Psychological Methods
2006, Vol. 11, No. 4, 363–385

1082-989X/06/$12.00 DOI: 10.1037/1082-989X.11.4.363

Sample Size Planning for the Standardized Mean Difference: Accuracy in Parameter Estimation Via Narrow Confidence Intervals

Ken Kelley
Indiana University

Joseph R. Rausch
University of Notre Dame

Methods for planning sample size (SS) for the standardized mean difference so that a narrow confidence interval (CI) can be obtained via the *accuracy in parameter estimation* (AIPE) approach are developed. One method plans SS so that the expected width of the CI is sufficiently narrow. A modification adjusts the SS so that the obtained CI is no wider than desired with some specified degree of certainty (e.g., 99% certain the 95% CI will be no wider than ω). The rationale of the AIPE approach to SS planning is given, as is a discussion of the analytic approach to CI formation for the population standardized mean difference. Tables with values of necessary SS are provided. The freely available Methods for the Behavioral, Educational, and Social Sciences (K. Kelley, 2006a) R (R Development Core Team, 2006) software package easily implements the methods discussed.

Keywords: sample size planning, standardized mean difference, accuracy in parameter estimation, power analysis, precision analysis

One of the simplest measures of effect is the difference between two independent group means. It is this difference that is evaluated with the two-group *t* test to infer whether the population difference between two group means differs from some specified null value, which is generally set to zero. However, in the behavioral, educational, and social sciences, units of measurement are often arbitrary, different researchers might measure the same phenomenon with different scalings of the same instrument, or different instruments altogether might be used. Because of the lack of standard measurement scales and procedures for most behavioral, educational, and social phenomena, the ability to compare measures of effect across different situations has led many researchers to use standardized measures of effect. Measures of effect, or *effect sizes*, that are standardized yield scale-free numbers that are not wedded to a specific instrument or scaling metric. Given the measurement issues in behavioral, educational, and social research, such standardized effect sizes provide what is arguably the optimal way to estimate the size of an effect, along with its corresponding confidence interval, for a more communal knowledge base to be developed and so that the results from different studies can be compared more readily.

A commonly used and many times intuitively appealing effect size is the standardized mean difference.[1] In fact, the standardized mean difference is the most widely used statistic in the context of meta-analysis for experimental and intervention studies (Hunter & Schmidt, 2004, p. 246). The population standardized mean difference is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}, \tag{1}$$

where μ_1 is the population mean of Group 1, μ_2 is the population mean of Group 2, and σ is the population standard deviation assumed to be equal across the two groups. Because the unstandardized (raw) mean difference may not be directly comparable across studies, the unstandardized difference between group means can be divided by the standard deviation to remove the particular measurement scale, yielding a pure number (Cohen, 1988, p. 20). A commonly used set of guidelines for the standardized mean

Ken Kelley, Inquiry Methodology Program, Indiana University; Joseph R. Rausch, Department of Psychology, University of Notre Dame.

Joseph R. Rausch is now at the Department of Psychology, University of Minnesota, Twin Cities Campus.

Correspondence concerning this article should be addressed to Ken Kelley, Inquiry Methodology Program, Indiana University, 201 North Rose Avenue, Bloomington, IN 47405. E–mail: kkiii@indiana.edu

[1] In some cases, the unstandardized difference between means is more intuitively appealing than is the standardized mean difference (e.g., Bond, Wiitala, & Richard, 2003). If the unstandardized mean difference is of interest, Kelley et al. (2003) discussed the methods analogous to those discussed in the present article.

difference in the behavioral, educational, and social sciences, although not without its critics (e.g., Lenth, 2001), is that δs of 0.2, 0.5, and 0.8 are regarded as small, medium, and large effects, respectively (Cohen, 1969, 1988).[2]

Suppose a researcher is interested in the effect of a particular treatment on the mean of some variable and would like to compare an experimental group with a control group. The researcher's review of the literature and a pilot study lead the researcher to believe that the effect of the treatment is of a "medium" magnitude, corresponding to a standardized mean difference of approximately 0.50. As is widely recommended in the literature, the researcher conducts a power analysis to determine the necessary sample size so that there will be a high probability of rejecting the presumed false null hypothesis. Basing the sample size calculations on a desired degree of power of 0.85, the researcher conducts the study with 73 participants per group.

The observed standardized mean difference was 0.53, giving some support to the researcher's assertion that the effect is of medium magnitude, and was shown to be statistically significant, $t_{(144)} = 3.20$, $p(t_{(144)} \geq |3.20|) = .002$. In accord with recent recommendations in the literature, the researcher forms a 95% confidence interval for δ, which ranges from 0.199 to 0.859. Although the researcher believed the effect was medium in the population, to the researcher's dismay, the lower limit of the confidence interval is smaller than "small" and the upper limit is larger than "large." The width of the researcher's confidence interval thus illustrates that even though the null hypothesis was rejected, a great deal of uncertainty exists regarding the value of δ, which is where the researcher's interest ultimately lies. Indeed, as Rosenthal (1993) argued, the results we are actually interested in from empirical studies are the estimate of the magnitude of the effect and an indication of its accuracy, "as in a confidence interval placed around the estimate" (p. 521).

The purpose of the present work is to offer an alternative to the power analytic approach to sample size planning for the standardized mean difference. This general approach to sample size planning is termed *accuracy in parameter estimation* (AIPE; Kelley, 2006b; Kelley & Maxwell, 2003, in press; Kelley, Maxwell, & Rausch, 2003), where what is of interest is planning sample size to achieve a sufficiently narrow confidence interval so that the parameter estimate will have a high degree of expected accuracy. A confidence interval consists of a set of plausible values that will contain the parameter with $(1 - \alpha)100\%$ confidence. Appropriately constructed confidence intervals will always contain the parameter estimate and will contain the parameter $(1 - \alpha)100\%$ of the time. The idea of the AIPE approach is that when the width of a $(1 - \alpha)100\%$ confidence interval decreases, the range of plausible values for the parameter decreases with the estimate necessarily contained within this set of plausible values. Provided that the confidence interval procedure is exact (i.e., the nominal coverage is equal to the empirical coverage) and holding constant the $(1 - \alpha)100\%$ confidence interval coverage, the expected difference between the estimate and the parameter decreases as the confidence interval width decreases.

In the context of parameter estimation, *accuracy* is defined as the square root of the mean square error, which is a function of both precision and bias. Precision is inversely related to the variance of the estimator, and bias is the systematic discrepancy between an estimate and the parameter it estimates. More formally, accuracy is quantified by the (square) root of the mean square error (RMSE) as

$$\begin{aligned} \text{RMSE} &= \sqrt{\text{E}[(\hat{\theta} - \theta)^2]} \\ &= \sqrt{\text{E}[(\hat{\theta} - \text{E}[\theta])^2] + (\text{E}[\hat{\theta} - \theta])^2} \\ &= \sqrt{\sigma_\theta^2 + \text{B}_\theta^2}, \end{aligned} \tag{2}$$

where $\text{E}[\cdot]$ represents expectation, θ is the parameter of interest, $\hat{\theta}$ is an estimate of θ, σ_θ^2 is the population variance of the estimator, and B_θ is the bias of the estimator. As the confidence interval width decreases, holding constant the confidence interval coverage, the estimate is contained within a narrower set of plausible parameter values and the expected accuracy of the estimate improves (i.e., the RMSE is reduced). Thus, provided that the confidence interval procedure is exact, when the width of the $(1 - \alpha)100\%$ confidence interval decreases, the expected accuracy of the estimate necessarily increases.

The effect of increasing sample size has two effects on accuracy. First, the larger the sample size, generally the more precise the estimate.[3] Second, estimates that are biased will generally become less biased as sample size increases, which must be the case for consistent estimators (regardless of whether the estimator is biased or unbiased; Stuart & Ord, 1994). Notice that when an estimate is unbiased (i.e., $\text{E}[\hat{\theta} - \theta] = 0$), precision and accuracy are equivalent. However, a precise estimator need not be an accurate estimator. Thus, precision is a neces-

[2] Of course, as with most rules of thumb, Cohen's (1988) guidelines have their limitations and should not be applied without first consulting the literature of the particular area. Overreliance on Cohen's guidelines can lead an investigator astray when planning sample size for a particular research question when the size of δ is misidentified, which is easy to do if the only possibilities considered are 0.2, 0.5, and 0.8.

[3] A counterexample is the Cauchy distribution, in which the precision of the location estimate is the same regardless of the sample size used to estimate it (Stuart & Ord, 1994, pp. 2–3).

sary but not a sufficient condition for accuracy.[4] Beyond the effect of improving precision, decreasing bias improves accuracy.[5] This usage of the term *accuracy* is the same as that used by Neyman (1937) in his seminal work on the theory of confidence intervals: "The accuracy of estimation corresponding to a fixed value of $1 - \alpha$ may be measured by the length of the confidence interval" (p. 358; we changed Neyman's original notation of α representing the confidence interval coverage to $1 - \alpha$ to reflect current usage).

One of the main reasons why researchers plan, conduct, and then analyze the data of empirical studies is to learn about some parameter of interest. One way in which researchers have attempted to learn about the parameter of interest historically has been by conducting null hypothesis significance tests. Null hypothesis significance testing allows researchers to reject the idea that the true value of the parameter of interest is some precisely specified value (usually zero for the standardized mean difference). By conducting a significance test that achieves statistical significance, researchers learn probabilistically what the parameter is not (e.g., δ is not likely zero) and possibly the direction of the effect. Another way in which researchers have attempted to learn about the parameter of interest is by forming confidence intervals for the population parameter of interest. By forming a confidence interval, not only does a researcher learn probabilistically what the parameter is not (i.e., those values outside the bounds of the interval) but also a researcher learns probabilistically a range of plausible values for the parameter of interest.[6]

As has been echoed numerous times in the methodological literature of the behavioral, educational, and social sciences (e.g., Nickerson, 2000, which along with the references contained therein provides a comprehensive historical review; see also Cohen, 1994; Meehl, 1997; Schmidt, 1996; among many others), there are serious limitations to null hypothesis significance tests. As Hunter and Schmidt (2004) and Cohen (1994) pointed out, the null hypothesis may almost never be exactly true in nature.[7] Regardless of whether the null hypothesis is true or false, what is often most informative is the value or size of the population effect. As recommended by Wilkinson and the American Psychological Association (APA) Task Force on Statistical Inference (1999), researchers should "*always present effect sizes for primary outcomes*" (p. 599), and they stressed that "interpreting effect sizes in the context of previously reported effects is essential to good research" (p. 599). Wilkinson and the APA Task Force on Statistical Inference also recommended that "*interval estimates should be given for any effect sizes involving principal outcomes*" (p. 599). It seems that there is general consensus in the methodological community of the behavioral, educational, and social sciences with regard to trying to understand various phenomena of interest, and that consensus is to report confidence intervals for effect sizes whenever possible; indeed, this strategy may be the future of quantitative methods in applied research (Thompson, 2002).

Even though the merits of significance testing have come under fire in the methodological literature, null hypothesis significance tests have played a major role in the behavioral, educational, and social sciences. Although reporting measures of effect is useful, reporting point estimates without confidence intervals to illustrate the uncertainty of the estimate can be misleading and cannot be condoned. Reporting and interpreting point estimates can be especially misleading when the corre-

---

[4] As an extreme example, suppose a researcher always ignores the data and estimates the parameter as a value that corresponds to a particular theory. Such an estimate would have a high degree of precision but potentially could be quite biased. The estimate would only have a high degree of accuracy if the theory was close to perfect.

