Running Head: Sample Size Planning

Sample Size Planning with Effect Size Estimates

Jeremy C. Biesanz

University of British Columbia

Sheree M. Schrager

Childrens Hospital Los Angeles

Abstract

The use of effect size estimates in planning the sample size necessary for a future study can introduce substantial bias in the sample size planning process. For instance, the uncertainty associated with the effect size estimate may result in average statistical power that is substantially lower than the nominal power specified in the calculation. The present manuscript examines methods for incorporating the uncertainty present in an effect size estimate into the sample size planning process for both statistical power and accuracy in parameter estimation (i.e., desired confidence interval width). Several illustrative examples are provided along with computer programs for implementing these procedures. Discussion focuses on the choices among different approaches to determining statistical power and accurate parameter estimation when planning the sample size for future studies.

Sample Size Planning with Effect Size Estimates

When designing a study or an experiment, a number of critical decisions need to be made based on incomplete or uncertain information before data collection begins. One of these critical decisions is planning *a priori* the sample size needed to achieve the researcher's goal. The goal of the sample size planning process may be adequate statistical power – the probability of correctly rejecting a false null hypothesis. Alternatively, the goal may be accurate parameter estimation – estimating the effect size with a specified level of precision. If a study has moderate to low statistical power, then there is a moderate to high probability that the time and resources spent on the study will yield a nonsignificant result. If a study results in a wide confidence interval for the effect size, then regardless of statistical significance, little information is gleaned regarding the actual magnitude of the effect. Consequently, it is good research practice – and indeed required by many granting agencies – to plan the sample size for a prospective study that will achieve the desired goal(s).

At first glance, study design is relatively simple, if computationally intensive, as there is a deterministic relationship among the criteria (i.e., statistical power or confidence interval width), sample size, the specified critical level (i.e., the Type I error rate α or the confidence level), and the population effect size.[1] If any three of these quantities are known, then the fourth can be calculated exactly. In practice, the sample size for a prospective study is often calculated by setting the desired level of statistical power at a particular value such as .80 (e.g., Cohen, 1988, 1992) or the width of the standardized mean difference confidence interval to be a certain level (e.g., .10 or .20) for a specified level of α. The necessary sample size for power may then be approximated from Cohen's (1988) tables or determined exactly using available software such as, for example, *G*Power* (Erdfelder, Faul, & Buchner, 1996), *Statistica* (Steiger, 1999), or *SAS* (O'Brien, 1998; SAS Institute Inc., 2003) among others. The necessary sample size for a

specified confidence interval for the standardized mean difference can be determined, for instance, from tables presented in Kelley & Raush (2006) or exactly from Kelley's (2007) MBESS program available in *R (R* Development Core Team, 2006). As straightforward as this may initially seem, the fine print on this process contains critical details that are often glossed over (e.g., see Lenth, 2001, for a practical discussion of the issues involved in study design). Both statistical power and accurate parameter estimation require an estimated or hypothesized population effect size (e.g., see Muller & Benignus, 1992, p. 217). The requisite sample size calculated in this manner is conditional on the specified population effect size. In other words, the logic of this manner of power calculation is as follows: *Assuming* the population effect size is a specified value, *then* with sample size *n*, power will be .80. This presents a certain irony – if the effect size is already known, why conduct the study? In practice, the effect size is not known precisely and exactly, but *estimates* of the effect size may be available.

The present manuscript examines the relationships among statistical power and accurate parameter estimation, sample size, and estimates of the effect size. Specifically, we first examine the impact of estimated effect sizes on statistical power and then discuss how to use prior information and probability distributions on the effect size to increase design efficiency, improve confidence intervals, and better achieve the desired level of statistical power or accurate parameter estimation. The manuscript is organized as follows: First we discuss traditional approaches to sample size planning and how the use of standard effect size estimates without incorporating information about uncertainty can bias statistical power. We then discuss the benefits and rationale for incorporating a Bayesian perspective in the study design process and illustrate how to use this approach for statistical power calculations given effect size estimates with (a) no prior information and (b) with prior information such as from a meta-analysis. We then discuss this approach when the criterion is accurate parameter estimation, i.e., a desired confidence

interval width. Finally, we discuss conceptual and practical issues related to sample size planning. Note that definitions and notation are summarized in Table 1A and expanded upon in footnote 2 and more extensive analytical details, as well as additional equations, are sequestered within footnotes.

Approaches to Specifying the Population Parameter

The population effect size parameter, for instance, δ, is a necessary input to the process of determining the sample size required for the desired level of statistical power or accurate parameter estimation. Since the parameter is not known, how then does one proceed? Consider how sample size planning is often initially taught. Two of the more widely adopted introductory statistics texts in psychology (Gravetter & Wallnau, 2006; Howell, 2007) present three approaches to determining the population effect size to use as the basis of planning sample size: (1) assessment of the minimal effect size that is important to detect, (2) Cohen's conventions, and (3) prior research.

1. Minimally important effect size. If the metric of the dependent variable is not arbitrary (e.g., blood pressure, cholesterol level, etc.) and there is a clear and well-defined clinical therapeutic level on that dependent variable, then sample size planning can be based around that clinical level. Mueller and colleagues present methods for power analysis to detect a specified level of change on the dependent variable that incorporates the uncertainty associated with estimates of the population standard deviation (e.g., Coffey & Muller, 1999; Muller, LaVange, Ramey, & Ramey, 1992; Taylor & Muller, 1995a).

In psychology, the dependent variable often is not measured on such clean ratio level scales, clearly demarked therapeutic levels of change are not known, and consequently standardized effect sizes may be the only available metric. The use of

standardized effect sizes in sample size planning is not without criticism (e.g, Lenth, 2001). In part, this criticism reflects concern about conflating the magnitude of an effect with actual importance – not unlike the confusion behind declaring that because two groups are statistically significantly different, that the difference between the two groups is therefore practically significant. Yet in the absence of any viable alternative, the use of a standardized effect size often is the only option. However, in this context, the choice of which standardized effect size is sufficiently important to detect is arbitrary and may vary across researchers. This naturally leads to considering qualitative interpretations of the magnitude of standardized effect sizes and Cohen's conventions.

2. Cohen's conventions. Cohen provided rough qualitative interpretations of standardized effect sizes corresponding to small, medium, and large effects. For the standardized mean difference these are .2, .5, and .8, and for the correlation these are .1, .3, and .5, respectively.[2] Examining statistical power for small, medium, and large effects is essentially equivalent to considering the entire power curve – the graph of how power changes as a function of effect size for a given sample size. Examining a power curve, although informative about the power-effect size relationship, does not provide a systematic or a formal basis for how to proceed. For example, a researcher examining a traditional power curve that displays statistical power as a function of the effect size for a given sample size may conclude that power is quite reasonable for a medium to largish effect size. Another researcher may look at the same curve and conclude that the study is grossly *over*powered given a large effect size. Yet another may conclude the study is grossly *under*powered given a medium effect. This is an extremely subjective decision-making process with little formal justification for the choice of the effect size on which to

base decisions. Indeed, many may not conduct power analyses at all given how subjective the process may appear.

3. Prior research. Following the recommendations of Wilkinson and the APA Task Force on Statistical Inference (1999), researchers have been encouraged to supplement the traditional *p*-values with effect size estimates and confidence intervals. Providing and examining effect sizes and corresponding confidence intervals helps shift the research question from solely asking, "Is the effect different from zero?" to inquiring as well, "What is the estimated magnitude of the effect and the precision of that estimate?" (see Ozer, 2007, for a discussion of interpreting effect sizes). As a consequence of this shift in reporting practice, effect size estimates are more readily accessible. When engaged in sample size planning for a future study, researchers often will have estimate(s) of the effect size at hand. These may come from previously published research, extensive internal pilot studies, conference presentations, unpublished manuscripts, or other sources. In this manuscript, we focus on this case – when there is *some* effect size estimate available that is relevant to the future study. That such estimates should be used in the sample size planning process is almost self-evident. For a researcher to assert the goal of achieving, for example, sufficient statistical power for a small to medium effect size (e.g., $\delta = .30$) rests on the premise that a small-medium effect is actually meaningful. Even if that premise is warranted, using that criterion may be grossly inefficient if there is evidence that the effect size is in reality larger. This criticism holds as well for dependent variables measured on well-defined scales with clear therapeutic levels change – if there is evidence that the effect is substantially larger than the minimum change needed to produce a therapeutic effect, designing a study to

detect that minimum change may be inefficient and costly. All available information should be used in the study design process.

The question that arises naturally is *how* to use that effect size estimate. As we will illustrate, naïvely using effect size estimates as their corresponding population parameters may introduce substantial bias into the sample size planning process.

How Naïve use of Effect Size Estimates Biases Statistical Power

We now consider the impact of using effect size estimates in power calculations in a straightforward manner and how this can lead to bias in the actual average level of power. Consider a hypothetical researcher who wishes to replicate a two-group study in which an intervention designed to change attitudes towards littering is implemented and littering behavior is subsequently measured. The original study involved a total of 50 participants (i.e., $n = 25$ per group) with an estimated standardized mean difference of $d = .50$ between the treatment and the control conditions. It seems quite reasonable to use this effect size estimate to conduct a power analysis to determine the sample size needed for the subsequent study to have adequate statistical power. Indeed, using this estimated effect size our researcher determines that 64 subjects per group are needed to have power of .80 under the assumption that $\delta = .50$.

At first glance it would seem logical to use the effect size estimates to guide power analyses in this manner. Although sometimes sample estimates are above the population parameter and sometimes below, shouldn't statistical power calculated on effect size estimates average to .80 across different sample realizations of the same population effect size? Interestingly, the answer is no. Even if the effect size estimator is unbiased with a symmetric sampling distribution, sample size calculations based on that effect size estimate can result in average statistical power that is substantially *lower* than the nominal level used in the calculations. Bias in estimated statistical power from the use of

estimated effect sizes emerges from the asymmetrical relationship between sample effect size estimates and actual statistical power (e.g., Gillett, 1994, 2002; Taylor & Muller, 1995b). This bias may in fact be quite substantial. Observed estimates below the population effect size will result in suggested sample sizes for future studies that result in power approaching 1. In contrast, effect size estimates above the population value suggest sample sizes for future studies that drop to power down to α, the Type I error rate, which is also the lower bound for power. This asymmetrical relationship results in *average* actual power across the sampling distribution of the effect size estimate that is less than the nominal power calculations based on each observed effect size estimate.

