardized scores) given in Table 11.2. The results of the regression model show that two regression coefficients are statistically significant, namely, the interaction of Core Self-Evaluation with Duty and the interaction of Core Self-Evaluation with Gratitude. Although Duty was found to have an interaction with Core Self-evaluation on Performance, the conditional effect of Duty when Core Self-Evaluation equals zero (i.e., at the mean) has a 95% confidence interval for the population value that ranges from −0.038 to 0.456. Notice that zero is contained in the confidence interval for the conditional Duty regression coefficient, which is necessarily the case because the corresponding NHST failed to reach statistical significance (the \( p \) value for Duty is .097).

Additionally, the interaction of Anticipated Gratitude and Core Self-Evaluation was also statistically significant. The squared multiple correlation coefficient for the entire model was .201 (\( p \) value is .01) with 95% confidence interval limits for the population value of \([.01, .31]\); the adjusted squared multiple correlation coefficient was .132.

Although the overall model was statistically significant and there were two statistically significant interactions, none of the four conditional effects

---

**TABLE 11.1**

Summary Statistics Taken (Means and Correlations) or Derived (Variances and Covariances)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Performance</td>
<td>3153.70</td>
<td>11923209.000</td>
<td>0.00</td>
<td>0.14</td>
<td>0.03</td>
<td>0.13</td>
<td>0.27</td>
<td>0.31</td>
<td>0.34</td>
</tr>
<tr>
<td>2. Core self-evaluation (CSE)</td>
<td>5.00</td>
<td>0.000</td>
<td>0.608</td>
<td>0.23</td>
<td>0.00</td>
<td>0.15</td>
<td>−0.10</td>
<td>−0.10</td>
<td>0.02</td>
</tr>
<tr>
<td>3. Duty</td>
<td>5.11</td>
<td>430.360</td>
<td>0.160</td>
<td>0.792</td>
<td>0.48</td>
<td>0.40</td>
<td>−0.05</td>
<td>−0.01</td>
<td>0.10</td>
</tr>
<tr>
<td>4. Guilt</td>
<td>3.97</td>
<td>152.318</td>
<td>0.000</td>
<td>0.628</td>
<td>2.160</td>
<td>0.33</td>
<td>−0.01</td>
<td>0.09</td>
<td>0.24</td>
</tr>
<tr>
<td>5. Gratitude</td>
<td>4.29</td>
<td>507.382</td>
<td>0.132</td>
<td>0.402</td>
<td>0.548</td>
<td>1.277</td>
<td>0.10</td>
<td>0.25</td>
<td>0.29</td>
</tr>
<tr>
<td>6. CSE x Duty</td>
<td>0.23</td>
<td>923.236</td>
<td>−0.077</td>
<td>−0.044</td>
<td>−0.015</td>
<td>0.112</td>
<td>0.980</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td>7. CSE x Guilt</td>
<td>0.00</td>
<td>1017.182</td>
<td>−0.074</td>
<td>−0.008</td>
<td>0.126</td>
<td>0.268</td>
<td>0.357</td>
<td>0.903</td>
<td>0.26</td>
</tr>
<tr>
<td>8. CSE x Gratitude</td>
<td>0.15</td>
<td>1139.11</td>
<td>0.151</td>
<td>0.086</td>
<td>0.342</td>
<td>0.318</td>
<td>0.461</td>
<td>0.240</td>
<td>0.941</td>
</tr>
</tbody>
</table>

*Note.* Variances are in the principle diagonal and are italicized, covariances are in the lower diagonal, and correlation coefficients are in the upper diagonal. Sample size is 86. Adapted from “I Won’t Let You Down . . . or Will I? Core Self-Evaluations, Other-Orientations, Anticipated Guilt and Gratitude, and Job Performance,” by A. M. Grant and A. Wrzesniewski, 2010, *Journal of Applied Psychology,* 95, p. 115. Copyright 2010 by the American Psychological Association.