[5] Some parameters have exact confidence interval procedures that are based on a biased point estimate of the parameter yet where an unbiased point estimate of the parameter also exists. A strategy in such cases is to report the unbiased estimate for the point estimate of the parameter in addition to the $(1 - \alpha)100\%$ confidence interval for the parameter (calculated on the basis of the biased estimate). Examples of parameters that have exact confidence interval procedures that are calculated on the basis of a biased estimate are the standardized mean difference (e.g., Hedges & Olkin, 1985), the squared multiple correlation coefficient (e.g., Algina & Olejnik, 2000), the standard deviation (see, e.g., Hays, 1994, for the confidence interval method and Holtzman, 1950, for the unbiased estimate), and the coefficient of variation (see, e.g., Johnson & Welch, 1940, for the confidence interval method and Sokal & Braumann, 1980, for its nearly unbiased estimate).

[6] Assuming that the assumptions of the model are met, the correct model is fit, and observations are randomly sampled, $1 - \alpha$ 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 value of the parameter. The procedure refers to the infinite number of confidence intervals that could theoretically be constructed and the $(1 - \alpha)100\%$ of those confidence intervals that correctly bracket the population parameter of interest (see Hahn & Meeker, 1991, for a technical review of confidence interval formation). Although the meaning of confidence intervals given is from a frequentist perspective, the methods discussed in the article are also applicable in a Bayesian context.

[7] This argument seems especially salient in the context of observational studies in which preexisting group differences likely exist. However, in unconfounded experimental studies with randomization, it seems plausible that a treatment might literally have no effect, which would of course imply that the null hypothesis is true.

sponding confidence interval is wide and thus little is known about the likely size of the population parameter of interest. Because confidence intervals provide a range of reasonable values that bracket the parameter of interest with some desired degree of confidence, confidence intervals provide a great deal of information above and beyond the estimated value of the effect size and the corresponding statistical significance test. Thus, effect sizes accompanied by their corresponding confidence intervals are perhaps the best way to illustrate how much information was learned about the parameter of interest from the study.

Suppose a very wide confidence interval is formed, and yet zero is excluded from the confidence interval. Such a confidence interval provides some but not much insight into the phenomenon of interest. What is learned in such a scenario is that the parameter is not likely zero and possibly the direction of the effect. Even in situations in which it is well established that the effect is not zero, providing statistical evidence that the effect is not zero is almost always a goal. The reason power analysis is so beneficial is because it helps to ensure that an adequate sample size is used to show that the effect is not zero in the population. However, the result of a significance test in and of itself does not provide information about the size of the effect.

The accuracy of parameter estimates is also important in another context when one wishes to show support for the null hypothesis (e.g., Greenwald, 1975) or in the context of equivalence testing (e.g., Steiger, 2004; Tryon, 2001). The "good enough" principle can be used and a corresponding "good enough belt" can be formed around the null value, where the limits of the belt would define what constituted a nontrivial effect (Serlin & Lapsley, 1985, 1993). Suppose that not only is the null value contained within the good enough belt but also the limits of the confidence interval are within the good enough belt. This would be a situation in which all of the plausible values would be smaller in magnitude than what has been defined as a trivial effect (i.e., they are contained within the good enough belt). In such a situation, the limits of the $(1 - \alpha)100\%$ confidence interval would exclude all effects of meaningful size. If the parameter is less in magnitude than what is regarded to be minimally important, then learning this can be very valuable. This information may or may not support the theory of interest, but what is important is that valuable information about the size of the effect and thus the phenomenon of interest has been gained.

Perhaps the ideal scenario in many research contexts is when the confidence interval for the parameter of interest is narrow (and thus a good deal is learned about the plausible value of the parameter) and does not contain zero (and thus the null hypothesis can be rejected). Accomplishing the latter, namely, rejecting the null hypothesis, has long been a central part of research design in the form of power analysis. However, accomplishing the former, namely, obtaining a narrow confidence interval, has not received much attention in the methodological literature of the behavioral, educational, and social sciences (cf. Algina & Olejnik, 2000; Bonett & Wright, 2000; Kelley & Maxwell, 2003; Kelley et al., 2003; Smithson, 2003).

Confidence intervals can be calculated for the standardized mean difference in two main ways. One method uses the bootstrap technique (e.g., Efron & Tibshirani, 1993) and does not require the assumption of homogeneity of variance or normality to obtain valid confidence intervals (Kelley, 2005), potentially using a robust estimator of standardized population separation in place of d (e.g., Algina, Keselman, & Penfield, 2005). The other method, which is optimal when the assumptions of normality, homogeneity of variance, and independence of observations are satisfied, is the analytic approach. The analytic approach requires specialized computer routines, specifically noncentral t distributions, to obtain the confidence limits for δ (e.g., Cumming & Finch, 2001; Kelley, 2005; Smithson, 2001; Steiger, 2004; Steiger & Fouladi, 1997). Throughout the remainder of the article, the focus is on the analytic approach to confidence interval formation.

The problem the present work addresses is that of obtaining an accurate estimate of the population standardized mean difference by planning sample size so that the observed $(1 - \alpha)100\%$ confidence interval will be sufficiently narrow with some specified probability. The following section provides an overview of confidence interval formation for the population standardized mean difference. Methods for planning sample size so that the expected width of the confidence interval is sufficiently narrow are then developed. The first procedure determines the sample size necessary for the expected width of the obtained confidence interval for the population standardized mean difference to be sufficiently narrow. Obtaining a large enough sample size so that the expected width will be sufficiently narrow does not guarantee that a computed interval will, in fact, be as narrow as specified. This method is extended into a follow-up procedure in which there will be some desired degree of certainty that the computed interval will be sufficiently narrow (e.g., 99% certain that the 95% confidence interval will be no wider than the specified width). Sample size tables are provided for a variety of situations on the basis of the premise that they will assist applied researchers in choosing an appropriate sample size given a particular goal within the AIPE framework for the standardized mean difference. Because a main goal of research is to learn about the parameter of interest, obtaining a narrow confidence interval may be the best way to fulfill this goal. It

is this premise, coupled with the usefulness of the standardized mean difference, that has motivated this article and the development of computer routines that can be used to carry out the methods discussed.[8]

Estimation and Confidence Interval Formation for the Standardized Mean Difference

Although δ is the ultimate quantity of interest, δ is unknown and must be estimated from sample data. The most common way in which δ is estimated is defined as

$$d = \frac{\overline{X}_1 - \overline{X}_2}{s}, \tag{3}$$

where $\overline{X}_j$ is the mean for the jth group ($j = 1, 2$) and s is the square root of the pooled variance (i.e., s is the square root of the unbiased estimate of the within-group variance).

As pointed out by Cumming and Finch (2001), there is inconsistency in the terminology and notation used when discussing the standardized and unstandardized effect sizes for mean differences (see also Grissom & Kim, 2005). Our proposal is to use Δ as $\mu_1 - \mu_2$ with $D = \overline{X}_1 - \overline{X}_2$ as its sample estimate, δ (Equation 1) as the population standardized mean difference and d (Equation 3) as its sample estimate, and δ_C as the population standardized mean difference using the control group standard deviation as the divisor and d_C as its sample estimate. Not discussed in this article but important nonetheless are the unbiased estimators of δ and δ_C (d and d_C are not unbiased), for which we suggest d_U and d_{C_U} as their notation (see Hedges, 1981, for its theoretical developments and Kelley, 2005, for some comparisons to the commonly used biased version). Discussed momentarily is the noncentral t distribution that has a noncentral parameter. There is also inconsistency in notation for this noncentral parameter, and we suggest λ as opposed to δ or Δ (both commonly used symbols) because of their use as population effect size measures.[9]

Part of this inconsistency in notation is a function of trying to attribute one or more versions of the standardized effect size to particular authors coupled with those same authors using different notation in different works. The estimated standardized mean difference, d, is often referred to as Cohen's d (even though Cohen used d as the population parameter and d_s as its sample analog; Cohen, 1988) because of Cohen's work on the general topic of effect size and power analysis (Cohen, 1969, p. 18) and sometimes as Hedges's g' (or g) because of Hedges's work on how the standardized effect size could be used in a meta-analysis context and its theoretical developments (Hedges, 1981, p. 110). The analogous standardized effect size based on the control group standard deviation is often called δ or Glass's g (Glass, 1976; Glass, McGaw, & Smith, 1981, p. 29; Hedges, 1981, p. 109). Furthermore, the Mahalanobis distance, which is the multivariate version of d (and for one variable is equal to d), was developed well before d was used as a standardized effect size in the behavioral, educational, and social sciences (Mahalanobis, 1936). Given all of the possible labelings of what is defined in Equation 3, we call this quantity the *standardized mean difference* without attempting to attribute this often used quantity to any one individual (recognizing that many have worked on its theoretical developments and others have encouraged its use) and use the notation d to represent the sample value (which is currently the most widely used notation).

Recall that the two-group t test is defined as

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{4}$$

where n_1 and n_2 are the sample sizes for Group 1 and Group 2, respectively. In the two-group situation, assuming homogeneity of variance, s is defined as

$$s = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}, \tag{5}$$

where s_1^2 and s_2^2 are the within-group variances for Group 1 and Group 2, respectively, and has $n_1 + n_2 - 2$ degrees of freedom. However, in an analysis of variance (ANOVA) context where more than two groups exist and the assumptions of ANOVA are satisfied, the estimate of s (as well as d) can be improved by pooling information across all J groups, even if what is of interest is the difference between

[8] Throughout the article, specialized software is used. Ken Kelley has developed an R package that contains, among other things, the necessary functions to form confidence intervals for the population standardized mean difference and to estimate sample size from the AIPE perspective for the standardized mean difference. The R package is titled Methods for the Behavioral, Educational, and Social Sciences (MBESS) and is an Open Source, and thus freely available, package available via the Comprehensive R Archival Network (CRAN; http://www.r-project.org/). The direct link to the MBESS page on CRAN, where the most up-to-date version of MBESS is available, is http://cran.r-project.org/src/contrib/Descriptions/MBESS.html (note that this Internet address is case sensitive).

[9] Much of the work contained in the present article can be applied to δ_C and d_C by modifying the degrees of freedom of the denominator to have degrees of freedom equal to that of s_C, the standard deviation of the control group.

the means of only two specific groups. Thus, more generally, s can be defined as

$$s = \sqrt{\frac{\sum_{j=1}^{J} s_j^2(n_j - 1)}{N - J}}, \tag{6}$$

where N is the total sample size ($N = \sum_{j=1}^{J} n_j$) and J is the number of groups ($j = 1, \ldots, J$). In situations where $J > 2$ and the ANOVA assumptions are satisfied, basing s on all groups leads to more degrees of freedom ($N - J$ degrees of freedom instead of $n_1 + n_2 - 2$). Holding everything else constant, the larger the degrees of freedom, the more powerful the significance test for the mean difference and the more accurate the estimate of the standardized (and unstandardized) mean difference. Thus, when information on $J \geq 3$ groups is available, making use of that information should be considered even if what is of interest is estimating δ for two specific groups. Of course, as J increases, the potential for the assumption of homogeneity of variance to be violated also increases, but if the assumption holds, more power and accuracy will be gained by using a pooled variance based on $J \geq 3$ groups.