To understand more clearly how average statistical power can differ from the nominal statistical power, consider the following thought experiment. A large number of researchers all examine the exact same effect using the same procedure, materials, and drawing random samples from the same population where the effect size is $\delta = .20$ with $n_1 = n_2 = 25$. Thus, each researcher has an independent sample from the sampling distribution of the standardized mean difference and uses this observed standardized mean difference to plan the required sample size necessary to achieve power of .80. Suppose one researcher observes $d = .30$ and uses this information as if it were the population effect size in a standard power analysis program, concluding that n should be 176 per group in the subsequent study to achieve power of .80. Another researcher observes $d = .15$ and determines that n should be 699 per group. Yet another researcher observes $d = .60$ and determines that n should be 45 per group, and so on. Researchers who observe a larger d will determine that they require a *smaller* sample size than those researchers who observe a smaller d. Figure 1 graphs the sampling distribution of the standardized mean difference based on $\delta = .20$ and $n = 25$, the sample size each

hypothetical researcher determines is needed for the subsequent study when the observed effect size (*d*) is used as the population parameter to plan sample size, and finally the actual statistical power for each researcher's subsequent study based on that sample size given that δ is actually .20. Only when the sample estimate is $|d| = \delta = .20$, the population standardized mean difference, does the actual power for a subsequent replication equal .80. Thus large *observed* standardized mean differences result in *low* statistical power since researchers will conclude that they require a relatively small sample size for the subsequent study.

On average, across the sampling distribution of the effect size estimate for this example, statistical power is only .61 – even though each sample size calculation was based on a nominal power of .80. Average statistical power is calculated by numerically integrating over the product of the sampling distribution of the standardized mean difference and the power curve in Figure 1. This bias in average statistical power is reduced both when the initial effect size estimate is measured with greater precision (e.g., based on larger sample sizes) and when the population effect size is larger. This can be seen in Figure 2, which graphs the average statistical power across the sampling distribution of the standardized mean difference as a function of the population standardized mean difference and the sample size.

The bias in statistical power is defined as the difference in the average statistical power across the sampling distribution and the nominal power used for each power calculation to determine sample size. The implications of blindly using effect size estimates in statistical power calculations and the resulting bias warrant incorporating information regarding the sampling variability of the effect size estimate into the study design process.

Clearly, the simple use of an effect size estimate in the sample size planning process is not justifiable. We now discuss how to use effect size estimates – and all of the information associated with the estimate – in the sample size planning process.

A Formal Basis for Sample Size Planning using Effect Size Estimates

A population effect size is a necessary input to the process when planning the sample size for a future study, whether the goal is a specified level of power or a specified level of precision for the effect size estimate. The present manuscript adopts a Bayesian perspective on the population effect size during the study design process; *however*, inferences and/or estimation *are based solely* on the data collected in the future study. Further discourse on amalgamating Bayesian and frequentist perspectives is deferred to the discussion.

Adopting the Bayesian perspective for considering the population effect size is a pragmatic solution to the vexing problem of how to use *estimates* of effect sizes in the sample size planning process. As we have seen, simply using the effect size estimate as a proxy for the parameter value results in levels of statistical power that are lower than specified in the planning process. In contrast to examining a single parameter value, the Bayesian perspective instead provides a probability distribution of parameter values known as the posterior distribution. The posterior distribution is the distribution of plausible parameter values given the observed effect size estimate and is a function of the likelihood of the observed data given a parameter value and the prior distribution of the parameter value.[3] In other words, the posterior distribution provides a whole distribution of parameter values to consider during the planning process.

Using the Bayesian framework, we can therefore perform a statistical power calculation or accuracy in parameter estimation calculation based on a given sample size and examine the consequent distribution of statistical power or interval precision as a function of the posterior distribution. In this way, the Bayesian framework provides a formal mechanism for incorporating the imprecision associated with the effect size estimate when planning sample size. The specific steps are as follows:

1. Determine the posterior distribution of the population effect size parameter given observed data (e.g., an effect size estimate). The posterior distribution can be thought of as representing the uncertainty associated with the observed effect size estimate as it is the distribution of plausible values of the parameter given the observed data.

2. The posterior distribution is used as input in the study design process to determine the posterior *predictive* distribution of the test-statistic for a specified future sample size. This represents the distribution of test-statistics for a given sample size across the plausible values for the population parameter.

3. The posterior predictive distribution of a test-statistic thus incorporates the uncertainty associated with estimated effect sizes. It is straightforward to then determine the sample size needed to determine expected (average) statistical power or desired confidence interval width. For instance, power is simply the proportion of the posterior predictive distribution that is larger in magnitude than the critical *t*-values.

Expected power (EP), determined by averaging across the posterior distribution, provides a formal basis for making definitive statements about the probability of the future study reaching the desired goal (i.e., significance or accurate parameter

estimation). However, by adopting a Bayesian perspective, there is an implicit change in the nature and interpretation of probabilities from conventional power calculations. To illustrate, consider the earlier example where a researcher has an effect size estimate of d = .50 based on n = 25. The traditional power calculation based on $\delta = d$ = .50 resulted in n = 64 to achieve power of .80. This is a probability statement about repeatedly conducting the exact same experiment an infinite number of times on samples from the same population: 80 percent of future studies based on n = 64 will be significant if δ = .50. In contrast, the Bayesian concept of expected power provides a different probability. As we illustrate shortly, with no additional information, using n = 64 results in expected power of only .67. This is not a statement about what would happen if the researcher repeated the experiment an infinite number of times. Instead, expected power is a statement about the proportion of researchers, examining different topics, in different populations, using different techniques, who, based on the same observed effect size estimate of .50 and no other information (i.e., different parameter values are all essentially equally likely), all conduct a future study based on n = 64. Sixty-seven percent of these researchers would obtain significant results in the future study. This is a subtle conceptual shift in the definition of power that we revisit and expand upon later after illustrating the actual mechanics and process of calculating expected power.

The difficulty in applying Bayes' Theorem and calculating expected power lies in determining the prior distribution of the parameter. Different choices of prior distributions yield different posterior distributions, resulting in the criticism that the researcher's subjectivity influences the Bayesian analysis. We first discuss and illustrate

the non-informative prior case before examining several techniques for incorporating additional information into the posterior distribution.

Power calculations based on an effect size estimate and a non-informative prior. Much work has been done to determine prior distributions that are not subjective, allow the observed data to dominate the calculation of the posterior distribution, and thereby minimize the impact of the prior distribution. These non-subjective priors (see Bernardo, 1997, for a deeper philosophical discussion) are also termed "probability matching priors" in that they ensure the frequentist validity of the Bayesian credible intervals based on the posterior distribution. In some cases this probability matching may be asymptotic (e.g., see Datta & Mukerjee, 2004, for a review) whereas, as we will demonstrate, for the effect size estimates *d* and *r* this probability match can be exact (Berger & Sun, 2008; Lecoutre, 1999, 2007; Naddeo, 2004). In other words, as discussed in more detail in Biesanz (2010), the Bayesian credible intervals considered in this manuscript under the non-informative prior distribution correspond exactly to confidence intervals for effect sizes calculated following the procedures outlined in Cumming & Finch (2001), Kelley (2007), Steiger & Fouladi (1997), and Smithson (2001). With an exact match between the traditional frequentist confidence interval and the Bayesian credible interval in this context, the posterior distribution represents exactly the same inferential information and uncertainty contained in traditional *p*-values. Differences between the two perspectives are solely philosophical and interpretational.

Suppose that a researcher has an effect size estimate *d*, as in our attitude-behavior example, or an observed correlation *r*, but no other sources of information to guide the power analysis such as relevant meta-analyses or comparable studies on the same topic.

Under a non-informative prior, the posterior distribution of the standardized mean difference (δ) is

$$(\delta \mid d) \sim \frac{z}{\sqrt{\frac{n_1 n_2}{n_1 + n_2}}} + \frac{d \times c_{(\upsilon)}}{\sqrt{\upsilon}}, \tag{4}$$

where z is a standard normal variate (i.e., $z \sim N(0,1)$), $c_{(\upsilon)} \sim \sqrt{\chi^2(\upsilon)}$ with $\upsilon = df_{obs} = (n_1 + n_2 - 2)$, and z and $c_{(\upsilon)}$ are independent with "~" interpreted as "*has the same distribution as*." The expression of the posterior distribution in (4) is a *randomly constructed distribution* (see Berger & Sun, 2008); all elements in this expression are either constants or standard reference distributions (normal and chi-square).

The posterior distribution of the effect size parameter represents the distribution of plausible values for the population effect size. The posterior distribution thus captures the imprecision associated with the effect size parameter given the observed data. However, for sample size planning, the distribution of interest is the posterior predictive distribution, $(t_{new} \mid t_{obs}, df_{new})$; see Table 1B. This represents the distribution of future hypothetical observed *t*-statistics based on a specified new sample size (df_{new}), which is a function of the posterior distribution of the effect size parameter. The posterior predictive distribution incorporates the uncertainty associated with the estimate of the effect size by integrating over the posterior distribution of the effect size parameter.

The posterior predictive distribution of the *t*-statistic is critical for sample size planning as it provides a direct route for determining expected statistical power (EP):

$$EP = 1 - \int_{t_{\alpha/2;df_{new}}}^{t_{1-\alpha/2;df_{new}}} (t_{new} \mid t_{obs}, df_{new}). \tag{5}$$

Expected statistical power is the proportion of the posterior predictive distribution

that is more extreme than the critical values based on the standard (central) *t*-distribution given specified α. Expected power depends on the choice of sample size for the future study. Increasing the sample size will increase statistical power; consequently, a statistical power calculation for planning sample size involves determining the requisite sample size (i.e., df_{new}) necessary to produce the desired expected power. The goal in the sample size planning process may be to determine df_{new} such that expected power is .80. A precise empirical solution to Equation (5) given df_{new} is straightforward, as the posterior predictive distribution is a known function of standard reference distributions (see Table 1B).

To illustrate, Figure 3 presents the posterior predictive *t*-distributions for two different sample sizes (n = 64 and n = 130 per group) based on the posterior distribution of effect sizes from our attitude-behavior littering example where we estimated d = .50 in a study where n = 25. A standard statistical power calculation based on the assumption that $\delta = d = .50$ suggest that the sample size of 64 per group will result in power of .80. However, on average across the distribution of plausible values for the population parameter, the actual statistical power is only .67. That is, on average across the posterior distribution of the effect size, only 67 percent of future studies based on a sample size of 64 will result in a rejection of the null hypothesis.

