Behav Res (2013) 45:955–967
DOI 10.3758/s13428-013-0320-7

Confidence intervals and sample size calculations for the standardized mean difference effect size between two normal populations under heteroscedasticity

G. Shieh

Published online: 7 March 2013
© Psychonomic Society, Inc. 2013

Abstract The use of effect sizes and associated confidence intervals in all empirical research has been strongly emphasized by journal publication guidelines. To help advance theory and practice in the social sciences, this article describes an improved procedure for constructing confidence intervals of the standardized mean difference effect size between two independent normal populations with unknown and possibly unequal variances. The presented approach has advantages over the existing formula in both theoretical justification and computational simplicity. In addition, simulation results show that the suggested one- and two-sided confidence intervals are more accurate in achieving the nominal coverage probability. The proposed estimation method provides a feasible alternative to the most commonly used measure of Cohen's *d* and the corresponding interval procedure when the assumption of homogeneous variances is not tenable. To further improve the potential applicability of the suggested methodology, the sample size procedures for precise interval estimation of the standardized mean difference are also delineated. The desired precision of a confidence interval is assessed with respect to the control of expected width and to the assurance probability of interval width within a designated value. Supplementary computer programs are developed to aid in the usefulness and implementation of the introduced techniques.

Keywords Behrens–Fisher problem · Cohen's *d* · Confidence interval · Precision · Welch's statistic

Electronic supplementary material The online version of this article (doi:10.3758/s13428-013-0320-7) contains supplementary material, which is available to authorized users.

G. Shieh (✉)
Department of Management Science, National Chiao Tung University, 1001 Ta Hsueh Road,
Hsinchu, Taiwan 30050
e-mail: gwshieh@mail.nctu.edu.tw

The reporting of effect sizes and associated confidence intervals for primary results in all empirical social science research has been recommended in Wilkinson and the American Psychological Association Task Force on Statistical Inference (1999), the American Educational Research Association Task Force on Reporting of Research Methods (2006), and the *Publication Manual of the American Psychological Association* (2010). Correspondingly, numerous practical guidelines and suggestions for selecting, calculating, and interpreting effect size indices for various types of statistical analyses have been provided in the literature, such as Alhija and Levy (2009), Breaugh (2003), Durlak (2009), Ferguson (2009), Fern and Monroe (1996), Grissom and Kim (2012), Huberty (2002), Kirk (1996), Kline (2004), Olejnik and Algina (2000), Richardson (1996), Rosenthal, Rosnow, and Rubin (2000), Rosnow and Rosenthal (2003), and Vacha-Haase and Thompson (2004). It has steadily become a general consensus in the methodological literature of behavior, education, management, and related disciplines that effect sizes accompanied by their corresponding confidence intervals are perhaps the best approach for conveying quantitative information in applied research.

According to the general review of Ferguson (2009), effect sizes can be categorized into four general classes: (1) group difference, (2) strength of association, (3) corrected estimates, and (4) risk estimates. The group difference indices estimate the magnitude of difference between two or more groups, and Cohen's *d* (Cohen, 1969) is the most commonly used measure across virtually all disciplines of the social sciences. Specifically, Cohen's *d* is an estimate of the standardized mean difference that reflects the difference between two sample means divided by their pooled sample standard deviation under homoscedasticity. For the purpose of measuring the size of effect between two treatment groups with unequal variances, Cohen's *d* is no longer a proper estimator, because its standardizer, the

pooled sample standard deviation, obscures whatever inherent differences in variance might have existed. Thus, the standardizer of the mean difference should be prudently modified to adequately scale actual characteristics of group discrepancies (Grissom & Kim, 2001). But there appears to be a lack of consensus in the literature on which standardizer is most appropriate under which circumstances for computing standardized mean difference. The adoption of different standardizers naturally leads to distinct effect size measures and ultimately results in varied target population counterparts. Therefore, the choice of a suitable effect size parameter and statistic is a difficult and substantive decision. In this regard, three primitive procedures are described in Glass, McGraw, and Smith (1981, p. 106) and Keselman, Algina, Lix, Wilcox, and Deering (2008, Equations 14, 16, and 17). First, a simple method is to apply the Glass (1976) formula with either one of the two sample standard deviations as the standardizer. This approach yields two conceptually different versions of the underlying target effect size, corresponding to the two heterogeneous variances. It is intuitive to perform the standardization of mean difference by the control group standard deviation. Then the experimental group standard deviation, supposedly with dissimilar magnitude, will have no influence on the resulting effect size (Grissom & Kim, 2001). To prevent this situation, one may attempt to report both standardized mean differences simultaneously. Unfortunately, the two choices of control group and experimental group standardizer can give substantial differences in effect size, and indiscriminate use of this strategy could result in some interpretive ambiguity.

Second, another popular alternative uses the square root of the average of the two sample variances as the standardizer. The synthesis of two variances with equal weights is intuitively straightforward. However, the group effect, no doubt, is estimated by the difference between two sample means, and the variance of the sample mean difference cannot be expressed in terms of the average variance. The only exception is when the group sizes are equal. Therefore, the simple average of two variances does not generally conform to the exact variance of the sample mean difference. In absence of a theoretical justification, this procedure is also vulnerable to the criticism against using the average standard deviation addressed in Glass et al. (1981, p. 106), in that it seems to reflect merely a statistical reaction to a perplexing choice. A similar concern is also provided in Grissom and Kim (2001, p. 136).

The third approach considers Welch's (1938) statistic for the well-known Behrens–Fisher problem of comparing the difference between two normal means that may have unequal population variances (Kim & Cohen, 1998). Due to the complexity in distributional properties, a variance stabilizing transformation of the Welch statistic is presented in Kulinskaya and Staudte (2007) for the estimation of a standardized effect size. In this case, the target parameter of the standardized mean difference is the mean difference divided by a modification of the exact standard deviation of the sample mean difference. Unlike the two previous procedures, the standardizer depends not only on the population variances, but also on the group size allocation ratios. It is noted in Keselman et al. (2008) that the particular formulation of Kulinskaya and Staudte (2007) raises a practical problem about its general use as an effect size measure. Specifically, the concern of Keselman et al. is about its dependence upon sample sizes. However, the standardizer in Keselman et al. (Equation 15) is not the same as those in Kulinskaya and Staudte (2006, Equation 8; 2007, Equation 1). Also, the resulting standardized mean difference actually depends not on the group sizes but, rather, on the allocation ratio between two groups.

In view of the practical importance of effect sizes and confidence intervals, this article proposes an alternative approach for interval estimation of the standardized mean difference between two normal populations under the assumption of possibly unequal variances. On the basis of the approximate noncentral t distribution for Welch's statistic, the inversion confidence interval principle (Steiger & Fouladi, 1997) is utilized to construct accurate confidence intervals for the standardized mean difference effect size. The accuracy of the suggested procedure is evaluated by the computed confidence interval corresponding to the nominal coverage probability and the actual probability of coverage it achieves. Extensive empirical investigations were conducted to demonstrate the advantages of the proposed approach over the variance stabilizing transformation method of Kulinskaya and Staudte (2007) under a variety of effect size configurations, variance patterns, and sample size structures. Moreover, sample size methodologies for precise interval estimation of standardized effect sizes are delineated in two distinct aspects. One method gives the minimum sample size, such that the expected confidence interval width is within the designated bound. The other method provides the sample size needed to guarantee, with a given assurance probability, that the width of a confidence interval will not exceed the planned range. Essentially, the suggested sample size procedures are direct and heteroscedastic extensions of those considered in Kelley and Rausch (2006) under a homogeneous variance setting. This investigation updates and expands the current work in such a way that the findings not only improve the fundamental limitations of the existing method, but also reinforce the practice of measuring effect size in the context of heteroscedastic situations. In addition, the SAS computer codes are available as supplemental materials to facilitate the recommended procedures for computing the confidence intervals of standardized mean difference and the necessary sample size in planning research designs.