Notice that the difference between d from Equation 3 and the two-group t statistic from Equation 4 is the quantity $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ contained in the denominator of the t statistic, which is multiplied by s to estimate the standard error. Because $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ can be rewritten as $\sqrt{\frac{n_2 + n_1}{n_1 n_2}}$, multiplying the inverse of this quantity by d leads to an equivalent representation of the t statistic:

$$t = d\sqrt{\frac{n_1 n_2}{n_2 + n_1}}. \tag{7}$$

Given Equation 7, it can be seen that Equation 3 can be written as

$$d = t\sqrt{\frac{n_2 + n_1}{n_1 n_2}}. \tag{8}$$

The usefulness of Equations 7 and 8 will be realized momentarily when discussing the formation of confidence intervals for δ.

The noncentral parameter in the two-group context indexes the magnitude of the difference between the null hypothesis of $\mu_1 = \mu_2$ and an alternative hypothesis of $\mu_1 \neq \mu_2$. The larger the difference between the null and alternative hypotheses, the larger the noncentral parameter. In the population, the degree to which $\mu_1 \neq \mu_2$ for $N - 2$ degrees of freedom is known as a noncentral parameter:

$$\lambda = \frac{\mu_1 - \mu_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \delta\sqrt{\frac{n_1 n_2}{n_2 + n_1}}. \tag{9}$$

The noncentral parameter λ is of the same form as a t statistic (for a technical discussion of the noncentral t distribution, see, e.g., Hogben, Pinkham, & Wilk, 1961; Johnson, Kotz, & Balakrishnan, 1995; Johnson & Welch, 1940). In fact, λ can be obtained by replacing the sample values in Equation 4 with their population values for the sample sizes of interest. Given the relationship between a t value and the corresponding noncentral parameter, λ can be estimated by the observed t statistic: $\hat{\lambda} = t$. Construction of confidence intervals for δ is indirect and proceeds by first finding a confidence interval for λ and then transforming those bounds via Equation 8 to the scale of δ using a combination of the confidence interval transformation principle and the inversion confidence interval principle (Cumming & Finch, 2001; Kelley, 2005; Steiger & Fouladi, 1997; Steiger, 2004).

Let $t'_{(q,\nu,\lambda)}$ be the critical value from the qth quantile from a noncentral t distribution with ν degrees of freedom and noncentral parameter λ. The degrees of freedom parameter is based on the sample size used to calculate s. To find the confidence bounds for δ, first find the confidence bounds for λ. Because of the confidence interval transformation principle, the one-to-one monotonic relation between λ and δ given n_1 and n_2 (Equations 7 and 8) implies that the $(1 - \alpha)100\%$ confidence bounds for λ provides, after transformation via Equation 8, the $(1 - \alpha)100\%$ confidence bounds for δ.

The confidence bounds for λ are determined by finding the noncentral parameter whose $1 - \alpha/2$ quantile is t (for the lower bound of the confidence interval) and by finding the noncentral parameter whose $\alpha/2$ quantile is t (for the upper bound of the confidence interval). Thus, the lower confidence bound for λ, λ_L, is the noncentral parameter that leads to $t'_{(1-\alpha/2,\nu,\lambda_L)} = t$ and the upper confidence bound for λ, λ_U, is the noncentral parameter that leads to $t'_{(\alpha/2,\nu,\lambda_U)} = t$.[10] For the lower and upper confidence bounds for λ, given α, ν, and t, the only unknown values are λ_L and λ_U. It is λ_L and λ_U that are of

[10] It is assumed here that the confidence interval will use the same rejection region in both tails. Although convenient, this is not necessary. Rejection regions could be defined so that $\alpha = \alpha_L + \alpha_U$ for the lower and upper rejection regions, respectively. It is assumed in this article that $\alpha/2 = \alpha_L = \alpha_U$ (i.e., equal probability in each rejection region). The MBESS package does not make this assumption, and thus varying values of α_L and α_U are possible when determining the confidence interval for the standardized mean difference.

interest when forming confidence intervals for λ and that have, until recently, been difficult to obtain. However, λ_L and λ_U from $t'_{(1-\alpha/2,\nu,\lambda_L)}$ and $t'_{(\alpha/2,\nu,\lambda_U)}$, respectively, are now easily obtainable with several software titles, making the formation of confidence limits for λ and ultimately for δ easy to find:

$$p[\lambda_L \leq \lambda \leq \lambda_U] = 1 - \alpha, \tag{10}$$

where p represents the probability of λ_L and λ_U bracketing λ at the $1 - \alpha$ level.

As an example, suppose two groups of 10 participants each have a standardized mean difference of 1.25 with the corresponding t value of 2.7951. The noncentral t distribution with noncentral parameter of 0.6038 has at its .975 quantile 2.7951, whereas the noncentral t distribution with noncentral parameter of 4.9226 has at its .025 quantile 2.7951, both with 18 degrees of freedom. Thus,

$$\text{CI}_{.95} = [0.6038 \leq \lambda \leq 4.9226], \tag{11}$$

where $\text{CI}_{.95}$ represents a 95% confidence interval. The relation between the two noncentral distributions and the observed t value is illustrated in Figure 1, where the shaded regions represent the areas of the distributions that are beyond the confidence limits. As can be seen in Figure 1, the noncentral t distribution on the left has a noncentral parameter of 0.6038, and at its .975 quantile is the observed t value, which is denoted with the bold vertical line near the center of the abscissa. As can also be seen, the noncentral t distribution on the right has a noncentral parameter of 4.9226, and at its .025 quantile is the observed t value, which is denoted with the same bold vertical line.

The shaded lines to the left and right of the λ_L and λ_U, respectively, illustrate the area of these distributions outside of the confidence bounds for λ. Furthermore, because of the one-to-one relation between λ and δ, the upper abscissa shows values of δ. Notice also that the shapes of the distributions are different, with the one on the right more variable and more positively skewed than the one on the left (because of the larger noncentral parameter and all other things being equal). Of special importance are the two outer vertical lines that represent the noncentral parameters of the two distributions. As can be seen, the noncentral parameters are not only the confidence limits for λ, but after the noncentral parameters have been rescaled with Equation 8, they yield the confidence limits for δ,

$$\text{CI}_{.95} = \left[0.6038\sqrt{\frac{10+10}{10\times 10}} \leq \delta \leq 4.9226\sqrt{\frac{10+10}{10\times 10}}\right], \tag{12}$$

which equals

$$\text{CI}_{.95} = [0.2700 \leq \delta \leq 2.2015]. \tag{13}$$

Notice that although 2.5% of the distribution of d is beyond the lower and upper limits, the distance between d and the limits is not the same. As Stuart and Ord (1994) discussed, "in general, the confidence limits are equidistant from the sample statistic only if its [i.e., the statistic's] sampling distribution is symmetrical" (p. 121). Furthermore, the bold vertical line in the center identifies the estimated noncentral parameter (on the lower abscissa) and the estimated standardized mean difference (on the upper abscissa).

As Vaske, Gliner, and Morgan (2002) stated, "large confidence intervals make conclusions more tentative and weaken the practical significance of the findings" (p. 294). In an effort to obtain narrower confidence intervals for significant effects, Vaske et al. (2002) suggested researchers report two confidence intervals, one based on the α value used to conduct the significance test and one that has a much larger α value and thus a much narrower confidence interval width, such as $\alpha = .30$ or $\alpha = .20$ (p. 299). Although some researchers may be willing to pay the price for such a trade-off (narrow confidence interval but low level of confidence interval coverage), readers may not be so willing to accept it (Grissom & Kim, 2005, pp. 61–62). Although such an approach is not advocated here, the desire to obtain narrow confidence intervals because of the benefits they provide is understandable.[11] Using the methods developed here will help researchers avoid obtaining confidence intervals whose widths are "embarrassingly large" (Cohen, 1994, p. 1002).

In some situations, the required sample size might be too large for a researcher to reasonably collect the method-implied sample size. As a reviewer pointed out, this could imply trading "embarrassingly large" confidence intervals for "distressingly large" sample sizes. The methods used here are still beneficial because it will be known a priori that the confidence interval will likely be wider than desired, alleviating any unrealistic expectations about the width of the confidence interval a priori. Furthermore, authors who are only able to obtain smaller sample sizes could use the methods to show that it would be difficult or impossible to obtain the required sample size for the confidence interval for δ to be as narrow as desired, even if the sample size provides sufficient statistical power (e.g., for a large effect size). In situations in which a single study cannot produce (e.g., because of insufficient resources) a sufficiently narrow confidence interval, the use of meta-analysis might be especially useful (Hedges & Olkin, 1985; Hunter & Schmidt, 2004).

Kelley et al. (2003) discussed methods for planning sam-

[11] We agree with those that state there is nothing magical about $\alpha = .05$. However, regardless of what the a priori α value is, the methods discussed in the next section are applicable because the α value is specified by the researcher when planning sample size.

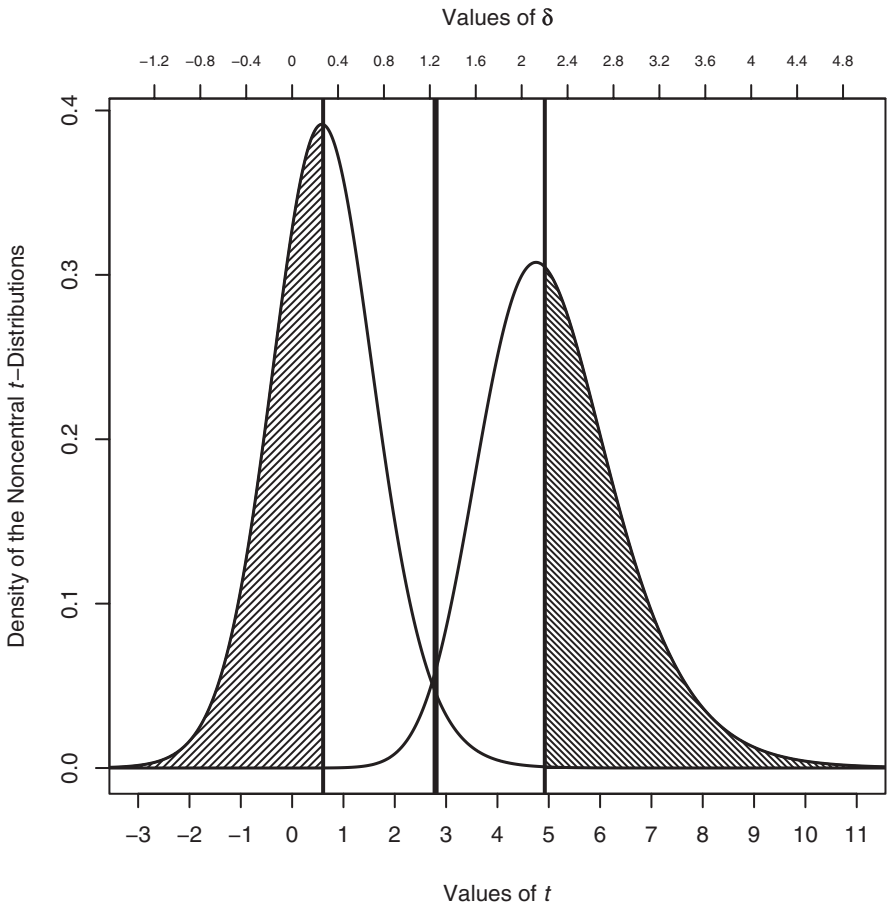

Figure 1. Density of the noncentral *t* distribution with 18 degrees of freedom and noncentral parameter 0.6038 (distribution on the left) and for the noncentral *t* distribution with 18 degrees of freedom and noncentral parameter 4.9226 (distribution on the right). Note that $t'_{(.975,18,0.6038)} = t'_{(.025,18,4.9226)} = t = 2.7951$. Thus, the 95% confidence interval for λ (shown on the lower abscissa) given the observed *t* value (2.7951) has lower and upper confidence bounds of 0.6038 and 4.9226, respectively. Transforming the confidence limits to the scale of δ (shown on the upper abscissa) leads to lower and upper 95% confidence bounds for δs of 0.2700 and 2.2015, respectively.

ple size so that the expected width of the confidence interval for the population unstandardized mean difference would be equal to some specified value. A modified method was also developed so that a desired degree of certainty (i.e., a probability) could be incorporated into the sample size procedure such that the obtained interval would be no wider than desired. However, planning sample size so that the obtained confidence interval is sufficiently narrow has not been discussed in the context of the standardized mean difference. The next section addresses this issue formally for the standardized mean difference and provides solutions so that the necessary sample size can be determined for the expected width to be sufficiently narrow, optionally with a desired degree of certainty that the obtained interval will be no wider than desired.

