Psychological Methods
2017, Vol. 22, No. 1, 94–113

1082-989X/17/$12.00 http://dx.doi.org/10.1037/met0000089

Estimating the Standardized Mean Difference With Minimum Risk: Maximizing Accuracy and Minimizing Cost With Sequential Estimation

Bhargab Chattopadhyay
University of Texas at Dallas

Ken Kelley
University of Notre Dame

The standardized mean difference is a widely used effect size measure. In this article, we develop a general theory for estimating the population standardized mean difference by minimizing both the mean square error of the estimator and the total sampling cost. Fixed sample size methods, when sample size is planned before the start of a study, cannot simultaneously minimize both the mean square error of the estimator and the total sampling cost. To overcome this limitation of the current state of affairs, this article develops a purely sequential sampling procedure, which provides an estimate of the sample size required to achieve a sufficiently accurate estimate with minimum expected sampling cost. Performance of the purely sequential procedure is examined via a simulation study to show that our analytic developments are highly accurate. Additionally, we provide freely available functions in R to implement the algorithm of the purely sequential procedure.

Keywords: sequential procedure, research design, effect size, sample size planning, accuracy

Supplemental materials: http://dx.doi.org/10.1037/met0000089.supp

One of the most commonly used effect sizes in psychology and related disciplines is the standardized mean difference. In the PsycINFO database from the beginning of 2000 until the end of 2015 there have been 1,687 peer-reviewed and scholarly articles that have included the terms standardized mean difference (or equivalent) in the abstract.[1] We searched only in the abstract because if a specific effect size is mentioned in the abstract, it would seem to be a highly important outcome of a study. Of course, many more articles mention the term somewhere in the body of the text, such as in the Results section. Our point in performing this search is simply to show the relevance of the standardized mean difference and its growing importance. As can be seen in Figure 1, there has been a steady increase in usage of the standardized mean difference as a primary discussion point of articles, in that the term is included in the abstract. We surmise that the rapid increase since the year 2000 is likely due, at least in part, to the recommendations of Wilkinson and the American Psychological Association (APA) Task Force on Statistical Inference, who stated that researchers should "*always present effect sizes for primary outcomes*" (emphasis in original), and the subsequent changes to the publishing guidelines of journals in psychology, education, and related fields (e.g., American Psychological Association, 2001), and especially its successor (American Psychological Association, 2010; see also American Educational Research Association, 2006; Association for Psychological Science, 2014), and based on the recommendation of methodologists.

Recommendations long made by methodologists within psychology and related disciplines, among others, about an overreliance on null hypothesis significance tests and the corresponding *p* value, the need to focus on effect sizes, and the importance of confidence intervals for population effect sizes, among other suggestions, has now been echoed by the American Statistical Association (ASA) in what is "the first time the ASA has spoken so publicly about a fundamental part of statistical theory and practice" (American Statistical Association, 2016). In an editorial by Ron Wasserstein, the ASA's Executive Director, on behalf of the ASA Board of Directors (Wasserstein & Lazar, 2016), six principles are addressed that could "improve the conduct or interpretation of quantitative science" (p. 131). The editorial goes on to say that "in view of the prevalent misuses of and misconceptions concerning *p*-values, some statisticians prefer to supplement or even replace *p*-values with other approaches" (p. 132). The suggestions for supplementing or replacing *p*-values are "methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discov-

This article was published Online First September 8, 2016.

Bhargab Chattopadhyay, Department of Mathematical Sciences, University of Texas at Dallas; Ken Kelley, Mendoza College of Business, University of Notre Dame.

Both authors contributed equally and authorship is alphabetical.

Correspondence concerning this article should be addressed to Bhargab Chattopadhyay, Department of Mathematical Sciences, University of Texas at Dallas, 800 West Campbell Rd, FO 2.402A, Richardson, TX 75080. E-mail: bhargab@utdallas.edu; or to Ken Kelley, Department of Information Technology, Analytics, and Operations, Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556. E-mail: kkelley@nd.edu

[1] We searched specifically for "standardized mean difference," "Cohen's *d*," "Cohens *d*," "Hedges' *g*," "Hedges *g*," or "standardized difference between means" in the abstract of peer-reviewed and scholarly articles. This search was performed on March 7, 2016 using PsycINFO.

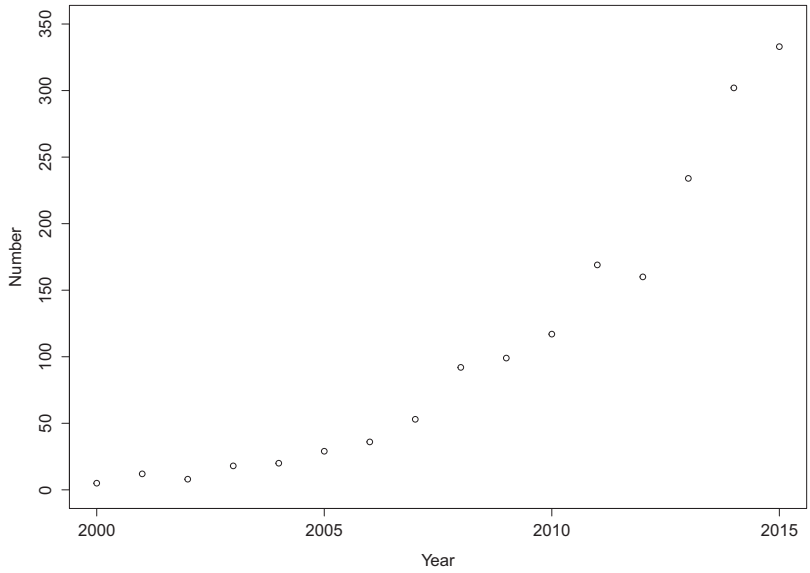

Figure 1. Number of times "standardized mean difference" (or equivalent) appeared in the abstract of peer-reviewed scholarly journal articles from 2000 to 2015 (obtained using PsycINFO).

ery rates. All these measures and approaches rely on further assumptions, but they may more directly address the size of an effect (and its associated uncertainty) or whether the hypothesis is correct" (p. 132). Our work directly addresses the size of the effect, here for the standardized mean difference, and does so with accuracy, as well as cost, in mind. Thus, we believe that our work is both timely and important for helping to advance psychology and related disciplines by focusing so explicitly on estimation of an important quantity.

The reason the standardized mean difference, and standardized effect sizes in general, are of such interest and widely used in psychology and related disciplines is that the metrics used in these disciplines are often arbitrary. Thus, when examining the difference between two groups with an arbitrary scale, the raw difference itself is often times not easily interpretable. For example, a 5-unit difference between estimated means on a survey that is on a 1–7 scale may mean one thing, whereas a 5-unit difference might mean something else on a 0–100 scale. Thus, by dividing the raw mean difference by the (usually) pooled standard deviation, the raw difference is rescaled by the within group variability leading to an effect size that is standardized (see Kelley & Preacher, 2012, for a discussion of effect sizes).

Methods of designing studies that consider the standardized mean difference as the primary outcome are numerous. For example, Cohen (1988) details how to plan sample size for statistical power when interest concerns the standardized mean difference. Raudenbush (1997) focuses on cost-effective sampling for cluster randomized designs. Kelley and Rausch (2006) detail how to plan sample size for accurate estimates when interest concerns the standardized mean difference accompanied by a narrow confidence interval. Pornprasertmanit and Schneider (2014) extend the accuracy approach to cluster randomized designs and also consider the sampling cost of data collection, which is something most articles on sample size planning in psychology fail to do. In practice, of course, the cost of collecting data is a major issue and is arguably the biggest obstacle researchers face when collecting an appropriate sample size. Unfortunately, many sample size planning developments have overlooked this important issue. Although in many cases it may be reasonable to focus on the literal sample size needed to satisfy a particular goal, it is certainly also reasonable to consider the cost as well and not overlook this important factor of a research project.

Our approach simultaneously considers sampling cost and estimation accuracy of the standardized mean difference. It is well known that as sample size increases, more information is obtained about the parameter of interest. As the sample size increases, $E[(\hat{\theta} - \theta)^2]$, which is the mean square error, decreases, where θ is an arbitrary parameter and $\hat{\theta}$ is an estimator of θ. When the mean of the squared distances between the estimator and the parameter it estimates is small (i.e., a small mean square error), it implies the estimator is producing accurate estimates. It is highly desirable to obtain accurate parameter estimates (e.g., Kelley, Maxwell, & Scott, 2003). Of course, increasing sample size also increases the sampling cost. Thus, researchers find themselves in a conundrum in which there is a necessity to keep both the mean square error small (i.e., a high degree of accuracy) and the sampling cost at a minimum. These are opposing goals: Increasing sample size leads to higher accuracy but also higher sampling cost, yet reducing sample size reduces sampling cost but leads to lower accuracy.

In the power analytic framework, a fixed sample size procedure is often used. A fixed sample size procedure is one in which sample size is planned a priori (i.e., before sampling begins) based on one or more specified values. Such "specified values" might be for a known, hypothesized, previously obtained value (e.g., from another study), or theoretically interesting parameter value (such as the minimum effect size of substantive interest). For example, the methods discussed in sample size planning books that are popular in psychology and related disciplines are "fixed" sample size procedures (e.g., Cohen, 1988; Davey & Savla, 2010; Kraemer & Thiemann, 1987; Murphy & Myors, 2004). However, a

fixed sample size procedure cannot achieve a compromise between sampling cost and mean square error (e.g., see Chattopadhyay & Kelley, in press; Ghosh & Sen, 1991; Sen, 1981). This is the case because once the sample size is selected, based on the particular scenario, sample size is, by definition, a fixed value without any consideration of sampling cost. Of course, if the specified input values used in a fixed sample size planning procedure is inappropriate, the procedure implied sample size may yield too many participants or too few. With a fixed sample size procedure, no data collected as part of the study can help inform any sort of "stopping rule" (i.e., a rule that states when the data collection will end), as the sample size is necessarily set a priori and any data collected is not, by definition, able to provide information on stopping the data collection. Although one could collect data until the financial resources run out, such a strategy is a poor substitute for a study planned by considering goals and sampling cost simultaneously and arguably has ethical implications (e.g., see Maxwell & Kelley, 2011).

To be clear, our method does not address statistical power or null hypothesis significance testing in any way. Our method supports the move toward an estimation based literature instead of a literature based on null hypothesis significance testing. Our work goes further than the Task Force on Statistical Inference suggested by not simply *presenting* a primary effect size, but rather by obtaining a sufficiently accurate estimate of the effect size, while considering study cost. Importantly, our method answers the call by professional organizations to focus research on the size of effects rather than a dichotomous outcome from a null hypothesis significance test (e.g., reject or fail-to-reject). For example, the executive committee of the Society for Personality and Social Psychology recently convened a Presidential Task Force on Publication and Research Practices (Funder et al., 2014) to make recommendations on how to "improve the dependability and replicability of research findings in personality and social psychology" (p. 3). The guidelines from the Society for Personality and Social Psychology Task Force on Publication and Research Practices states that "the problems with focusing exclusively on the observed p level are exacerbated when researchers overrely on the dichotomous distinction between 'significant' and 'non-significant' results" (p. 5). They go on to state, with regards to treating results as dichotomous based on a p value being less than the Type I error rate (e.g., .05) that this common practice risks treating nearly equivalent findings as if they were importantly different, especially if one finding barely attains the p-value < 0.05 threshold whereas the other barely misses it. Our sequential method that we develop here avoids the "significant" and "nonsignificant" verbiage of the null hypothesis testing framework and approaches research from a purely estimation perspective. Although our approach does not preclude one from also testing a null hypothesis or forming a confidence interval, these inferential procedures should only be done after the stopping rule has been met. Correspondingly, even though the procedure is evaluated again and again, only one confidence interval and only one null hypothesis significance test should be calculated, and these calculations should only be done after the stopping rule has been satisfied (and thus sampling has stopped).

We realize that our approach is not the only way that a study could be designed and conducted and that null hypothesis testing may be important in a particular application. However, we want to make clear that our method addresses estimation, which has taken on an increasingly important role in psychological research in recent years. As the APA publication manual states: "It is almost always necessary to include some measure of effect size in the Results section" (American Psychological Association, 2010, p. 34). For our purposes, we are working toward the accurate estimation of the standardized mean difference while considering cost. Similarly, the Association for Psychological Science (APS) implemented a change effective beginning in 2014: "*Psychological Science* recommends the use of the 'new statistics'—effect sizes, confidence intervals, and meta-analysis—to avoid problems associated with null-hypothesis significance testing (NHST)" (Association for Psychological Science, 2014). In an editorial for *Psychological Science*, the flagship journal of the Association for Psychological Science (2014), Eich (2014) states strong support for the use of estimation and basing interpretation of results on point and interval estimates, which he notes harkens back to the APA publication manual: "Psychologists should, whenever possible, use estimation and base their interpretation of research results on point and interval estimates" (p. 5). We believe that there is a shift in the field and there is a clear need for methodologists to focus on better ways of estimating effect sizes, which is our aim.

We are certainly not the first to suggest sequential methods as a way to improve estimation. In fact, the idea of sequential sampling methods comes from Mahalanobis (1940) for estimating acreage of jute crop in the whole state of Bengal. In 1943 Wald used sequential methods in military applications and the methods were thought to be so valuable there were classified "restricted" until 1945 (Statistical Research Group, 1945; see also Wald, 1945). A modern example is from Petrie, Bulman, and Osborn (2002), who proposed the application of sequential methods in dentistry. Leroux, Mancl, and DeRouen (2005) used such sequential methods for longitudinal clinical trials within dentistry and developed a sequential testing procedure for multiple endpoint trials and applied the sequential methods for studying the safety of dental amalgam fillings. In pharmacology, Todd, Whitehead, Stallard, and Whitehead (2001) advocated the use of sequential methods for stopping additional data collection as soon as there is sufficient evidence to reach a firm conclusion in phase III clinical trials over fixed sample size procedures citing economic and ethical reasons. Donaire et al. (2009) conducted sequential analysis of the fMRI (functional MRI) data obtained during epileptic seizures in order to study the temporal development of BOLD (blood oxygenation level dependent) signal changes in patients. For other examples and applications, we refer interested readers to Armitage (1969), Jennison and Turnbull (2010), and Miladinovic et al. (2013), among others.