Confidence interval procedures

The well-known Behrens–Fisher problem is to compare the difference between two normal means when the variances are different. Accordingly, Welch's (1938) approximate t procedure has been considered as a satisfactory and robust solution over the two-sample t under the heterogeneous variances assumption of the Behrens–Fisher problem. To facilitate exposition, consider independent random samples from two normal populations with the following formulation:

$$V = \frac{\overline{X}_1 - \overline{X}_2}{\left(S_1^2/N_1 + S_2^2/N_2\right)^{1/2}}, \tag{1}$$

where μ_1, μ_2, σ_1^2, σ_2^2 are unknown parameters, $j = 1, \ldots, N_i$, and $i = 1$ and 2. The widely recognized Welch t statistic is of the form

$$X_{ij} \sim N(\mu_i, \sigma_i^2),$$

$$\text{where} \quad \overline{X}_1 = \sum_{j=1}^{N_1} X_{1j}/N_1 \ , \quad \overline{X}_2 = \sum_{j=1}^{N_2} X_{2j}/N_2 \ , \quad S_1^2 = \sum_{J=1}^{N_1} \left(X_{1j} - \overline{X}_1\right)^2/(N_1 - 1), \text{ and } S_2^2 = \sum_{j=1}^{N_2} \left(X_{2j} - \overline{X}_2\right)^2/(N_2 - 1).$$

Although the underlying normality assumption in the above-mentioned two-sample location problem is a convenient and useful setup, the exact distribution of Welch's statistic V is comparatively complicated. The practical importance and methodological complexity of the problem has led to numerous attempts to develop various procedures and algorithms for resolving the issue (Kim & Cohen, 1998, and references therein). To extend the notion of effect size within the heteroscedastic framework, Kulinskaya and Staudte (2007) suggested considering the following measure of standardized mean difference:

$$\delta* = \frac{\mu_1 - \mu_2}{\left(\sigma_1^2/q_1 + \sigma_2^2/q_2\right)^{1/2}}, \tag{2}$$

where $q_i = N_i/N$, $i = 1$ and 2 is the group size allocation ratio, and $N = (N_1 + N_2)$. It is important to note that the effect size $\delta*$ is a function of mean difference, variance components, and allocation ratios. Then, if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, it can be easily seen that $\delta* = \delta/(1/q_1 + 1/q_2)^{1/2}$, where $\delta = (\mu_1 - \mu_2)/\sigma$ is the prevailing Cohen's (1969) population effect size index. The particular relation between two effect sizes $\delta*$ and δ reveals that the heteroscedastic adaptation of $\delta*$ relative to the homoscedastic counterpart of δ relies on the design factor through the sample size distributional proportions q_1 and q_2 ($q_1 + q_2 = 1$). For ease of reference, the effect sizes suggested by Glass et al. (1981) with either one of the standard deviations as the standardizer are $\delta_1 = (\mu_1 - \mu_2)/\sigma_1$ and $\delta_2 = (\mu_1 - \mu_2)/\sigma_2$, respectively. On the other hand, the above-mentioned simple alternative involving average variance is expressed as $\delta_A = (\mu_1 - \mu_2)/\left\{\left(\sigma_1^2 + \sigma_2^2\right)/2\right\}^{1/2}$. All these distinct standardized mean difference effect sizes are unique and have their merits in appropriate situations. However, the expected value and variance of the mean difference are $E\left(\overline{X}_1 - \overline{X}_2\right) = \mu_1 - \mu_2$ and $Var\left(\overline{X}_1 - \overline{X}_2\right) = \sigma_1^2/N_1 + \sigma_2^2/N_2 = \left(\sigma_1^2/q_1 + \sigma_2^2/q_2\right)/N$, respectively. Therefore, to eliminate the potential influence of sample size on the net group effect $\mu_1 - \mu_2$ through sample mean difference $\overline{X}_1 - \overline{X}_2$, one may use $\left(\sigma_1^2/q_1 + \sigma_2^2/q_2\right)$ as the standardizer by factoring out the magnitude of total sample size from $Var\left(\overline{X}_1 - \overline{X}_2\right)$. Unlike the homogeneous variance cases, the group variances intertwine with the sample size allocation ratios in the standardizer $\left(\sigma_1^2/q_1 + \sigma_2^2/q_2\right)$. It suggests that the standardized mean difference needs to accommodate the design characteristic of group allocation scheme under the more sophisticated situation of heteroscedasticity. Note that the squared multiple correlation coefficient is the prevailing strength of association effect size in linear regression. Despite its usefulness, applied researchers may not notice that it is a function of both the model (coefficient and variance) parameters and the distribution properties (variance and covariance matrix) of the designated covariates (for further details, see Gatsonis & Sampson, 1989; Shieh, 2006). In this two-sample case, the group assignment is the covariate variable and represents the extent to which knowledge of group membership improves prediction of outcomes for the criterion variable in the sample. In addition, it is demonstrated in Kulinskaya and Staudte (2006, pp. 99–101) that the strength of association effect size or weighted coefficient of determination is directly related to the standardized mean difference $\delta*$ under the heteroscedastic ANOVA models when there are two populations. According to the statistical properties of Welch's statistic under heteroscedasticity, it does not appear possible to define a proper standardized effect size without accounting for the relative group size of subpopulations in a sampling scheme. Since Welch's approach to the Behrens–Fisher problem is so entrenched, we restrict our attention to the estimation procedures of $\delta*$ defined in Equation 2.

Using a chi-square approximation to the distribution of a positive linear combination of two chi-square variables, Welch (1938) proposed the approximate distribution for V when $\mu_1 = \mu_2$:

$$V \dot{\sim} t(\nu),$$

where $t(\nu)$ is the t distribution with degrees of freedom ν and $\nu = \nu\left(N_1, N_2, \sigma_1^2, \sigma_2^2\right)$ with

$$\nu = \frac{\left\{\sigma_1^2/N_1 + \sigma_2^2/N_2\right\}^2}{\left\{\sigma_1^2/N_1\right\}^2/(N_1 - 1) + \left\{\sigma_2^2/N_2\right\}^2/(N_2 - 1)}.$$

For making statistical inferences, Kulinskaya and Staudte (2007) derived an approximate confidence interval procedure for δ* by stabilizing the variance of the Welch statistic. However, three potential limitations to their method should be pointed out. First, their theoretical presentations and algebraic expressions are complicated. The determination of confidence limits entails lengthy and tedious calculations. Therefore, their result is of limited usage in application. Second, the numerical examinations in Kulinskaya and Staudte (2007) show that although their method seems to provide useful results for small values of standardized mean difference, accuracy of the confidence intervals deteriorates substantially as the true standardized mean difference deviates from zero. Third, the distribution of V is generally skewed, and this essential feature gives rise to asymmetric confidence intervals for δ* as in the corresponding interval estimation of δ. Even though it is not obvious at first sight, the transformed equidistant confidence interval of Kulinskaya and Staudte (2007) is therefore presumably inappropriate and is not likely to be accurate when one-sided confidence intervals are considered. In an effort to improve the quality of research analysis and design, an alternative approach is presented next to construct confidence intervals of the standardized mean difference.