Sample Size Planning From an AIPE Perspective for the Standardized Mean Difference

There are presumably two reasons why sample size planning for the standardized mean difference from an AIPE perspective has not been formerly considered. First, sample size planning has been almost exclusively associated with power analysis, and thus planning sample size in order to obtain parameter estimates with a high degree of expected accuracy (i.e., a narrow confidence interval) has only recently been considered in much of the behavioral, educational, and social sciences. Second, working with noncentral *t* distributions has proven quite difficult because of the additional complexity of the probability function of the *t* distribution when the noncentral parameter is not zero.

Specialized computer algorithms are necessary to determine quantiles at desired probability values and probability values at desired quantiles. With the focus of sample size planning for power at the neglect of accuracy and the inability to readily work with noncentral *t* distributions, it is no wonder that sample size planning from an accuracy perspective has not yet been developed for the standardized mean difference.

Nevertheless, the solution to this problem is of interest to substantive researchers who want to estimate the sample size necessary to obtain narrow confidence intervals and for methodologists who study the properties of point estimates and their corresponding confidence intervals. There are also potential uses in the context of meta-analytic work.[12]

When attempting to plan sample size, for the expected width of the obtained confidence interval to be sufficiently narrow for the population standardized mean difference, it is necessary to use an iterative process. Because the confidence interval width for δ is not symmetric, the desired width can pertain to the full confidence interval width, the lower width, or the upper width. Let δ_U be defined as the upper limit and δ_L be defined as the lower limit of the observed confidence interval for δ. The full width of the obtained confidence interval is thus given as

$$w = \delta_U - \delta_L, \tag{14}$$

the lower width of the obtained confidence interval is given as

$$w_L = d - \delta_L, \tag{15}$$

and the upper width of the obtained confidence interval is given as

$$w_U = \delta_U - d. \tag{16}$$

The goals of the research study will dictate the confidence interval width for which sample size should be planned. In general, w is the width of interest. Although the methods discussed are directly applicable to determining sample size for the lower or the upper limit, we focus exclusively on the full confidence interval width. Let ω be defined as the desired confidence interval width, which is specified a priori by the researcher, much like the desired degree of statistical power is chosen a priori when determining necessary sample size in a power analytic context (e.g., Cohen, 1988; Kraemer & Thiemann, 1987; Lipsey, 1990; Murphy & Myors, 1998).

The idea of determining sample size so that $E[w] = \omega$ is analogous to other methods of planning sample size when a narrow confidence interval is desired (e.g., Guenther, 1981; Hahn & Meeker, 1991; Kelley & Maxwell, 2003; Kupper & Hafner, 1989). The goal is to determine the sample size so that $E[w] = \omega$. However, because the theoretical sample size where $E[w] = \omega$ is almost always a fractional value, $E[w]$ is almost always just less than ω for the necessary sample size to be some whole number. The population values are used in the confidence interval as if the population values were sample values, and then the necessary sample size is solved for so that $E[w] \leq \omega$. In general, sample size can be solved analytically or computationally. Solving sample size computationally, which is especially convenient when the confidence interval does not have a convenient closed-form expression, begins by finding a minimal sample size so that $E[w] > \omega$. The minimal sample size can then be incremented by 1 until $E[w] \leq \omega$.

Because the noncentral *t* distribution is used for confidence intervals for δ, sample size is solved for computationally. The initial value of the sample size used in the algorithm is based on the standard normal distribution, which guarantees that the initial sample size will not be too large. If σ is known and is common for the two groups, a confidence interval for the standardized mean difference is given as

$$p\left[\frac{\bar{X}_1 - \bar{X}_2}{\sigma} - z_{(1-\alpha/2)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \frac{\mu_1 - \mu_2}{\sigma} \leq \frac{\bar{X}_1 - \bar{X}_2}{\sigma} + z_{(1-\alpha/2)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] = 1 - \alpha. \tag{17}$$

Because $z_{(1-\alpha/2)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ is subtracted and added to the observed standardized difference in means, the width of the confidence interval is given as

$$2z_{(1-\alpha/2)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

When $n_1 = n_2 = n$, the confidence interval width can be simplified to

$$2z_{(1-\alpha/2)}\sqrt{\frac{2}{n}}.$$

Solving analytically for the necessary sample size so thatthe expected width of the confidence interval is equal to ω is given as

$$n_{(0)} = \text{ceiling}\left[8\left(\frac{z_{(1-\alpha/2)}}{\omega}\right)^2\right],$$

[12] It should be noted that whenever planning sample size, regardless of the perspective one is planning from, if the assumptions the procedure is based on are not satisfied, then the sample size estimate may not be appropriate. The degree of the inappropriateness of the estimated sample size will depend strongly on the specifics of the situation.

where $n_{(0)}$ is the initial value of sample size that will be used in the algorithm for determining the necessary sample size and ceiling[·] rounds the value in brackets to the next largest integer.

Of course, in practice, the use of the confidence interval given in Equation 17 is not appropriate because σ is almost never known and its estimate is a random variable necessitating a noncentral confidence interval, as discussed in the previous section. However, to obtain an initial value of sample size that is guaranteed to be no larger than the necessary sample size, the standard normal distribution is used in place of the noncentral *t* distribution. The use of the critical value from the standard normal distribution ensures that the starting value for the sample size used in the remainder of the algorithm is not initially overestimated, as replacing the critical value with a noncentral *t* value at the same α level is guaranteed to increase the width of the confidence interval.

Given $n_{(0)}$, the expected confidence interval can be calculated using the noncentral method previously discussed by replacing *d* in the confidence interval procedure with δ. The value of δ is used in the sample size procedure because δ is (essentially) the expected value of *d*, and thus the procedure is based on the value that is expected to be obtained in the study.[13] Next, increment sample size by one, yielding $n_{(1)}$, and then determine the expected width of the confidence interval, which is now based on $n_{(1)}$ ($n_{(1)} = n_{(0)} + 1$). If the expected width using $n_{(1)}$ is equal to or narrower than the desired width, the procedures can be stopped and the necessary sample size can be set to $n_{(1)}$. If the expected confidence interval width is wider than the desired width, sample size can be incremented by one and the expected width determined again. This process continues until the expected width is equal to or just narrower than the desired width. At the iteration where this happens, set $n_{(i)}$ to the necessary sample size. The idea of the algorithm is fairly straightforward: (a) Use δ as if it were *d*, (b) increase necessary sample size until the expected width of the confidence interval is sufficiently narrow, and (c) set the value of sample size to the necessary value so that $E[w] \leq \omega$.

Although in some situations ensuring that the expected width of a confidence interval for δ is sufficiently narrow is satisfactory, in most situations the desire is for *w* to be no larger than ω. The procedure just discussed in no way implies that the observed confidence interval width in any particular study will be no larger than ω, as *w* is a random variable that will fluctuate from study to study or from replication to replication of the same study.[14] Thus, it is important to remember that the algorithm just presented provides the sample size such that $E[w] \leq \omega$. A modified sample size procedure can be performed so that there is a desired degree of certainty that *w* will not be larger than ω.

Ensuring a Confidence Interval No Wider Than Desired With a Specified Degree of Certainty

As a function of the properties of the noncentral *t* distribution, as the magnitude of δ gets larger, holding the confidence interval coverage and sample size constant, the expected width of the confidence interval becomes wider.[15] However, the width of the observed confidence interval is a function of *d* and the per-group sample sizes. When determining the necessary sample size given δ, the variability of *d* is also important, because if the sample collected yields a *d* smaller in magnitude than the δ specified when determining the sample size, then *w* will be narrower than ω. However, when the sample collected yields a *d* larger in magnitude than δ, *w* will be wider than ω. Although the former situation might be desirable, the latter situation might be disappointing because the confidence interval width was larger than desired.[16]

To avoid obtaining a *d* larger in magnitude than the value the sample size procedure is based on with some specified degree of certainty and thus a *w* wider than ω, a modified sample size procedure can be used. Let γ be this desired probability, such that γ represents the probability that *d* will not be larger in magnitude than δ_γ, where δ_γ is the point that *d* will exceed in magnitude only $(1 - \gamma)100\%$ of the time. Thus,

$$p(|d| \leq |\delta_\gamma|) = \gamma, \tag{18}$$

implying that *d* will be contained within the limits $-\delta_\gamma$ and δ_γ $\gamma 100\%$ of the time. Notice that $|d| > |\delta_\gamma|$, when holding everything else constant, will yield a confidence interval wider than ω because confidence intervals for δ become wider as the magnitude (absolute value) of *d* increases.

Because δ can be transformed to λ (using the population

[13] Actually, *d* is a biased estimate of δ. However, for even moderate sample sizes (e.g., 30), the discrepancy between E[*d*] and δ is trivial (Hedges & Olkin, 1985, chapter 5). Although the expected value of *d* given δ and *n* could be substituted for δ in the method, doing so leads to no difference in sample size estimates for almost all realistic situations and will potentially lead to differences only in situations where the procedures yield a very small necessary sample size.

[14] Although the expected value of *w* is ω, this does not imply that 50% of the distribution of *w* will be narrower than specified. In fact, the distribution of *w* can be quite skewed and it is generally the case that more than 50% of the distribution is less than ω.

[15] This is not necessarily true with all effect sizes. For example, the confidence interval width for the squared multiple correlation coefficient is generally at its maximum for values of the sample squared multiple correlation coefficient that are around .30–.40 (Algina & Olejnik, 2000; Kelley, 2006b), depending on the particular condition.