In this article, we propose a solution to choosing an appropriate sample size for the standardized mean difference. In particular, we develop a procedure which simultaneously reduces both the mean square error for the standardized mean difference and sampling cost associated with collecting data. We are able to do this with a novel application of sequential analysis. Our specific proposal is a purely sequential procedure that yields an estimate of the population standardized mean difference using a sample size that optimizes the mean square error (i.e., accuracy) of estimating the unknown population standardized mean difference and the sam-

pling cost. In a purely sequential procedure, preliminary information about the parameter of interest is obtained by first collecting a small sample (called the pilot sample). Then, in successive stages, the same number of additional observations (e.g., 1, 2, 5, 10) are collected and used to simultaneously update the estimate of the parameter. This is done repeatedly until a prespecified condition (i.e., a stopping rule) has been satisfied.

We believe that our work has the potential to fundamentally shift the manner in which studies are designed in psychology and related disciplines, as the ideas we discuss extend beyond the standardized mean difference. Although most sample size planning methods are developed under the fixed sample size perspective, sequential methods are arguably more appropriate in many instances. A priori sample size planning methods, that is, "fixed *n*" designs, generally do not acknowledge that the value(s) of the parameter(s) used for the sample size planning procedure may be different from the actual value of the parameters in the population of interest. When there is a nontrivial difference between the input and the actual values of the population parameters, the sample size used in a study can be severely under or overestimated. Our proposal here for the standardized mean difference, a very important and commonly used effect size in psychology and related fields, seeks to obtain an accurate estimate while also considering the sampling cost; ignoring the sampling cost is an unrealistic constraint of practical research.

Estimation of the Standardized Mean Difference

We now begin to formalize our ideas, beginning with estimating the population standardized mean difference for two independent groups. The population standardized mean difference is defined as

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

where μ_1 and μ_2 are the population means from Groups 1 and 2, respectively, and σ is the population standard deviation of scores within the two groups under the homogeneity of variance assumption ($\sigma_1^2 = \sigma_2^2 = \sigma^2$). Thus, if the population standardized mean difference is 0.50, then the population mean for Group 1 is 0.50 standard deviations larger than the population mean for Group 2.

In practice, the population values of the means, μ_1 and μ_2, and common standard deviation, σ, are unknown. As a result, δ itself is unknown. However, this population value is of primary interest in many studies. Because the value of the population standardized mean difference is often a primary outcome in studies (recalling Figure 1), it is important to estimate the population standardized mean difference using sample estimates. We use $\bar{X}_{1n_1}$ and $\bar{X}_{2n_2}$ to denote the sample mean of scores on an outcome of interest from Groups 1 and 2, respectively. Groups 1 and 2 have n_1 and n_2 individuals, respectively, in the group. We use $s^2_{1n_1}$ and $s^2_{2n_2}$ to represent the usual unbiased estimator of the common variance from Group 1 and Group 2, respectively. Notice that our sample means and sample variances have an n_j ($j = 1, 2$) subscript. We note the sample size in the subscript of the estimators explicitly to denote the sample sizes on which these estimators are based. This notation is useful when we discuss properties of estimators at different sample sizes. With this notation, the commonly used estimator of the population standardized mean difference defined in Equation 1 is

$$d_{\tilde{n}} = \frac{\bar{X}_{1n_1} - \bar{X}_{2n_2}}{s}, \tag{2}$$

where s is the pooled sample standard deviation defined as

$$s = \sqrt{\frac{(n_1 - 1)s^2_{1n_1} + (n_2 - 1)s^2_{2n_2}}{n_1 + n_2 - 2}}. \tag{3}$$

Here, the total sample size from both groups is $\tilde{n} = n_1 + n_2$. Now, let us assume that the scores of n_1 individuals belonging to Group 1 are sampled from a normal distribution with population mean μ_1 and population variance σ^2. Further, assume that scores of n_2 individuals belonging to Group 2 are sampled from a normal distribution with population mean μ_2 and population variance σ^2. Using notation, we write the normal distribution assumption for Groups 1 and 2 as

$$X_{i1} \sim N(\mu_1, \sigma^2) \tag{4}$$

and

$$X_{i2} \sim N(\mu_2, \sigma^2), \tag{5}$$

respectively. We use $N(\mu_j, \sigma^2)$ to denote that the jth group follows a normal distribution with population mean μ_j ($j = 1, 2$) and variance σ^2. We will assume a normal distribution within each of the two groups for the remainder of the article.

Suppose we want to not only estimate δ, but also to *accurately* estimate δ, such that the estimated standardized mean difference, $d_{\tilde{n}}$, is sufficiently close to δ. Further suppose that $d_{\tilde{n}'}$ is an estimator of δ based on a total sample of size $\tilde{n}'$, such that $\tilde{n}' < \tilde{n}$. Following Rao (1973, p. 315), $d_{\tilde{n}}$ is preferred over $d_{\tilde{n}'}$ as an estimator of δ because the probability that $d_{\tilde{n}}$ lies between $[\delta - \epsilon, \delta + \epsilon]$ is higher than the probability that $d_{\tilde{n}'}$ lies between $[\delta - \epsilon, \delta + \epsilon]$ for all $\epsilon > 0$ and δ, with ε representing the maximum probable error. Holding everything else constant, statistically an estimate that is based on a larger sample size is preferred to an estimate based on a smaller sample size. This can be proven using Chebyshev's inequality for any type of distribution (see, e.g., Lord, 1953, or Lim & Leek, 2012).[2] By application of Chebyshev's inequality, the probability that the estimate of $d_{\tilde{n}}$ will lie outside the interval $[\delta - \epsilon, \delta + \epsilon]$ will be bounded above by (i.e., no larger than)

$$P(|d_{\tilde{n}} - \delta| \geq \epsilon) \leq \frac{\mathrm{E}[(d_{\tilde{n}} - \delta)^2]}{\epsilon^2}. \tag{6}$$

Thus at most, $\mathrm{E}[(d_{\tilde{n}} - \delta)^2]/\epsilon^2 \times 100\%$ of the values of $d_{\tilde{n}}$ lie outside the interval $[\delta - \epsilon, \delta + \epsilon]$ and ϵ is known as the maximum probable error.

Suppose that the experimenter is willing to pay \$1,000 so that the maximum probable error in estimating the true standardized mean difference, δ, using the sample standardized mean difference, $d_{\tilde{n}}$ is ϵ. In other words, the researcher is willing to pay \$1,000 so that the squared maximum probable error will be ϵ^2, that is, the squared difference between the point estimate $d_{\tilde{n}}$ and δ will be at

[2] For unbiased estimators, no more than $1/k^2 \times 100\%$ of the values of the distribution can be more than k standard deviations away from the parameter.

most ϵ^2. But holding everything else constant, due to the sampling error the amount will be, $A\mathrm{E}[(d_{\tilde{n}} - \delta)^2]$, where, $A = \$1{,}000/\epsilon^2$. Thus, A has a unit "dollar per square unit of ϵ". The value of A depends not only on the amount of money (e.g., U.S. dollars, Euros, British pounds) the researcher is willing to pay for a sufficiently small deviation from the parameter (i.e., the maximum absolute difference desired between the population value and its estimate), but also on the value of ϵ. Hence, holding everything else constant, the experimenter will pay less for a larger value of ϵ, and pay more for a smaller value of ϵ. One may note that the value of ϵ specified by the researcher is based on the study's goals and for the specified maximum probable error, ϵ, the amount that the experimenter will be paying is context specific. In exploratory studies, for example, one may allow ϵ to be larger than, for example, a confirmatory study in which important decisions will follow based on the size of the effect. Further details of the interpretation of A and the method more generally in the context of the coefficient of variation, are considered in Chattopadhyay and Kelley (in press). Another interpretation of A from a decision theoretic perspective, yet one that is conceptually similar, is given in Mukhopadhyay and De Silva (2009).

General statistical theory shows that, holding everything else fixed, an estimate of a parameter based on a larger sample size is preferable to an estimate of the same population parameter that is based on a smaller sample size. This is the case because the probability is higher that an estimate is closer to the population value when the sample size is larger as compared with the estimate being based on a smaller sample size. In our case, an estimate of δ based on a larger sample size is preferable to an estimate based on a smaller sample size, holding everything else constant, because the sampling distribution of $d_{\tilde{n}}$ has a higher concentration (density) around δ and a lower concentration in the tails of the sampling distribution of $d_{\tilde{n}}$ as compared with the sampling distribution of $d_{\tilde{n}'}$ ($\tilde{n}' < \tilde{n}$). As a result, holding everything else constant, the mean square error of $d_{\tilde{n}}$, $\mathrm{E}[(d_{\tilde{n}} - \delta)^2]$ is smaller than the mean square error of $d_{\tilde{n}'}$.

Statistically, the accuracy of an estimator is defined as the MSE, which is equal to the sum of its variance and the square of its bias (e.g., Rozeboom, 1966; see also Kelley et al., 2003). For an unbiased estimator, the precision, which is the reciprocal of variance, and accuracy are equivalent concepts. As has been discussed in the literature (e.g., Hedges, 1981; Hedges & Olkin, 1985), $d_{\tilde{n}}$ is a biased estimator of δ. Correspondingly, $\mathrm{E}[d_{\tilde{n}} - \delta] \neq 0$ and $\mathrm{E}[(d_{\tilde{n}} - \delta)^2]$ is the MSE of $d_{\tilde{n}}$. By applying Taylor's theorem in the approximate expression of MSE as deduced in Hedges (1981) and proved here in Lemma 1, the expression of the MSE of d_n is given by

$$\mathrm{E}[(d_{\tilde{n}} - \delta)^2] \approx \left(\frac{1}{n_1} + \frac{1}{n_2}\right) + \frac{\delta^2}{2(n_1 + n_2)}. \tag{7}$$

It can be seen from the first component of the right-hand-side of Equation 7 that the MSE is inversely related to the sample size. Therefore, an increase in total sample size will decrease the MSE and thus increase the probability that $d_{\tilde{n}}$ will be between $[\delta - \epsilon, \delta + \epsilon]$, holding everything else constant. However, in practice, increasing sample size increases the cost of collecting data (i.e., sampling cost), holding everything else constant. Thus, a larger sample size will increase the accuracy of the estimator but will require a higher sampling cost. This is an important issue and at the heart of our work here. The specific problem we solve is finding the minimum sample size required to estimate δ to a specified level of accuracy while taking into consideration the sampling cost.

Minimum Risk Point Estimation Problem

As discussed, the MSE of d_n (i.e., $\mathrm{E}[(d_n - \delta)^2]$) becomes smaller as the sample size grows larger; that is, as we get more and more information about the unknown population standardized mean difference, the accuracy improves. However, improving accuracy (i.e., reducing the MSE) leads to a larger sampling cost. If smaller sampling cost is desired, then a smaller sample size can be taken, but this in turn will increase the MSE. Our goal is to find an optimization procedure to ensure maximum accuracy while minimizing sampling cost. By sampling cost we mean the actual cost, such as dollars, associated with collecting each additional observation.

To account for (a) A (i.e., the price one is willing to pay so that the maximum probable error in estimating the true standardized mean difference is ϵ); (b) accuracy of an estimator (i.e., MSE); and (c) the sampling cost, we define a function, often called a risk function, that simultaneously considers these three factors. The risk function is defined as

$$R_{\tilde{n}}(\delta) = A\mathrm{E}[(d_{\tilde{n}} - \delta)^2] + c(n_1 + n_2), \tag{8}$$

where c is the cost of sampling a participant, $c(n_1 + n_2)$ is the cost of sampling n_1 observations from Group 1 and n_2 observations from Group 2, and A depends on ϵ^2. The risk function gives, on average, what can be called the *total expected cost* of estimating the population standardized mean difference using n_1 observations from Group 1 and n_2 observations from Group 2 with a maximum probable error ϵ. Note that this "total expected cost" is an "expected cost" due to the expectation in Equation 8. Further, note that the "total expected cost" is a "total" because it considers the price one is willing to pay for a sufficiently accurate estimate, not just the sampling cost (i.e., $c(n_1 + n_2)$). It is this risk function of Equation 8 that we seek to minimize, as minimization of this function is what leads to a research design framework that considers the three important factors: A, the mean square error, and sampling cost. For more details regarding this risk function, we refer readers to Sen (1981).

For example, if an interviewer spends 1 hr with a participant at a rate of \$15 per hr and a proprietary exam is used that costs \$5 per participant, the cost, c, is \$20 (assuming no other costs for sampling). In this scenario, if the researcher is willing to pay \$1,000 to have the maximum probable error between the unknown population standardized mean difference and its estimate be $\epsilon = .10$, then $A = \$1{,}000/0.1^2 = \$100{,}000$. Consider the funding for a grant in which the effectiveness of a treatment is to be evaluated. One could conceptualize the cost of obtaining (a) the point estimate; (b) hypothesis test; and (c) confidence interval (which are usually regarded as the primary outcomes of a study) as being worth the total cost of the study. In particular, to obtain these three values, it costs some amount, which is the total amount funded by the grant agency.