What sample size then will produce a desired level of power such as .80? The df_{new} needed to achieve a specified level of expected power as a function of an observed effect size can be determined by systematically examining a range of sample sizes and modeling the nonlinear relationship between expected power using a nonparametric smoother such as a loess function. This represents a power curve that incorporates the

uncertainty associated with the effect size estimate. Figure 4 illustrates such a power curve for the present example. If needed, this procedure can be further refined by adapting stochastic approximation (e.g., see Robbins & Monro, 1951; Tierney, 1983) to solve Equation (5) with a specified degree of precision.[4] In the present example, only when the sample size is increased to $n = 130$ will 80% of future studies result in a rejection of the null hypothesis given the uncertainty associated with the effect size estimate.

Non-informative prior for the correlation. Distributions based on the correlational metric often present computational difficulties (see Naddeo, 2004, for the development and expression of the posterior distribution of the correlation under a non-informative prior). Consequently, the correlation is often re-expressed through the Fisher *r*-*z* normalizing transformation to simplify matters considerably (e.g., see Fouladi & Steiger, 2008, for more analytical details). However, for the present purposes it is both desirable as well as feasible to keep all analytical results in the original correlational metric. By adapting Shieh's (2006) expression for the sampling distribution for the noncentral *t*-distribution for the correlation of a randomly sampled predictor as a two stage distribution, the distributions for the correlation presented in Tables 1A and 1B follow (see also Berger & Sun, 2008; Biesanz, 2010). The difference in the expression of the noncentrality parameters for the correlation versus the standardized mean difference arises from considering the predictor to be randomly sampled as opposed to fixed. This introduces extra variability into the sampling distribution of the test-statistic that must be incorporated into the power calculation, and consequently, the posterior predictive distribution for the *t*-test for the correlation is different from the standardized mean

difference (see Table 1B and Biesanz, 2010, for further discussion of the implications of fixed versus random predictors). The logic and process of using Equation (5) remain unchanged – the goal is still to determine the sample size needed in a future study for a specified level of expected power.

The process of using a non-informative prior is relatively straightforward. Routines in *R* to estimate the sample size needed to achieve a given level of statistical power based on an observed effect size estimate and all of the examples presented in this manuscript are available from the authors. We also provide a rough guide to use to adjust the results from traditional power analysis software. Table 2 presents the adjustment (multiplicative factor) needed for sample size planning when power of .80 is desired. For example, given an observed standardized mean difference of .30 based on $n = 20$, traditional power analysis programs suggest $n = 176$ is required for power to be .80 in a subsequent study. However, after incorporating the uncertainty associated with the effect size estimate, sample size of $n = 176 \times 2.83 = 498$ is required to achieve average power of .80. Note that for very small and imprecise effect sizes (e.g., $d = .10$ with $n = 10$), the multiplicative factor is less than 1. This occurs when most of the posterior distribution is greater in magnitude than the observed effect size.

Power calculations based on effect size estimates and prior information

It is rare to have an effect size estimate without any additional prior information. For example, one may have conducted several other relevant studies which may or may not already be published. Alternately, with the increasing use of meta-analysis as a quantitative tool to review a literature, substantial information regarding effect sizes for a particular field is now commonly available. To illustrate how additional information may be used to better estimate the posterior distribution of the effect size, we first consider the

ideal case where the population distribution of effect sizes is already known before examining methods to incorporate the uncertainty associated with the estimate of the distribution of effect sizes in addition to that associated with the effect size estimate(s) at hand.

Known population distribution of effect sizes. Consider the ideal and hypothetical situation where we know that the distribution of effect sizes in a particular literature to be $N(\mu,\tau^2)$ and we have an observed effect size estimate $d_0 = \hat{\delta}$ with sampling variance σ_0^2. We present this case simply to familiarize readers with the computational mechanics before considering the usual case where distributions of effect sizes are estimated, not known. It follows from Bayes' Theorem that the posterior distribution of plausible effect sizes for the effect size estimate, $g(\delta \mid d_0)$, is $N((1-\gamma)d_0 + \gamma\mu,\ \sigma_0^2(1-\gamma))$, where

$$\gamma = \frac{\sigma_0^2}{\tau^2 + \sigma_0^2}. \qquad (6)$$

Incorporating prior information from a meta-analysis changes the posterior distribution from the case where there is no existing prior information and a non-informative prior is used in two important ways. First, the mean of the posterior distribution is shifted or "shrunk" towards the mean of the prior distribution (μ). Second, the variance of the posterior distribution is reduced relative to the sampling distribution of the effect size estimate. The net result of incorporating prior information is generally a substantial increase in the precision of the posterior distribution as we illustrate shortly. This has immediate benefits in sample size planning as the probability of extreme effect sizes, both large and small, is reduced when calculating expected power.

However, incorporating prior information from relevant studies into the power analysis rests on the presumption that the effect size estimate under consideration is exchangeable with those from the prior studies. This assumption of exchangeability is

met when our opinion of each effect size *prior* to observing the data is exactly the same (de Finetti, 1974). Formally, a sample of effect sizes is exchangeable if their joint prior probability distribution remains invariant under permutation of the specific studies. In other words: If, *prior* to conducting the study or observing its effect size, we have no reason to believe that this particular study should have a larger or smaller effect size than those from the meta-analysis, the assumption of exchangeability will be reasonable (see Draper, 1987; Draper, Hodges, Mallows, & Pregibon, 1993; Gelman, Carlin, Stern, & Rubin, 2004, pp. 121-124). If there is specific information that leads one to believe that this particular study is not essentially a random sample from the distribution of effect sizes estimated from the meta-analysis (e.g., this study involves a population of participants who tend to generate larger effect sizes), then this information can be incorporated directly into the meta-analysis as a predictor of the mean effect size resulting in the assumption of partial or conditional exchangeability after incorporating this information. Note that this assumption of exchangeability does not imply that the effect sizes are all equal – there can be systematic differences between studies resulting in substantial random effects variance – just that we have no information available before conducting our study as to where on the random effects distribution a particular study is likely to lie.

Assuming that the assumption of exchangeability is reasonable, actually incorporating prior information in practice is slightly more complex than the ideal case where the prior distribution is known. A meta-analysis within a particular research context provides only *estimates* of the mean effect size of a literature at $\hat{\mu}$ with random effects variance $\hat{\tau}^2$. Naïvely using the estimates from a meta-analysis to calculate the posterior distribution ignores the uncertainty associated with the meta-analytic estimates

of $\hat{\mu}$ and $\hat{\tau}^2$ and therefore underestimates the variance in the posterior distribution (Morris, 1983). We consider two approaches to estimating the posterior distribution of an effect size estimate given only estimates of the prior distribution: Empirical Bayes (EB) and Hierarchical Bayes (HB).

Empirical Bayesian posterior distributions. Following Morris (1983), there is a long research tradition of correcting the variance of the posterior distribution to reflect the uncertainty associated with using estimates of the prior distribution to provide confidence intervals (e.g., Carlin & Gelfand, 1990, 1991; Datta, Ghosh, Smith, & Lahiri, 2002; Laird and Louis, 1987, 1989; see also Cox, 1975). Laird and Louis (1987, 1989) developed a parametric bootstrap approach to estimating the posterior distribution where samples are randomly generated based on the estimates of the prior distribution, a naïve posterior distribution is calculated for each bootstrap, and then the posterior distribution is calculated as the mixture of each of the bootstrapped naïve posterior distributions. Carlin and Gelfand (1990, 1991) modified this approach to provide estimates of the posterior distribution that are conditional for an observed effect size as well as a correction to calibrate credible intervals if these are desired. Implementing this approach, although computationally intensive, requires only the estimates of the prior information (i.e., the random effects estimates obtained from the meta-analysis) and the precision of the individual studies that were involved in the meta-analysis. The latter is provided by either the average study sample size or, ideally, the full set of sample sizes of the studies included in the meta-analysis.

To illustrate this approach, first determine the mean and random effects variance $(\hat{\mu}, \hat{\tau}^2)$ from a meta-analysis based on *m* studies. Define d_0 as the observed effect size

estimate that we wish to use as input for the sample size planning process. The specific steps are:

1. Draw a random sample of size *m* from $N(\hat{\mu},\hat{\tau}^2)$ to produce δ_j^* (j=1, …, m). These are the latent effect sizes.

2. For each latent effect size draw an "observed" test statistic from the noncentral *t*-distribution with noncentrality parameter $\lambda_j^* = \delta_j^* / w_j$ where $w_j = \sqrt{\frac{1}{n_{1_j}} + \frac{1}{n_{2_j}}}$. Next, convert the noncentral *t* to a *d* to produce the "observed" effect sizes d_j^* (j=1, …, m) for the parametric bootstrap. Note that n_{1_j} and n_{2_j} are the sample sizes associated with study *j* in the meta-analysis. If the exact sample sizes are unavailable, simply use the range of sample sizes present in the meta-analysis, ensuring that the median sample size roughly corresponds to that reported in the meta-analysis.

3. For the set of m+1 effect sizes $\{d_0, d_j^*\}$ estimate $\hat{\mu}^*$ and $\hat{\tau}^{2*}$. This then allows the computation of the posterior distribution for the effect size estimate based on the bootstrap as $\pi(\delta_0 \mid d_0, \hat{\mu}^*, \hat{\tau}^{2*}) = a^{-1} f(d_0 \mid \delta_0) g(\delta_0 \mid \hat{\mu}^*, \hat{\tau}^{2*})$ (see Equation (3) in Footnote 3).

4. Following Laird & Louis (1987, Equation 11), repeat steps 1-3 a total of *k* times and estimate $\pi(\delta_0 \mid d_0) \approx \sum_{i=1}^{k} \pi(\delta_0 \mid d_0, \hat{\mu}^*, \hat{\tau}^{2*}) / k$. This represents a mixture of the posterior distributions derived in step 3, which may be approximated by drawing random samples of the same size from each of the *k* different bootstrapped posterior distributions to create an empirical approximate parametric bootstrapped posterior distribution.[5] With

a sufficiently large value of *k* (e.g., 1000 or more), the empirical approximation of $\pi(\delta_0 \mid d_0)$ will be quite stable and the computational feasibility of this approach is reasonable on a desktop computer.