---

12Grant and Wrzesniewski (2010) fit the unstandardized model given in Table 11.2. Confidence intervals for regression coefficients and the squared multiple correlation coefficient were not given in Grant and Wrzesniewski but are provided here. The standardized solution is used here because the standardized regression coefficients quantify the expected change in the dependent variable in standard deviation units for a one standard deviation change in the particular regressor. For example, a standardized regression coefficient of .20 means that a 1 unit (corresponding to 1 standard deviation) change in the regressor will have a .20 unit (corresponding to .20 standard deviation) change on the dependent variable. Standardized regression coefficients have a straightforward interpretation and interest is in relative performance, not literally modeling the total 3-week productivity of employees.
Sample Size Planning

(Duty, Anticipated Gratitude, Core Self-Evaluation, and Gratitude) reached statistical significance. Because of the importance of understanding how other-oriented attributes are associated with Performance as well as understanding how much Performance can be explained by such a model, suppose that a follow-up study based on Grant and Wrzesniewski (2010) needs to be planned. How might sample size be planned in such a context? We briefly review multiple regression so that notation can be defined and used in the sample size planning procedure. All sample sizes are planned with the MBESS R package (R Development Core Team, 2011). We will provide the necessary code to plan sample size, but because of space restrictions, we are unable to provide a detailed explanation of the MBESS functions used for sample size planning. Kelley (2007a), however, has reviewed methods of estimating effect sizes and forming confidence intervals for the corresponding population value, and Kelley and Maxwell (2008) have provided a more thorough treatment of using the MBESS R package for sample size planning in a regression context.

The Multiple Regression Model

The multiple regression model in the population is

$$Y_i = \beta_0 + X_{ik}\beta_k + \epsilon_i,$$

(2)

where $Y_i$ is the dependent variable of interest for the $i$th individual ($i = 1, \ldots, N$), $\beta_0$ is the intercept, $X_{ik}$ is the value of the $k$th ($k = 1, \ldots, K$) regressor variable for the $i$th individual, and $\epsilon_i$ is the error for the $i$th individual (i.e., $\epsilon_i = Y_i - \bar{Y}_i$, where $\bar{Y}_i$ is the model implied value of the dependent variable for the $i$th individual). The $K$ length vector of regression coefficients is obtained from the following expression:

$$\beta = \Sigma^{-1}_{X} \sigma_{XY},$$

(3)

where $\beta$ is the $K$ length vector of regression coefficients, $\Sigma_{XX}$ is the population $K \times K$ covariance matrix of the regressor variables, with a “$^{-1}$” exponent denoting the inverse of the matrix, and $\sigma_{XY}$ is a $K$ length column vector of the population covariances of the dependent variable with each of the regressors. The population intercept is obtained from

$$\beta_0 = \mu_Y - \mu_X^T \beta,$$

(4)

where $\mu_Y$ is the population mean of the dependent variable and $\mu_X$ is the $K$ length vector of population means for the regressors with the prime superscript denoting matrix–vector transposition (where columns are interchanged with rows).

The population squared multiple correlation coefficient is obtained as follows:

$$P_{YY}^2 = \sigma_Y^2 \Sigma^{-1}_{XX} \sigma_{XY},$$

(5)

where $P$ is the uppercase Greek letter rho, with the subscript representing $Y$ being regressed on (the centered dot) the set of regressors of interest.

Equation 5 has assumed population values for an unstandardized regression model. Sample values of the means, variances, and covariances can be substituted for the population values to obtain the usual ordinary least squares regression estimates. Should a standardized regression model be of interest, the means of the outcome variable and the regressors are all zero, thus eliminating the intercept from the model because of its value being zero, and the $K \times K$

---

**TABLE 11.2**

<table>
<thead>
<tr>
<th>Regressors</th>
<th>Coefficient</th>
<th>95% confidence interval limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Core self-evaluation (CSE)</td>
<td>-0.016</td>
<td>[-.224, .192]</td>
</tr>
<tr>
<td>2. Duty</td>
<td>0.210</td>
<td>[-.038, .456]</td>
</tr>
<tr>
<td>3. Guilt</td>
<td>-0.052</td>
<td>[-.285, .182]</td>
</tr>
<tr>
<td>4. Gratitude</td>
<td>0.057</td>
<td>[-.185, .299]</td>
</tr>
<tr>
<td>5. CSE x Duty</td>
<td>0.242</td>
<td>[.016, .467]</td>
</tr>
<tr>
<td>6. CSE x Guilt</td>
<td>0.278*</td>
<td>[.032, .521]</td>
</tr>
<tr>
<td>7. CSE x Gratitude</td>
<td>0.203</td>
<td>[.01, .31]</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>.312</td>
<td></td>
</tr>
</tbody>
</table>