To put into perspective the idea of the cost associated with the outcomes that relate directly to the effectiveness of a treatment, consider the 13 grants funded by the Institute of Education Sciences (IES; the research arm of the U.S. Department of Education)

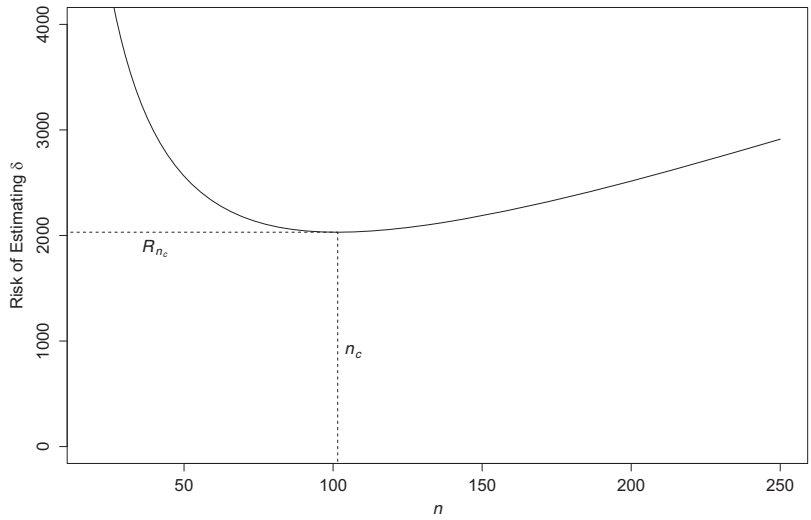

Figure 2. Risk (Equation 8) as a function of sample size for $\delta = 0.5$, $A = 1{,}125/\epsilon^2$, $\epsilon = 0.15$, and $c = 5$.

in 2013 that are classified as "Efficacy and Replication" (using http://ies.ed.gov/funding/grantsearch with the appropriate options selected). These 13 grants totaled $42,751,921 in funds granted by IES. The total sample size of the primary unit of interest (e.g., students, teachers, families) was 20,781, for a cost of $2,057.26 per research participant. Our point here is to say that it is not unreasonable to consider a price one is willing to pay, in the literal sense, for an accurate measurement of the primary outcome of interest. Note that if the researcher wants an estimate with expectation closer to the true value δ, ϵ needs to be smaller. Thus, a higher cost would be necessary because it will lead to a larger sample size, holding everything else constant.

In some ways, we might have conceptualized the goals of the studies too broadly, as the main outcome of interest for some researchers is a dichotomous variable, namely, "was the null hypothesis of no effect rejected?" Our work is consistent with recommendations from professional organizations, methodologists, editors, et cetera, which is to go beyond the results of a null hypothesis and consider estimation and accuracy of the effect sizes that drive research questions (see, e.g., Kelley & Preacher, 2012 for a review and references).

Using the approximate expression of MSE of $d_{\bar{n}}$ from Equation 7, the approximate risk function defined in Equation 8 becomes

$$R_{\bar{n}}(\delta) \approx A\left\{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)+\frac{\delta^2}{2(n_1+n_2)}\right\} + c(n_1+n_2). \quad (9)$$

This gives the approximate total expected cost to estimate the unknown population standardized mean difference using estimator $d_{\bar{n}}$ from n_1 observations from Group 1 and n_2 observations from Group 2 with a maximum probable error ϵ (i.e., it is the approximate risk). Our specific objective in this article is to find the sample size for which the approximate risk function defined in Equation 9 is minimized. This problem is known as the *minimum risk point estimation* problem (e.g., see Chattopadhyay & Kelley, in press; Ghosh & Sen, 1991; Sen, 1981).

For not too small sample sizes, provided δ^2 is known, the approximate risk function in Equation 9 is minimized if

$$n_c = \sqrt{\frac{A}{2c}}\xi \quad (10)$$

individuals are selected from Groups 1 and 2, where,

$$\xi^2 = \left(2+\frac{\delta^2}{4}\right). \quad (11)$$

(this is proved in Lemma 2 in Appendix A).[3] That is, n_c is the theoretically optimal (true) sample size that should be collected from each of the two groups in order to minimize the total expected cost to estimate δ if, in fact, the true value of δ was used. By "theoretically optimal" we mean that, if the true parameter(s) were known and all assumptions met, n_c is the sample size that satisfies the goal.

Using Equation 9, the approximate total expected cost of estimating the population standardized mean difference (δ) using a total sample of size $2n_c$, is denoted as,

$$R_{n_c}(\delta) = \frac{A}{n_c}\xi^2 + 2cn_c = 4cn_c, \quad (12)$$

where, ξ^2 is as defined in Equation 11. Thus, n_c is the theoretically optimal sample size from each group that is required to achieve a minimum risk if δ were known. See Figure 2, which shows how risk is a function of sample size for a specified situation.

Here, $n_c = \sqrt{\frac{A}{2c}}\xi$ is a nearly exact analytic solution and would be the required sample size from both groups if δ was known. However, we note that we are talking about estimating δ; if one knew the value of δ it would not need to be estimated from a sample. Hence, for estimating the unknown population standardized mean difference, δ, it will not be a reasonable choice to use for planning sample size, as it will almost always be unknown. Our method minimizes the total expected cost with the use of a purely sequential procedure in which the required sample size will not be

[3] By "not too small sample sizes" here and elsewhere we mean a sample size that is large enough so that the noted properties hold. The exact value of "not too small" is context specific.

fixed in advance. The theoretically optimal sample size will necessarily be unknown because we assume that a researcher does not know δ before the start of the study. We prove statistically and demonstrate with a Monte Carlo simulation study in a later section that our method yields a sample size that closely approximates n_c in applied situations (i.e., when δ is unknown) and also that the cost of estimating the population standardized mean difference using sequential procedure is close to the theoretical total expected cost given in Equation 12.

Sequential Optimization Procedure

As opposed to fixed-sample procedures, in sequential procedures, the sample size is not fixed in advance. No fixed sample-size procedure can provide a solution to the minimum risk point estimation problem (e.g., see Chattopadhyay & Kelley, in press; Dantzig, 1940; De & Chattopadhyay, 2015). Here, we propose a purely sequential procedure to estimate the population standardized mean difference.

In a sequential procedure, the estimation of parameter(s) continues in stages. In the first stage, a small sample called a pilot sample is observed, and then the parameters are estimated to check a predefined condition in a predefined rule, which is known as the stopping rule. Further sampling of observations is carried out if the predefined condition is not met, with further sampling stopped once the predefined condition is satisfied. At a particular stage, if the predefined condition is not met, the researcher collects one or more additional observations and then estimates the parameter of interest. This process is repeated until the predefined condition is met. For details about the general theory of sequential estimation procedures, we refer interested readers to Sen (1981), Ghosh and Sen (1991), Mukhopadhyay and Chattopadhyay (2012), and Chattopadhyay and Mukhopadhyay (2013).

As discussed, the theoretically optimal sample size, n_c, required to minimize the function that considers the approximate mean square error and the sampling cost is unknown because it depends on δ, which is itself unknown in practice. Thus, in order to estimate n_c, an estimator of δ is desired. For the ease of notation, we will henceforth denote d_n as the estimator of the population standardized mean difference and s_n as the pooled sample standard deviation when $n_1 = n_2 = n$ observations are drawn from each of the two groups.

Recall from Equation 11 that ξ^2 depends on δ. Because δ is unknown, to get an estimator of ξ based on n observations drawn from both groups, we replace δ with the estimator $d_n = (\bar{X}_{1n} - \bar{X}_{2n})/s_n$, rewritten Equation 2. We define an estimator of ξ^2 as

$$V_n^2 = \left(2 + \frac{d_n^2}{4}\right). \tag{13}$$

We now develop an algorithm to find an estimate of the optimal sample size via the purely sequential estimation procedure.

Stages of Implementing the Methods

Stage I. First, scores of m randomly selected individuals are collected from each of the two groups. Thus there are m observations collected from Group 1 and m observations collected from individuals belonging to Group 2. Following Chattopadhyay and Kelley (in press) we recommend using the pilot sample size m given as

$$m = \max\{m_0, \lceil (A/(2c))^{1/(2+2\gamma)} \rceil\}, \tag{14}$$

where $m_0 (\geq 4)$ is the least possible sample size required to estimate δ^2 and $\lceil \cdot \rceil$ is the ceiling function of the term—the ceiling being the smallest integer not less than $(A/(2c))^{1/(2+2\gamma)}$. Based on this pilot sample of size m, an estimate of ξ^2 is obtained by computing V_m^2. If $m < \lceil \sqrt{\frac{A}{2c}}(V_m + m^{-\gamma}) \rceil$, then proceed to the next step. Otherwise, if $m \geq \lceil \sqrt{\frac{A}{2c}}(V_m + m^{-\gamma}) \rceil$, stop sampling and set the final sample size equal to $2m$. We will discuss the use of the term $m^{-\gamma}$ and choice of γ momentarily.

Stage II. Obtain m' additional scores from each group, where we set $m' = 1$ in general (and here specifically) randomly selected individuals (different from those who were selected during Stage I) belonging to a particular group. Thus, there are $m + m'$ observations from each group (i.e., the pilot sample size and an additional m' observations per group, for a total sample size of $2[m + m']$). If $m + m' \geq \lceil \sqrt{\frac{A}{2c}}(V_{m+m'} + (m + m')^{-\gamma}) \rceil$ stop further sampling and set the final sample size equal to $2(m + m')$. If $m + m' < \lceil \sqrt{\frac{A}{2c}}(V_{m+m'} + (m + m')^{-\gamma}) \rceil$, then continue the sampling process by sampling m' more individuals per group.

This process of collecting the same number of observations in each stage after Stage I continues until there are N_c observations from each group such that $N_c \geq \lceil \sqrt{\frac{A}{2c}}(V_{N_c} + N_c^{-\gamma}) \rceil$. At this stage, we stop further sampling and report that the final sample size is $2N_c$. In other words, the final sample size for each group is N_c.

At each stage of the algorithm, we check whether the sample size collected up to that stage is at least as large as the estimated value of n_c using observations collected until that stage. We recommend researchers use software, such as that we provide using the R language, to implement our procedure.

Based on the algorithm just outlined, a sampling stopping rule can be defined as follows:

N_c is the smallest integer $n (\geq m)$ such that

$$n \geq \sqrt{\frac{A}{2c}}(V_n + n^{-\gamma}), \tag{15}$$

where $\gamma \in (0, 1/2)$ and the term $n^{-\gamma}$ is a correction term which ensures that the sampling process does not stop too early for the optimal sample size because of the use of the approximate expression. For details about the correction term, refer to De and Chattopadhyay (2015), Sen and Ghosh (1981), or Chattopadhyay and Kelley (in press). Note that for not too small sample sizes, $(V_n + n^{-\gamma})$ converges to ξ. We suggest, for practical purposes, $\gamma = 0.49$. However, we note that γ can take on other values.[4] Figure 3 presents a flowchart which describes the sequential procedure that we developed.

Characteristics of Our Sequential Procedure: A Summary

If observations are collected using Equation 15, sampling is guaranteed to be eventually terminated, which is proved in Lemma

[4] For not too small sample size, $\sqrt{2 + \frac{d_n^2}{4}} + n^{-\gamma}$ converges to $\sqrt{2 + \frac{\delta^2}{4}}$. Thus, the convergence rate increases as γ increases. So a higher value of γ, for example $\gamma = 1$, is a choice. Now, if one uses a value of γ higher than 0.5, Theorem 2 will not be satisfied.

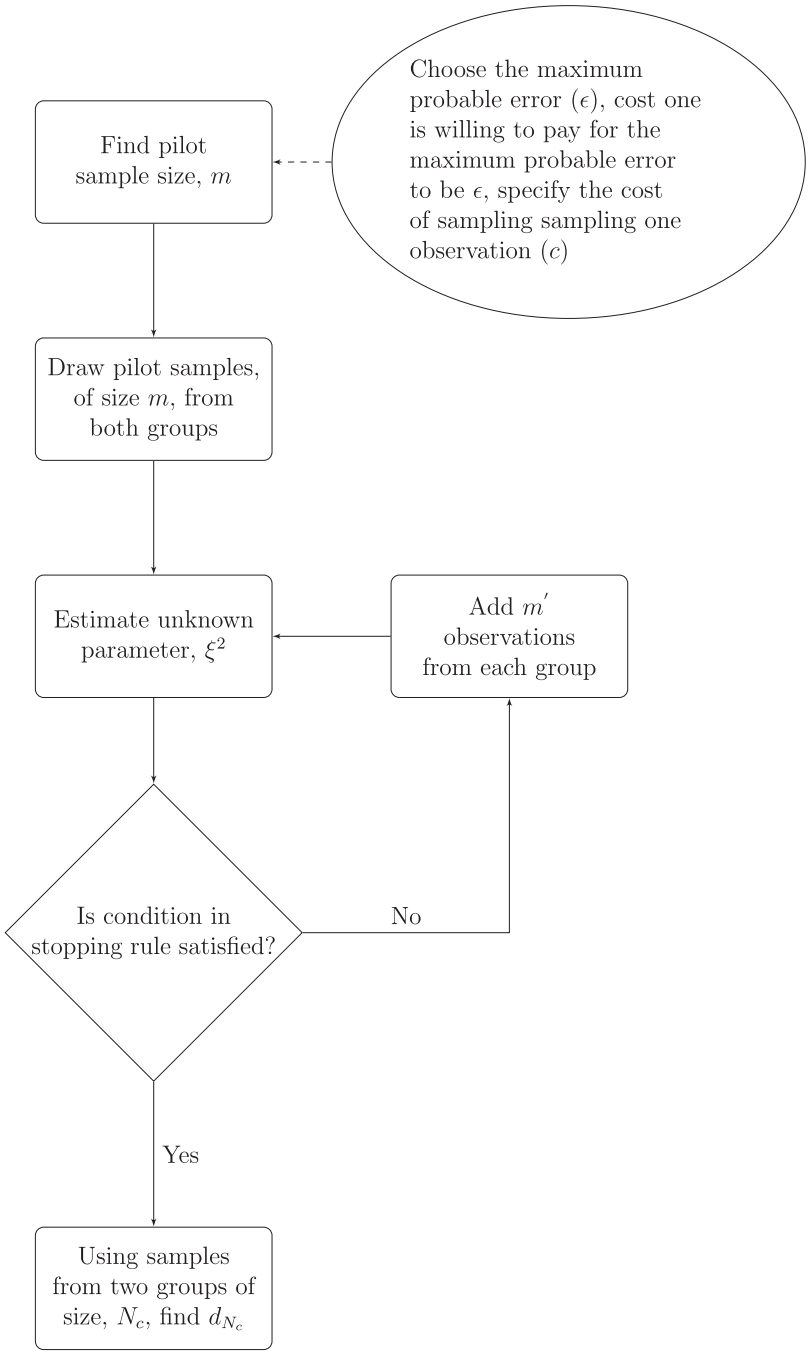

Figure 3. Flowchart that describes the sequential procedure developed.