It can be shown with the same theoretical arguments and analytic derivations as those in Welch (1938) that the statistic V has the general approximate distribution

$$V \dot{\sim} t(\nu, \Delta*), \tag{3}$$

where $t(\nu, \Delta^*)$ is a noncentral t distribution with degrees of freedom ν and noncentrality parameter $\Delta^* = N^{1/2}\delta^*$. Noting that degrees of freedom ν depends on the unknown variances, an approximate version can be obtained by substituting the respective sample estimates for the variances. For inferential purposes, the term of degrees of freedom ν in Equation 3 is replaced by its counterpart $\hat{\nu}$ with direct substitution of $\left(S_1^2, S_2^2\right)$ for $\left(\sigma_1^2, \sigma_2^2\right)$ in ν:

$$\hat{\nu} = \frac{\left\{S_1^2/N_1 + S_2^2/N_2\right\}^2}{\left\{S_1^2/N_1\right\}^2/(N_1-1) + \left\{S_2^2/N_2\right\}^2/(N_2-1)}.$$

Hence, the adjustment gives the following modified distribution:

$$V \dot{\sim} = t(\hat{\nu}, \Delta*). \tag{4}$$

The noncentral t distribution provides a feasible approximation to the underlying distribution of V in terms of degrees of freedom $\hat{\nu}$ and noncentrality parameter Δ^*. It is noteworthy that Welch's V statistic and suggested approximate noncentral t distribution given in Equation 4 closely resemble the two-sample t and corresponding exact noncentral t distribution under a homogeneous variance setup. This can be readily obtained from the first moment of the approximate noncentral t distribution that $E[V] \doteq (\hat{\nu}/2)^{1/2} \times \Gamma\{(\hat{\nu}-1)/2\}\Delta* / \Gamma\{\hat{\nu}/2\}$, where $\Gamma\{\cdot\}$ is the gamma function. Hence, a nearly unbiased point estimator of the standardized mean difference δ* is

$$\hat{\delta}^*_{\mathrm{NU}} = \frac{\Gamma\{\hat{\nu}/2\}}{(N\hat{\nu}/2)^{1/2}\Gamma\{(\hat{\nu}-1)/2\}} V.$$

In addition, the noncentrality inversion procedure of Steiger and Fouladi (1997) can immediately be applied to find the confidence intervals of δ* with the noncentral distribution $t(\hat{\nu}, N^{1/2}\delta*)$. Explicitly, the upper $100(1-\alpha_1)$% confidence interval of δ* is of the form $\left(\hat{\delta}^*_{\mathrm{L}}, \infty\right)$, in which $\hat{\delta}^*_{\mathrm{L}}$ satisfies

$$P\left\{t\left(\hat{\nu}, N^{1/2}\hat{\delta}^*_{\mathrm{L}}\right) < V_{\mathrm{O}}\right\} = 1-\alpha_1, \tag{5}$$

where V_{O} is the observed value of Welch's statistic V defined in Equation 1. Likewise, the lower $100(1-\alpha_2)$% confidence interval of δ* is of the form $\left(-\infty, \hat{\delta}^*_{\mathrm{U}}\right)$, in which $\hat{\delta}^*_{\mathrm{U}}$ satisfies

$$P\left\{t\left(\hat{\nu}, N^{1/2}\hat{\delta}^*_{\mathrm{U}}\right) > V_{\mathrm{O}}\right\} = 1-\alpha_2. \tag{6}$$

Furthermore, a $100(1-\alpha)$% two-sided confidence interval $(\hat{\delta}^*_{\mathrm{L}}, \hat{\delta}^*_{\mathrm{U}})$ of δ* can be obtained by jointly applying Equations 5 and 6 with $\alpha = \alpha_1 + \alpha_2$. The most common practice is to assume $\alpha_1 = \alpha_2 = \alpha/2$, and this leads to the $100(1-\alpha)$% two-sided confidence interval for δ* with equal tail confidence probability. In short, with the desired confidence level, observed value V_{O}, and estimated degrees of freedom , the numerical computation of confidence limits $\hat{\delta}^*_{\mathrm{L}}$ and $\hat{\delta}^*_{\mathrm{U}}$ requires the evaluation of the noncentrality distribution function of a noncentral t variable, such as the SAS noncentrality function TNONCT. The SAS/IML (SAS Institute, 2011) program employed to perform the suggested confidence interval calculations is available as electronic supplementary material.

Numerical comparison of confidence interval procedures

To clarify the issues surrounding the adequacy of statistical analysis and quality of research findings, we next perform numerical investigations to evaluate and compare the accuracy of the suggested procedure and Kulinskaya and Staudte's (2007, Equation 8) formula for computing confidence intervals of standardized mean difference under various model configurations likely to occur in practice.

For the purposes of assessing the behavior of interval estimation procedures with potentially diverse situations, we consider the model characteristics with variances $\left(\sigma_1^2, \sigma_2^2\right)$ = (1, 1) and (1, 4) and sample sizes (N_1, N_2) = (15, 15), (10, 20), and (20, 10). These settings not only include both homoscedastic/heteroscedastic and balanced/unbalanced designs, but also create direct and inverse pairing between variance and sample size structures. To prevent repetition in the homogeneous variance case of $\left(\sigma_1^2, \sigma_2^2\right)$ = (1, 1), the sample sizes setting (N_1, N_2) = (20, 10) is omitted from the investigation. Thus, only the two combinations of (N_1, N_2) = (15, 15) and (10, 20) are examined. Overall, these considerations result in a total of five different joined configurations. Although these sample sizes are smaller than would be likely in many two-sample location and effect size studies, it is plausible that if problems or deficiencies were to be seen with confidence interval calculations, they would be most apparent with small group sizes. Without loss of generality, the second group mean is fixed as $\mu_2 = 0$, and the first group mean μ_1 is chosen such that the standardized mean difference $\delta^* = 0$, 1, 2, and 3 for each combined structure of $\left(\sigma_1^2, \sigma_2^2\right)$ and (N_1, N_2).

With the given sample sizes and parameter configurations, estimates of the true coverage probability are computed through Monte Carlo simulation of 10,000 independent data sets. For each replicate, the confidence limits associated with one-sided upper and lower 100(1 − α)% confidence intervals are computed for both (1 − α) = 0.95 and 0.975. These confidence limits are also employed to construct the two-sided 90% and 95% confidence intervals. Accordingly, a total of six different sets of confidence intervals are obtained. Thus, our simulations cover a much broader range of situations than those considered in the previous study of Kulinskaya and Staudte (2007), where they examined only the performance of two-sided 95% confidence intervals. In each case, the simulated coverage probability is the proportion of the 10,000 replicates whose intervals contain the population effect size δ*. The accuracy of the examined procedure is determined by the difference between the simulated coverage probability and designated coverage probability as error = simulated coverage probability − nominal coverage probability. The actual coverage probabilities and errors corresponding to the cross settings of $\left(\sigma_1^2, \sigma_2^2\right)$ and (N_1, N_2) are presented in Tables 1, 2, 3, 4 and 5.