[16] Alternatively, one could use the largest value of δ that would seem plausible in the particular situation for the obtained confidence interval not to be larger than some specified value.

analog of Equation 7) and vice versa (using the population analog of Equation 8), if the λ_γ can be found that satisfies

$$p(|t| \leq |\lambda_\gamma|) = \gamma, \tag{19}$$

then λ_γ can be transformed into δ_γ. The value of λ_γ is thus the value that satisfies the expression

$$\int_{-\lambda_\gamma}^{\lambda_\gamma} f(t_{(\lambda;\nu)})dt = \gamma, \tag{20}$$

where $f(t_{(\lambda;\nu)})$ is the probability density function of the noncentral t distribution, t is the random t variate, and ν is the degrees of freedom ($n_1 + n_2 - 2$ in the present context). Thus, λ_γ is the noncentral value along with its opposite (i.e., its negative value) that excludes $(1 - \gamma)100\%$ of the sampling distribution of t values. Excluding $(1 - \gamma)100\%$ of the sampling distribution of t values that have the widest confidence limits and then using λ_γ in place of λ in the procedure will ensure that no more than $(1 - \gamma)100\%$ of the confidence intervals will be wider than desired, as confidence intervals will be wider than desired if and only if, holding everything else constant, $|t| > |\lambda_\gamma|$, which will occur only $(1 - \gamma)100\%$ of the time because of the definition of λ_γ. The noncentral nature of d, as explained below, makes the development of a sample size planning procedure more difficult than the development of analogous procedures for effects that follow central distributions (e.g., Guenther, 1981; Hahn & Meeker, 1991; Kelley & Maxwell, 2003; Kelley et al., 2003; Kupper & Hafner, 1989).

It is first helpful to compare Equation 20 with the integral form for confidence intervals. The two-sided $(1 - \alpha)100\%$ confidence limits for a noncentral t distribution are defined as

$$\int_{-\infty}^{\lambda_{L_2}} f(t_{(\lambda;\nu)})dt = \alpha/2 \tag{21}$$

and

$$\int_{\lambda_{U_2}}^{\infty} f(t_{(\lambda;\nu)})dt = \alpha/2, \tag{22}$$

where λ_{L_2} and λ_{U_2} are the lower and upper two-sided $(1 - \alpha)100\%$ confidence limits for λ, respectively. Finding λ_{L_2} and λ_{U_2} from Equations 21 and 22 would lead to a $(1 - \alpha)100\%$ confidence interval for λ. (Notice that λ_{L_2} and λ_{U_2} are the values in Figure 1 in which the lower and upper vertical lines, respectively, define the confidence limits.) The one-sided confidence limits for a noncentral t distribution are defined as

$$\int_{-\infty}^{\lambda_{U_1}} f(t_{(\lambda;\nu)})dt = \alpha \tag{23}$$

for a lower $(1 - \alpha)100\%$ confidence interval for λ or as

$$\int_{\lambda_{L_1}}^{\infty} f(t_{(\lambda;\nu)})dt = \alpha \tag{24}$$

for an upper $(1 - \alpha)100\%$ confidence interval for λ, where λ_{U_1} and λ_{L_1} are the upper and lower one-sided confidence limits. Notice how the form of the confidence limits for λ (Equations 21–24) differs from the form of Equation 20. Equations 21–24 each have limits that stretch to positive or to negative infinity, where the lower and the upper limits contain $(\alpha/2)100\%$ of the distribution on each side (for the two-sided confidence intervals) or $\alpha 100\%$ on either side (for the one-sided confidence intervals.) Equation 20 is defined such that there is $(1 - \gamma)100\%$ of the distribution beyond the confidence interval limits as a typical two-sided confidence interval, with the nontypical requirement that the confidence limits are of the same magnitude. Because of the nonsymmetric properties of the noncentral t distribution, there is not an equal proportion beyond each of the limits.

For a given sample size and level of confidence interval coverage, the width of the confidence interval for λ (or δ) is based only on $\hat{\lambda}$ (or d). The rationale for determining λ_γ via Equation 19 is based on this fact, as a negative value or a positive $\hat{\lambda}$ larger in magnitude than λ, the value on which the sample size procedure is based, will lead to a confidence interval wider than desired. Equation 20 can be solved for the value that will ensure only $(1 - \gamma)100\%$ of the distribution of the noncentral parameter will be larger in magnitude than λ_γ. The width of the confidence interval for the noncentral parameter is of the same width regardless of sign. Ultimately, λ_γ will be converted to δ_γ so that δ_γ can be used in place of δ in the standard sample size procedure to ensure that w will be no larger than ω with $\gamma 100\%$ certainty.

Although Equation 20 does not have a straightforward analytic solution, lower and upper bounds can be determined such that a range of values can be searched to find the necessary value of λ_γ that satisfies Equation 20. The confidence limit from a one-sided confidence interval of the form

$$\int_{-\infty}^{\lambda_{U_1}} f(t_{(\lambda;\nu)})dt = \gamma, \tag{25}$$

where λ_{U_1} is the limit of the $\gamma 100\%$ confidence interval, is used as a lower bound for λ_γ. The reason that λ_{U_1} is a lower bound for λ_γ is that using λ_{U_1} in place of λ_γ would lead to more confidence intervals that are wider than desired. The proportion of confidence intervals wider than desired is not only equal to the area beyond λ_{U_1} in Equation 25 but also equal to the proportion of the noncentral distribution beyond $-\lambda_{U_1}$. Thus, the total proportion of confidence intervals wider than desired if λ_{U_1} was used in place of λ_γ when determining the modified sample size would be

$$\int_{-\infty}^{-\lambda_{U_1}} f(t_{(\lambda;\nu)})dt + \int_{\lambda_{U_1}}^{\infty} f(t_{(\lambda;\nu)})dt = p(|t| > |\lambda_{U_1}|), \qquad (26)$$

which is greater than $1 - \gamma$. The first integral is equal to some positive value and the second integral is equal to $1 - \gamma$, necessitating that $p(|t| > |\lambda_{U_1}|) > 1 - \gamma$.

The confidence limits from a $\gamma 100\%$ two-sided confidence interval are of the form

$$\int_{-\infty}^{\lambda_{L_2}} f(t_{(\lambda;\nu)})dt = (1 - \gamma)/2 \qquad (27)$$

and

$$\int_{\lambda_{U_2}}^{\infty} f(t_{(\lambda;\nu)})dt = (1 - \gamma)/2. \qquad (28)$$

Notice here that both confidence limits contain $[(1 - \gamma)/2]100\%$ of the distribution beyond each confidence limit. The upper confidence limit can be used as an upper bound for λ_γ, because (unless $\lambda = 0$) there will be less than $(1 - \gamma)100\%$ of the distribution that is more extreme than $-\lambda_{U_2}$ and λ_{U_2}. This is the case because $[(1 - \gamma)/2]100\%$ of the distribution is greater than λ_{U_2}, and because λ_{L_2} is smaller in magnitude than λ_{U_2}, there must be less than $[(1 - \gamma)/2]100\%$ more extreme than λ_{L_2}. Thus,

$$\int_{-\infty}^{-\lambda_{U_2}} f(t_{(\lambda;\nu)})dt + \int_{\lambda_{U_2}}^{\infty} f(t_{(\lambda;\nu)})dt = p(|t| > |\lambda_{U_2}|), \qquad (29)$$

which is less than $1 - \gamma$. This is the case when δ is positive because the first integral is necessarily smaller than the second integral, and the second integral is equal to $1 - \gamma/2$, necessitating that $p(|t| > |\lambda_{U_2}|) < 1 - \gamma$ (the opposite is true when δ is negative). Because $-\lambda_{U_1}$ and λ_{U_1} bound more than $(1 - \gamma)100\%$ of the distribution, λ_{U_1} must be smaller in magnitude than λ_γ. Because $-\lambda_{U_2}$ and λ_{U_2} bound less than $(1 - \gamma)100\%$ of the distribution, λ_{U_2} must be larger than λ_γ in magnitude. Thus, λ_γ lies somewhere between λ_{U_1} and λ_{U_2}. The closer λ is to zero, the closer λ_γ is to λ_{U_2} in magnitude (as the noncentral t distribution becomes more symmetric). The farther away λ is from zero, the closer λ_γ is to λ_{U_1} in magnitude (as the proportion of the distribution less than $-\lambda_{U_1}$ approaches zero). An optimization routine that iterates over the interval λ_{U_1} to λ_{U_2} searching for λ_γ such that

$$p(t < -\lambda_\gamma) + p(t > \lambda_\gamma) = 1 - \gamma \qquad (30)$$

will yield the λ_γ that can be substituted for λ in the standard procedure. Because the standard procedure is based on δ, λ_γ can be transformed into δ_γ so that δ_γ can replace δ from the standard procedure.

Given the detailed discussion above, a summary follows. First, recall (from Equation 19) that $p(|t| \leq |\lambda_\gamma|) = \gamma$ implies (from Equation 18) that $p(|d| \leq |\delta_\gamma|) = \gamma$. A d larger in magnitude than δ_γ implies $w \geq \omega$ (due to the definition of δ_γ, as it is the value that will be exceeded in magnitude only $[1 - \gamma]100\%$). Basing the sample size procedure on δ_γ will thus ensure that no less than $\gamma 100\%$ of the confidence interval widths will be greater than ω, because at least $\gamma 100\%$ of the sampling distribution of d is less than δ_γ. Because $w \leq \omega$ whenever $|d| \leq |\delta_\gamma|$, planning sample size on the basis of δ_γ will lead to no less than $\gamma 100\%$ certainty that $w \leq \omega$.

A brief conceptual overview. The discussion up to this point has thus far been rather technical. A very general review that is largely conceptual is provided. On the basis of the necessary sample size from the original procedure, where sample size was based on the expected confidence interval width being sufficiently narrow, determine δ_γ. Recall that the value of δ_γ is the value on the scale of δ that is expected to be exceeded in magnitude only $(1 - \gamma)100\%$ of the time. The value of δ_γ is found by solving iteratively for δ_γ (using the noncentral t distribution; see Equation 30) in the following equation:

$$p(d < -\delta_\gamma) + p(d > \delta_\gamma) = 1 - \gamma. \qquad (31)$$

Given δ_γ, substitute δ_γ for δ in the original procedure and solve for sample size as before, by incrementing sample size, beginning where the starting value is now the original sample size, until the $E[w] \leq \omega$. The effect of replacing δ with δ_γ leads to $\gamma 100\%$ of the sampling distribution of d being less than δ_γ. When sample size is based on δ_γ, any d less than δ_γ in magnitude, which will occur $\gamma 100\%$ of the time, will imply an observed confidence interval width less than ω.

Tables of Necessary Sample Size

Although the AIPE approach to sample size planning for the standardized mean difference can be readily carried out using MBESS, selected tables of necessary sample size are provided. The tables are not meant to supplant the computer routines; rather, they are designed so that researchers can quickly estimate the necessary sample size to obtain some desired confidence interval width, possibly with some degree of certainty. The necessary parameters manipulated in the tables are δ, ω, $1 - \alpha$, and γ.

The values of δ used in the tables are 0.05 and 0.10 through 1.00 by 0.10s. The values of the desired full width (ω) used in the tables are 0.10 through 0.50 by 0.05s, and 0.60 through 1.00 by 0.10s. The desired degree of certainty values used in the tables are no γ specification (i.e., $E[w] = \omega$) and γ values of .80 and .99. The confidence level ($1 - \alpha$) was specified at .90, .95, and .99 for the values in Tables 1, 2, and 3, respectively. There are thus a total of 1,386 cells in the tables representing a wide variety of situations. The tables can easily be consulted when considering sample size planning given the goals of AIPE for the standardized mean difference. Of course, not all interesting combinations of δ, ω, γ, and α are tabled. However, for situations not covered in the tables, the Appendix provides computer code using MBESS that show how sample size can be easily determined.