3 in Appendix A. For a given cost c per observation, the risk function for using the estimator of the population standardized mean difference as defined in Equation 2, based on the final sample size N_c, is given by

$$R_{N_c}(\delta) = A\mathrm{E}[(d_{N_c} - \delta)^2] + 2c\mathrm{E}[N_c]. \quad (16)$$

Theorems 1 and 2 proved in Appendix A are very important. Theorem 1 indicates that, under appropriate conditions, our purely sequential procedure samples on an average n_c observations from each group. Theorem 2 ensures that, on average, the cost of estimating population standardized mean difference using a total of $2N_c$ observations is close to the theoretically minimum cost, $R^*_{n_c}(\delta)$, defined in Equation 12. What this means from a practical perspective is that we were able to show that the procedure we developed will, on average, produce the (a) theoretically optimal sample size and (b) cost almost the same as the theoretically minimum total expected cost.

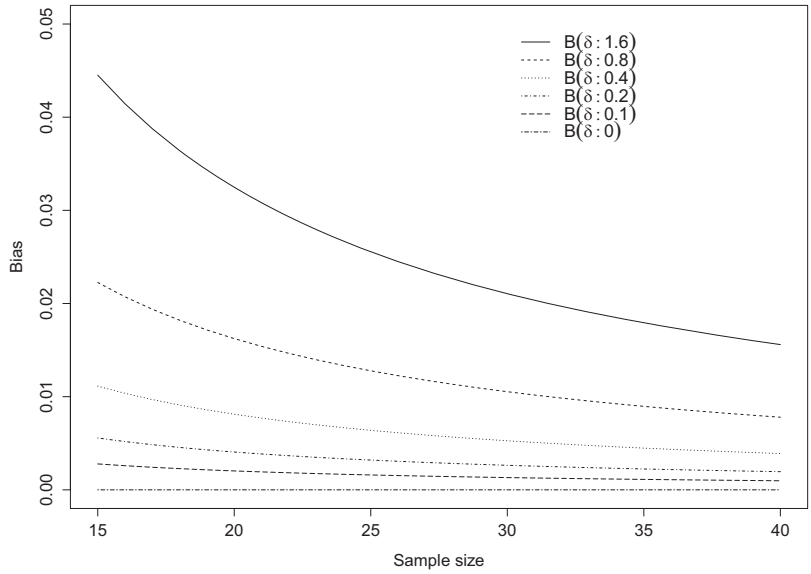

Figure 4. Bias in estimating δ using the estimator d_n for equal sample sizes.

Characteristics of the Final Sample Size: A Simulation Study

We now demonstrate the properties of our method using a Monte Carlo simulation study. The method, as discussed above, with important proofs and lemmas in Appendix A, produces what is statistically a nearly exact procedure. However, as our aim here is to illustrate the properties of the final sample and the distribution of final sample sizes, we provide the demonstration that follows.

To implement the sequential procedure in this Monte Carlo demonstration, we fix the cost of sampling each unit (e.g., person) in both populations to be $c = \$1$. Suppose that the researcher is willing to pay \$1,125 so that the absolute difference between the point estimate of the standardized mean difference, $d_{\tilde{n}}$ and the true value, δ, will be at most $\epsilon = 0.15$, and correspondingly $A = \$50{,}000 (= 1{,}125/0.15^2)$. Here, we use $\gamma = 0.49$ as suggested in the previous section. We compute the pilot sample size by using the pilot sample size formula given in the algorithm mentioned in the previous section: $m = \max\{4, \lceil(50{,}000/(2 \times 1))^{1/(2+2\times 0.49)}\rceil\} = 30$ (the pilot sample size is 30). In the other scenario, we used the values of $A = \$40{,}000/0.20^2$ and $c = \$500$ for $\epsilon = 0.20$ and then similarly computed the pilot sample size, m for both the combinations of A and c. We use several combinations of A, c and ϵ.

We then implement the purely sequential procedure and, for the sample size (N), we estimate the mean sample size ($\bar{N}$), the standard error (s($\bar{N}$)) of N, the standardized mean difference ($\bar{d}_N$), the proportion of times $|d_N - \delta| \leq \epsilon$ (p), the risk efficiency ($\bar{r}_N$) and its standard error (s($\bar{r}_N$)) based on 5,000 replications via Monte Carlo simulations by drawing random samples from several normal distributions. We summarize our findings in Tables 1–4. The eighth column gives OSR, which represents the oversampling rate computed by $(N - n_c) \times 100\%/n_c$. In each replication, we first draw m observations from the normal populations and then follow the algorithm of the purely sequential procedure by drawing m' observations from each group at each stage after the pilot stage. We summarize our findings in Tables 1–4. Please note that Table 1 and Table 3 describe scenarios in which after the pilot stage, scores from $m' = 1$ individual belonging to each group added. Table 2 and Table 4 describe, respectively, scenarios in which, after the pilot stage, scores from 20 and 10 individuals belonging to each group are added at each stage.

From the fifth column of Tables 1–4, we find that the ratio of the average final sample size and the optimal sample size, n_c, is close to 1, which is the true value for an exact procedure. The last column suggests that the ratio of the risk of estimating the standardized mean difference, δ, using the purely sequential procedure is close to the optimal sample size risk, R_{n_c}. Thus, we find that our purely sequential procedure works remarkably well in small to large sample size scenarios. In fact, for Table 1, none of the scenarios yielded more than 5% average on the sample size.

For small sample size scenarios, for Tables 3–4, the largest oversampling rate occurs in the situation in which the mean of the sample size from our procedure is 56.3680, whereas the theoretically optimal value is 47 (and thus the relative discrepancy is $(56.3680 - 47)/47 = 0.1993$, implying that there is an average of 19.93% oversampling in this small sample size condition). Thus, in the worst relative case, the average sample size from the procedure was about nine more (per group) compared with what was theoretically ideal.

Using our sequential method, as discussed, we found the sample size required to obtain an estimate that simultaneously considered accuracy and cost (both structural and sampling). We note that the quality of estimation was not an issue and was not affected by the sequential procedure, due to the consistency property of the estimator used for the standardized mean difference (i.e., Equation 2). Regarding the effectiveness of the procedure in terms of accuracy, which was only one of the two dimensions we use for the optimization, cost being the other, our procedure was quite effective. In the larger sample size scenarios (Table 1 and Table 2), around 90% of the replications in each of the conditions had the absolute

Table 1
Estimated Average Final Sample Size With m′ = 1, A = $50,000, c = $1

Distribution	δ	$\bar{N}$ s($\bar{N}$)	n_c	$\bar{N}/n_c$	$\bar{d}_N$ p	$\bar{r}_N$ s($\bar{r}N$)	$\frac{\bar{r}_N}{R_{n_c}}$	OSR
Group 1: $N(5, 4)$	0	235.0418	224	1.0493	−.0018	895.9990	1.0000	4.93%
Group 2: $N(5, 4)$		.0029			.8918	.0098		
Group 1: $N(5.2, 4)$	0.1	235.1774	224	1.0499	.1022	896.5703	1.0006	4.99%
Group 2: $N(5, 4)$		.0056			.8930	.0172		
Group 1: $N(5.4, 4)$	0.2	235.6276	225	1.0472	.1975	898.2326	.9980	4.73%
Group 2: $N(5, 4)$		.0093			.8846	.0311		
Group 1: $N(5.8, 4)$	0.4	237.3000	226	1.0500	.4002	904.8876	1.0010	5.00%
Group 2: $N(5, 4)$		.0150			.8928	.0590		
Group 1: $N(6.6, 4)$	0.8	243.7522	233	1.0461	.8015	931.1255	.9991	4.61%
Group 2: $N(5, 4)$		.0287			.8832	.1158		
Group 1: $N(8.2, 4)$	1.6	267.8542	257	1.0422	1.6025	1,029.2360	1.0012	4.23%
Group 2: $N(5, 4)$		.0535			.8750	.2168		

Note. $N(\cdot, \cdot)$ represents a normal distribution with the first parenthetical value the population mean and the second the population variance; δ is the population standardized mean difference; $\bar{N}$ is the mean final sample size; n_c is the theoretically optimal sample size is the population parameters were known; $\bar{d}_N$ is the mean standardized mean difference; p represents proportion of times $|d_N - \delta| \leq \epsilon$; $\bar{r}_N$ is the mean risk; OSR is the oversampling rate (i.e., $(N - n_c)/n_c \times 100$).

difference between the estimated standardized mean difference and the population standardized mean difference less than the specified value of ϵ. For the smaller sample size scenarios (Table 3 and Table 4), the percentage of replications in each of the conditions had the absolute difference between the estimated standardized mean difference and the population standardized mean difference less than the specified value of ϵ was around the 60–70 percent mark. In fact, across Tables 1–4 in the document and Tables S1–S6 in the supplemental materials, the condition in which desired accuracy was smallest was for very large delta (δ = 1.60), which had the theoretically optimal sample size of 52. In this scenario 64.10% of the replications were smaller than desired. However, to be clear, we did not optimize estimation based on accuracy alone, as our procedure optimized estimation based on accuracy and cost.

Application

Here we provide an example for illustrative purposes based on a recent study on the effect of *same language subtitling* (SLS) on reading ability (see Kothari, 2008; Kothari & Bandyopadhyay, 2014). SLS is a concept of subtitling the dialogue in movies or TV programs to the same language. This is literally "closed captioning," but with a different purposes. Whereas closed captioning displays the spoken language as text, such as for those that have hearing impairments or when the sound cannot be heard in a particular environment, SLS is meant to facilitate learning the written representation of a language that is already known verbally. That is, consider the case in which someone understands spoken English but understands little written English. SLS is meant to increase knowledge of the written representation of

Table 2
Estimated Average Final Sample Size With m′ = 20, A = $50,000, c = $1

Distribution	δ	$\bar{N}$ s($\bar{N}$)	n_c	$\bar{N}/n_c$	$\bar{d}_N$ p	$\bar{r}_N$ s($\bar{r}N$)	$\frac{\bar{r}_N}{R_{n_c}}$	OSR
Group 1: $N(5, 4)$	0	250	224	1.1161	.0020	900.4087	1.0049	11.61%
Group 2: $N(5, 4)$		.0000			.9038	.0082		
Group 1: $N(5.2, 4)$	0.1	250	224	1.1161	.1001	900.8966	1.0055	11.61%
Group 2: $N(5, 4)$		.0000			.9058	.0148		
Group 1: $N(5.4, 4)$	0.2	250	225	1.1111	.1971	902.3313	1.0026	11.11%
Group 2: $N(5, 4)$		.0000			.9124	.0258		
Group 1: $N(5.8, 4)$	0.4	250	226	1.1062	.3999	908.3985	1.0049	10.62%
Group 2: $N(5, 4)$		.0043			.9016	.0135		
Group 1: $N(6.6, 4)$	0.8	250.0280	233	1.0731	.8026	932.6694	1.0007	7.31%
Group 2: $N(5, 4)$		.0080			.8868	.0274		
Group 1: $N(8.2, 4)$	1.6	274.6400	257	1.0686	1.599	1,030.3503	1.0023	6.86%
Group 2: $N(5, 4)$		.1194			.8806	.2247		

Note. $N(\cdot, \cdot)$ represents a normal distribution with the first parenthetical value the population mean and the second the population variance; δ is the population standardized mean difference; $\bar{N}$ is the mean final sample size; n_c is the theoretically optimal sample size is the population parameters were known; $\bar{d}_N$ is the mean standardized mean difference; p represents proportion of times $|d_N - \delta| \leq \epsilon$; $s(\cdot)$ is the standard deviation of the parenthetical value; $\bar{r}_N$ is the mean risk; OSR is the oversampling rate (i.e., $(N - n_c)/n_c \times 100$).

Table 3
Estimated Average Final Sample Size With $m' = 1$, $A = \$1{,}000{,}000$, $c = \$500$

Distribution	δ	$\bar{N}$ $\mathbf{s}(\bar{N})$	n_c	$\bar{N}/n_c$	$\bar{d}_N$ p	$\bar{r}_N$ $\mathrm{s}(\bar{r}N)$	$\frac{\bar{r}_N}{R_{n_c}}$	OSR
Group 1: $N(5, 4)$	0	50.0210	45	1.1116	.0010	90,204.8500	1.0023	11.16%
Group 2: $N(5, 4)$		.0020			.6818	4.2475		
Group 1: $N(5.2, 4)$	0.1	50.0392	45	1.1119	.1072	90,268.4700	1.0030	11.19%
Group 2: $N(5, 4)$		.0028			.6822	5.5749		
Group 1: $N(5.4, 4)$	0.2	50.0964	45	1.1133	.2026	90,425.9100	1.0047	11.33%
Group 2: $N(5, 4)$		.0043			.6918	7.9303		
Group 1: $N(5.8, 4)$	0.4	50.3882	46	1.0954	.4027	91,074.5500	.9899	9.54%
Group 2: $N(5, 4)$		.0078			.6906	13.1353		
Group 1: $N(6.6, 4)$	0.8	51.6882	47	1.0997	.8004	93,670.5200	.9965	9.97%
Group 2: $N(5, 4)$		.0129			.6766	25.1307		
Group 1: $N(8.2, 4)$	1.6	56.4876	52	1.0863	1.6121	103,555.9000	.9957	8.63%
Group 2: $N(5, 4)$		.0239			.6410	48.0637		

Note. $N(\cdot, \cdot)$ represents a normal distribution with the first parenthetical value the population mean and the second the population variance; δ is the population standardized mean difference; $\bar{N}$ is the mean final sample size; n_c is the theoretically optimal sample size is the population parameters were known; $\bar{d}_N$ is the mean standardized mean difference; p represents proportion of times $|d_N - \delta| \leq \epsilon$; $\bar{r}_N$ is the mean risk; OSR is the oversampling rate (i.e., $(N - n_c)/n_c \times 100$).