An examination of the numerical results of both approaches reveals the general pattern that when all other factors remain constant, the discrepancy between simulated and nominal coverage probabilities of two-sided confidence intervals tends to increase for larger effect size δ*, smaller confidence level (1 − α), heteroscedastic structure, and unbalanced design. The differences are substantially prominent for the cases with inverse pairing of variances and sample sizes in Table 5. In addition, the prescribed overall phenomenon is much more noticeable for Kulinskaya and Staudte's (2007) method than for the proposed approach. As demonstrated in Kulinskaya and Staudte (2007), their two-sided confidence interval procedure appears to have reasonably good coverage probabilities for small δ* (≤1). However, the behavior of two-sided confidence intervals can be distorted to a remarkable degree when δ* > 1, as is shown here. For illustration, the simulated coverage probabilities of one-sided upper and lower 95% confidence intervals and two-sided 90% confidence intervals in Table 5 are plotted for the proposed approach and Kulinskaya and Staudte's (2007) method in Fig. 1, respectively.

Table 1 Simulated coverage probability and error of the approximate confidence intervals for standardized mean difference effect size when $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $N_1 = 15$, and $N_2 = 15$

δ*	The proposed approach						Kulinskaya and Staudte (2007)					
	Upper 95% CI	Error	Lower 95% CI	Error	Two-sided 90% CI	Error	Upper 95% CI	Error	Lower 95% CI	Error	Two-sided 90% CI	Error
0	0.9478	−0.0022	0.9508	0.0008	0.8986	−0.0014	0.9500	0.0000	0.9520	0.0020	0.9020	0.0020
1	0.9515	0.0015	0.9516	0.0016	0.9031	0.0031	0.9602	0.0102	0.9466	−0.0034	0.9068	0.0068
2	0.9539	0.0039	0.9509	0.0009	0.9048	0.0048	0.9737	0.0237	0.9201	−0.0299	0.8938	−0.0062
3	0.9556	0.0056	0.9544	0.0044	0.9100	0.0100	0.9814	0.0314	0.8953	−0.0547	0.8767	−0.0233
	Upper 97.5% CI	Error	Lower 97.5% CI	Error	Two-sided 95% CI	Error	Upper 97.5% CI	Error	Lower 97.5% CI	Error	Two-sided 95% CI	Error
0	0.9752	0.0002	0.9741	−0.0009	0.9493	−0.0007	0.9752	0.0002	0.9745	−0.0005	0.9497	−0.0003
1	0.9795	0.0045	0.9759	0.0009	0.9554	0.0054	0.9821	0.0071	0.9759	0.0009	0.9580	0.0080
2	0.9785	0.0035	0.9748	−0.0002	0.9533	0.0033	0.9869	0.0119	0.9635	−0.0115	0.9504	0.0004
3	0.9791	0.0041	0.9784	0.0034	0.9575	0.0075	0.9911	0.0161	0.9499	−0.0251	0.9410	−0.0090

Table 2 Simulated coverage probability and error of the approximate confidence intervals for standardized mean difference effect size when $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $N_1 = 10$, and $N_2 = 20$

δ*	The proposed approach						Kulinskaya and Staudte (2007)					
	Upper 95% CI	Error	Lower 95% CI	Error	Two-sided 90% CI	Error	Upper 95% CI	Error	Lower 95% CI	Error	Two-sided 90% CI	Error
0	0.9478	−0.0022	0.9502	0.0002	0.8980	−0.0020	0.9506	0.0006	0.9524	0.0024	0.9030	0.0030
1	0.9409	−0.0091	0.9510	0.0010	0.8919	−0.0081	0.9667	0.0167	0.9310	−0.0190	0.8977	−0.0023
2	0.9381	−0.0119	0.9562	0.0062	0.8943	−0.0057	0.9845	0.0345	0.8134	−0.1366	0.7979	−0.1021
3	0.9364	−0.0136	0.9564	0.0064	0.8928	−0.0072	0.9922	0.0422	0.6548	−0.3042	0.6380	−0.2620
	Upper 97.5% CI	Error	Lower 97.5% CI	Error	Two-sided 95% CI	Error	Upper 97.5% CI	Error	Lower 97.5% CI	Error	Two-sided 95% CI	Error
0	0.9727	−0.0023	0.9745	−0.0005	0.9472	−0.0028	0.9740	−0.0010	0.9750	0.0000	0.9490	−0.0010
1	0.9686	−0.0064	0.9774	0.0024	0.9460	−0.0040	0.9831	0.0081	0.9712	−0.0038	0.9543	0.0043
2	0.9667	−0.0083	0.9795	0.0045	0.9462	−0.0038	0.9922	0.0172	0.9035	−0.0715	0.8957	−0.0543
3	0.9679	−0.0071	0.9801	0.0051	0.9480	−0.0020	0.9957	0.0207	0.7733	−0.2017	0.7690	−0.1810

In addition, the coverage probability of their two-sided $100(1 - \alpha)\%$ confidence interval may be acceptable, but the one-sided upper and lower $100(1 - \alpha/2)\%$ confidence intervals constructed with the two confidence limits do not necessarily maintain the selected level. As is shown in Tables 1, 2, 3, 4 and 5, when $\delta^* \geq 1$, the coverage probabilities of their 95% upper and lower confidence intervals are typically higher and lower, respectively, than the nominal level. Specifically, it can be found in Table 2 that the coverage probability of the two-sided 90% confidence interval is 0.8977 with error −0.0023 when $\delta^* = 1$. However, the excellent coverage performance of the two-sided 90% confidence interval may be misleading because the corresponding upper and lower confidence limits are problematic. The corresponding coverage probabilities for the upper and lower 95% confidence intervals are 0.9667 and 0.9310, respectively. The associated errors of 0.0167 and −0.0190 present a sizable amount of discrepancy, and they suggest that the two finite confidence limits are both smaller than the respective exact value. Thus, a mere coverage probability evaluation of two-sided confidence intervals may obscure systematic underestimation in confidence limits that might have existed in Kulinskaya and Staudte's (2007) variance stabilizing transformation.

In contrast, the performances of the suggested one- and two-sided interval procedures appear to be fairly effective for the range of model specifications considered in the present article. The only exceptions are associated with the extreme settings for inverse pairing of variances and sample sizes in Table 5. Although one of the errors is as much as −0.0215, the results are still comparable to or outperform those of Kulinskaya and Staudte (2007). Nonetheless, additional numerical investigations showed that the coverage properties are substantially improved with errors −0.0082

Table 3 Simulated coverage probability and error of the approximate confidence intervals for standardized mean difference effect size when $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $N_1 = 15$, and $N_2 = 15$