Note. Values in brackets represent the lower and upper 95% confidence interval limits. All confidence intervals are based on noncentral methods because the effect sizes are standardized and are calculated with the MBESS R package. Data from Grant and Wrzesniewski (2010). $^* p < .05.$
correlation matrix of the regressor variables is then substituted for \( \Sigma_{XX} \) and the vector of correlations between the \( K \) regressors and \( Y \) is substituted for \( \sigma_{XY} \). Because a correlation matrix is simply a standardized covariance matrix, Equations 2 through 5 hold for the point estimates from the multiple regression model in standardized form.\(^1\) Using this notation, we will move forward with the examples of how sample size can be planned in the context of multiple regression from both the statistical power and AIPE perspectives, for targeted (the Duty regression coefficient) and omnibus (the squared multiple correlation coefficient) effect sizes.

### The Power Analytic Approach to Sample Size Planning

When interest concerns the existence of an effect, the power analytic approach to sample size planning should be used. The following two subsections illustrate power analysis for a regression coefficient and for the squared multiple correlation coefficient using the MBESS R package.

#### Statistical Power for a Targeted Effect (\( \beta_h \))

Although the regression coefficient for the conditional effect of Duty was not statistically significant in the Grant and Wrzesniewski (2010) study, theory suggests that it is a nonzero positive value. It is possible that the Grant and Wrzesniewski (2010) study did not have sufficient power to detect this specific effect. Evidence that the conditional Duty regression coefficient in the population is a positive value would come from a statistically significant positive regression coefficient, signifying that when the other predictors are at their mean (i.e., zero), Duty has a linear impact on Performance. Using a Type I error rate of .05 and a two-sided NHST, we seek to answer the question, “What size sample is necessary in order to have statistical power of .95 to reject the null hypothesis that the conditional Duty regression coefficient equals zero?” Note that we plan sample size for statistical power of .95 to weight the probability of a Type I and Type II error the same, implying that neither type of error has differential preference.

As is almost always the case with sample size planning, input parameters are necessary to define the alternative distribution. For lack of better information, we will use the sample values from Grant and Wrzesniewski (2010). Planning sample size in this context involves three parameters, which can be obtained from \( \Sigma_{XX} \) and \( \sigma_{XY} \). In particular, the three necessary values are \( \beta_h \), the value of the population regression coefficient of interest, \( \mathbf{P}_{YX}^2 \), the value of the population squared multiple correlation coefficient, and \( \mathbf{P}_{X_kX_h}^2 \) which is the value of the population squared multiple correlation coefficient when the regressor of interest is regressed on the remaining \( K-1 \) regressors. That is, \( \mathbf{P}_{X_kX_h}^2 \) is the squared multiple correlation coefficient for a model where \( X_h \) is the outcome variable and the other \( K-1 \) regressors continue to be used as regressors. \( \mathbf{P}_{X_kX_h}^2 \) can be obtained using the equations given above (if \( X_h \) is substituted for \( Y \) and regressed on the remaining \( K-1 \) regressors) or via the following expression:

\[
P_{X_kX_h}^2 = 1 - \left( \frac{\sigma_{YX}^2}{\sigma_{XX}^2} \right)^{\frac{1}{2}}, \tag{6}
\]

where \( c_{kk} \) is the \( k \)th diagonal element of \( \Sigma_{XX}^{-1} \) (Harris, 2001). In the Grant and Wrzesniewski (2010) study, the estimated \( \mathbf{P}_{X_kX_h}^2 = .34 \).