English, ultimately to improve reading ability. The studies on the impact of SLS showed that continuous exposure to SLS results in improved reading skills (e.g., see Kothari, 2008). The title of Kothari (2008) helps to illustrate the potential impact: *Let a Billion Readers Bloom: Same Language Subtitling (SLS) on Television for Mass Literacy*.

In a study on SLS effectiveness, Kothari and Bandyopadhyay (2014) divided schoolchildren in India who were between 6 and 14 years of age into two groups: High-SLS and Low-SLS. The students were exposed to their assigned SLS program either regularly for the High-SLS group, or rarely for the Low-SLS group. In each case the children knew spoken Hindi but may not have known much written Hindi. For details of the SLS implementation, see Kothari and Bandyopadhyay (2014). For purposes of our example, we focus only on a single reading measure, which is decoding 22 simple two-syllable words. Each student is measured on a scale between 0 to 22 based on their performance, with 0 indicating all words missed and 22 indicating a perfect score. The standardized mean difference is most useful here, as compared with the (unstandardized) mean difference, because the researchers wish to obtain a measure not specifically tied to the 22 word scale. That is, the mean difference scaled in terms of the common standard deviation is the outcome of interest (i.e., d).

Suppose that the research goal is to obtain an accurate estimation of the impact of SLS on reading ability. Understanding and communicating the magnitude of the effect is thought to be important because tax revenue is being spent on implementing SLS to increase literacy and also to increase reading skills in Hindi in India. The study was facilitated by national TV broadcaster Doordarshan (e.g., see Kothari et al., 2002, 2004). As is typical when dealing with policy issues, there are competing priorities and thus funded projects generally need to demonstrate the size of their effect in order to continue funding priorities.

Table 4
Estimated Average Final Sample Size With $m' = 10$, $A = \$1{,}000{,}000$, $c = \$500$

Distribution	δ	$\bar{N}$ $\mathbf{s}(\bar{N})$	n_c	$\bar{N}/n_c$	$\bar{d}_N$ p	$\bar{r}_N$ $\mathrm{s}(\bar{r}N)$	$\frac{\bar{r}_N}{R_{n_c}}$	OSR
Group 1: $N(5, 4)$	0	51	45	1.1333	−.0010	90,408.4326	1.0045	13.33%
Group 2: $N(5, 4)$		.0000			.6878	3.7785		
Group 1: $N(5.2, 4)$	0.1	51.0020	45	1.1334	.1052	90,470.3474	1.0052	13.34%
Group 2: $N(5, 4)$		.0020			.6838	5.0445		
Group 1: $N(5.4, 4)$	0.2	51.0280	45	1.1340	.2007	90,607.7000	1.0067	13.40%
Group 2: $N(5, 4)$		.0075			.6980	7.5948		
Group 1: $N(5.8, 4)$	0.4	51.3180	46	1.1156	.3956	91,271.1800	.9921	11.56%
Group 2: $N(5, 4)$		.0248			.6882	16.2145		
Group 1: $N(6.6, 4)$	0.8	56.3680	47	1.1993	.7905	95,000.9000	1.0106	19.93%
Group 2: $N(5, 4)$		.0071			.6934	38.5825		
Group 1: $N(8.2, 4)$	1.6	61.0520	52	1.1741	1.61513	104,676.5000	1.0065	17.41%
Group 2: $N(5, 4)$		.0102			.6530	41.0473		

Note. $N(\cdot, \cdot)$ represents a normal distribution with the first parenthetical value the population mean and the second the population variance; δ is the population standardized mean difference; $\bar{N}$ is the mean final sample size; n_c is the theoretically optimal sample size is the population parameters were known; $\bar{d}_N$ is the mean standardized mean difference; p represents proportion of times $|d_N - \delta| \leq \epsilon$; $s(\cdot)$ is the standard deviation of the parenthetical value; $\bar{r}_N$ is the mean risk; OSR is the oversampling rate (i.e., $(N - n_c)/n_c \times 100$).

Although the size of the effect is important for such studies, there are only limited resources available for the data collection and thus consideration of the sample size is important. Collecting more data than necessary could be argued to be a waste of resources (e.g., obtained from taxes). However, a sample that is not large enough to produce an estimate with as much accuracy as is needed is also a waste of resources if the study is found to be "inconclusive." Thus, the research team seeks a balance between the estimation accuracy (i.e., small mean square error) of the standardized mean difference and the study cost by considering sampling cost (i.e., the total amount of money spent on data collection).

First, we need to consider the sampling cost. The assessment of the decoding of the 22 words is performed by a surveyor on in-school visits. Suppose on any school day, the surveyor is allowed to interview students for two hours. In two hours suppose it is possible to interview 10 students individually who are between 6 and 14 years of age. For each day, the surveyor will be given $24 including travel cost and an hourly wage (that is, $24 total for the 2 hr of work and travel costs). Thus, the sampling cost per student is estimated to be $2.40 (i.e., $c = \$2.40$).

Second, we need to consider the maximum probable error (ϵ) in estimating the population standardized mean difference. Because accurate estimation of the population standardized mean difference is desired, suppose that the researcher team selects a value of epsilon of .50, as they do not wish the estimated standardized mean difference to be more than 0.50 units from the population value. Thus, $\epsilon = 0.5$.

Third, the research team needs to consider the price they are willing to pay in order for the estimate to be sufficiently accurate. This is an important consideration because, along with ϵ, the amount of investment in the study's success determines A. The cost the research team is willing to pay in order to have a sufficiently narrow estimate (i.e., in order for ϵ to not exceed 0.5) is $2,500. The value of $2,500 is based on the amount of funds that they are willing to invest in the success of the study. Thus, the value of A is the price they are willing to pay for a sufficiently small ϵ and ϵ itself: $A = \$10{,}000 (= \$2{,}500/0.5^2)$. In words, A can be described as the price willing to pay per squared unit of the maximum probable difference (i.e., ϵ).

Using the values of $c (= 2.40)$ and $A (= 10{,}000)$, we first obtain a pilot sample size, m, to be drawn from both groups. Here, the pilot sample size from each group will be $m = \max\{4, \lceil (10{,}000/(2 \times 2.4))^{1/(2+2\times 0.49)} \rceil\} = 13$. Using the MBESS R package, the `mr.smd()` function can be used to obtain the pilot sample as follows:

```
require(MBESS)
mr.smd(pilot=TRUE, A = 10000, sampling
  .cost=2.4, gamma=.49)
```

where R code is represented in typewriter font to distinguish it from regular text and punctuation has been removed so as to not confuse it with code. Note that the code shown is submitted directly into the R console at the prompt, which is ">" by default. Further, the code requires that the MBESS R package (Kelley, 2007a; 2007b; 2016) be installed, which can be done with the following code on most systems: `install.packages("MBESS")`. To be clear here, the pilot sample size of 13 implies that there are 26 observations total, 13 per group.

Consider the hypothetical data collected as part of the pilot sample, which can be entered into R as a vector for analysis purposes as follows:

```
High.SLS <- c(11, 7, 22, 13, 6, 9, 11, 16,
  12, 17, 14, 8, 16)
Low.SLS <- c(3, 6, 10, 8, 14, 5, 12, 10, 6,
  8, 13, 5, 9)
```

Note that one can apply the standardized mean difference function, `smd()`, in order to obtain the estimated value of the standardized mean difference as follows,

```
smd(Group.1=High.SLS, Group.2=Low.SLS)
```

which returns a d of 1.021484.

The function `mr.smd()` implements a check to determine if the criterion specified by the method (i.e., the stopping rule) we proposed has been satisfied (i.e., Equation 15), which requires the user to specify d, the sample size upon which d was calculated (i.e., n, which is assumed equal across group), A, c, and γ (which, recall, we suggest be .49). Thus, for our situation, in which we have $d = 1.021484$, $n_1 = n_2 = n = 13$, $A = \$10{,}000$, c = sampling.cost = $2.40, and $\gamma = .49$, the code can be implemented as:

```
mr.smd(d=1.021484, n=13, A=10000,
  sampling.cost=2.40)
```

which in this case returns FALSE, indicating that the stopping rule of Equation 15 was not satisfied. An alternative approach is to use the `smd()` function within the `mr.smd()` function as:

```
mr.smd(d=smd(Group.1=High.SLS, Group.2=
  Low.SLS), n=13, A=10000, sampling.cost=
  2.40)
```

At this point we have not said anything about m', which is the number of observations added at each stage of the sequential procedure. For now, suppose that $m' = 1$, which means that we will add a single observation and then check the stopping rule again.[5] Because the stopping rule is not satisfied, another observation per-group is collected yielding $n = 14$, and the stopping rule is evaluated again. For example, if a new observation from the High.SLS was collected as 10 and a new observation from the Low.SLS group was collected as eight, these values would be included in the data vector for each group:

```
High.SLS <- c(11, 7, 22, 13, 6, 9, 11, 16,
  12, 17, 14, 8, 16, 10)
Low.SLS <- c(3, 6, 10, 8, 14, 5, 12, 10, 6,
  8, 13, 5, 9, 8)
```

The stopping rule could then be applied again with the new data:

```
mr.smd(d=smd(Group.1=High.SLS, Group.2=
  Low.SLS), n=14, A=10000, sampling.cost=
  2.40)
```

[5] We realize that in some situations it is as easy to collect multiple observations as it is a single observation, such as might be the case in a classroom. In other situations, however, such as in online data collection sites, each observations is done separately. We illustrate here using $m' = 1$ but discuss the implications if, for example, $m' = 5$ or some other value.

which is also FALSE (again, meaning that the stopping rule is not satisfied). This process continues until the stopping rule is satisfied.

Suppose that data are continued to be collected and the stopping rule evaluated after each observation per group is collected. Further suppose that $n = 75$ (implying the total sample size is 150) and calculate $d = 1.00$. Then, when we implement the function we get

```
mr.smd(d=1.00, n=75, A=10000,
sampling.cost=2.40)
```

TRUE, which was FALSE for all smaller sample sizes. Thus, our sampling stops at this point and we have obtained our sequential estimate of the standardized mean difference based on both its accuracy as well as cost considerations.

At the conclusion of the study we also suggest, as is widely recommended in the literature, that the confidence interval for the population standardized mean difference be given, which can be obtained using MBESS as

```
ci.smd(smd=1.00, n.1=75, n.2=75,
  conf.int = .95),
```

which returns a confidence interval with limits of 95%: CI .95 = [.6588, 1.338].

Suppose, however, that instead of using $m' = 1$ as above, the sample size at each stage of sampling is five (i.e., $m' = 5$), meaning that at each stage of sampling an additional five observations are taken. In so doing, the pilot sample remains the same (as m and m' are two separate entities and there is no requirement that one is larger than the other. Other than the sample size changing by 5 each time, the method proceeds in exactly the same way as before. One may wonder which m' to choose. We are noncommittal on this, as it is researcher specific; from a statistical perspective m' is arbitrary. If sampling is being done in group context, there would be no reason to set $m' = 1$. However, in studies where the sampling can easily be with a single observation at each state, it may pose little additional effort for a researcher to input the datum value into a file and run the `mr.smd()` each time data is input. In cases where the data collection procedure is automated (e.g., online surveying tools), once data has been collected, as part of the data storage process an R script can be run to evaluate if sampling should continue.

Note that the study that served as our motivating example was conducted in India. The cost of sampling and overall study cost may seem quite different than study costs for studies conducted elsewhere. Imagine that the sampling cost per student (i.e., c) were 10 times larger; A would also be 10 times larger. The final sample size is the same because the multiplier cancels. Thus, the methods we have developed are general and the researcher supplied values (i.e., A, c, and ϵ) can be specified as needed for the particular context and goals.

A General Scenario

There are situations in which the cost of collecting data or the sampling cost per participant is different for two different groups and due to this, the number of observations to be sampled from each group is also different.

In a situation of unequal cost per participant across the groups, we extend the methods from above to a general scenario. More specifically, we now develop a method to find optimal sample sizes for two groups such that the total expected cost of estimating δ (by using $d_{\tilde{n}}$) with a maximum probable error ϵ is minimized, as before, but now when the cost of sampling one observation from both groups is different. By following the same technique, we present an outline of the procedure to estimate the optimal sample sizes for two groups without proof.

In this case, the total expected cost of estimating δ (by using $d_{\tilde{n}}$) using n_1 observations from Group 1 and n_2 observations from Group 2 with a maximum probable error ϵ or the risk function is given as,

$$R_{\tilde{n}}(\delta) = A\mathrm{E}[(d_{\tilde{n}} - \delta)^2] + (c_1 n_1 + c_2 n_2) \approx A\left\{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) + \frac{\delta^2}{2(n_1 + n_2)}\right\} + (c_1 n_1 + c_2 n_2), \tag{17}$$

where c_1 and $c_2 (\geq c_1)$ are the cost of sampling each observation belonging to Groups 1 and 2, respectively. Here, without loss of generality, we assume that the cost of sampling one observation in Group 2 is more in Group 1.

The approximate total expected cost or the risk function in Equation 17 can be minimized if n_1^* individuals are selected from Group 1 and n_2^* individuals are selected from Group 2, where n_1^* and n_2^* can be found by solving

$$n_1^* = \sqrt{\frac{A}{c_1}}\left(1 + \frac{\delta^2}{2\left(1 + \sqrt{\frac{A}{(c_2 - c_1)(n_1^*)^2 + A}}\right)}\right)^{1/2} \tag{18}$$

and

$$n_2^* = n_1^*\left(1 + \frac{(c_2 - c_1)(n_1^*)^2}{A}\right)^{-1/2}, \tag{19}$$

respectively. The above is proved in Lemma 6 in Appendix B. Because δ is unknown, we cannot use Equations 18 and 19 to find the optimal sample sizes n_1^* and n_2^*. So, as before, we use a purely sequential procedure to find an estimate of optimal sample sizes n_1^* and n_2^*. Here we present an algorithm to find an estimate of the optimal sample sizes. First, we define m_{ij} as the number of observations to be sampled for the ith group at the jth stage and we also define $\tilde{m}_j$ as the total number of observations sampled until jth stage.