As can be seen from the tables, necessary sample size can become very large for very narrow desired confidence interval widths (e.g., $\omega = 0.10$ and $\omega = 0.15$). Few behavioral, educational, or social scientists will likely have such resources at their disposal to achieve a confidence interval for δ whose expected value is close to 0.10 units wide. Thus, the expectation is that almost all confidence intervals for δ will be wider than 0.10 in practice. Even when the value shown on one of the tables for a particular condition may be distressingly large, the tables will help to illustrate that obtaining a confidence interval less than some desired width may not be practical for a particular situation. Furthermore, because the ultimate goal might be to obtain accurate estimates of the parameters of interest, when this cannot be done satisfactorily in a single study, the use of meta-analysis should be considered. Of course, when an investigator is entering into a new area of research or performs the study in a fundamentally different way compared with previous studies, the use of meta-analysis may be inapplicable. Another possibility is multiple-site studies, an idea that has recently been reproposed (Maxwell, 2004, p. 161), in which several collaborative research teams collect the same type of data under the same (or realistically similar) conditions. The idea of such multisite studies is to spread the burden but reap the benefits of estimates that are accurate and/or statistically significant.

The way in which the tables are used is to first identify the table that corresponds to the confidence level of interest (the $1 - \alpha$ values for Tables 1, 2, and 3 are .90, .95, and .99, respectively). After identifying the correct confidence level, one of the three γ values must be specified (each table consists of three subtables where the particular γ is specified at the top of each subtable). Next, base the sample size calculation on δ (δ is specified in the column headings). Finally, the desired ω must be specified (ω is given in the first column of each subtable). The combination of each of the required values leads to a particular cell in the table that corresponds to the per-group sample size. The total sample size is thus twice the value on the table because the procedure assumes equal per-group sample sizes.

As an example of the use of the tables, suppose that a researcher wishes to obtain a confidence interval with an expected width of 0.50 units when $\delta = 0.80$ at the 95% confidence level. Determining the necessary sample size requires the first subtable (where $E[w]$ is the subtable heading) of Table 2 (where $\alpha = .05$), where $\omega = 0.50$ (the 9th row) and $\delta = 0.80$ (the 10th column). The necessary sample size in this situation is shown to be 133 participants per group (266 total).

Further suppose that the researcher wishes to be 99% certain that the 95% confidence interval will be no larger than 0.50 units wide. Using the third subtable of Table 2 (where $\gamma = 0.99$) and the same procedure just discussed, a sample size of 142 per group (284 total) is necessary. As is demonstrated here, increasing the sample size from the expected value being sufficiently narrow to a narrow confidence interval with a high degree of certainty generally does not necessitate a large increase in sample size relative to the initial value of sample size. This phenomenon is discussed in the next section.

Why Such a Small Change in Sample Size?

In some cases, modifying the sample size so that there is a high probability of obtaining a confidence interval no wider than desired adds a surprisingly small increase in necessary sample size. From the previous example, recall that a 95% confidence interval when $\delta = 0.80$ and $\omega = 0.50$ requires a necessary sample size of 133 per group. When the desired degree of certainty is specified at .99, the necessary sample size required increases to 142 per group (an increase in total sample size of 18; 6.767%). Thus, in this situation, a fairly small increase in sample size has a fairly large effect on the probability of obtaining a sufficiently narrow confidence interval.

Small increases in sample size when going from the expected width being sufficiently narrow to having a degree of certainty that the width will be sufficiently narrow arise for several reasons. First, with reasonably large sample sizes, δ_γ will not be much larger than δ. Recall that the upper $\gamma 100\%$ limit from a one-sided confidence interval is the lower bound

Table 1
Necessary Sample Size per Group for 90% Confidence Intervals for the Population Standardized Mean Difference for Selected Situations

	δ										
ω	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
						γ = E[*w*]					
0.10	2166	2168	2176	2189	2208	2233	2262	2298	2338	2384	2436
0.15	963	964	967	973	982	993	1006	1021	1039	1060	1083
0.20	542	542	544	548	552	559	566	575	585	596	609
0.25	347	347	349	351	354	358	362	368	375	382	390
0.30	241	241	242	244	246	249	252	256	260	265	271
0.35	177	177	178	179	181	183	185	188	191	195	199
0.40	136	136	136	137	138	140	142	144	147	150	153
0.45	107	108	108	109	110	111	112	114	116	118	121
0.50	87	87	88	88	89	90	91	92	94	96	98
0.60	61	61	61	61	62	63	63	64	65	67	68
0.70	45	45	45	45	46	46	47	47	48	49	50
0.80	34	34	34	35	35	35	36	36	37	38	39
0.90	27	27	27	28	28	28	28	29	29	30	31
1.00	22	22	22	22	23	23	23	24	24	24	25
						γ = .80					
0.10	2166	2169	2179	2194	2214	2240	2271	2307	2349	2397	2450
0.15	963	965	969	976	986	997	1012	1028	1047	1068	1092
0.20	542	543	546	550	555	562	570	580	591	603	617
0.25	347	348	350	353	356	361	366	372	379	387	396
0.30	242	242	243	245	248	251	255	259	264	270	276
0.35	178	178	179	181	183	185	188	191	195	199	204
0.40	136	136	137	139	140	142	144	147	150	153	157
0.45	108	108	109	110	111	113	114	116	119	121	124
0.50	88	88	88	89	90	91	93	95	97	99	101
0.60	61	61	62	62	63	64	65	66	68	69	71
0.70	45	45	45	46	47	47	48	49	50	51	53
0.80	35	35	35	35	36	37	37	38	39	40	41
0.90	28	28	28	28	29	29	30	30	31	32	32
1.00	23	23	23	23	23	24	24	25	25	26	27
						γ = .99					
0.10	2169	2173	2185	2202	2225	2253	2287	2326	2370	2420	2475
0.15	965	968	974	982	993	1007	1023	1041	1062	1085	1110
0.20	544	546	550	555	562	570	579	590	602	615	630
0.25	349	350	353	357	361	367	373	380	388	397	407
0.30	243	244	246	249	252	256	261	266	272	279	286
0.35	179	180	182	184	187	190	193	197	202	207	212
0.40	138	138	140	142	144	146	149	152	156	160	164
0.45	109	110	111	113	114	117	119	122	125	128	131
0.50	89	89	90	92	93	95	97	100	102	105	108
0.60	62	63	64	65	66	67	69	71	72	74	77
0.70	47	47	47	48	49	51	52	53	55	56	58
0.80	36	36	37	38	39	40	41	42	43	44	46
0.90	29	29	30	30	31	32	33	34	35	36	37
1.00	24	24	25	25	26	27	27	28	29	30	31

Note. δ is the population standardized mean difference, γ is the desired degree of certainty of achieving a confidence interval for δ no wider than desired, ω is the desired confidence interval width, and E[*w*] is the expected confidence interval width.

Table 2
Necessary Sample Size per Group for 95% Confidence Intervals for the Population Standardized Mean Difference for Selected Situations

	δ										
ω	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
						γ = E[*w*]					
0.10	3075	3078	3089	3108	3135	3170	3212	3262	3320	3385	3458
0.15	1367	1368	1373	1382	1394	1409	1428	1450	1476	1505	1537
0.20	769	770	773	777	784	793	803	816	830	847	865
0.25	492	493	495	498	502	508	514	522	532	542	554
0.30	342	342	344	346	349	353	357	363	369	377	385
0.35	251	252	253	254	256	259	263	267	272	277	283
0.40	193	193	194	195	196	199	201	204	208	212	217
0.45	152	152	153	154	155	157	159	162	164	168	171
0.50	123	124	124	125	126	127	129	131	133	136	139
0.60	86	86	86	87	88	89	90	91	93	95	97
0.70	63	63	64	64	64	65	66	67	68	70	71
0.80	49	49	49	49	49	50	51	52	52	53	55
0.90	38	38	39	39	39	40	40	41	42	42	43
1.00	31	31	31	32	32	32	33	33	34	34	35
						γ = .80					
0.10	3076	3079	3093	3113	3142	3178	3222	3274	3333	3400	3475
0.15	1368	1369	1376	1385	1398	1415	1435	1458	1485	1515	1548
0.20	770	771	774	780	788	797	809	822	837	854	873
0.25	493	494	496	500	505	511	519	527	537	548	561
0.30	343	343	345	348	351	356	361	367	374	382	391
0.35	252	252	254	256	258	262	266	270	276	281	288
0.40	193	193	195	196	198	201	204	208	212	216	221
0.45	153	153	154	155	157	159	162	164	168	171	175
0.50	124	124	125	126	127	129	131	134	136	139	142
0.60	86	86	87	88	89	90	92	93	95	97	100
0.70	64	64	64	65	66	67	68	69	70	72	74
0.80	49	49	49	50	51	51	52	53	54	56	57
0.90	39	39	39	40	40	41	42	42	43	44	45
1.00	32	32	32	32	33	33	34	35	35	36	37
						γ = .99					
0.10	3078	3083	3100	3123	3155	3194	3241	3295	3358	3428	3505
0.15	1370	1372	1381	1392	1407	1426	1448	1473	1502	1534	1569
0.20	772	773	779	786	795	806	819	833	850	869	889
0.25	495	496	500	505	511	518	527	537	548	560	574
0.30	344	345	348	352	356	362	368	375	383	392	402
0.35	253	254	257	260	263	267	272	278	284	290	298
0.40	195	195	197	200	202	206	210	214	219	224	230
0.45	154	155	156	158	161	164	167	170	174	179	183
0.50	125	126	127	129	131	133	136	139	142	146	150
0.60	88	88	89	91	92	94	96	98	101	103	106
0.70	65	65	66	67	69	70	72	73	75	77	80
0.80	50	51	51	52	53	55	56	57	59	61	62
0.90	40	41	41	42	43	44	45	46	47	49	50
1.00	33	33	34	35	35	36	37	38	39	41	42

Note. δ is the population standardized mean difference, γ is the desired degree of certainty of achieving a confidence interval for δ no wider than desired, ω is the desired confidence interval width, and E[*w*] is the expected confidence interval width.