To illustrate the empirical Bayesian approach, consider the attitude-behavior littering example. A recent meta-analysis of the attitude-behavior relationship by Webb and Sheeran (2006), based on 47 studies, provides a mean SMD of $\hat{\mu} = .325$ with variance $\hat{\tau}^2 = .069$.[6] Our observed effect size estimate $d = .50$ has an estimated variance of .086 under the non-informative prior. In contrast, the parametric bootstrap empirical Bayesian approach results in an approximate posterior distribution with mean $\hat{\mu}_{\delta_{EB}} = .405$ with variance = .037. Incorporating the prior information provided by the meta-analysis shifts the mean effect size estimate towards the average provided by the meta-analysis as well as substantially reducing the variability in the posterior distribution.

With this empirical approximate posterior distribution, we can first generate the posterior predictive distribution, $\left(t_{new} \mid t_{obs}, df_{new}, \hat{\mu}, \hat{\tau}^2\right)$, which incorporates the prior information from the meta-analysis. This represents the mixture of noncentral *t*-distributions based on the posterior distribution calculated in step four (i.e., for every δ_0 generated in Step 4, estimate $\left(t_{new} \mid \delta_0, df_{new}\right)$ and then average over the resulting predicted *t*-distributions). The resulting estimate of the posterior predictive distribution is then used to solve (5) to estimate the sample size required for a given power. In the present example, a sample size of $n = 160$ per group is needed to achieve average statistical power of .80. The required sample size increased in this example as the posterior distribution has shifted towards smaller effect sizes, albeit with smaller

variability, through the inclusion of prior information. Note that if our observed effect size estimate had been instead below the meta-analytic mean effect size, the suggested sample size would generally have decreased instead.

Hierarchical Bayesian posterior distributions. A full Bayesian analysis places a prior distribution on the mean and variance of the effect sizes in the literature (i.e., a hyperprior). Then, using Markov Chain Monte Carlo (MCMC) methods, the posterior distribution of the effect size can be numerically estimated (see Casella & George, 1992; Gelman et al., 2004, for introductions to MCMC methods).[7]

Just like the empirical Bayes methods, the hierarchical Bayesian approach provides a large sample of values from the posterior distribution of the effect size of interest that can be used to calculate the posterior predictive distribution and thus expected power. The hierarchical Bayesian approach results in a mean estimate of $\widehat{\mu}_{\delta_{HB}}$ = .404 with variance .0397 for the littering initial study across 10,000 draws from the posterior distribution. Using this estimated posterior distribution to determine the posterior predictive distribution (see Table 1B) to solve Equation (5) results in a sample size of 172 individuals per group need to achieve adequate power of .80 across the posterior distribution. Figure 5 presents the posterior distributions for the attitude-littering example based on (a) no prior information and (b) estimated posterior distributions derived from empirical and hierarchical Bayesian approaches. In this example there is striking convergence between the empirical and hierarchical Bayesian approaches to the posterior distribution.

The hierarchical Bayesian approach can also be easily implemented when there are only several other relevant studies that can be used for prior information. To

illustrate, consider the small meta-analysis presented in Dunn, Biesanz, Human, and Finn (2007) on the effects of self-presentation on positive affect within social interactions. Across the 5 studies in this manuscript, the mean effect size was $\hat{\mu}_d = .44$. Suppose one were interested in replicating Study 2b, where self-presentation was manipulated directly for romantic partners. The observed effect size was $d = .77$, with $CI_{.95} = [.15, 1.35]$, and a naïve power calculation suggests that 28 participants are needed in each condition to achieve adequate statistical power of .80. Simply incorporating the effect size uncertainty by using the algorithm presented earlier for the non-informative prior suggests increasing this sample size to 40 per group. However, by using the other 4 studies as prior information within a hierarchical Bayesian analysis, we obtain an estimate of the posterior distribution of $\hat{\mu}_{\delta_{HB}} = .576$ with variance .073 for Study 2b and determine instead that 80 participants per group are required to ensure adequate average statistical power for this study. Incorporating the information from the other 4 studies dramatically reduces the uncertainty associated with the effect size for Study 2b. Indeed, after incorporating the prior information we can estimate that the expected power under the naïve power calculation that resulted in $n = 28$ is instead only .53. Figure 6 presents the expected power as a function of the sample size based on the posterior distribution of the effect size under the hierarchical Bayesian model.

Note that in conducting a hierarchical Bayesian analysis, a number of decisions need to be made up front, even in a relatively straightforward model such as a hierarchical normal model. For instance, the properties of the Markov chain are asymptotic and it may take many iterations to converge to the proper posterior distribution. Consequently the first number of iterations (e.g., 10,000 in the present

examples) are discarded. Sequential draws from the chain will not be independent, so the chain may be "thinned" by taking only every i^{th} observation where i=100 in the present examples. Distributions other than normal may be presumed for the posterior and prior distributions. For instance, a robust hierarchical Bayesian analysis may place a *t*-distribution on the prior random effects (e.g., $\delta_j \sim t_v\left(\mu,\ \tau^2\right)$ with v df). Such a robust Bayesian analysis places a "fatter" tail on the prior distribution of random effect sizes and consequently does not "shrink" the observed effect size down to the mean effect size as much as the normal distribution; in the present examples, this will lead to an estimate of fewer participants to achieve adequate statistical power. Given all of these decisions, the hierarchical Bayesian analysis requires careful attention to ensure that the model has converged on the posterior distribution of interest and that the conclusions are not inordinately sensitive to necessary decisions such as the nature of the prior and posterior distributions. Nonetheless, a full hierarchical Bayesian analysis represents a well-justified and attractive approach to estimating the posterior distribution when feasible.

Accurate Parameter Estimation based on Effect Size Estimates

Instead of ensuring an adequate probability of rejecting a false null hypothesis, estimating the effect size with a specified degree of precision may be the primary impetus for conducting the future study. Frameworks for determining the sample size needed to achieve a certain expected level of precision within a subsequent study have been developed and are a natural parallel to a focus on estimating effect sizes and presenting confidence intervals (e.g., see Jiroutek, Muller, Kupper, & Stewart, 2003; Kelley, 2008; Kelley, Maxwell, & Rausch, 2003; Kelley & Maxwell, 2003; Kelley & Rausch, 2006). Prior information can be included within this question as well (e.g., see Santis, 2007).

Since confidence interval width depends as well on the population effect size, uncertainty in effect size estimates can impact sample size planning in this context. We now illustrate how to plan the sample size for a future study to achieve a specified degree of precision while incorporating the uncertainty associated with an initial effect size estimate. Define ω as the width of the 1-α equal-tailed confidence interval based on an observed *t*-statistic as follows:

$$\omega(t_{obs}) = g(\lambda_{upper}) - g(\lambda_{lower}), \tag{10}$$

where $g(\lambda)$ converts the noncentrality parameter back to the effect size of interest (e.g., δ or ρ) and λ_{upper} and λ_{lower} are defined as

$$\int_{-\infty}^{\lambda_{lower}} f(\lambda \mid t_{obs}) d\lambda = \alpha / 2 \tag{11}$$

and

$$\int_{\lambda_{upper}}^{\infty} f(\lambda \mid t_{obs}) d\lambda = \alpha / 2 . \tag{12}$$

In other words, λ_{upper} and λ_{lower} represent the (1 - α/2) and the α/2 quantiles, respectively, of the posterior distribution of the noncentrality parameter under a non-informative prior distribution as defined in Tables 1A and 1B. As discussed in Biesanz (2010; see also Berger & Sun, 2008; LeCoutre, 2007), this approach results in the same confidence intervals as those derived through pivoting the cumulative distribution function (e.g., see Cumming and Finch, 2001; Kelley, 2007; Steiger and Fouladi, 1997; Smithson, 2003; Steiger, 2004).

The expected confidence interval width for a future study given a specified sample size can be determined through the posterior predictive distribution of the test-

statistic. For the non-informative prior distribution, the expected width is

$$E(\omega \mid t_{obs}, df_{new}) = \int_{-\infty}^{+\infty} \omega(t_{new} \mid t_{obs}, df_{new}) dt_{new} \quad (13)$$

and when prior information is incorporated into the posterior distribution, the expected with is

$$E(\omega \mid t_{obs}, df_{new}, \hat{\mu}, \hat{\tau}^2) = \int_{-\infty}^{+\infty} \omega(t_{new} \mid t_{obs}, df_{new}, \hat{\mu}, \hat{\tau}^2) dt_{new} \; . \quad (14)$$

For each *t*-value in the posterior predictive distribution, we can determine the resulting effect size confidence interval based on that (unobserved) *t*-statistic. The expected width is the average of the resulting widths across the posterior predictive distribution. Because the expected confidence interval width is a monotonic function of sample size – intervals become more precise with larger samples – equations (13) or (14) can be solved to determine the sample size necessary to achieve a specified expected width.

Quantiles for the future confidence intervals can easily be obtained as well. The width of a confidence interval has a monotonic relationship with the magnitude of the estimated noncentrality parameter (more specifically, holding the sample size constant, ω increases as $\hat{\delta}$ increases and decreases as $\hat{\rho}$ increases). As a consequence of this relationship, the quantiles of the absolute value of the posterior predictive distribution of the test statistic (i.e., $|t_{new}|$), when converted to interval width (ω), provide the quantiles for the future study's effect size confidence interval width.

These quantiles can be used to specify the degree of certainty that the future confidence interval will be no wider than desired (e.g., see Kelley & Rausch, 2006).

Figure 6 illustrates the median confidence interval width for different sample sizes based on the effect size from Study 2b from Dunn, Biesanz, Human, and Finn (2007) and using the informative prior. The standardized mean difference is relatively insensitive to imprecision in the parameter estimate (e.g., see as well Kelley & Rausch, 2006, Tables 1-3). For instance, for $n = 80$, which is when average power is .80, the median 90% confidence interval width $\hat{\omega} = 0.530$ which itself has $CI_{.80} = [.521, .552]$. Thus, for this sample size, we are 90% confident that a future study will result in a 90% confidence interval whose width ($\hat{\omega}$) is less than .552. This would suggest that if the goal of the study was to more precisely estimate the effect size, a larger sample size would be required. Equations (13) for the non-informative prior or equation (14) for the informative prior can be solved for expected confidence interval width or adapted to examine quantiles. For instance, a sample size of 151 in this example would result in a 90% confidence interval around the effect size estimate whose width ($\hat{\omega}$) is less than .40 with confidence of 90%.