Stage I

First, scores of m_{11} randomly selected individuals are collected from Group 1 and m_{21} observations from Group 2, such that,

$$m_{11} = \max\{m_0, \lceil (A/(c_1))^{1/(2+2\gamma)} \rceil\} \tag{20}$$

and

$$m_{21} = \max\left\{m_0, \left\lceil m_{11}\left(1 + \frac{(c_2 - c_1)(m_{11})^2}{A}\right)^{-1/2}\right\rceil\right\}, \tag{21}$$

where $\gamma = 0.49$ and $m_0 (\geq 4)$ is the least possible sample size required to estimate δ, where $\lceil\ \rceil$ is the ceiling function of the term defined before. Based on this pilot sample of sizes m_{11} of Group 1 and m_{21} of Group 2, an estimate of δ is obtained from $\tilde{m}_1 = m_{11} + m_{21}$ observations by computing $d_{\tilde{m}_1}$. If $m_{11} < \left\lceil \sqrt{\frac{A}{c_1}}\left(1 + \frac{d_{\tilde{m}_1}^2}{2}\left(1 + \sqrt{\frac{A}{(c_2 - c_1)(m_{11})^2 + A}}\right)^{-1}\right)^{1/2} + \sqrt{\frac{A}{c_1}} m_{11}^{-\gamma}\right\rceil$, then go to the next step. Otherwise, if $m_{11} \geq \left\lceil \sqrt{\frac{A}{c_1}}\left(1 + \frac{d_{\tilde{m}_1}^2}{2}\left(1 +\right.\right.\right.$

$\sqrt{\frac{A}{(c_2 - c_1)(m_{11})^2 + A}}\Big)^{-1}\Big)^{1/2} + \sqrt{\frac{A}{c_1}}m_{11}^{-\gamma}\Big\rceil$, then stop sampling and set the final sample size equal to $\tilde{m}_1 = m_{11} + m_{21}$.

Stage II

Obtain scores from m' randomly selected individuals (different from those who were selected during Stage I) belonging to Group 1. Thus, there are $m_{12} = m_1 + m'$ observations from Group 1. Now, the sample size for Group 2 will be $m_{22} = \max\Big\{m_{21}, \Big\lceil m_1\Big(1 + \frac{(c_2 - c_1)(m_{12})^2}{A}\Big)^{-1/2}\Big\rceil\Big\}$. So, obtain scores from $m_{22} - m_{21}$ randomly selected individuals from Group 2 in Stage II. Thus, at the second stage, there is a total of $\tilde{m}_2 = m_{12} + m_{22}$ observations in the study. If $m_{12} \geq \Big\lceil\sqrt{\frac{A}{c_1}}\Big(1 + \frac{d^2_{\tilde{m}_2}}{2}\Big(1 + \sqrt{\frac{A}{(c_2 - c_1)(m_{12})^2 + A}}\Big)^{-1}\Big)^{1/2} + \sqrt{\frac{A}{c_1}}m_{12}^{-\gamma}\Big\rceil$ stop, further sampling and set the final sample size equal to $\tilde{m}_2 = m_{12} + m_{22}$. If $m_{12} < \Big\lceil\sqrt{\frac{A}{c_1}}\Big(1 + \frac{d^2_{\tilde{m}_2}}{2}\Big(1 + \sqrt{\frac{A}{(c_2 - c_1)(m_{12})^2 + A}}\Big)^{-1}\Big)^{1/2} + \sqrt{\frac{A}{c_1}}m_{12}^{-\gamma}\Big\rceil$, then continue the sampling process by sampling m' more individuals for Group 1 and then compute sample size needed for Group 2.

This process of collecting the same number of observations in each stage after Stage I continues until there are N_1 observations from Groups 1 and N_2 observations from Group 2 such that $N_1 \geq \Big\lceil\sqrt{\frac{A}{c_1}}\Big(1 + \frac{d^2_{N_1+N_2}}{2}\Big(1 + \sqrt{\frac{A}{(c_2 - c_1)(N_1)^2 + A}}\Big)^{-1}\Big)^{1/2} + \sqrt{\frac{A}{c_1}}N_1^{-\gamma}\Big\rceil$ and N_2 computed using the value of N_1. At this stage, we stop further sampling and report that the final sample size is $N_1 + N_2$.

Thus N_1 and N_2 are the final estimated sample sizes that should be drawn from Groups 1 and 2, respectively. This will minimize the total expected cost of estimating δ within a maximum probable error, ϵ, in cases when the sample sizes and the sampling cost per unit observation differs in both groups. If observations are collected using the sequential procedure described in this section, sampling from both groups is guaranteed to be eventually terminated. This is proved in Lemma 7 in Appendix B.

Bounds on Necessary Sample Size

Upon careful examination of our tables, one may notice that as δ gets larger, n_c increases, but it does not increase by much. Kelley and Rausch (2006) discussed how, from the accuracy in parameter estimation approach, the width of the confidence was affected by the size of δ, but only to a small extent. This characteristic has a useful implication when considering sample size before a study is undertaken.

We note that the larger the true population standardized mean difference, the larger the mean square error. Suppose a researcher is interested in using our approach but is uncertain what the necessary sample size from this sequential procedure might be, especially to consider if a sample size that large might be obtainable. By plugging in a hypothetical value for δ, A, and c, one can estimate the final sample size, momentarily treating that the hypothetical value of δ as if it were the obtained sample estimate. This hypothetical value should be the *maximum absolute effect size* that theory, practice, or the literature would support. In so doing, this hypothetically "maximum effect size" serves as an upper bound on the theoretically optimal sample size. That is, if the true value of δ is smaller than the hypothetically "maximum" absolute effect size, then the optimal sample size will be smaller. Correspondingly, although the necessary sample size might be smaller, one can obtain an upper bound on sample size when, along with A, and c, a maximum absolute effect size is plugged into the procedure.

Similarly, one could find a lower bound on the sample size by plugging into the procedure for a *minimum absolute effect size*, likely zero. Thus, by using minimum and maximum hypothetical absolute effect sizes in the procedure, one can obtain bounds on the sample size. Such a procedure can be used to help approximate or choose sample size in research proposals, such as grant applications. Our experience is such that the sample size that is to be used is included in the proposal. From a sequential perspective the final sample size is unknown and thus cannot be specified because the population parameters are not known. Nevertheless, by using the procedure we outline above, in which the maximum and minimum absolute effect sizes that the theory, practice, or the literature would support, one can find what can be considered as a functional lower and upper bounds on the study sample size. It should be clear, however, that the proposal given for the lower and upper bounds will likely not be the same sample size that is obtained in a study in which the sample effect size is used to perform a check to evaluate if optimization has occurred.

Discussion

The population standardized mean difference is a very popular measure of difference between means in a standardized metric, which quantifies the number of standard deviations the population mean of Group 1 is away from the population mean of Group 2. The accuracy of the estimator of the population standardized mean difference increases as the mean square error (MSE) decreases. Also, MSE decreases as the sample size increases holding everything else constant, but increasing sample size from each group will increase the total sampling cost. A cost function for estimating the population standardized mean difference is defined, which depends on both the MSE and the total cost of sampling. This cost function needs to be minimized.

If the true value of δ is known or hypothesized, our procedure can be used as a fixed-n a priori sample size planning method, in that if the population value of δ is known, then so too is n_c, which can be the sample size a researcher uses, as would be done with an a priori sample size planning method. That is, one could plan, in a priori fashion, a sample size based on the known or, more likely, hypothesized value of δ. The planned sample size would be what we have referred to as the theoretically optimal sample size. Of course, in any given situation an obtained d, estimate of δ, will be smaller or larger than δ, and thus the procedure may stop earlier or later than would application of the sequential method as intended.

One nuance that we did not discuss is how to assign the 2 (groups) × m' (participants sampled at each step) to the two groups. Sometimes, such as when groups are naturally formed (male/female; republican/democrat; third grader/fourth grader), the participants are simply placed into the appropriate group. However, in a randomized design scenario, it is not obvious how the $2 \times m'$ participants should be assigned to a group. The most straightforward approach is by (simple) randomization to group. Alternatively, the participants can be matched, and then the matched pairs randomly assigned. Depending on the context, one of these or other approaches may be appropriate,

whereas other might be infeasible. For example, a matched pairs approach may be impractical in a classroom environment in which a researcher has only a limited time to collect data and no time to assess the variable that would typically be used for matching. We leave the nuance of assignment of the $2 \times m'$ participants to researchers based on their circumstance and methodologists who might more fully explore assignment to group in sequential stages of a study. However, one suggestion is to group the $2 \times m'$ participants in a manner analogous to how it would be done in a nonsequential procedure.

A potential limitation is that we regard the cost of data collection per participant as a fixed value. There are some scenarios in which the cost of sampling changes over time. We have ignored such a possibility to focus on known and fixed cost. That being said, one could modify our procedures so as to make the way in which cost is considered more flexible. Additionally, our procedure only considers the "cost of sampling," rather than "cost to conduct the study." That is, there are cost factors that are beyond "cost of sampling."

One question that some researchers may pose is "What value should be chosen for ϵ?" There is not a simple answer to this question, as it is based on the goals of the researcher. One value we suggest researchers consider is $\epsilon = .10$, which, for example, would allow d_n to not likely differ by more than .10 units from δ. Consider a supposed value of $\delta = .50$. Having an estimate that is between .40 and .60 is, likely, informative and useful in many situations. This yields a sample estimate that is not unreasonably far from the population value. Of course, $\epsilon = .05$ is better from an accuracy perspective (though more expensive due to extra sampling cost), and $\epsilon = .01$ better still. One could consider a percentage of a supposed true value of δ. For example, once again supposing that $\delta = .50$, obtaining an estimate within 10% of this (supposed) value would suggest $\epsilon = .05$ as a reasonable value. Of course, one may want to be within 5% of the true value, supposed to be $\delta = .50$, which would then suggest that $\epsilon = .025$. None of these values are wrong. We suspect that different research areas will tend to develop their own norms for ideal or at least typically chosen values of $\epsilon = .05$.

A fixed sample size procedure, which is widely recommended in psychology and related disciplines, cannot be implemented to minimize the total expected cost ($4cn_c$), because it depends on the population standardized mean difference, which is the same parameter that needs to be estimated. In this article, we develop a purely sequential procedure which provides an estimate of the sample size required to achieve sufficient accuracy with minimum total expected cost. The purely sequential procedure developed here ensures that the sampling procedure stops at an average sample size that is very close to the theoretically optimal sample size for each group. The theoretically optimal sample size, recall, is the sample size that satisfied the procedure had the population standardized mean difference been known. Our findings from the Monte Carlo simulation study showed that our developments of the purely sequential procedure led to a method that is very effective. Further, we implement the methods in the freely available MBESS R package so that researchers can easily implement the methods we have discussed. Thus, our sequential procedure overcomes the biggest obstacle in sample size planning, which is specifying the population value of the effect size. The idea of choosing the population parameter before a study in a power analysis has been referred to as the "problematic parameter" (Lipsey, 1990, Chapter 3), because of the difficulty of choosing an appropriate value upon which to base sample size.

As a reminder, even though our approach does not prevent one from also testing a null hypothesis or forming a confidence interval, these inferential procedures should only be done *after* the stopping rule has been satisfied and more data is not being collected. The "sample-evaluate-sample-evaluate" method poses no problems for a sequential estimation procedure when a stopping rule based on the accuracy of the estimate is being used. However, if one were to approach hypothesis testing from a "sample-evaluate-sample-evaluate" method, where the researcher's stopping rule was "stop when statistical significance is obtained" the *p*-value would not be correct as there would be capitalization on chance and there would be more Type I errors than the nominal value of the Type I error rate (e.g., 5%) would suggest. Thus, our stopping rule is fundamentally different than one that seeks the obtainment of *p*-values less than the designated Type I error rate. Using a sequential approach in which a hypothesis test is performed at each step would yield *p*-values that capitalize on chance and are therefore biased, making the results of such a hypothesis test unusable for valid inference. For more details, we refer to Brannath, Gutjahr, and Bauer (2012), Graf and Bauer (2011), and Timmesfeld, Schäfer, and Müller (2007), among others. We note, however, that although these methods can be used to mitigate the inflation of the Type I error rate, to our knowledge, no such methods exist for sequential stopping rules in a hypothesis testing framework that simultaneously considers sampling costs.

Our work on sequential estimation for the standardized mean differences does not consider the population parameter in isolation, but rather it simultaneously considers cost. From a very practical perspective, all studies should be concerned with the financial cost of sampling. Without considering the actual cost, the sample size planning procedure used may require a sample size that exhausts the resources available to the researchers. One workaround that has been used is to forgo conducting a study in which the sample size planning procedure calls for a sample that is too large for the available financial resources. This approach implies that cost is given 100% of the weight. However, we believe that it is more informative to consider both cost and accuracy simultaneously due to their necessarily intertwined relationship. We believe that our work has the potential to shift the way some studies are designed and grants written, as cost considerations are often a major part of the way in which empirical grants are evaluated.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications: American Educational Research Association. *Educational Researcher, 35,* 33–40. http://dx.doi.org/10.3102/0013189X035006033

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

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

American Statistical Association. (2016). *American statistical association releases statement on statistical significance and p-values: Provides principles to improve the conduct and interpretation of quantitative science*. Retrieved from http://www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf

Armitage, P. (1969). Sequential analysis in therapeutic trials. *Annual Review of Medicine, 20,* 425–430. http://dx.doi.org/10.1146/annurev.me.20.020169.002233

Association for Psychological Science. (2014). *Submission guidelines*. Retrieved from http://www.psychologicalscience.org/index.php/publications/journals/psychological_science/ps-submissions

Brannath, W., Gutjahr, G., & Bauer, P. (2012). Probabilistic foundation of confirmatory adaptive designs. *Journal of the American Statistical Association, 107,* 824–832. http://dx.doi.org/10.1080/01621459.2012.682540

Cedilnik, A., Kosmelj, K., & Blejec, A. (2006). Ratio of two random variables: A note on the existence of its moments. *Metodoloski Zvezki, 3,* 1–7.

Chattopadhyay, B., & Kelley, K. (in press). Estimation of the coefficient of variation with minimum risk: A sequential method for minimizing sampling error and cost with a stopping rule. *Multivariate Behavioral Research*.