Table 3
Necessary Sample Size per Group for 99% Confidence Intervals for the Population Standardized Mean Difference for Selected Situations

	δ										
ω	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
						γ = E[*w*]					
0.10	5310	5315	5335	5368	5415	5474	5547	5633	5733	5846	5972
0.15	2360	2363	2371	2386	2407	2433	2466	2504	2548	2599	2654
0.20	1328	1329	1334	1342	1354	1369	1387	1409	1434	1462	1493
0.25	850	851	854	859	867	876	888	902	918	936	956
0.30	590	591	593	597	602	609	617	626	638	650	664
0.35	434	434	436	439	442	447	453	460	469	478	488
0.40	332	333	334	336	339	343	347	353	359	366	374
0.45	263	263	264	266	268	271	274	279	284	289	295
0.50	213	213	214	215	217	219	222	226	230	234	239
0.60	148	148	149	150	151	153	155	157	160	163	166
0.70	109	109	109	110	111	112	114	116	118	120	122
0.80	83	84	84	84	85	86	87	89	90	92	94
0.90	66	66	66	67	67	68	69	70	71	73	74
1.00	54	54	54	54	55	55	56	57	58	59	60
						γ = .80					
0.10	5311	5317	5339	5375	5423	5485	5561	5649	5751	5866	5994
0.15	2361	2364	2374	2391	2413	2441	2475	2514	2560	2612	2669
0.20	1329	1330	1336	1346	1359	1375	1394	1417	1443	1472	1505
0.25	851	852	856	862	870	881	893	908	925	944	965
0.30	591	592	595	599	605	613	621	632	644	657	672
0.35	434	435	437	441	445	451	457	465	474	484	495
0.40	333	333	335	338	341	346	351	357	363	371	379
0.45	263	264	265	267	270	273	278	282	288	294	301
0.50	213	214	215	217	219	222	225	229	234	239	244
0.60	148	149	150	151	153	155	157	160	163	166	170
0.70	109	109	110	111	112	114	116	118	120	123	126
0.80	84	84	85	85	86	88	89	91	93	95	97
0.90	66	67	67	68	69	70	71	72	73	75	77
1.00	54	54	54	55	56	57	58	59	60	61	63
						γ = .99					
0.10	5314	5322	5348	5388	5440	5506	5585	5677	5783	5902	6034
0.15	2364	2368	2381	2400	2424	2455	2491	2534	2582	2636	2696
0.20	1331	1334	1341	1353	1367	1385	1407	1431	1459	1491	1525
0.25	853	855	860	868	878	890	904	920	939	959	982
0.30	593	594	599	604	612	620	630	642	655	670	686
0.35	436	437	441	445	451	457	465	474	484	495	507
0.40	334	336	338	342	346	352	358	365	373	381	391
0.45	265	266	268	271	275	279	284	290	296	303	311
0.50	215	216	218	220	223	227	231	236	241	247	253
0.60	150	151	152	154	156	159	162	166	170	174	178
0.70	111	111	113	114	116	118	121	123	126	129	133
0.80	85	86	87	88	90	91	93	96	98	101	103
0.90	68	68	69	70	72	73	75	76	79	81	83
1.00	55	56	57	58	59	60	61	63	65	66	68

Note. δ is the population standardized mean difference, γ is the desired degree of certainty of achieving a confidence interval for δ no wider than desired, ω is the desired confidence interval width, and E[*w*] is the expected confidence interval width.

for δ_γ and the upper limit from a two-sided confidence interval is the upper bound for δ_γ. In the example, the upper limit of a 99% one-sided confidence interval for δ is 1.0959. The upper limit from a 99% two-sided confidence interval for δ is 1.1277. Substituting these values for 0.80 as if they were δ in the standard procedure leads to necessary sample sizes of 142 and 144 for the upper one-sided and two-sided confidence intervals, respectively. The actual δ_γ value in this case is 1.1073, which leads to the necessary sample size of 142 per group. Holding constant δ, the larger the required sample size, the closer δ_γ will be to δ.

When the common population variance is unity and thus $\delta = \mu_1 - \mu_2$, the standardized and unstandardized confidence intervals for the mean difference estimate the same quantity. Confidence intervals for the population quantities thus try to bracket the same population value. Comparing the confidence interval widths between the two methods of confidence interval formation shows that for the same sample size, the width is much less variable for the standardized mean difference than it is for the unstandardized mean difference. Figure 2 illustrates the standard deviation of confidence interval widths calculated in a population in which the common variance is unity with three different values of $\mu_1 - \mu_2$ and for per-group samples sizes of 3 to 25. The three curves show that the standard deviation of the confidence interval width for the standardized mean difference changes as a function of $\mu_1 - \mu_2$, holding constant σ at 1. The unstandardized mean difference is unaffected by changes in the mean difference because the observed mean difference does not determine the confidence interval width. This is the case because δ changes as a function of $\mu_1 - \mu_2$, holding constant σ, but (for normally distributed data) the unstandardized mean difference is independent of σ.

It is well-known that the width of the confidence interval for δ becomes larger, holding everything else constant, for larger values of *d*. What does not seem to be well-known, however, is that the confidence interval does not become much wider as δ becomes larger over what is thought to be the typical range of *d* in the behavioral, educational, and social sciences. For example, the 95% confidence interval when *d* is 0.05 with $n_1 = n_2 = 30$ is −0.4564 to 0.5559, whereas it is 0.5052 to 1.5868 when *d* is 1.05. Although the limits are much different, their widths are relatively close. The width of the former example is 1.0123 and for the latter

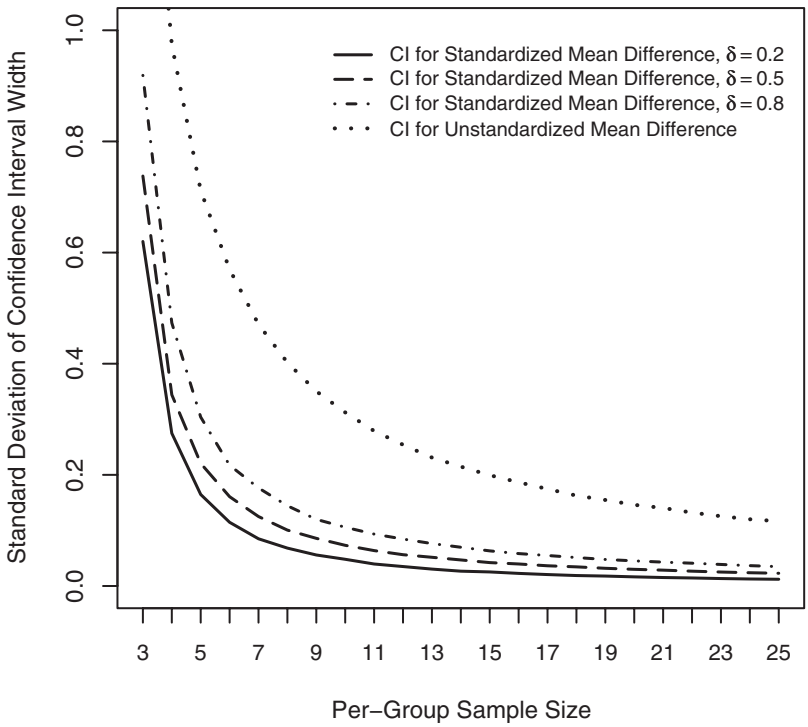

Figure 2. Standard deviation of confidence interval widths for the standardized and unstandardized mean difference when the common within-group standard deviation is unity. Regardless of the size of the mean difference, the width of the confidence interval (CI) for the unstandardized mean difference does not change because the width depends only on sample size and the estimated common standard deviation. The standard deviation of the CI widths was calculated from a Monte Carlo simulation study with 25,000 replications. All of the assumptions of the CI procedures were satisfied, and thus the variability of the CI widths represent the theoretical CI variability.

example is 1.0816. When the per-group sample size is 30, values of d between 0.05 and 1.05 have a 95% confidence interval width between 1.0123 and 1.0816, a remarkably small range given the large range of d (the confidence interval width has a range of 0.0693 whereas d has a range of 1).

Because the expected confidence interval width does not change much as δ gets larger, substituting δ_γ from the modified sample size procedure when a degree of certainty is incorporated for δ in the standard procedure will thus not generally lead to a much larger sample size. Furthermore, changes in sample size follow a step function (because sample size can subsume only whole numbers), whereas values of δ change following a smooth continuous function. Thus, a range of δ values will have the same necessary sample size, holding everything else constant. For example, δs between 0.7659 and 0.8070 all lead to a necessary sample size of 133 when $\omega = 0.5$ for a 95% confidence interval. When incorporating a desired degree of certainty of $\gamma = .99$, where δ_γ is 1.1073, δs between 1.0814 and 1.1106 all lead to a necessary sample size of 142. These ideas taken together explain why sample size does not increase as much as might be expected when incorporating a desired degree of certainty.

Given the present discussion, a note of caution is warranted. Because the sample size may not change much from the standard approach even when a large degree of certainty parameter is specified, some researchers may get the impression that using the modified sample size is unnecessary. Although this may be largely true for very small δ values, ignoring γ and the modified sample size procedure cannot be recommended. This is the case because even though the sample size may not change much, the proportion of confidence intervals that are sufficiently narrow may be much less than desired when only the standard sample size procedure is used. This is the case because the confidence intervals are not very variable (and thus from sample to sample they tend to be close in value) and because the expected width will increase (even if it does so by only a small amount). Combining the relatively small variability for the confidence interval widths and the expected width being wider for larger values of δ has the effect that even a small increase in sample size can lead to a much larger proportion of the sampling distribution of confidence interval widths being less than the value specified. Furthermore, as δ gets larger, the difference between the standard and modified sample sizes becomes more pronounced and can lead to very large differences in necessary sample size.

Sample Size Planning for Power Versus Accuracy

As has been implicit in the previous discussion, there are fundamental differences in the goals of sample size planning for power and sample size planning for accuracy. The power analytic approach has as its goal rejecting a false null hypothesis with some specified probability. The AIPE approach has as its goal obtaining an accurate estimate, operationalized by a narrow confidence interval. An accurate estimate need not be significant and a significant estimate need not be accurate. Although each of the approaches is valuable, each is designed to answer a different question. As is shown in the following paragraphs, necessary sample size can be very different depending on the particular question asked.

Kelley and Maxwell (2003) compared power and accuracy for a regression coefficient, and Kelley et al. (2003) compared power and accuracy for the unstandardized mean difference. Figure 3 shows the necessary sample sizes for power levels of .50, .80, and .95 and desired confidence interval widths of 0.35, 0.25, and 0.15 for δ values between 0.10 and 0.50. Although the desired levels of power and desired confidence interval widths are arbitrary, they are thought to be reasonable values for comparison purposes. Notice that the abscissa begins shifted 0.10 units from 0. Values of δ closer to 0 lead to very large necessary sample sizes for the power analytic approach. Similarly, as desired power increases arbitrarily close to 1 (especially for small δ values) and as desired confidence interval width decreases arbitrarily close to 0, necessary sample sizes also become very large.

As can be seen in the figure, as the size of the effect increases, necessary per-group sample size for a desired degree of power decreases, holding everything else constant. Not obvious from a casual glance at the figure is the fact that the necessary per-group sample size for AIPE increases as δ gets larger, holding everything else constant. However, the rate of decreasing sample size for power as the effect gets larger is much faster than the rate of increasing sample size for AIPE as the effect gets larger for the standardized mean difference. For example, when the desired confidence interval width is 0.25 and $\delta = 0.10$ for a 95% confidence interval, the necessary per-group sample size is 493, yet when $\delta = 0.50$, the necessary per-group sample size is 508. When the desired power is .80 and $\delta = 0.10$ when the Type I error rate is .05, the necessary per-group sample size is 1,571, yet when $\delta = 0.50$, the necessary per-group sample size is 64. Planning sample size from a power analytic approach is a fundamentally different task than planning sample size from an AIPE approach. As is illustrated in Figure 3, the power analytic approach and the AIPE can lead to very different answers to the question "What size sample do I need?"