Conclusion and Discussion

The present approach provides a formal basis for planning sample size. By focusing on prior research when available and the uncertainty surrounding their effect size estimates, subjectivity in the sample size planning process is greatly minimized. If effect size estimates are routinely used in power calculations without accounting for their uncertainty, statistical power will be generally substantially lower than the nominal statistical power. However, by incorporating information on the distribution of plausible effect sizes – either with or without prior information – sample size planning can be substantially improved and researchers may increase their confidence that, on average,

they may better achieve their desired level of statistical power or confidence interval width.

The procedures outlined in the present manuscript may lead to suggested sample sizes that are substantially larger than those from the naïve use of the effect size estimate (e.g., see Table 2). The imprecision in the initial effect size estimate may have a strong influence on the suggested sample size. As effect size estimates become more precise, the procedures outline in the present manuscript converge to those from standard statistical power packages treating the effect size estimate as the population parameter. If the initial effect size estimate is large and imprecise, although it may be tempting to run only the few participants suggested by the naïve power calculation, many more participants may be needed given the paucity of information present on the actual effect size.

Mixing Frequentist and Bayesian Thinking in the Sample Size Planning Process

The present manuscript adopts a Bayesian perspective on the population effect size during the study design process; however, inferences and/or estimation *are based solely* on the data collected in the future study. This represents an amalgam of Bayesian and frequentist perspectives and follows the spirit of the recent American Statistical Association Presidential addresses from both Efron (2005) and Little (2006). Similar blended perspectives are found as well in Bayarri and Berger (2006), Casella (2007), Gelman, Meng, and Stern (1996), Rubin (1984), and Rubin and Stern (1998). The present manuscript is focused on determining the sample size for a single prospective study where inferences will be made solely based on that future study. This still represents the modal inferential design in psychology.

Adopting the Bayesian perspective allows the determination of the probability of achieving the desired study goals without the necessity of specifying the exact population parameter. Instead we consider the entire distribution of plausible population parameter values given an observed effect size estimate to compute statistical power and accurate parameter estimation. The benefit of not having to specify the exact parameter value during the sample size process comes at the cost of having to specify a prior distribution. The use of non-informative priors when there really is no other information or informative priors such as from relevant meta-analyses both provide reasonable objective and formal bases for planning future studies.

Empirical Bayes (EB) versus Hierarchical Bayes (HB) and Considerations of Exchangeability

When prior information such as a meta-analysis or a series of related studies is available, both EB and HB approaches are viable options for incorporating this prior information into study design. The HB approach is theoretically elegant and the preferred approach when feasible. However, implementing the HB analysis requires individual effect size estimates from each of the studies in the meta-analysis as well as the variance estimates from each individual study. This raw information is not always available in published meta-analyses. Even when this information is available, the HB approach may require recreating a complex analytical model to account for publication bias as well as monitoring for convergence; consequently, this approach may require experience, and some expertise in the use of these estimation techniques (e.g., censoring parameters to account for publication bias; see Eberly & Casella, 1999) may be needed. For example, Shadish and Baldwin (2005) examined the effectiveness of behavioral

marital therapy and estimated the effectiveness at ($\hat{\mu} = .585,\ \hat{\tau}^2 = .005$) across 30 studies. However, after accounting for publication bias, these estimates changed considerably to ($\hat{\mu} = .498,\ \hat{\tau}^2 = .313$). The HB approach in this case would require including censoring parameters and may present challenges in the estimation. In contrast, the EB approach requires only the bottom-line output from the meta-analysis after accounting for the publication bias (that is, the mean and random variance estimates of studies included in the meta-analysis) along with rough study design data to provide an approximation of the posterior distribution. The EB approach does not require monitoring for convergence and is readily extendable to generalized linear models which may require considerable expertise to implement in the HB approach. The EB approach may often be much easier to implement and consequently preferable from a practical perspective.

Limited empirical research comparing these approaches has found that they provide comparable results (e.g., MacNab, Farrell, Gustafson, & Wen, 2004). Indeed, in the present attitude-littering example these approaches yield essentially identical results using the Webb and Sheeran (2006) meta-analysis. In sum, the HB approach is to be preferred when practical given the theoretical justification for the approach, but the EB approach may provide an excellent and easily implementable approximation.

Alternative Approaches and Design Considerations

Although the focus of the present manuscript has been on statistical power and confidence interval width for inferences based on a single prospective study, alternative approaches and frameworks have been explored in the literature. For example, Muller and Pasour (1997) explore the impact on power when the effect size is considered fixed

and the variance is estimated (see also Coffey & Muller, 1999). Furthermore, if the costs of running participants and the costs and benefits of different outcomes can be determined, Equation (5) is easily expanded to determine the sample size required to maximize expected utility (e.g., Lindley, 1997). This approach is conceptually elegant and worthy of serious consideration for applied research when estimates of the utility of an effective treatment can be determined. For basic research, the utility of detecting a significant effect may be much less readily quantifiable and thus this approach may not be easily implemented.

In conclusion, when determining sample sizes for subsequent studies, all of the information available regarding effect sizes should enter into the calculation. Most importantly, this includes the uncertainty associated with effect size estimates. We provide accessible routines and software to help make these calculations that illustrate all of the examples presented in this manuscript. We hope that the disparity between the expected power under naïve power calculations and the nominal statistical power provides the impetus to re-examine standard approaches to sample size determination and to implement formal approaches for incorporating existing information in the sample size planning process.

References

Bayarri, M.J., and Berger, J.O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, 19, 58-80.

Berger, J. O., & Sun, D. (2008). Objective priors for the bivariate normal model. *The Annals of Statistics, 36*, 963-982.

Bernardo, J. M. (1997). Non-informative priors do not exist: A discussion. *Journal of Statistics Planning and Inference, 65*, 159-189.

Biesanz, J. C. (2010). Confidence distributions for standardized effect sizes. *Manuscript under revision*.

Carlin, B. P., & Gelfand, A. E. (1990). Approaches for empirical Bayes confidence intervals. *Journal of the American Statistical Association, 85*, 105-114.

Carlin, B. P., & Gelfand, A. E. (1991). A sample reuse method for accurate parametric empirical Bayes confidence intervals. *Journal of the Royal Statistical Society, Series B (Methodological), 53*, 189-200.

Casella, G. (2007). Why (Not) Frequentist Inference (Too)? *Boletín de la Sociedad de Estadística e Investigación Operativa, 23*, 1-5.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *American Statistician, 46*, 167-174.

Cohen, J. (1988) *Statistical Power Analysis tor the Behavioral Sciences*, 2nd ed., Hillsdale, N.J., L. Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Coffey, C. S., & Muller, K. E. (1999). Exact test size and power of a Gaussian error linear model for an internal pilot study. *Statistics in Medicine, 18*, 1199-1214.

Cox, D. R. (1975). Prediction intervals and empirical Bayes confidence intervals. In J. Gani (Ed.), *Perspectives in probability and statistics, papers in honor of M. S. Bartlett*. (pp. 47-55). New York, NY: Academic Press.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 530-572.

Datta, G. S., & Ghosh, J. K. (1995a). Non-informative priors for maximal invariant in group models. *Test, 4*, 95-114.

Datta, G. S., & Ghosh, M. (1995b). Some remarks on non-informative priors. *Journal of the American Statistical Association, 90,* 1357-1363.

Datta, G. S., Ghosh, M., Smith, D. D., & Lahiri, P. (2002). On an asymptotic theory of conditional and unconditional coverage probabilities of empirical Bayes confidence intervals. *Scandinavian Journal of Statistics, 29*, 139-152.

Datta, G. S., & Mukerjee, R. (2004). *Probability matching priors: Higher order asymptotics*. Springer-Verlag: New York, NY.

Datta, G. S., & Sweeting, T. J. (2005). Probability matching priors. Research report No. 252, Department of Statistical Science, University College London.

de Finetti, B. (1974). *Theory of Probability*. London: Wiley.

Draper, D. (1987). Comment: On exchangeability judgments in predictive modeling and the role of data in statistical research. *Statistical Science, 2*, 454-461.

Draper, D., Hodges, J. S., Mallows, C. L., & Pregibon, D. (1993). Exchangeability and data analysis. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 156*, 9-37.

Dunn, E. W., Biesanz, J. C., Human, L., & Finn, S. (2007). Misunderstanding the affective consequences of everyday social interactions: The hidden benefits of putting one's best face forward. *Journal of Personality and Social Psychology 92*, 990-1005.

Eberly, L. E. , & Casella, G. (1999). Bayesian estimation of the number of unseen studies in a meta-analysis. *Journal of Official Statistics, 15* , 477-494.

Efron, B. (2005). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association, 100,* 1–5.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*, 1-11.

Fouladi, R. T., & Steiger, J. H. (2008). The Fisher transform of the Pearson product moment correlation coefficient and its square: Cumulants, moments, and applications. *Communications in Statistics – Simulation and Computation, 37*, 928-944.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York, NY: Chapman & Hall/CRC.

Gelman, A., Meng, X. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733-807.

Ghosh, M. & Yang, M. C. (1996). Non-informative priors for the two-sample normal problem. *Test, 5,* 145-157.

Gillett, R. (1994). An average power criterion for sample size estimation. *The Statistician, 43*, 389-394.

Gillett, R. (2002). The unseen power loss: Stemming the flow. *Educational and Psychological Measurement*, *62*, 960-968.

Gravetter, F. J., & Wallnau, L. B. (2006). *Statistics for the Behavioral Sciences (7th ed).* Wadsworth Publishing, Belmont, CA.

Howell, D. C. (2007). *Statistical methods for psychology (6th ed)*. USA: Thompson/Wadsworth.

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

Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20,* 1-23.

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

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

Kelley, K. & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Confidence interval width via narrow confidence intervals. *Psychological Methods, 11*, 363-385.

Laird, N. M., & Louis, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association, 82*, 739-750.

Laird, N. M., & Louis, T. A. (1989). Empirical Bayes confidence intervals for a series of related experiments. *Biometrics, 45*, 481-495.

Lecoutre, B. (1999). Two useful distributions for Bayesian predictive procedures under normal models. *Journal of Statistical Planning and Inference, 77*, 93-105.

Lecoutre, B. (2007). Another look at confidence intervals for the noncentral T distribution. *Journal of Modern Applied Statistical Methods, 6*, 107-116.

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

Lindley, D. V. (1997). The choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46, 129–138.

Little, R. J. (2006). Calibrated Bayes: A Bayes/Frequentist Roadmap. *American Statistician, 60,* 213-223.

MacNab, Y. C., Farrell, P. J., Gustafson, P., & Wen, S. (2004). Estimation in Bayesian disease mapping. *Biometrics, 60*, 865-873.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association, 78*, 47-55.