Given the preceding, the sample size needed so that the conditional Duty regression coefficient will have statistical power of .95 can be planned using the `ss.power.rc()` MBESS function is as follows:

```
ss.power.rc(β_x = .21, P_{YX}^2 = .20, P_{YX}^2 = .34, K = 7, desired.power = .95, α.level = .05, σ_Y = 1, σ_Y = 1)
```

where \( \beta_x \) is the hypothesized population value of the regression coefficient (standardized here), \( P_{YX}^2 \) is the hypothesized population value of the squared multiple correlation coefficient, \( P_{YX}^2 \) is the hypothesized

---

\(^1\)Although multiple regression model in a standardized form can offer interpretational ease, exact confidence intervals for standardized regression coefficients are more difficult than their unstandardized counterparts. A discussion of this is provided in Kelley (2007b). We used software to compute the confidence intervals, which are based on noncentral t-distributions, and not delve into the details of computation here. Additionally, computing confidence intervals for the squared multiple correlation coefficient is also difficult, and we do not discuss the details. Kelley (2008) has reviewed methods of confidence interval formation for the squared multiple correlation coefficient, which is unchanged regardless of an unstandardized or standardized model that is fitted. In practice, the confidence intervals in these situations can be obtained easily with the MBESS software.
population value of the squared multiple correlation coefficient when the regressor of interest is regressed on the remaining \(k-1\) regressors, \(K\) is the number of regressors, \texttt{desired.power} is the desired level of statistical power, \texttt{alpha.level} is the Type I error rate, and \texttt{sigma.X} and \texttt{sigma.Y} are the population standard deviations of the regressor of interest and the dependent variable. Implementation of the function as shown in Equation 7 returns a necessary sample size of 380. Thus, if a researcher wanted to have statistical power of .95 for the test of the conditional Duty regression coefficient, given our assumptions, a sample size of 380 would be necessary.

**Statistical Power for the Omnibus Effect (\(P^2_{YX}\))**

In some applications of multiple regression, no single regressor is of primary importance. Correspondingly, it is sometimes of interest only to have a model that, overall, accounts for a statistically significant portion of the variance of the outcome variable. In particular, of interest is obtaining a statically significant test of the squared multiple correlation coefficient. To plan sample size for statistical power of the squared multiple correlation coefficient, the only population value that must be known or estimated is \(P^2_{YX}\).

The \texttt{ss.power.R2()} MBESS function can be used to plan sample size for a specified value of power for the test of the squared multiple correlation coefficient. The way in which the function is used is as follows:

\[
\texttt{ss.power.R2(Population.R2=.20, alpha.level = 0.05, desired. power = 0.95, K = 7)}
\]

where \texttt{Population.R2} is the hypothesized population value of the squared multiple correlation coefficient, with the other parameters of the function as defined previously. Implementation of the above function returns a necessary sample size of 95. Thus, to have statistical power of .95 that some of the variability in the dependent variable (here Performance) is being accounted for in the population by the set of regressors, where the presumed squared multiple correlation coefficient is .20 with seven regressors and a Type I error rate of .05, a sample size of 95 is necessary.

**THE AIPE APPROACH TO SAMPLE SIZE PLANNING**

When interest concerns the magnitude of an effect size, the AIPE approach to sample size planning should be used.

**AIPE for a Targeted Effect (\(\beta_k\))**

The 95% confidence interval from the Grant and Wrzesniewski (2010) study for the standardized conditional Duty regression coefficient is \([- .038, .456]\). This confidence interval is so wide that it illustrates the lack of knowledge about the population value of the standardized conditional Duty regression coefficient. Correspondingly, it really is not clear how much of an impact, if any, the conditional effect of Duty has on Performance when the other regressors are zero. Although theory suggests that the conditional Duty should have a positive impact on Performance, the magnitude of that impact is important. Suppose there is a desire to have a 95% confidence interval for the conditional Duty regression coefficient that is .20 units wide. The population parameters that must be known or estimated for an application of AIPE to a standardized regression coefficient are \(P^2_{YX}\), \(P^2_{XkX}\), and \(\beta_k\), as in the context of statistical power.\(^{14}\)

The function \texttt{ss.aipe.src()} from the MBESS R package can be used as follows:

\[
\texttt{ss.aipe.src(Rho2.Y_X=.20, Rho2.k_X.without.k=.34, K = 7, beta.k=.21, width=.20, conf. level=.95, assurance = NULL)}
\]

where width is the desired confidence interval with and the other function parameters are as defined in the \texttt{ss.power.rc()} function. Implementation of the above function returns a necessary sample size of 482.