δ*	The proposed approach						Kulinskaya and Staudte (2007)					
	Upper 95% CI	Error	Lower 95% CI	Error	Two-sided 90% CI	Error	Upper 95% CI	Error	Lower 95% CI	Error	Two-sided 90% CI	Error
0	0.9445	−0.0055	0.9494	−0.0006	0.8939	−0.0061	0.9472	−0.0028	0.9515	0.0015	0.8987	−0.0013
1	0.9461	−0.0039	0.9516	0.0016	0.8977	−0.0023	0.9564	0.0064	0.9446	−0.0054	0.9010	0.0010
2	0.9456	−0.0044	0.9510	0.0010	0.8966	−0.0034	0.9664	0.0164	0.9129	−0.0371	0.8793	−0.0207
3	0.9420	−0.0080	0.9526	0.0026	0.8946	−0.0054	0.9742	0.0242	0.8775	−0.0725	0.8517	−0.0483
	Upper 97.5% CI	Error	Lower 97.5% CI	Error	Two-sided 95% CI	Error	Upper 97.5% CI	Error	Lower 97.5% CI	Error	Two-sided 95% CI	Error
0	0.9721	−0.0029	0.9757	0.0007	0.9478	−0.0022	0.9724	−0.0026	0.9761	0.0011	0.9485	−0.0015
1	0.9739	−0.0011	0.9764	0.0014	0.9503	0.0003	0.9780	0.0030	0.9758	0.0008	0.9538	0.0038
2	0.9704	−0.0046	0.9736	−0.0014	0.9440	−0.0060	0.9818	0.0068	0.9577	−0.0173	0.9395	−0.0105
3	0.9691	−0.0059	0.9780	0.0030	0.9471	−0.0029	0.9850	0.0100	0.9383	−0.0367	0.9233	−0.0267

Table 4 Simulated coverage probability and error of the approximate confidence intervals for standardized mean difference effect size when $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $N_1 = 10$, and $N_2 = 20$

δ*	The proposed approach						Kulinskaya and Staudte (2007)					
	Upper 95% CI	Error	Lower 95% CI	Error	Two-sided 90% CI	Error	Upper 95% CI	Error	Lower 95% CI	Error	Two-sided 90% CI	Error
0	0.9493	−0.0007	0.9498	−0.0002	0.8991	−0.0009	0.9507	0.0007	0.9517	0.0017	0.9024	0.0024
1	0.9505	0.0005	0.9509	0.0009	0.9014	0.0014	0.9713	0.0213	0.9420	−0.0080	0.9133	0.0133
2	0.9537	0.0037	0.9518	0.0018	0.9055	0.0055	0.9899	0.0399	0.8863	−0.0637	0.8762	−0.0238
3	0.9579	0.0079	0.9542	0.0042	0.9121	0.0121	0.9963	0.0463	0.7844	−0.1656	0.7807	−0.1193
	Upper 97.5% CI	Error	Lower 97.5% CI	Error	Two-sided 95% CI	Error	Upper 97.5% CI	Error	Lower 97.5% CI	Error	Two-sided 95% CI	Error
0	0.9738	−0.0012	0.9763	0.0013	0.9501	0.0001	0.9744	−0.0006	0.9769	0.0019	0.9513	0.0013
1	0.9764	0.0014	0.9781	0.0031	0.9545	0.0045	0.9862	0.0112	0.9765	0.0015	0.9627	0.0127
2	0.9779	0.0029	0.9756	0.0006	0.9535	0.0035	0.9953	0.0203	0.9450	−0.0300	0.9403	−0.0097
3	0.9807	0.0057	0.9772	0.0022	0.9579	0.0079	0.9986	0.0236	0.8823	−0.0927	0.8809	−0.0691

for larger sample sizes $(N_1, N_2) = (40, 20)$. In short, the variance stabilizing transformation of Kulinskaya and Staudte (2007) may be useful for small effect size $\delta^* < 1$. In view of the unknown nature of the effect sizes, technical complexity, and computational requirements, it is worthwhile to consider an alternative procedure that yields reliable results with fewer limitations and difficulties. Therefore, the proposed approach is recommended over the current method of Kulinskaya and Staudte (2007) for its overall performance, methodological transparency, and computational ease.

Sample size calculations

From an advance study design viewpoint, researchers may wish to credibly address specific research questions and confirm meaningful effect sizes, so that the resulting confidence interval will meet the designated precision requirements. Hence, it is of both practical and theoretical importance to develop sample size procedures for precise interval estimation of the standardized mean difference.

It follows from the suggested approach that an approximate $100(1-\alpha)\%$ two-sided confidence interval $(\widehat{\delta}_L^*, \widehat{\delta}_U^*)$ of δ^* can be obtained from Equations 5 and 6 with equal tail confidence probability, $\alpha_1 = \alpha_2 = \alpha/2$. Unlike the common confidence intervals constructed with the standard pivotal method, the interval $(\widehat{\delta}_L^*, \widehat{\delta}_U^*)$ does not have an explicit analytic form, and the width of a confidence interval $(\widehat{\delta}_L^*, \widehat{\delta}_U^*)$, denoted by $W = \widehat{\delta}_U^* - \widehat{\delta}_L^*$, cannot be expressed as a multiple of the estimated standard deviation. Instead, the confidence limits $\widehat{\delta}_L^*$ and $\widehat{\delta}_U^*$

Table 5 Simulated coverage probability and error of the approximate confidence intervals for standardized mean difference effect size when $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $N_1 = 20$, and $N_2 = 10$

δ*	The proposed approach						Kulinskaya and Staudte (2007)					
	Upper 95% CI	Error	Lower 95% CI	Error	Two-sided 90% CI	Error	Upper 95% CI	Error	Lower 95% CI	Error	Two-sided 90% CI	Error
0	0.9469	−0.0031	0.9539	0.0039	0.9008	0.0008	0.9517	0.0017	0.9579	0.0079	0.9096	0.0096
1	0.9408	−0.0092	0.9511	0.0011	0.8919	−0.0081	0.9753	0.0253	0.9172	−0.0328	0.8925	−0.0075
2	0.9347	−0.0153	0.9535	0.0035	0.8882	−0.0118	0.9884	0.0384	0.7176	−0.2324	0.7060	−0.1940
3	0.9285	−0.0215	0.9522	0.0022	0.8807	−0.0193	0.9894	0.0394	0.4999	−0.4501	0.4893	−0.4107
	Upper 97.5% CI	Error	Lower 97.5% CI	Error	Two-sided 95% CI	Error	Upper 97.5% CI	Error	Lower 97.5% CI	Error	Two-sided 95% CI	Error
0	0.9727	−0.0023	0.9750	0.0000	0.9477	−0.0023	0.9739	−0.0011	0.9767	0.0017	0.9506	0.0006
1	0.9663	−0.0087	0.9760	0.0010	0.9423	−0.0077	0.9859	0.0109	0.9649	−0.0101	0.9508	0.0008
2	0.9611	−0.0139	0.9780	0.0030	0.9391	−0.0109	0.9929	0.0179	0.8393	−0.1357	0.8322	−0.1178
3	0.9543	−0.0207	0.9775	0.0025	0.9318	−0.0182	0.9923	0.0173	0.6406	−0.3344	0.6329	−0.3171