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

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

Muller, K. E., & Pasour, V. B. (1997). Bias in linear model power and sample size due to estimating variance. *Communications in Statistics- Theory and Methods, 26*, 839-851.

Naddeo, S. (2004). Exact Bayesian higher posterior density interval for the correlation coefficient of a normal bivariate distribution. *Communications in Statistics: Simulation and Computation, 33*, 983-990.

O'Brien R.G., (1998). *A tour of UnifyPow: A SAS module/macro for sample-size analysis*. Pages 1346-1355 in Proceedings of the 23rd SAS Users Group International Conference, Cary (NC): SAS Institute.

Ozer, D. J. (2007). Evaluating effect size in personality research. In R.W. Robins, R.C. Fraley, and R.F. Krueger (Eds.). *Handbook of Research Methods in Personality Psychology*. New York: Guilford.

R Development Core Team (2006). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics, 22*, 400-427.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Russell Sage Foundation.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*, 1151-1172.

Rubin, D. B., & Stern, H. S. (1998). Sample size determination using posterior predictive distributions. *Sankhya: The Indian Journal of Statistics, Series B, 60*, 161-175.

Santis, F. De. (2007). Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170*, 95–113.

SAS Institute Inc. (2003). *SAS Technical Report P-243, SAS/STAT Software: The POWER Procedure (Experimental), Release 6.0*. Cary, NC: SAS Institute Inc.

Shadish, W. R., & Baldwin, S. A. (2005). The effects of behavioral marital therapy: A meta-analysis of randomized controlled trials. *Journal of Consulting and Clinical Psychology, 73,* 6-14.

Shieh, G. (2006). Exact interval estimation, power calculation, and sample size determination in normal correlation analysis. *Psychometrika, 71*, 529-540.

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.* Sage Publications, Thousand Oaks, CA.

Steiger, J. H. (1999). *STATISTICA Power Analysis*. Tulsa, OK: Statsoft, Inc.

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 models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.). *What if there were no significance tests?* (pp. 221-257). Mahwah, NJ: Lawrence Erlbaum Associates.

Taylor, D. J., & Muller, K. E. (1995a). Computing confidence bounds for power and sample size of the general linear model. *The American Statistician, 49*, 43-47.

Taylor, D. J., & Muller, K. E. (1995b). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics- Theory and Methods, 25*, 1595-1610.

Tierney, L. (1983). A space-efficient recursive procedure for estimating a quantile of an unknown distribution. *SIAM Journal on Scientific and Statistical Computing, 4*, 706-711

Wilkinson, L., & the Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist, 54*, 594-604.

Webb, T. L., & Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin, 132*, 249-268.

Author Note

Jeremy C. Biesanz, Department of Psychology, University of British Columbia; Sheree M. Schrager, The Saban Research Institute, Childrens Hospital Los Angeles. We thank Carl Falk, Paul Gustafson, and Victoria Savalei for helpful comments on previous versions of this manuscript. This research was partially supported by Social Sciences and Humanities Research Council of Canada Grant SSHRC 410-2005-2287 to Jeremy C. Biesanz. Code in *R* implementing the noninformative prior and EB procedures and illustrating the empirical examples is available at http://www.psych.ubc.ca/~jbiesanz/.

Correspondence regarding this manuscript should be addressed to Jeremy C. Biesanz, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, BC, Canada V6T 1Z4. E-mail: jbiesanz@psych.ubc.ca

Table 1A. Definitions and distributions for the standardized mean difference (SMD) and the correlation.

Term	SMD	Correlation
Definitions, Notation, and Relationships		
Effect size parameter	$\delta = \frac{\mu_1 - \mu_2}{\sigma} = \lambda / \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$	$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\lambda}{\sqrt{\lambda^2 + df_{obs}}}$
Effect size estimate	$d = \hat{\delta} = \frac{\bar{Y}_1 - \bar{Y}_2}{s}$	$r = \hat{\rho} = \frac{s_{XY}}{s_X s_Y}$
Degrees of freedom (υ)	$\upsilon = df_{obs} = n_1 + n_2 - 2$	$\upsilon = df_{obs} = n - 2$
Observed *t*-value (t_{obs})	$t_{obs} = \hat{\lambda} = d\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$	$t_{obs} = \hat{\lambda} = \frac{r \times \sqrt{df_{obs}}}{\sqrt{1 - r^2}}$
Noncentrality parameter (λ)	$\lambda = \delta\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$	$\lambda = \frac{\rho \times \sqrt{df_{obs}}}{\sqrt{1 - \rho^2}}$

Table 1B. Distributions for the standardized mean difference (SMD) and the correlation.

Distribution	SMD	Correlation
Sampling Distribution:	$(t_{obs} \mid \lambda) \sim \frac{z+\lambda}{c_{(\upsilon)}/\sqrt{\upsilon}}$	$(t_{obs} \mid \lambda) \sim \frac{z+\lambda\left(c_{(\upsilon+1)}/\sqrt{\upsilon}\right)}{c_{(\upsilon)}/\sqrt{\upsilon}}$
Posterior Distribution:	$(\lambda \mid t_{obs}) \sim z + t_{obs}\left(c_{(\upsilon)}/\sqrt{\upsilon}\right)$	$(\lambda \mid t_{obs}) \sim \frac{z+t_{obs}\left(c_{(\upsilon)}/\sqrt{\upsilon}\right)}{c_{(\upsilon+1)}/\sqrt{\upsilon}}$
Posterior distribution of the effect size	$(\delta \mid t_{obs}) \sim \frac{z+t_{obs}\left(c_{(\upsilon)}/\sqrt{\upsilon}\right)}{\sqrt{\frac{n_{1_{obs}} n_{2_{obs}}}{n_{1_{obs}}+n_{2_{obs}}}}}$	$(\rho \mid t_{obs}) \sim u(\lambda \mid t_{obs}), u(y) = \frac{y}{\sqrt{y^2+\upsilon}}$
Posterior predictive: $\left(t_{new} \mid t_{obs}, df_{new}\right)$	$\sim \frac{z+(\delta \mid t_{obs})\sqrt{\frac{n_{1_{new}} n_{2_{new}}}{n_{1_{new}}+n_{2_{new}}}}}{\left(c_{(\upsilon_{new})}/\sqrt{\upsilon_{new}}\right)}$	$\sim \frac{z+\frac{(\lambda \mid t_{obs}) \times c_{(\upsilon_{new}+1)}}{\sqrt{\upsilon_{obs}}}}{\left(c_{(\upsilon_{new})}/\sqrt{\upsilon_{new}}\right)}$

Note: $z \sim N(0,1)$ is a standard normal variate, $c_{(\upsilon)} \sim \sqrt{\chi^2(\upsilon)}$, υ are the *df* $\left(df_{obs}\right)$ associated with the observed test-statistic (t_{obs}), υ_{new} are the *df* associated with a potential future study, and z and $c_{(\upsilon)}$ are independent with "~" interpreted as "*has the same distribution as*."

Table 2. Sample size multipliers based on a non-informative prior for traditional power analysis calculations at power of .80 when an effect size estimate treated as the population parameter.

Observed	Sample Size for the Observed Effect Size									
Effect Size	10	20	30	40	50	60	70	80	90	100
SMD										
.10	0.36	0.68	0.97	1.23	1.46	1.67	1.85	2.02	2.16	2.29
.20	1.23	2.02	2.50	2.75	2.86	2.87	2.82	2.73	2.63	2.52
.30	2.16	2.83	2.85	2.64	2.39	2.17	2.00	1.86	1.75	1.66
.40	2.77	2.77	2.34	2.00	1.77	1.62	1.52	1.45	1.39	1.34
.50	2.91	2.32	1.85	1.61	1.46	1.38	1.29	1.25	1.22	1.19
.60	2.73	1.91	1.56	1.40	1.31	1.25	1.21	1.18	1.16	1.14
.70	2.41	1.65	1.40	1.29	1.22	1.18	1.16	1.14	1.12	1.11
.80	2.10	1.49	1.30	1.21	1.17	1.14	1.12	1.11	1.08	1.08
.90	1.87	1.38	1.24	1.17	1.14	1.11	1.10	1.08	1.07	1.06
1.00	1.70	1.30	1.19	1.15	1.11	1.10	1.07	1.06	1.06	1.05
Correlation										
.10	0.53	1.03	1.30	1.99	2.27	2.48	2.64	2.75	2.83	2.87
.20	1.88	2.75	2.92	2.78	2.56	2.33	2.12	1.97	1.85	1.74
.30	2.84	2.72	2.21	1.86	1.65	1.52	1.43	1.37	1.32	1.28
.40	2.77	1.88	1.45	1.32	1.25	1.20	1.17	1.15	1.13	1.12
.50	1.87	1.36	1.24	1.17	1.14	1.11	1.09	1.08	1.07	1.06
.60	1.58	1.26	1.16	1.11	1.09	1.07	1.06	1.05	1.05	1.04
.70	1.53	1.22	1.14	1.10	1.08	1.06	1.05	1.05	1.04	1.04

Note: Sample size is *n* per group for the standardized mean difference. If the observed effect size is $d = .50$ based on $n = 20$ per group, traditional power analyses suggest $n = 64$. Instead, use $n = 2.32 \times 64 = 149$ per group (rounded) under the non-informative prior.