Chattopadhyay, B., & Mukhopadhyay, N. (2013). Two-stage fixed-width confidence intervals for a normal mean in the presence of suspect outliers. *Sequential Analysis, 32,* 134–157. http://dx.doi.org/10.1080/07474946.2013.774609

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

Dantzig, G. B. (1940). On the non-existence of tests of "student's" hypothesis having power functions independent of σ. *The Annals of Mathematical Statistics, 11,* 186–192. Retrieved from http://www.jstor.org/stable/2235875

Davey, A., & Savla, J. (2010). *Statistical power analysis with missing data: A structural equation modeling approach*. New York, NY: Routledge.

De, S. K., & Chattopadhyay, B. (2015). *Minimum risk point estimation of Gini Index*. Retrieved from http://arxiv.org/abs/1503.08148

Donaire, A., Falcón, C., Carreno, M., Bargallo, N., Rumià, J., Setoain, J., . . . Fernadnez, S. (2009). Sequential analysis of fmri images: A new approach to study human epileptic networks. *Epilepsia, 50,* 2526–2537. http://dx.doi.org/10.1111/j.1528-1167.2009.02152.x

Eich, E. (2014). Business not as usual. *Psychological Science, 25,* 3–6. http://dx.doi.org/10.1177/0956797613512465

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review, 18,* 3–12. http://dx.doi.org/10.1177/1088868313507536

Ghosh, B. K., & Sen, P. K. (1991). *Handbook of sequential analysis*. New York, NY: CRC Press.

Graf, A. C., & Bauer, P. (2011). Maximum inflation of the type 1 error rate when sample size and allocation rate are adapted in a pre-planned interim look. *Statistics in medicine, 30,* 1637–1647.

Gut, A. (2009). *Stopped random walks: Limit theorems and applications*. Cambridge, UK: Springer.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6,* 107–128. http://dx.doi.org/10.3102/10769986006002107

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

Jennison, C., & Turnbull, B. W. (2010). *Group sequential methods with applications to clinical trials*. Boca Raton, FL: CRC Press.

Kelley, K. (2007a). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20,* 1–24. http://dx.doi.org/10.18637/jss.v020.i08

Kelley, K. (2007b). Methods for the behavioral, educational, and educational sciences: An R package. *Behavior Research Methods, 39,* 979–984. http://dx.doi.org/10.3758/BF03192993

Kelley, K. (2016). MBESS 4.0.0 (or greater). [Computer software and manual]. Retrieved from http://www.cran.r-project.org

Kelley, K., Maxwell, S. E., & Scott, E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8,* 305–321. http://dx.doi.org/10.1037/1082-989X.8.3.305

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17,* 137–152. http://dx.doi.org/10.1037/a0028086

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11,* 363–385. http://dx.doi.org/10.1037/1082-989X.11.4.363

Kothari, B. (2008). Let a billion readers bloom: Same language subtitling (sls) on television for mass literacy. *International review of education, 54,* 773–780.

Kothari, B., & Bandyopadhyay, T. (2014). Same language subtitling of Bollywood film songs on TV: Effects on literacy. *Information Technologies & International Development, 10,* 31–47. Retrieved from http://itidjournal.org/index.php/itid/article/view/1307

Kothari, B., Pandey, A., & Chudgar, A. R. (2004). Reading out of the idiot box: Same-language subtitling on television in India. *Information Technologies and International Development, 2,* 23–44. http://dx.doi.org/10.1162/1544752043971170

Kothari, B., Takeda, J., Joshi, A., & Pandey, A. (2002). Same language subtitling: A butterfly for literacy? *International Journal of Lifelong Education, 21,* 55–66. http://dx.doi.org/10.1080/02601370110099515

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

Leroux, B. G., Mancl, L. A., & DeRouen, T. A. (2005). Group sequential testing in dental clinical trials with longitudinal data on multiple outcome variables. *Statistical Methods in Medical Research, 14,* 591–602. http://dx.doi.org/10.1191/0962280205sm421oa

Lim, I. S., & Leek, E. C. (2012). Curvature and the visual perception of shape: Theory on information along object boundaries and the minima rule revisited. *Psychological Review, 119,* 668. http://dx.doi.org/10.1037/a0025962

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

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*. http://dx.doi.org/10.1037/h0063675

Mahalanobis, P. C. (1940). A sample survey of the acreage under jute in Bengal, with discussion on planning of experiments. *Snakhyā, 4,* 511–531. Retrieved from http://www.jstor.org/stable/40383954

Maxwell, S. E., & Kelley, K. (2011). Ethics and sample size planning. In A. T. Panter & S. K. Sterba (Eds.), *Handbook of ethics in quantitative methodology* (pp. 159–184). New York, NY: Taylor & Francis.

Miladinovic, B., Mhaskar, R., Hozo, I., Kumar, A., Mahony, H., & Djulbegovic, B. (2013). Optimal information size in trial sequential analysis of time-to-event outcomes reveals potentially inconclusive results because of the risk of random error. *Journal of clinical epidemiology, 66,* 654–659. http://dx.doi.org/10.1016/j.jclinepi.2012.11.007

Mukhopadhyay, N., & Chattopadhyay, B. (2012). A tribute to Frank Anscombe and random central limit theorem from 1952. *Sequential Analysis, 31,* 265–277. http://dx.doi.org/10.1080/07474946.2012.694344

Mukhopadhyay, N., & Chattopadhyay, B. (2013). On a new interpretation of the sample variance. *Statistical Papers, 54,* 827–837. http://dx.doi.org/10.1007/s00362-012-0465-y

Mukhopadhyay, N., & Chattopadhyay, B. (2014). A note on the construction of a sample variance. *Sri Lankan Journal of Applied Statistics, 15,* 71–80. http://dx.doi.org/10.4038/sljastats.v15i1.6795

Mukhopadhyay, N., & De Silva, B. M. (2009). *Sequential methods and their applications*. Boca Raton, FL: CRC Press.

Murphy, K. R., & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd ed.). Mahwah, NJ: Erlbaum.

Petrie, A., Bulman, J., & Osborn, J. (2002). Further statistics in dentistry Pt. 4: Clinical trials 2. *British Dental Journal, 193,* 557–561. Retrieved from http://discovery.ucl.ac.uk/id/eprint/121963

Pornprasertmanit, S., & Schneider, W. J. (2014). Accuracy in parameter estimation in cluster randomized designs. *Psychological Methods, 19,* 356–379. http://dx.doi.org/10.1037/a0037036

Rao, C. R. (1973). *Linear statistical inference and its applications*. New York, NY: Wiley.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2,* 173. http://dx.doi.org/10.1037/1082-989X.2.2.173

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

Sen, P. K. (1981). *Sequential nonparametrics: Invariance principles and statistical inference*. New York, NY: Wiley.

Sen, P. K., & Ghosh, M. (1981). Sequential point estimation of estimable parameters based on u-statistics. *The Indian Journal of Statistics, Series A,* 331–344. Retrieved from http://www.jstor.org/stable/25050282

Statistical Research Group. (1945). *Sequential analysis in inspection and experimentation: Introduction* (Vols. SRG Report 255, Section 1). New York, NY: Columbia University Press.

Timmesfeld, N., Schäfer, H., & Müller, H.-H. (2007). Increasing the sample size during clinical trials with *t*-distributed test statistics without inflating the type I error rate. *Statistics in Medicine, 26,* 2449–2464. http://dx.doi.org/10.1002/sim.2725

Todd, S., Whitehead, A., Stallard, N., & Whitehead, J. (2001). Interim analyses and sequential designs in phase iii studies. *British Journal of Clinical Pharmacology, 51,* 394–399. http://dx.doi.org/10.1046/j.1365-2125.2001.01382.x

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics, 16,* 117–186. Retrieved from http://www.jstor.org/stable/2235829

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician, 70,* 129–133. http://dx.doi.org/10.1080/00031305.2016.1154108

Appendix A

Lemmas and Proofs Justifying Developments

Lemma 1

The approximate expression of the mean square error (MSE) of d_n is

$$\mathrm{E}[(d_{\tilde{n}} - \delta)^2] \approx \left(\frac{1}{n_1} + \frac{1}{n_2}\right) + \frac{\delta^2}{2(n_1 + n_2)}. \tag{22}$$

Proof

Let $n_{12} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. Suppose, $T_{\tilde{n}} = \frac{(\bar{X}_{1n} - \bar{X}_{2n})}{\sqrt{n_{12}}s}\left(= \frac{d_n}{\sqrt{n_{12}}}\right) \sim t_{\Delta,\nu}$, where, $t_{\Delta,\nu}$ represents a t distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom (d.f.) and noncentrality parameter $\Delta = \frac{(\mu_1 - \mu_2)}{\sqrt{n_{12}}\sigma}\left(= \frac{\delta}{\sqrt{n_{12}}}\right)$. Using Equation (6d) in Theorem 1 in Hedges (1981), we have,

$$\begin{aligned}\mathrm{E}[(d_{\tilde{n}} - \delta)^2] &= n_{12}(T_{\tilde{n}} - \Delta)^2 = n_{12}\left(\mathrm{E}\left[T_{\tilde{n}}^2\right] - 2\Delta\mathrm{E}[T_{\tilde{n}}] + \Delta^2\right) \\ &= n_{12}\left(\frac{\nu(1 + \Delta^2)}{\nu - 2} - 2\Delta^2\sqrt{\frac{\nu}{2}}\frac{\Gamma\left(\frac{\nu - 1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} + \Delta^2\right).\end{aligned} \tag{23}$$

We note that $\nu = n_1 + n_2 - 2$ (same as m_i in Hedges, 1981) and $n_{12} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$ (same as $\tilde{n}_i$ in Hedges, 1981). As per Equation 6e of Hedges (1981), $c(\nu) = \frac{\Gamma\left(\frac{\nu}{2}\right)}{\sqrt{\frac{\nu}{2}}\Gamma\left(\frac{\nu-1}{2}\right)}$ and as per our notation, $\Delta^2 = \frac{\delta^2}{n_{12}}$. Hedges (1981, p. 114) gave an approximation of $c(\nu) \approx 1 - \frac{3}{4\nu - 1}$. Our derivations lead to a rewrite of Hedges' expression but they also include the quantity of the remainder term in order to derive what follows. In particular, using Taylor's theorem, we get,

$$\frac{1}{c(\nu)} \approx \left(1 - \frac{3}{4\nu - 1}\right)^{-1} = 1 + \frac{3}{4\nu} + O(\nu^{-2}) \tag{24}$$

Plugging the value of $1/c(\nu)$ of Equation 24 in Equation 23 leads to

$$\begin{aligned}\mathrm{E}[(d_{\tilde{n}} - \delta)^2] &= n_{12}\left(\frac{\nu}{\nu - 2}(1 + \Delta^2) + \Delta^2\left[1 - 2\left(1 + \frac{3}{4\nu}\right)\right]\right) + O(\nu^{-2}) \\ &= n_{12}\left(\frac{\nu}{\nu - 2}(1 + \Delta^2) - 2\left(1 + \frac{3}{4\nu}\right)\Delta^2 + \Delta^2\right) + O(\nu^{-2})\end{aligned} \tag{25}$$

For not too small sample size, the $O(\nu^{-2})$ is negligible and hence can be ignored. □

If sample sizes for both groups are same, that is, $n_1 = n_2 = n$, then,

$$\mathrm{E}[(d_n - \delta)^2] \approx \frac{\left(2 + \frac{\delta^2}{4}\right)}{n}. \tag{26}$$

and the approximate expression of the MSE of d_n will be $\mathrm{E}[(d_n - \delta)^2] \approx \frac{\xi^2}{n}$. Figure 4 shows the bias in estimating the δ using the estimator d_n. B represents Bias. For example, B(δ:1.6) represents the bias of estimating δ using d_n when true value of δ is 1.6. Thus, we find that as the sample size increases, the bias is negligible and this decreases as the sample size increases.[6]

[6] Our sequential methodology works well as shown in Tables 1–4 in which the required theoretical sample size was between 45 to 52.

(*Appendices continue*)

Lemma 2

The approximate risk function given in Equation 9 is minimized if $n_c\left(=\sqrt{\frac{A}{c}}\xi\right)$ individuals are selected from groups 1 and 2, where, $\xi^2 = \left(2 + \frac{\delta^2}{4}\right)$.

Proof

For proof, please refer to Lemma 6. □

Lemma 3

Under the assumption that $\delta < \infty$, for any $c > 0$, the stopping time N_c is finite, that is, $P(N_c < \infty) = 1$.

Proof

Note that d_n^2 is a consistent estimator of δ^2. Hence, the result can be obtained from the fact that $d_n^2 \to \delta^2$ almost surely as $n \to \infty$. □

Theoretical Results

In our notation $c \downarrow 0$ means "c converges to a small positive number greater than 0" that is c is always positive and cannot take the value of exactly 0.

We formally state the results in the following theorems, with the proofs of the two theorems being similar to the proof given in Chattopadhyay and Kelley (in press) and De and Chattopadhyay (2015).

Theorem 1

For the minimum sample size $m_0 \geq 4$, the stopping rule in Equation 15 yields that on average, the final sample size of our procedure is asymptotically the same as the optimal sample size. Mathematically, $E(N_c/n_c) \to 1$ as $c \downarrow 0$.

Proof

First, we introduce notation. Note from Equation 15 That $N_c \geq m(=(A/(2c))^{\frac{1}{2(1+\gamma)}}) \geq m_0$ with probability 1.