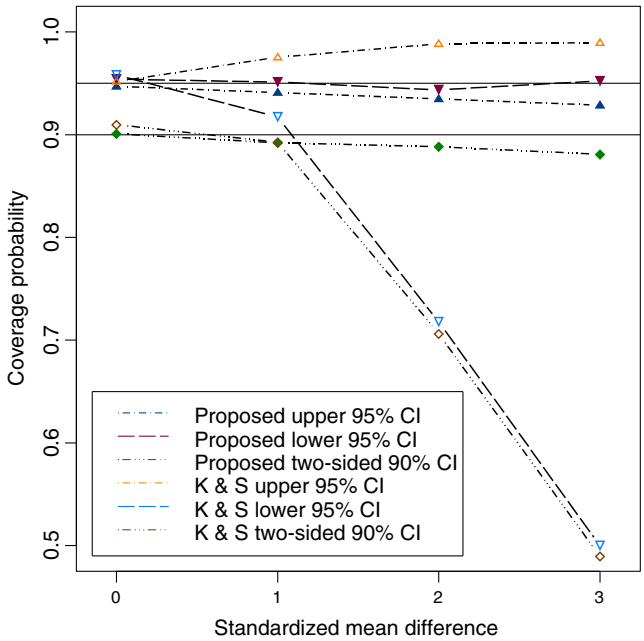

Fig. 1 Simulated coverage probability of the proposed and Kulinskaya and Staudte's (2007) confidence intervals

are derived by the inversion confidence interval principle with the observed value of Welch's statistic V and estimated degrees of freedom $\hat{\nu}$. Although there is no convenient closed-form expression for the interval $(\hat{\delta}_L^*, \hat{\delta}_U^*,)$, the confidence limits $\hat{\delta}_L^*$ and $\hat{\delta}_U^*$ and the width W still depend on the statistic V, variance estimates S_1^2 and S_2^2, and sample sizes N_1 and N_2, as well as the confidence coefficient $1 - \alpha$. To ensure that the confidence interval is narrow enough to produce meaningful findings, when planning a study, researchers must consider the stochastic nature of confidence intervals. Consequently, the inherent randomness in Welch's statistic V and degrees of freedom $\hat{\nu}$ should be considered in determining the sample sizes required to achieve the specified precision properties of a confidence interval.

Two useful principles concerning the control of the expected width and the assurance probability of the width within a preassigned value are considered here. First, it is necessary to determine the required sample size such that the expected width $E[W]$ of a $100(1 - \alpha)\%$ confidence interval $(\hat{\delta}_L^*, \hat{\delta}_U^*)$ is within the given bound

$$E[W] \leq b, \tag{7}$$

where b (>0) is a constant. Second, one may compute the sample size needed to guarantee, with a given assurance probability, that the width W of a $100(1 - \alpha)\%$ confidence interval $(\hat{\delta}_L^*, \hat{\delta}_U^*)$ will not exceed the planned value

$$P\{W \leq \omega\} \geq 1 - \gamma, \tag{8}$$

where $(1 - \gamma)$ is the specified assurance level and ω (>0) is a constant. Although the two notions of expected width and assurance probability have been considered for sample size determination, the exact computations of $E[W]$ and $P\{W \leq \omega\}$ are more involved than those in Kelley and Rausch (2006) under a homogeneous variance framework. Naturally, the underlying exact distributional property of Welch's statistic V and estimated degrees of freedom $\hat{\nu}$ should be incorporated into the sample size calculations as much as possible. The exact distribution of Welch's test statistic V is comparatively complicated and may be expressed in different forms (see Wang, 1971; Lee and Gurland, 1975; Nel, van der Merwe, & Moser 1990 for technical derivation and related details). In order to facilitate the implementation of sample size procedures, we consider the following alternative expression of V described in Jan and Shieh (2011) for its ease of numerical computation:

$$V = T/H^{1/2},$$

where $T \sim t(N_1 + N_2 - 2, \Delta*), H = \left[(\sigma_1^2/N_1)\{B/p\} + (\sigma_2^2/N_2)\{(1-B)/(1-p)\}\right]/\sigma^2, B \sim \text{Beta}\{(N_1 - 1))/2, (N_2 - 1)/2\}$, $\sigma^2 = \sigma_1^2/N_1 + \sigma_1^2/N_2$, and $p = (N_1 - 1)/(N_1 + N_2 - 2)$. It is important to note that T and B are independent. Also,

$$\hat{\nu} = 1/\left\{B_1^2/(N_1 - 1) + B_2^2/(N_2 - 1)\right\},$$

where $B_2 = 1 - B_1$ and $B_1 = \left[(\sigma_1^2/N_1)\{B/p\}\right]/\left[(\sigma_1^2/N_1)\{B/p\} + (\sigma_2^2/N_2)\{(1-B)/(1-p)\}\right]$. Essentially, the exact distribution of V involves a Beta mixture of noncentral t distributions, and both H and $\hat{\nu}$ are functions of the Beta random variable B. It is easily seen that the confidence limits $\hat{\delta}_L^*$, $\hat{\delta}_U^*$ and are functions of V and $\hat{\nu}$, as is the interval width W. With the prescribed alternative distributional formulations for V and $\hat{\nu}$ through T and B, the interval width W is denoted by $W = W(T, B)$ to emphasize its dependence on the two random variables of T and B. Hence, the exact evaluation of $E[W]$ described in Equation 7 is performed with

$$E[W] = E_T\{E_B[W(T,B)]\}, \tag{9}$$

where the expectations E_T and E_B are taken with respect to the distribution of T and B, respectively. Likewise, the calculation of $P\{W \leq \omega\}$ presented in Equation 8 is conducted by

$$P\{W \leq \omega\} = E_T\{E_B[g(T,B)]\}, \tag{10}$$

where $g\{\cdot\}$ is an indicator function such that $g(T, B) = 1$ if $W(T, B) \leq \omega$ and $g(T, B) = 0$ if $W(T, B) > \omega$. It is interesting to note that the approximate noncentral t distribution for V given

in Equation 3 may be considered to avoid theoretical and computational complications, while maintaining methodological simplicity in the assessment of expected width and assurance probability. Our numerical computations show that this approach generally gives identical or very similar sample sizes as the exact approach. However, the corresponding simulation results show that the small difference in sample sizes can cause significant inferior performance in assurance probability. Therefore, this simplified method is not considered further in this article.

With the exact computational formulas of expected width and assurance probability given in Equations 9 and 10, the sample sizes (N_1, N_2) needed to attain the specified precision can be found by a simple iterative search for the chosen confidence level ($1 - \alpha$), parameter values (μ_1, μ_2, σ_1^2, σ_2^2), and sample size allocation ratio q_1. Actually, the task is simplified to deciding the optimal sample size N_1 required to achieve the desired precision level, because the other sample size $N_2 = N_1 (1 - q_1)/q_1$ is a function of N_1 and q_1. Accordingly, the sample sizes (N_{EW1}, N_{EW2}) needed for the expected width of a $100(1 - \alpha)\%$ confidence interval ($\widehat{\delta}_L^*$, $\widehat{\delta}_U^*$) to fall within the designated bound b are the minimum integers (N_1, N_2) such that $E[W] \leq b$. On the other hand, the sample sizes (N_{TP1}, N_{TP2}) required to guarantee with a given assurance probability ($1 - \gamma$) that the width of a $100(1 - \alpha)\%$ confidence interval ($\widehat{\delta}_L^*$, $\widehat{\delta}_U^*$) will not exceed the planned range ω are the smallest integers (N_1, N_2) such that $P\{W \leq \omega\} \geq 1 - \gamma$. The numerical computation of expected width and assurance probability requires the evaluation of the noncentrality inversion function of a noncentral t distribution and the two-dimensional integration with respect to a Beta and a noncentral t probability distribution function. To enhance the applicability of these sample size methodologies, supplementary SAS/IML (SAS Institute, 2011) computer programs have been written to aid users of the suggested techniques.