Discussion

In the context of comparing the means of two groups, the confidence interval for the population group mean difference is often of interest. In many cases in the

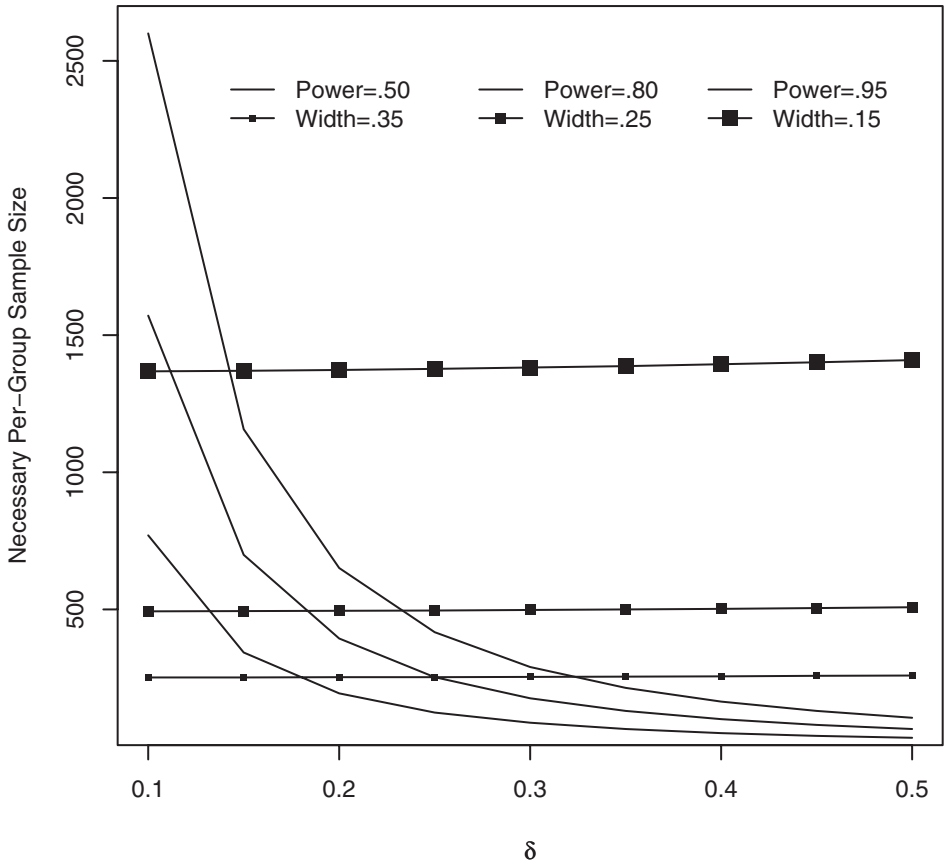

Figure 3. Comparison of the power analytic and the accuracy in parameter estimation approach to sample size planning for desired power of .50, .80, and .95 and for desired confidence interval widths of 0.15, 0.25, and 0.35 when the Type I error rate is .05.

behavioral, educational, and social sciences, the standardized mean difference is more appropriate than the unstandardized mean difference because of the arbitrariness of many measurement scales. Using point estimates in the construction of confidence intervals for population parameters that lead to wide intervals does not shed much light on the population parameter of interest. The value of the population parameter is often the driving force of research, and thus an accurate estimate of the parameter is generally the most useful information that can be obtained. Holding the confidence interval coverage constant, the narrower the confidence interval, the more information about the population parameter of interest is obtained. Learning the value of the parameter, whatever it may be, is generally more informative than the results of a null hypothesis significance test. Even when what is of interest is the direction of the effect, something null hypothesis significance tests are especially helpful at discerning, learning the value of the parameter is informative because knowing the value of the parameter implies you know its direction. Given that, sample size planning should often be considered from the AIPE perspective (Kelley & Maxwell, 2003; Kelley et al., 2003), which has as its goal obtaining narrow confidence intervals corresponding to accurately estimated parameters.

In the present article, we developed methods that can be used to determine necessary sample size so that the expected width of the confidence interval for the standardized mean difference will be sufficiently narrow, optionally with some desired degree of certainty that the obtained interval will be sufficiently narrow. The methods discussed were implemented in the freely available MBESS (Kelley, 2006a) package for the R software program (R Development Core Team, 2006). Tables have been provided so that researchers can quickly determine or approximate the necessary sample size when the goal is to obtain a narrow confidence interval for δ. It is our hope that those planning sample size will consider the AIPE approach, either instead of or in addition to the power analytic perspective. Embracing the AIPE perspective of sample size planning will lead to a better understanding of the particular phenomenon of

interest than will approaching sample size planning solely from a power analytic perspective.

References

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods, 10,* 317–328.

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

Bond, C. F., Jr., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods, 8,* 406–418.

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

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

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

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

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 532–574.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

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

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* Mahwah, NJ: Erlbaum.

Guenther, W. C. (1981). Sample size formulas for normal theory *t* tests. *The American Statistician, 35,* 243–244.

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

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

Hedges, L. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128.

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hogben, D., Pinkham, R. S., & Wilk, M. B. (1961). The moments of the non-central *t*-distribution. *Biometrika, 48,* 465–468.

Holtzman, W. H. (1950). The unbiased estimate of the population variance and standard deviation. *American Journal of Psychology, 63,* 615–617.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed., Vol. 2). New York: Wiley.

Johnson, N. L., & Welch, B. L. (1940). Applications of the non-central *t*-distribution. *Biometrika, 31,* 362–389.

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrapping as an alternative to parametric confidence intervals. *Educational and Psychological Measurement, 65,* 51–69.

Kelley, K. (2006a). MBESS: Methods for the behavioral, educational, and social sciences. (Version 0.0–3) [Computer software and manual]. Retrievable from http://www.cran.r-project.org/

Kelley, K. (2006b). *Sample size planning for the squared multiple correlation coefficient: Accuracy in parameter estimation via narrow confidence intervals*. Manuscript submitted for publication.

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

Kelley, K., & Maxwell, S. E. (in press). Sample size planning for multiple regression: Power and accuracy for omnibus and targeted effects. In J. Brannon, P. Alasuutari, & L. Bickman (Eds.), *Sage handbook of social research methods.* Thousand Oaks, CA: Sage.

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

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects?* Beverly Hills, CA: Sage.

Kupper, L. L., & Hafner, K. B. (1989). How appropriate are popular sample size formulas? *The American Statistician, 43,* 101–105.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician, 55,* 187–193.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Mahalanobis, P. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India, 12,* 49–55.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–426). Mahwah, NJ: Erlbaum.

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Erlbaum.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236,* 333–380.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

R Development Core Team. (2006). R: A language and environment for statistical computing [Computer software and manual]. Retrieved from http://www.r-project.org

Rosenthal, R. (1993). Cumulative evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 519–559). Hillsdale, NJ: Erlbaum.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73–83.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of methodological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *Methodological and quantitative issues in the analysis of psychological data* (pp. 199–228). Mahwah, NJ: Erlbaum.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61,* 605–632.

Smithson, M. (2003). *Confidence intervals.* Thousand Oaks, CA: Sage.

Sokal, R. R., & Braumann, C. A. (1980). Significance tests for coefficients of variation and variability profiles. *Systematic Zoology, 29,* 50–66.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9,* 164–182.

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. (1994). *Kendall's advanced theory of statistics: Distribution theory* (6th ed., Vol. 1). New York: Wiley.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31,* 25–32.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6,* 371–386.

Vaske, J. J., Gliner, J. A., & Morgan, G. A. (2002). Communicating judgments about practical significance: Effect size, confidence intervals and odds ratios. *Human Dimensions of Wildlife, 7,* 287–300.

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

(Appendix follows)

Appendix

Applying the Methods With the MBESS R Package

All of the methods and procedures discussed and the algorithms presented can easily be implemented in the Methods for the Behavioral, Educational, and Social Sciences (MBESS) R package (Kelley, 2006a). This Appendix provides a brief overview of the way in which the functions can be used. Those not familiar with R will see that R is a command-driven language, in which various functions are input directly into the R program. Before using the functions contained within MBESS, one must load the MBESS package into the current R session. Loading MBESS is accomplished with the command `library` at the command prompt (>) after the package has been installed: `library(MBESS)`. The easiest way to install a package is to use the Install Package(s) feature under the Packages menu.

Confidence Intervals for Noncentral *t* Parameters

For constructing confidence intervals for the noncentral parameter from a noncentral *t* distribution, the `conf.limits.nct()` function can be used. The lower and upper critical values from the noncentral *t* distribution for the example used in Figure 1 are returned by specifying the following arguments in the `conf.limits.nct()` function:

```
conf.limits.nct(ncp=2.7951, df=18,
conf.level=0.95),
```

where `ncp` is the (estimated) noncentral parameter, `df` is the degrees of freedom, and `conf.level` is the desired level of confidence ($1 - \alpha$). Execution of this function yields 0.6038 and 4.9227 for the lower and upper 95% confidence limits for λ, respectively.

Confidence Intervals for the Standardized Mean Difference

Given the one-to-one relation between λ and δ and the confidence interval transformation principle previously discussed, the confidence limits for δ can be found by transforming the confidence limits of λ given the relation specified in Equation 8. Alternatively, the `ci.smd()` function can be used directly to determine the confidence limits for δ. The lower and upper critical value from the example used in Figure 1 are returned using the following specifications:

```
ci.smd(smd=1.25, n.1=10, n.2=10,
conf.level=0.95),
```

where `smd` is the standardized mean difference (i.e., *d*), `n.1` and `n.2` are the per-group sample sizes for Groups 1 and 2, respectively, and `conf.level` is the desired level of confidence. Application of this function yields 0.2700 and 2.2015 for the lower and upper 95% confidence limits for δ, respectively.

Computing Necessary Sample Size for the Standardized Mean Difference From the AIPE Perspective

The function `ss.aipe.smd()` determines the necessary sample size so that the expected value of $w \leq \omega$ for the standardized mean difference. An example call to the `ss.aipe.smd()` function is given as follows:

```
ss.aipe.smd(delta=.50,
conf.level=.95, width=.30),
```

which yields $n_1 = n_2 = 353$, where `delta` is the population standardized mean difference, `conf.level` is the level of confidence, and `width` is the desired confidence interval width. Thus, if $\delta = .50$, the confidence interval coverage is set to .95 and the width of the interval is specified as .30, a per-group sample size of 353 (706 total) is necessary.

The `degree.of.certainty` parameter can be specified in the `ss.aipe.smd()` function so that there will be some desired degree of certainty (i.e., γ) that the observed confidence interval is sufficiently narrow. Setting the degree of certainty to .99 and using the `ss.aipe.smd()` function as

```
ss.aipe.smd(delta=.50, conf.level=.95, width=.30,
degree.of.certainty=.99)
```

yields a necessary sample size of 362 (724 total).

Sensitivity Analysis for the Standardized Mean Difference

Sensitivity analysis to assess the effect of misspecifying δ on the width of the confidence interval can be performed with the `ss.aipe.smd.sensitivity()` function. The function `ss.aipe.smd.sensitivity()` allows one to specify the true population δ and an estimated but incorrect δ, so that the effect of misspecifying δ on the width of the confidence interval can be empirically determined. The function performs a simulation whereby the empirical findings regarding the width of the confidence interval can be determined. Visualization of the results of the simulation can be very helpful for determining how discrepant the assumed value can be from δ to still have an acceptably narrow confidence interval for δ. The

`ss.aipe.smd.sensitivity()` function can be specified as

```
ss.aipe.smd.sensitivity(true.delta=1.00,
estimated.delta=1.25, desired.width=.50,
certainty=.85, conf.level=0.95, G=10000),
```

where `true.delta` and `estimated.delta` are the true and the estimated δ values, `desired.width` is the desired confidence interval width, `certainty` is the desired degree of certainty, `conf.level` is the desired confidence level ($1 - \alpha$), and `G` is the number of replications that take place within the simulation study. Instead of specifying `estimated.delta`, one can select a particular sample size using `selected.n`, so that the properties of the confidence interval can be readily determined with a specified δ value and a specific per-group sample size.

Received December 5, 2005
Revision received August 9, 2005
Accepted August 24, 2006 ■