\(^{14}\)The value of the regression coefficient is necessary in this case because we are working with a standardized solution. In the case of an unstandardized solution, however, the confidence interval for the (unstandardized) regression coefficient is independent of the confidence interval. Correspondingly, it is unnecessary to know or estimate an unstandardized regression coefficient in the context of AIPE. See Kelley (2007b) and Kelley and Maxwell (2008) for more details. The power analytic approach, for both unstandardized and standardized regression coefficients, requires the known or estimated population value of the regression coefficient.
Notice that assurance = NULL is specified above, which implies that an assurance parameter is not incorporated into this sample size planning procedure. Thus, approximately half of the time the confidence interval will be wider than the desired value of .20. By specifying some assurance value (a value greater than .50 but less than 1), however, the modified sample size can be obtained. For example, if the desired assurance is .99, the above code can be modified as follows:

```r
ss.aipe.src(Rho2.Y_X=.20,
Rho2.k_X.without.k=.34, K = 7,
beta.k=.21, width=.20, conf.
level=.95, assurance=.99)
```

where assurance is the desired level of assurance and conf.level is the confidence level of interest. Implementation of this function returns a necessary sample size of 528, which then ensure that 99% of confidence intervals formed using this procedure will be no wider than the desired value of .20.

**AIPE for the Omnibus Effect (P^2_{YX})**

The 95% confidence interval from the Grant and Wrzesniewski (2010) study for the population squared multiple correlation coefficient was [.01, .31]. Such a wide range for the plausible values of the population squared multiple correlation coefficient is undesirable, as the proportion of variance that is accounted for the model may be close to 0 or close to 0.33, corresponding with very different interpretations of how well the seven regressors are able to account for the variance of Performance in the population.

Suppose one desires a width of .15 for the confidence interval for the population squared multiple correlation coefficient. The only population parameter that must be presumed or estimated for an application of AIPE to the squared multiple correlation coefficient is $P^2_{YX}$ like in the context of statistical power. The function `ss.aipe.R2()` from the MBESS R package can be used as follows:

```r
ss.aipe.R2(Population.R2=.20,
width=.15, conf.level = 0.95,
K = 7, assurance = NULL)
```

where the parameters are the same as noted previously. Implementation of the above yields a necessary sample size of 361. This code does not incorporate an assurance parameter. Modifying the code to incorporate an assurance parameter of .99 leads to

```r
ss.aipe.R2(Population.R2=.20,
width=.15, conf.level = 0.95,
K = 7, assurance=.99)
```

which returns a necessary sample size of 403.

It is important to remember here that the appropriate sample size depends on the research question. For illustrative purposes, we planned sample size for four different goals, which would ordinarily not be the case. If, however, multiple goals are of interest, then we recommend using the largest of the planned sample sizes.

**SOFTWARE FOR SAMPLE SIZE PLANNING**

We used the MBESS R package in our example of sample size planning. Many sample size planning software packages are available for a wide variety of research designs, effect sizes, and goals. Kelley (in press) has provided a table of such programs. Sample size planning for the majority of the most widely used statistical methods in psychology can be implemented with the programs listed in Table 11.3. We are not able to provide a review of exactly what each of the programs is capable of doing, as doing so is beyond the scope of this chapter.

**DISCUSSION**

Sample size planning has been discussed in numerous book length treatments (e.g., Aberson, 2010; Bausell & Li, 2002; Chow, Shao, & Wang, 2003; Cohen, 1988; Dattalo, 2008; Davey & Savla, 2010; Kraemer & Thiemann, 1987; Lipsey, 1990; Machin, Campbell, Tan, & Tan, 2009; Murphy, Myors, & Wolach, 2008). Despite the push for the use of effect sizes and confidence intervals, these works do not contain as much emphasis on the AIPE approach as they do on power analysis. Because a single chapter cannot compete with book-length treatments for depth (e.g., the how-tos) or breadth (e.g., the number of designs), we hope our chapter provides a big picture view of the overarching principles that should be considered when designing a study and planning an appropriate sample size.
### TABLE 11.3
Software Titles Useful for Planning Sample Size