For fixed $\epsilon,\gamma > 0$, note the following definitions:

$$n_{1c} = \left(\frac{A}{2c}\right)^{\frac{1}{2(1+\gamma)}}, \tag{27}$$

$$n_{2c} = n_c(1 - \epsilon), \tag{28}$$

and

$$n_{3c} = n_c(1 + \epsilon) \tag{29}$$

where $n_c = \sqrt{\frac{A}{2c}}\xi$. the definition of stopping rule N_c in (15) yields

$$\sqrt{\frac{A}{2c}}\, V_{N_c} \leq N_c \leq mI(N_c = m) + \sqrt{\frac{A}{2c}}\left(V_{N_c-1} + (N_c - 1)^{-\gamma}\right). \tag{30}$$

Because $N_c \to \infty$ almost surely as $c \downarrow 0$ and $V_n \to \xi$ almost surely as $n \to \infty$, by Theorem 2.1 of Gut (2009), $V_{N_c} \to \xi$ almost surely. Hence, dividing all sides of (30) by n_c and letting $c \downarrow 0$, $N_c/n_c \to 1$ almost surely as $c \downarrow 0$.

Now, $N_c \geq m$ almost surely and $n_c \geq 1$, dividing (30) by n_c yields

$$N_c/n_c \leq \frac{mI(N_c = m)}{n_c} + \frac{1}{\xi}\sup_{c>0}\left(V_{N_c-1} + (N_c - 1)^{-\gamma}\right) \text{ almost surely.} \tag{31}$$

Here, $P(N_c \geq m) = 0$ and $N_c \geq m \geq m_0$ with probability 1. Consider the inequality in Equation 31. Note that, $V_{N_c-1} \leq \sup_{c>0} V_{N_c-1}$ and $E[\sup_{c>0} V_{N_c-1}] < \infty \Leftrightarrow E[\sup_{n>m_0} V_n] < \infty$, with $\Leftrightarrow$ meaning "implies and is implied by." Thus, to prove, $E[\sup_{c>0} V_{N_c-1}] < \infty$ it is enough to show $E[\sup_{n>m_0} V_n] < \infty$ or $E[\sup_{n>m_0} V_n^2] < \infty$. We know that sample mean and sample variance are both U-statistics (e.g., see Mukhopadhyay & Chattopadhyay, 2013, 2014). Using Cauchy-Schwartz inequality and Cedilnik et al. (2006), we can say that for $t > 1$ and $m_0 > 4$

$$E\left[\sup_{n>m_0} V_n^2\right] = 2 + \left\{\frac{1}{(t-1)}E[(\bar{X}_{m_0} - \bar{Y}_{m_0})^4]E[(S_{m_0})^{-4t}]\right\}^{1/2} (< \infty), \tag{32}$$

Because $N_c/n_c \to 1$ almost surely as $c \downarrow 0$, by the dominated convergence theorem, we conclude that $\lim_{c\to 0} E[N_c/n_c] = 1$. Hence, the proof of Theorem 1 is complete. □

Theorem 2

For the minimum sample size $m_0 \geq 4$, the stopping rule in Equation 15 yields that the ratio regret is asymptotically 1. Mathematically, if $\gamma \in \left(0, \frac{1}{2}\right)$, $R_{N_c}(\delta)/R_{n_c}(\delta) \to 1$ as $c \downarrow 0$.

We need to show $\lim_{c\downarrow 0} R_{N_c}(\delta)/R^*_{n_c}(\delta) = \lim_{c\downarrow 0}(A/(4cn_c))E[d_{N_c} - \delta]^2 + \frac{1}{2}\lim E[N_c/n_c] = 1$. Thus, it is enough to show that $\lim_{c\to 0}(A/(2cn_c))E[d_{N_c} - \delta]^2 = 1$, i.e., $\lim_{c\downarrow 0} n_c E[d_{N_c} - \delta]^2 = \xi^2$. Because we know that $n_c E[d_{n_c} - \delta]^2 = \xi^2$, it is sufficient to show that

$$\lim_{c\downarrow 0} n_c\{E[(d_{N_c} - \delta)^2 - (d_{n_c} - \delta)^2]\} = 0. \tag{33}$$

Using lemmas and arguments in Chattopadhyay and Kelley (in press) and De and Chattopadhyay (2015) as needed, Theorem 2 can be proved. Here we will just prove required important lemmas.

(*Appendices continue*)

Lemma 4

If nonnegative i.i.d. random variables $X_1, \ldots, X_n$ are from the distribution F such that $E(X_1^{\max(2r,p)}) < \infty$ for some positive integers r and p, then for any $k > 0$,

$$P\Big(\max_{n_{1c}\le n\le n_{2c}} \left|(\bar{X}_{1n} - \bar{X}_{2n})^2 - (\mu_{1n} - \mu_{2n})^2\right| \ge k\Big) \le O(n_{1c}^{-r/2}) + O(n_{1c}^{-p/2}) \quad \text{as } c \downarrow 0.$$

Proof

Define, $\hat{\Delta}_n = \bar{X}_{1n} - \bar{X}_{2n}$ and $\Delta = \mu_1 - \mu_2$. Then the proof will be similar to the proof of Lemma 7.1 in De and Chattopadhyay (2015). □

Lemma 5

Suppose that nonnegative i.i.d. random variables $X_1, \ldots, X_n$ are observed from the distribution F such that $E(X_1)^{\max\{4p,\ 2p(r-1)\}}$ exist for some positive integers r and p. Then, for any $k > 0$,

$$P\Big(\max_{n_{1c}\le n\le n_{2c}} \left|\frac{1}{s_n^{2r}} - \frac{1}{\sigma^{2r}}\right| \ge k\Big) \le O(n_{1c}^{-p/2}) \text{ as } c \downarrow 0. \tag{34}$$

Proof

By Taylor expansion of $s_n^{-2r} = \frac{1}{\sigma^{2r}}(1 + (s_n^2 - \sigma^2)/\sigma^2)^{-r}$, we have

$$\left|\left(\frac{1}{s_n^{2r}} - \frac{1}{\sigma^{2r}}\right) I\left(\frac{1}{s_n^2} < \frac{1}{\sigma^2}\right)\right| = \frac{1}{\sigma^{2r}}\left|\left\{-\frac{r}{\sigma^2}(s_n^2 - \sigma^2) + \frac{r(r+1)}{2\sigma^4}\frac{(s_n^2 - \sigma^2)^2}{z^{r+2}}\right\} I\left(\frac{1}{s_n^2} < \frac{1}{\sigma^2}\right)\right|, \tag{35}$$

where $z \in [1, s_n^2/\sigma^2]$. Because $z^{-r+2} I(s_n^{-2} < \sigma^{-2}) \le 1$, we get

$$\left|\frac{1}{s_n^{2r}} - \frac{1}{\sigma^{2r}}\right| = \left|\left(\frac{1}{s_n^{2r}} - \frac{1}{\sigma^{2r}}\right) I\left(\frac{1}{s_n^2} \ge \frac{1}{\sigma^2}\right) + \left(\frac{1}{s_n^{2r}} - \frac{1}{\sigma^{2r}}\right) I\left(\frac{1}{s_n^2} < \frac{1}{\sigma^2}\right)\right| \le \left(\frac{1}{s_n^{2r}} - \frac{1}{\sigma^{2r}}\right)^+ + \frac{r}{\sigma^{2r+2}}|s_n^2 - \sigma^2| + \frac{r(r+1)}{2\sigma^{2r+4}}(s_n^2 - \sigma^2)^2. \tag{36}$$

Let $U_{1n} = \left(\frac{1}{s_n^{2r}} - \frac{1}{\sigma^{2r}}\right)^+$, $U_{2n} = \frac{r}{\sigma^{2r+2}}|s_n^2 - \sigma^2|$, and $U_{3n} = \frac{r(r+1)}{2\sigma^{2r+4}}(s_n^2 - \sigma^2)^2$. Using (36), we can write

$$P\Big(\max_{n_{1c}\le n\le n_{2c}} \left|\frac{1}{s_n^{2r}} - \frac{1}{\sigma^{2r}}\right| \ge k\Big) \le P\Big(\max_{n_{1c}\le n\le n_{2c}} U_{1n} \ge \frac{k}{3}\Big) + P\Big(\max_{n_{1c}\le n\le n_{2c}} U_{2n} \ge \frac{k}{3}\Big) + P\Big(\max_{n_{1c}\le n\le n_{2c}} U_{3n} \ge \frac{k}{3}\Big). \tag{37}$$

Because $\left(\frac{1}{s_n^{2r}} - \frac{1}{\sigma^{2r}}\right)$ is a reverse submartingale and $f(x) = x^+$ is a non-decreasing convex function of x, U_{1n} is a reverse submartingale. Therefore, using maximal inequality for reverse submartingales

$$\begin{aligned} P\Big(\max_{n_{1c}\le n\le n_{2c}} U_{1n} \ge \frac{k}{3}\Big) &\le \Big(\frac{3}{k}\Big)^p \mathrm{E}\left[\left(\frac{1}{s_{n_{1c}}^{2r}} - \frac{1}{\sigma^{2r}}\right)^+\right]^p \\ &\le \Big(\frac{3}{k}\Big)^p \mathrm{E}\left[\left(\frac{1}{s_{n_{1c}}^{2}} - \frac{1}{\sigma^{2}}\right)\left(\frac{1}{s_{n_{1c}}^{2(r-1)}} + \frac{1}{\sigma^2 s_{n_{1c}}^{2(r-2)}} + \ldots + \frac{1}{\sigma^{2(r-1)}}\right) I\big(s_{n_{1c}}^2 < \sigma^2\big)\right]^p \\ &\le \Big(\frac{3}{k}\Big)^p r^p \mathrm{E}\left[\left(\frac{1}{s_{n_{1c}}^{2}} - \frac{1}{\sigma^{2}}\right)^p s_{n_{1c}}^{-2p(r-1)}\right] \\ &\le \Big(\frac{3r}{k}\Big)^p \left\{\mathrm{E}\big[(s_{n_{1c}}^2 - \sigma^2)^{4p}\big] \mathrm{E}\left[\left(\frac{1}{\sigma^2 s_{n_{1c}}^2}\right)^{4p}\right]\right\}^{\frac{1}{4}} \left\{\mathrm{E}\left(\frac{1}{s_{n_{1c}}^{2p(r-1)}}\right)\right\}^{\frac{1}{2}} \le O(n_{1c}^{-p/2}). \end{aligned} \tag{38}$$

The last two inequalities are obtained by using Cauchy–Schwarz inequality and Lemma 2.2 of Sen and Ghosh (1981) and also due to the existence of $\mathrm{E}\left[\left(\frac{1}{s_{n_{1c}}}\right)^{4p}\right]$ and $\mathrm{E}\left[\left(\frac{1}{s_{n_{1c}}}\right)^{2p(r-1)}\right]$. Because $|s_n^2 - \sigma^2|$ and $(s_n^2 - \sigma^2)$ are reverse submartingales, we can write

$$P\Big(\max_{n_{1c}\le n\le n_{2c}} U_{2n} \ge \frac{k}{3}\Big) \le \Big(\frac{3r}{k\sigma^{2r+2}}\Big)^{2p} E\big(s_{n_{1c}}^2 - \sigma^2\big)^{2p} \le O(n_{1c}^{-p}), \tag{39}$$

$$P\Big(\max_{n_{1c}\le n\le n_{2c}} U_{3n} \ge \frac{k}{3}\Big) \le \Big(\frac{3r(r+1)}{2k\sigma^{2r+4}}\Big)^{p} E\big(s_{n_{1c}}^2 - \sigma^2\big)^{2p} \le O(n_{1c}^{-p}). \tag{40}$$

Applying Equations (38), (39), and (40) in (37) completes the Proof. □

(*Appendices continue*)

Appendix B

Lemmas and Proofs for General Scenario

This appendix supports our discussion in the text with the technical underpinnings of the equations presented.

Lemma 6

The approximate risk function given in Equation 17 is minimized if $n_c\left(=\sqrt{\frac{A}{2c}}\xi\right)$ individuals are selected from Groups 1 and 2, where, $\xi^2 = \left(2 + \frac{\delta^2}{4}\right)$.

Proof

Let us start with an approximate risk function in which cost of sampling per observation in both groups are different. Suppose c_1 and c_2 be the sampling cost of observing each participant in Groups 1 and 2, respectively. So, the approximate risk function given in Equation 9 is

$$R_{\tilde{n}}(\delta) \approx A\left\{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) + \frac{\delta^2}{2(n_1 + n_2)}\right\} + (c_1 n_1 + c_2 n_2)$$

Minimizing $R_{\tilde{n}}(\delta)$ with respect to n_1 and n_2, we get,

$$\begin{aligned} -\frac{A}{n_1^2} - \frac{A\delta^2}{2(n_1 + n_2)^2} + c_1 = 0 \\ -\frac{A}{n_2^2} - \frac{A\delta^2}{2(n_1 + n_2)^2} + c_2 = 0 \end{aligned} \tag{41}$$

Solving the above, we can get the optimal sample sizes n_1^* for Group 2 and n_2^* for Group 2 by using,

$$n_1^* = \sqrt{\frac{A}{c_1}}\left(1 + \frac{\delta^2}{2\left(1 + \sqrt{\frac{A}{(c_2 - c_1)(n_1^*)^2 + A}}\right)}\right)^{1/2} \tag{42}$$

and

$$n_2^* = n_1^*\left(1 + \frac{(c_2 - c_1)(n_1^*)^2}{A}\right)^{-1/2}. \tag{43}$$

Differentiating both in Equation 41 and consequently finding the Hessian matrix by plugging in optimal sample sizes that can be found in Equations 42 and, it can be shown that these optimal sample sizes minimizes the approximate risk function.

Now, in Equation 9, the sampling cost per unit observation for both groups are same, so, $c_1 = c_2 = c$, say. Using, that we get $n_1^* = n_2^*$, let it be n_c. Plugging in n_c, we get $n_c\left(=\sqrt{\frac{A}{2c}}\xi\right)$, where, $\xi^2 = \left(2 + \frac{\delta^2}{4}\right)$. □

Lemma 7

Under the assumption that $\delta < \infty$, for any $c_1 > 0$, $c_2 > 0$, the sample sizes N_1 and N_2 are both finite.

Proof

Note that d_n^2 is a consistent estimator of δ^2. Hence, the result for the sample size N_1 can be obtained from the fact that $d_n^2 \to \delta^2$ almost surely as $n \to \infty$. Because N_2 depends on N_1, so, N_2 is also finite. □

Received January 27, 2015
Revision received March 17, 2016
Accepted March 20, 2016 ■