Numerical illustration of sample size calculations

In order to demonstrate the features and evaluate the performance of the proposed sample size procedures, numerical investigations are performed for precise interval estimation of the standardized mean difference. The empirical study was carried out in two stages. The first stage involved extensive sample size calculations for the two precision measures of expected width and assurance probability across a wide range of model configurations. In the second stage, a Monte Carlo simulation study was conducted to gain understanding of the precision outcome for the suggested sample size formulas under the design characteristics described in the first stage.

The determination of sample sizes needed for the chosen precision of the confidence interval procedures requires detailed specifications of the confidence level, sample size allocation structure, and the magnitudes of mean effects and variance components. It is evident that the influence of each of these components on the precision behavior not only differs, but also depends on the concurrent impact of other factors. To provide a systematic explication, the numerical trials are specified by fixing all but one of the following factors and varying that single factor in the assessment: (1) error variances, (2) sample size allocation ratio, and (3) effect size. Specifically, the appraisals include three different settings for each of the two factors of error variances and sample size allocation ratio with (σ_1^2, σ_2^2) = (1, 1) and (1, 4) and q_1 = 1/2, 1/3, and 2/3, respectively. For brevity, the setup of (σ_1^2, σ_2^2) = (1, 1) and q_1 = 2/3 is omitted from the six crossed combinations of error variances and sample size ratios as in the previous numerical study. Moreover, the population standardized mean difference effect size δ* is set to have four different values, δ* = 0, 1, 2, 3. The remaining key features corresponding to interval assurance and precision criteria are chosen as confidence level ($1 - \alpha$) = 0.95, interval bound $b = \omega = 0.5$,

Table 6 Computed sample size, expected width, and assurance probability performance for 95% two-sided confidence interval of standardized mean difference effect size δ* when interval bound $b = \omega = 0.5$, assurance probability $1 - \gamma = 0.9$, variances (σ_1^2, σ_2^2) = (1, 1), and sample size allocation ratio $q_1 = 1/2$ ($N_2/N_1 = 1$)

δ*	Expected width					Assurance probability				
	N_1	N_2	Simulated $E[W]$	Actual $E[W]$	Error	N_1	N_2	Simulated $P\{W \leq \omega\}$	Actual $P\{W \leq \omega\}$	Error
0	32	32	0.4921	0.4921	0.0000	32	32	0.9723	0.9704	0.0019
1	48	48	0.4952	0.4954	−0.0002	53	53	0.9207	0.9214	−0.0007
2	95	95	0.4976	0.4973	0.0003	105	105	0.9210	0.9186	0.0024
3	172	172	0.4990	0.4988	0.0002	187	187	0.9084	0.9079	0.0005

Table 7 Computed sample size, expected width, and assurance probability performance for 95% two-sided confidence interval of standardized mean difference effect size δ* when interval bound $b = \omega = 0.5$, assurance probability $1 - \gamma = 0.9$, variances $\left(\sigma_1^2, \sigma_2^2\right) = (1, 1)$, and sample size allocation ratio $q_1 = 1/3$ ($N_2/N_1 = 2$)

δ*	Expected width					Assurance probability				
	N_1	N_2	Simulated $E[W]$	Actual $E[W]$	Error	N_1	N_2	Simulated $P\{W \le \omega\}$	Actual $P\{W \le \omega\}$	Error
0	21	42	0.4969	0.4970	−0.0001	22	44	0.9847	0.9852	−0.0005
1	37	74	0.4963	0.4966	−0.0003	42	84	0.9278	0.9282	−0.0004
2	83	166	0.4996	0.4997	−0.0001	92	184	0.9079	0.9069	0.0010
3	160	320	0.4999	0.4997	0.0002	173	346	0.9010	0.9030	−0.0020

and assurance probability $(1 - \gamma) = 0.90$. Accordingly, the computed sample sizes (N_{EW1}, N_{EW2}) and (N_{TP1}, N_{TP2}) with respect to the selected precision requirements of expected width and assurance probability, respectively, are listed in Tables 6, 7, 8, 9 and 10 for five combined error variances and sample size allocation patterns. As expected, the sample sizes vary with the error variance and sample size allocation configurations in these tables. But it is evident from the reported results in these tables that the sample sizes increase with an increasing value of δ* when all other factors are fixed. This particular phenomenon differs from the common sample size procedures for precise interval estimation of mean differences, in which the required sample sizes do not vary with the magnitude of population mean difference, as shown in Kupper and Hafner (1989) and Wang and Kupper (1997). Although the results are not completely comparable, it typically requires a larger sample size to meet the necessary precision of assurance probability than the control of a designated expected width. Hence, the sample sizes computed by the expected width approach tend to be inadequate to guarantee the desired assurance level of interval width. Consequently, Kupper and Hafner (1989) recommended the assurance probability approach over the expected width criterion for sample size determination, although the notion of expected width is widely covered in standard texts for sample size determination of precise interval estimation. It is noteworthy that the expected width and assurance probability principles are closely related to the two distinct principles of unbiasedness and consistency in statistical point estimation. Therefore, the two measures impose unique and distinct precision characteristics on the resulting confidence intervals. Each arguably has theoretical grounds and implications in its own right. More important, researchers need to justify the appropriate precision consideration for sample size determination on the basis of their knowledge of a research area and the specific circumstances they face, rather than assume a convenient value offered by a simple guideline or rule of thumb without reference to all of the critical factors in a study design.

The results here enable researchers to better understand the essential relationship that exists between the planned sample size and interval precision conditional on the critical information of model configurations. The associated accuracy issue of the sample size methodology with respect to the precision considerations of expected width and assurance probability is considered in the following simulation study. In this case, under the computed sample sizes, parameter configurations and precision settings described in Tables 6, 7, 8, 9 and 10, estimates of the true expected width or assurance probability are computed through Monte Carlo simulation of 10,000 independent data sets. For each replicate, the confidence limits and corresponding interval width of the two-sided 95% confidence intervals are calculated. Then the simulated expected width is the

Table 8 Computed sample size, expected width, and assurance probability performance for 95% two-sided confidence interval of standardized mean difference effect size δ* when interval bound $b = \omega = 0.5$, assurance probability $1 - \gamma = 0.9$, variances $\left(\sigma_1^2, \sigma_2^2\right) = (1, 4)$, and sample size allocation ratio $q_1 = 1/2$ ($N_2/N_1 = 1$)

δ*	Expected width					Assurance probability				
	N_1	N_2	Simulated $E[W]$	Actual $E[W]$	Error	N_1	N_2	Simulated $P\{W \le \omega\}$	Actual $P\{W \le \omega\}$	Error
0	32	32	0.4927	0.4927	0.0000	32	32	0.9466	0.9467	−0.0001
1	53	53	0.4971	0.4974	−0.0003	59	59	0.8964	0.9022	−0.0058
2	116	116	0.4991	0.4990	0.0001	128	128	0.9023	0.9028	−0.0005
3	221	221	0.4993	0.4990	0.0003	239	239	0.9070	0.9089	−0.0019

Table 9 Computed sample size, expected width, and assurance probability performance for 95% two-sided confidence interval of standardized mean difference effect size δ* when interval bound $b = \omega = 0.5$, assurance probability $1 - \gamma = 0.9$, variances $\left(\sigma_1^2, \sigma_2^2\right) = (1, 4)$, and sample size allocation ratio $q_1 = 1/3$ ($N_2/N_1 = 2$)