<table>
<thead>
<tr>
<th>Software title*</th>
<th>Author(s)/publisher</th>
<th>Operating system(s)</th>
<th>Free?</th>
<th>Web resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>G*Power</td>
<td>E. Erdfeider, F. Faul, &amp; A. Buchner</td>
<td>Windows/Mac</td>
<td>Yes</td>
<td><a href="http://www.psycho.uni-duesseldorf.de/aap/projects/gpower">http://www.psycho.uni-duesseldorf.de/aap/projects/gpower</a></td>
</tr>
<tr>
<td>Optimal Design</td>
<td>J. Spybrook, S. W. Raudenbush, R. Congdon, &amp; A. Martinez</td>
<td>Windows</td>
<td>Yes</td>
<td><a href="http://www.wtgrantfoundation.org/resources/overview/research_tools">http://www.wtgrantfoundation.org/resources/overview/research_tools</a></td>
</tr>
<tr>
<td>PinT</td>
<td>T. Snijders, R. Bosker, &amp; H. Guldemond</td>
<td>Windows</td>
<td>Yes</td>
<td><a href="http://www.power-analysis.com">http://www.power-analysis.com</a></td>
</tr>
<tr>
<td>Power and Precision</td>
<td>Biostat</td>
<td>Windows</td>
<td>No</td>
<td><a href="http://www.power-analysis.com">http://www.power-analysis.com</a></td>
</tr>
<tr>
<td>R</td>
<td>Package: asypow</td>
<td>Windows/Mac/Unix</td>
<td>Yes</td>
<td><a href="http://cran.r-project.org/web/packages/asypow/index.html">http://cran.r-project.org/web/packages/asypow/index.html</a></td>
</tr>
<tr>
<td></td>
<td>Package: MBESS</td>
<td>Windows</td>
<td>No</td>
<td><a href="http://cran.r-project.org/web/packages/MBESS/index.html">http://cran.r-project.org/web/packages/MBESS/index.html</a></td>
</tr>
<tr>
<td></td>
<td>Package: pamm</td>
<td>Windows</td>
<td>No</td>
<td><a href="http://cran.r-project.org/web/packages/pamm/index.html">http://cran.r-project.org/web/packages/pamm/index.html</a></td>
</tr>
<tr>
<td></td>
<td>Package: pwr</td>
<td>Windows</td>
<td>No</td>
<td><a href="http://cran.r-project.org/web/packages/pwr/index.html">http://cran.r-project.org/web/packages/pwr/index.html</a></td>
</tr>
<tr>
<td>SIZ</td>
<td>Cytel</td>
<td>Windows</td>
<td>No</td>
<td><a href="http://www.cytel.com/Software/SiZ.aspx">http://www.cytel.com/Software/SiZ.aspx</a></td>
</tr>
</tbody>
</table>

**Note.** Software titles are listed in alphabetical order. The failure to list a sample size planning software does not imply that it should not be considered. Purposely not included, for example, are narrowly focused sample size planning software titles. Also not included are web resources (e.g., online calculators), some of which can be helpful. Additionally, general software titles that could be made to plan sample size with the appropriate programming are not included, as the listed software titles are those that were developed specifically to plan sample size or contain specialized functions or procedures for planning sample size. From *Oxford Handbook of Quantitative Methods* by T. Little (Ed.), in press, New York, NY: Oxford University Press. Reprinted with permission of Oxford University Press, Inc.
Altman and Gardner (1988) encouraged medical studies to move away from NHST and to focus on confidence intervals. They also cautioned, however, that "confidence intervals convey only the effects of sampling variation on the precision of the estimated statistics and cannot control for non-sampling errors such as biases in design, conduct, or analysis" (p. 747). Correspondingly, if biases creep into a research design, not only can the population value be estimated in a biased fashion, but also the wrong quantity can be estimated precisely. That is to say, a narrow confidence interval could be obtained that brackets the wrong population quantity. The importance of considering the various types of validity when planning a study needs to be taken seriously (e.g., Shadish, Cook, & Campbell, 2002).

Without careful consideration of the issues we have discussed, answering one of the most common questions asked when planning research, namely, “What sample size do I need?” cannot adequately be addressed. Having a solid grasp of the foundational issues and appropriately linking the question of interest to the most appropriate type of sample size planning will facilitate the process of sample size planning. Once these issues are addressed, a software program can be used to plan the appropriate sample size, if available. A critical evaluation of the process of sample size planning will lead to better designed studies that we believe will facilitate a more cumulative and productive research literature.

References