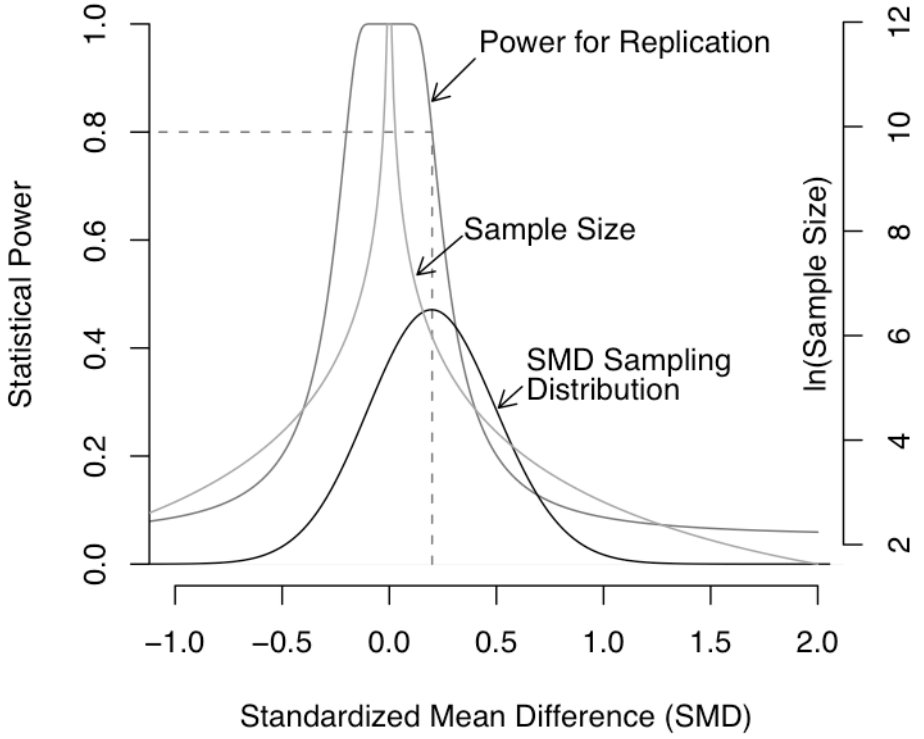

Figure 1. Statistical power as a function of using an observed standardized mean different to plan sample size for a future study. Graphed are (a) the sampling distribution of a small standardized mean difference (population SMD $\delta = .20$, $n = 25$), (b) the natural log of the sample size required to achieve power of .80 when the observed sample SMD is treated as the population effect size, and finally (c) the actual statistical power for replication using that sample size given that in fact $\delta = .20$. Larger observed values of the SMD result in lower sample sizes, which, in turn, results in power less than .80. Only when the observed sample estimate corresponds exactly to the population parameter (i.e., $d = \delta = .20$) does the sample size planning process using solely the observed effect size result in statistical power of .80.

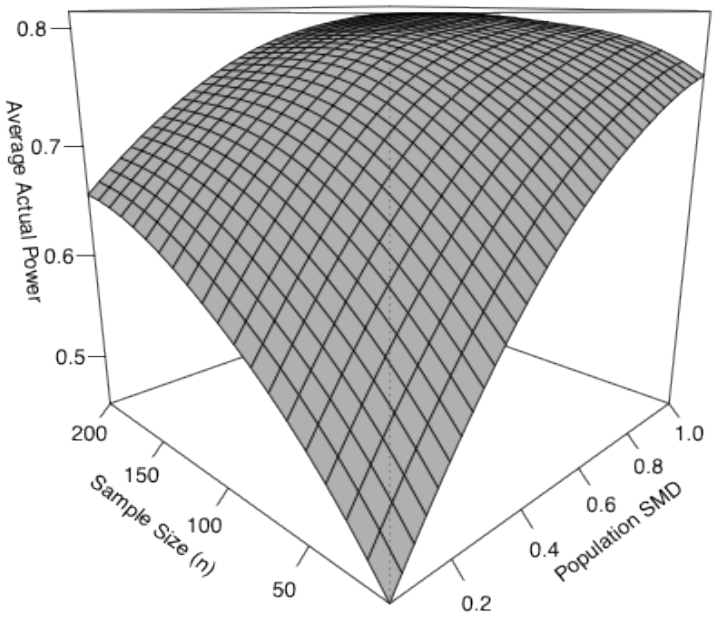

Figure 2. Average statistical power across the sampling distribution given a specified population standardized mean difference (SMD; δ) and sample size per group (*n*) when the observed standardized mean difference from the sampling distribution is used as the basis for planning sample size for the next study specifying power of .80 with $\alpha = .05$, two-tailed. Average statistical power was empirically estimated based on a simulation of 20,000 draws from the sampling distribution for each combination of sample size and SMD – *n* ranged from 10 to 200 by increments of 10 and δ ranged from .05 to 1.0 by increments of .05. The graphed response surface is the nonparametric loess relationship between *n*, δ, and average power across the sampling distribution as in Figure 1.

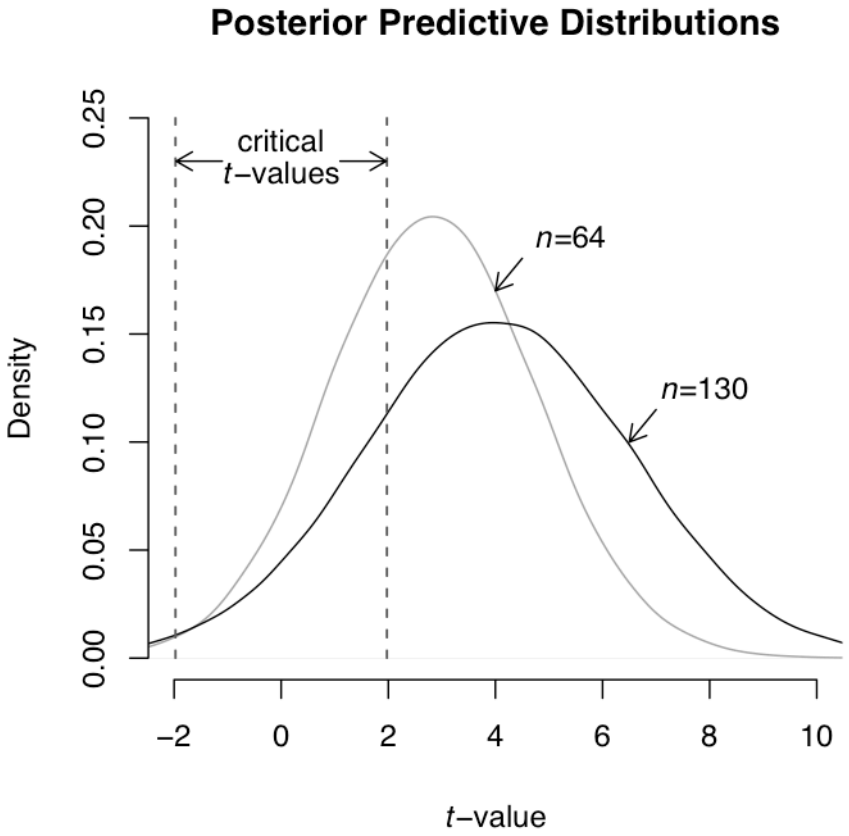

Figure 3. Posterior predictive distributions using sample sizes of n = 64 and n = 130 per group based on an observed d = .50 with n = 25 and a non-informative prior. Note that each posterior distribution has imperceptibly different critical t-values as they differ in their degrees of freedom (+/-1.97897 versus +/-1.96920).

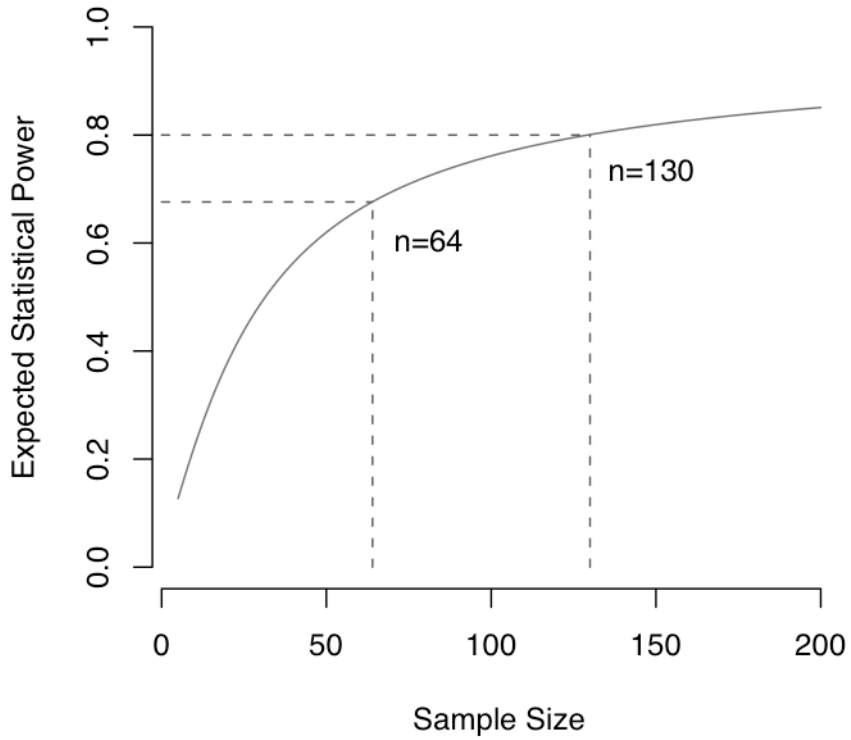

Figure 4. Expected power as a function of sample size under a non-informative prior distribution for the attitude-behavior example ($d = .50, n = 25$). A sample size of $n = 130$ is required to achieve expected power of .80 whereas $n = 64$ results in expected (average) power of .67.

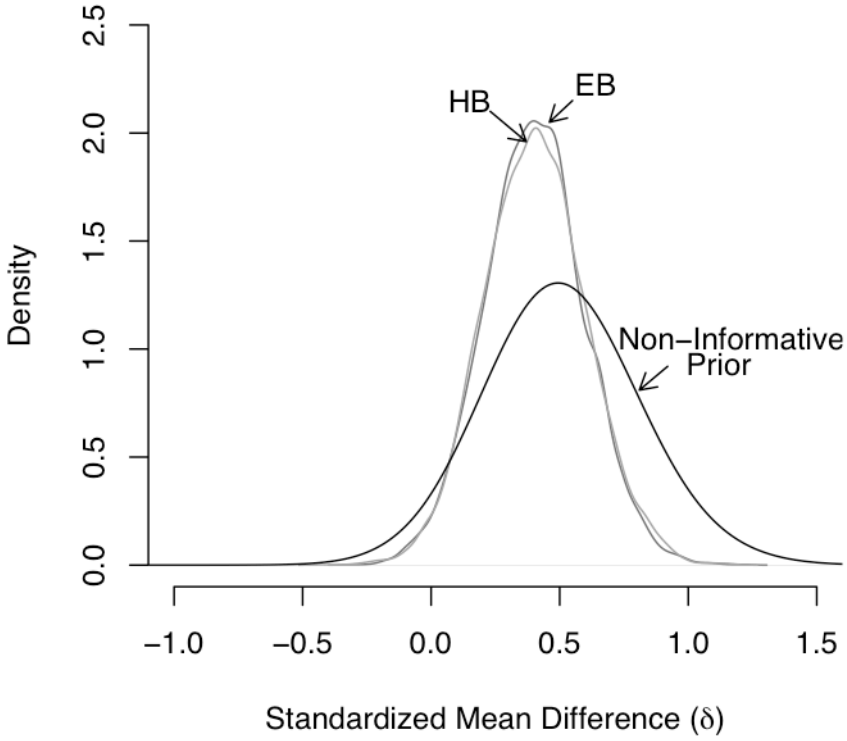

Figure 5. Estimated posterior distributions based on a non-informative prior, empirical Bayes (EB), and hierarchical Bayes (HB) for the attitude-behavior example effect size (d = .50, n = 25) where prior information is obtained from Webb and Sheeran (2006; Table 1).