δ*	Expected width					Assurance probability				
	N_1	N_2	Simulated $E[W]$	Actual $E[W]$	Error	N_1	N_2	Simulated $P\{W \le \omega\}$	Actual $P\{W \le \omega\}$	Error
0	21	42	0.4960	0.4960	0.0000	21	42	0.9120	0.9041	0.0079
1	32	64	0.4953	0.4955	−0.0002	35	70	0.8983	0.9035	−0.0052
2	63	126	0.4989	0.4987	0.0002	70	140	0.9168	0.9171	−0.0003
3	115	230	0.4983	0.4981	0.0002	125	250	0.9125	0.9185	−0.0060

mean of the 10,000 replicates of interval widths, whereas the simulated assurance probability is the proportion of the 10,000 replicates whose values of interval width are less than or equal to the specified bound $\omega = 0.5$. Due to the underlying metric of integer sample sizes, the achieved precision levels associated with the presented sample sizes (N_{EW1}, N_{EW2}) and (N_{TP1}, N_{TP2}) should be less than or greater than the nominal level for width bound $b = 0.5$ and assurance probability $(1 - \gamma) = 0.90$, respectively. The differences between the actual precision performance and the target level are more pronounced for the cases of $\delta^* = 0$ with comparatively small sample sizes. In order to provide a rigorous assessment, the exact values of the actual expected width and the actual assurance probability are also calculated on the basis of Equations 7 and 8, respectively. The adequacy of the sample size procedure for precise interval estimation is determined by one of the following formulas: error = simulated expected width − actual expected width or error = simulated assurance probability − actual assurance probability. Both the simulated and actual values of expected width and assurance probability, along with the associated errors, are summarized in Tables 6, 7, 8, 9 and 10 as well. It can be seen from the results that the performance of the proposed methods appears to be remarkably good for the range of model specifications considered here. Specifically, the absolute errors of the expected width are less than 0.001 for the 20 cases examined here. On the other hand, all the absolute discrepancies in assurance probability are smaller than 0.01 throughout these tables. In view of these comprehensive empirical evaluations, the proposed methods are accurate enough to compute required sample sizes for precise interval estimation of standardized mean difference in practical applications.

Conclusions

The standardized mean difference with heterogeneous variances is one of the advocated effect size indices used across a variety of research disciplines. Unfortunately, the diversity of suggested measures indicates that there is no firm consensus as to the unified definition of a standardized mean difference effect size. In this article, we confine ourselves to the normal theory framework and purport to demonstrate comprehensive treatment to the effect size measure of standardized mean difference considered in Kulinskaya and Staudte (2007). Accordingly, the well-known Welch's statistic plays an important role in finding useful estimation procedures for the particular effect size. The purpose of this article is twofold. The first goal is to provide an

Table 10 Computed sample size, expected width, and assurance probability performance for 95% two-sided confidence interval of standardized mean difference effect size δ* when interval bound $b = \omega = 0.5$, assurance probability $1 - \gamma = 0.9$, variances $\left(\sigma_1^2, \sigma_2^2\right) = (1, 4)$, and sample size allocation ratio $q_1 = 2/3$ ($N_2/N_1 = 1/2$)

δ*	Expected width					Assurance probability				
	N_1	N_2	Simulated $E[W]$	Actual $E[W]$	Error	N_1	N_2	Simulated $P\{W \le \omega\}$	Actual $P\{W \le \omega\}$	Error
0	42	21	0.4989	0.4989	0.0000	44	22	0.9463	0.9465	−0.0002
1	92	46	0.4995	0.4995	0.0000	108	54	0.9195	0.9239	−0.0044
2	240	120	0.4994	0.4992	0.0002	266	133	0.9016	0.9016	0.0000
3	486	243	0.4991	0.4990	0.0001	526	263	0.9169	0.9139	0.0030

interval estimation procedure more efficient than the current variance stabilizing transformation method of Kulinskaya and Staudte (2007). The suggested approach utilizes the convenient noncentrality inversion technique for constructing the confidence limits, and it has the advantages of overall accuracy, methodological transparency, and computational ease. The second objective of this report is to study the corresponding sample size determination problem for precise interval estimation in advance research planning. Despite the underlying distributional complexity, the exact statistical properties of Welch's statistic are analytically described and computationally employed in the developed algorithms for accurate sample size calculations. The presented procedures provide a feasible solution for determining optimal sample sizes to ensure the desired precision of expected width and assurance probability when critical information of the model configurations is available. It is hoped that the proposed interval estimation and sample size procedures will increase the practice of reporting confidence intervals of the standardized mean difference effect size in future research practice.

Author Note The author thanks the editor, Gregory Francis, and the two anonymous reviewers for their helpful comments.

References

Alhija, F. N. A., & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement, 69,* 245–265.

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

American Educational Research Association Task Force on Reporting of Research Methods. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Research, 35,* 33–40.

Breaugh, J. A. (2003). Effect size estimation: Factors to consider and mistakes to avoid. *Journal of Management, 29,* 79–97.

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

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology, 34,* 917–928.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practices, 40,* 532–538.

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research, 23,* 89–105.

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

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 10,* 3–8.

Glass, G. V., McGraw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Newbury Park, CA: Sage.

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods, 6,* 135–146.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and Multivariate Applications* (2nd ed.). New York, NY: Routledge.

Huberty, C. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62,* 227–240.

Jan, S. L., & Shieh, G. (2011). Optimal sample sizes for Welch's test under various allocation and cost considerations. *Behavior Research Methods, 43,* 1014–1022.

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.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13,* 110–129.

Kim, S. H., & Cohen, A. S. (1998). On the Behrens-Fisher problem: A review. *Journal of Educational and Behavioral Statistics, 23,* 356–377.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kulinskaya, E., & Staudte, R. G. (2006). Interval estimates of weighted effect sizes in the one-way heteroscedastic ANOVA. *British Journal of Mathematical and Statistical Psychology, 59,* 97–111.

Kulinskaya, E., & Staudte, R. G. (2007). Confidence intervals for the standardized effect arising in the comparison of two normal populations. *Statistics in Medicine, 26,* 2853–2871.

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

Lee, A. F. S., & Gurland, J. (1975). Size and power of tests for equality of means of two normal populations with unequal variances. *Journal of the American Statistical Association, 70,* 933–941.

Nel, D. G., van der Merwe, C. A., & Moser, B. K. (1990). The exact distribution of the univariate and multivariate Behrens-Fisher statistics with a comparison of several solutions in the univariate case. *Communications in Statistics-Theory and Methods, 19,* 279–298.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25,* 241–286.

Richardson, J. T. E. (1996). Measures of effect size. *Behavior Research Methods, Instruments, & Computers, 28,* 12–22.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect size in behavioral research: A correlational approach*. New York: Cambridge University Press.

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology, 57,* 221–237.

SAS Institute. (2011). *SAS/IML User's Guide, Version 9.3*. Cary, NC: SAS Institute Inc.

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

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 222–257). Mahwah, NJ: Lawrence Erlbaum.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51,* 473–481.

Wang, Y. Y. (1971). Probabilities of the type I errors of the Welch tests for the Behrens-Fisher problem. *Journal of the American Statistical Association, 66,* 605–608.

Wang, Y., & Kupper, L. L. (1997). Optimal sample sizes for estimating the difference in means between two normal populations treating confidence interval length as a random variable. *Commemorations in Statistics-Theory and Methods, 26,* 727–741.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29,* 350–362.

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