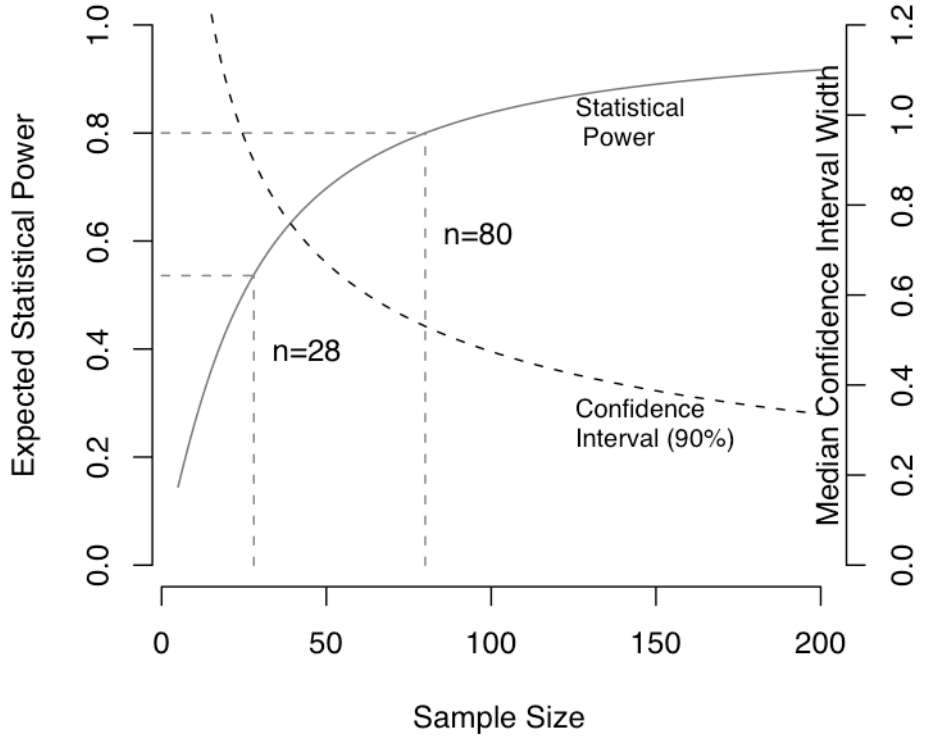

Figure 6. Expected (average) power as a function of sample size for Study 2b (d = .77, n = 23; Dunn, Biesanz, Human, & Finn, 2007) using the posterior distribution produced under a hierarchical Bayesian analysis based on the other 4 study effect sizes. A sample size of n = 80 is required to achieve expected (average) power of .80 whereas the n of 28 from the naïve power calculation results in expected (average) power of .53. Graphed as well is the median 90% confidence interval width as a function of sample size across the posterior predictive distribution.

Footnotes

[1] We will assume that $\alpha = .05$ and that tests are two-tailed unless otherwise specified. Sample size (n_1 and n_2) refers to the number of observations per group for the standardized mean difference, whereas n refers to the number of observations when discussing the correlation.

[2] Psychological constructs are often measured on arbitrary scales; consequently, it is helpful to express the magnitude of a treatment or relationship between variables in an effect size metric that is free of the actual measurement scale used in a particular study. We consider in detail two of the more commonly used standardized effect size measures, the standardized mean difference (SMD) and the correlation. The population standardized mean difference is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}, \qquad (1)$$

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 which is assumed here to be equal in each group. The population standardized mean difference is estimated by

$$d = \hat{\delta} = \frac{\bar{Y}_1 - \bar{Y}_2}{s}, \qquad (2)$$

where $\bar{Y}_1$ and $\bar{Y}_2$ are the means of Groups 1 and 2 with sample sizes n_1 and n_2, respectively, and s is the pooled standard deviation. For the correlation between two variables X and Y, we consider the standard Pearson correlation whose population value is defined as $\rho = \sigma_{XY} / (\sigma_X \sigma_Y)$ and is estimated by $r = s_{XY}/(s_X s_Y)$ with sample size n where σ_{XY} (s_{XY}) is the population (sample) covariance between the X and Y with standard

deviations σ_X (s_X) and σ_Y (s_Y), respectively.

Although it is possible to transform the standardized mean difference into the correlational metric and vice versa (e.g., see Rosenthal, 1994), we assume that the standardized mean difference represents a fixed effects model in that the levels of Groups 1 and 2 are determined by the experimenter. In contrast, for the correlation we assume that the levels of the predictor *X* are randomly sampled. This distinction results in effect size measures that are not interchangeable for the purposes of sample size planning. All formulae, definitions, and distributions are presented for ease of reference in Tables 1A and 1B.

[3] Specifically, the posterior distribution of the standardized mean difference, given observed standardized mean difference ($\hat{\delta}$), is formally derived from Bayes' Theorem as follows:

$$g(\delta \mid \hat{\delta}) = a^{-1} f(\hat{\delta} \mid \delta)\pi(\delta), \tag{3}$$

where $g(\delta \mid \hat{\delta})$ is the posterior distribution of the parameter given our observed standardized mean difference, $f(\hat{\delta} \mid \delta)$ is the likelihood function of the standardized mean difference given the population value δ, $\pi(\delta)$ is the prior distribution of the population standardized mean difference, and *a* is the normalizing constant where $a = \int_{-\infty}^{\infty} f(\hat{\delta} \mid \delta)\pi(\delta)d\delta$.

[4] For the non-informative prior it is possible to directly solve expected power through a more traditional Bayesian analysis. Define $H(\delta)$ as the statistical power to detect effect size δ with fixed sample size *n*. Then expected power (EP; Gillett, 1994;

Lindley, 1997) is $EP = \int_{-\infty}^{+\infty} H(\delta)\pi(\delta \mid d)\mathrm{d}\delta$ where $\pi(\delta \mid d)$ is defined in (3). We reflect the lack of prior knowledge with the non-informative prior $g(\delta) \propto \left[8 + \delta^2\right]^{-1/2}$ (Datta & Ghosh, 1995a, 1995b; Datta & Sweeting, 2005; Ghosh & Yang, 1996). Note that this particular non-informative prior distribution is known as a Jeffrey's prior, which is proportionally equivalent to the square root of the Fisher information for the standardized mean difference. This approach can be solved analytically using numerical integration and provides essentially equivalent answers to (5) for the examples examined in the present manuscript. The direct Bayesian solution, although appealing in its precision, may be difficult to estimate for small effect sizes and/or small sample sizes whereas the stochastic approximation solution will always provide an answer and is a more stable method of estimation.

[5] Samples from the posterior distribution $\pi\left(\delta \mid d_0, \hat{\mu}^*, \hat{\tau}^{2*}\right)$ can be determined through direct approximation (i.e., see Gelman, Carlin, Stern, & Rubin, 2004, pp. 283-4) using the following steps. First, determine a large and very fine grid of evenly spaced values on the noncentrality parameter λ (e.g., from -25 to 25 by increments of .02). Next, determine the density of the prior distribution for each point on the grid, which is given by $\lambda / w_0 = \delta \sim N\left(\hat{\mu}^*, \hat{\tau}^{2*}\right)$ where $w_0 = \sqrt{1/n_{1_0} + 1/n_{2_0}}$ and reflects the sample sizes for the initial effect size estimate. Next, the density of the observed *t*-value given λ is readily obtained using the noncentral *t*-distribution for each point on the grid of λ values (e.g., the likelihood function $f(\hat{\lambda}_0 \mid \lambda) = t(\hat{\lambda}_0 \mid \lambda, df = n_{1_0} + n_{2_0} - 2)$ where $\hat{\lambda}_0 = d_0 w_0$). The

density of the posterior distribution is then the product of the prior density, the likelihood, and the scaling constant a^{-1} defined by Equation (3). The grid of noncentrality parameter values is then rescaled into the metric of standardized mean differences ($\delta = \lambda / w_0$) and random samples are drawn from the grid using the density of the posterior distribution as sampling weights. For the examples presented we use the faster normal approximation which provide essentially equivalent results.

[6] Because an estimate of the random effects variance was not originally provided, these estimates were generated by reanalysis of the data presented in Webb & Sheeran (2006; Table 1) under restricted maximum likelihood. In the interest of simplifying the presentation, both the empirical and hierarchical Bayesian analyses treat all of the values in Table 1 as the observed estimates.

[7] MCMC methods involve first specifying the posterior distributions of the parameters of interest in the model. After providing initial starting values, random draws are made from these posterior distributions. The parameters in the model are estimated using these values, and then the entire process is repeated iteratively a large number of times. Under some minimal assumptions, the process converges to the target posterior distribution. The three equations necessary in the present context for a hierarchical Bayesian analysis, assuming a normally distributed prior distribution and a non-informative hyperprior (i.e., the prior distribution of the prior distribution), are illustrated below (e.g., see Gelman, Carlin, Stern, & Rubin, 2004).

$$\delta_j \left| \mu, \tau^2, \sigma, d_j \sim \pi\left(\delta_j \mid d_j, \mu, \tau^2\right) = c^{-1} f\left(d_j \mid \delta\right) g\left(\delta \mid \mu, \tau^2\right) \right. \qquad (7)$$

$$\mu|\delta_j,\tau^2,\sigma \sim N\left(\sum_{j=1}^{m+1}\delta_j/(m+1),\ \tau^2/(m+1)\right) \tag{8}$$

$$\tau^2|\delta_j,\mu,\sigma,d_j \sim inv-\chi^2\left(m,\ \frac{1}{m}\sum_{j=1}^{m+1}(\delta_j-\mu)^2\right). \tag{9}$$

Iterating through these three posterior distribution equations will eventually result in convergence to the posterior distribution of interest. Note that *m* is the number of studies in the meta-analysis; the observed effect size estimate is also included in the analysis, resulting in *m*+1 effect sizes.