Behav Res (2012) 44:202–212
DOI 10.3758/s13428-011-0139-z

Optimal sample sizes for precise interval estimation of Welch's procedure under various allocation and cost considerations

Gwowen Shieh · Show-Li Jan

Published online: 30 July 2011

Abstract Welch's (*Biometrika* 29: 350–362, 1938) procedure has emerged as a robust alternative to the Student's *t* test for comparing the means of two normal populations with unknown and possibly unequal variances. To facilitate the advocated statistical practice of confidence intervals and further improve the potential applicability of Welch's procedure, in the present article, we consider exact approaches to optimize sample size determinations for precise interval estimation of the difference between two means under various allocation and cost considerations. The desired precision of a confidence interval is assessed with respect to the control of expected half-width, and to the assurance probability of interval half-width within a designated value. Furthermore, the design schemes in terms of participant allocation and cost constraints include (a) giving the ratio of group sizes, (b) specifying one sample size, (c) attaining maximum precision performance for a fixed cost, and (d) meeting a specified precision level for the least cost. The proposed methods provide useful alternatives to the conventional sample size procedures. Also, the developed programs expand the degree of generality for the existing statistical software packages and can be accessed at brm.psychonomic-journals.org/content/ supplemental.

Keywords Behrens-Fisher problem · Precision · Study design

Introduction

The fundamental results and associated usages of standard parametric procedures—such as Student's *t*, ANOVA *F*, and ordinary least squares regression—are well documented in the literature. One important assumption underlying the prescribed traditional methods is that of equal population variances. Although the homogeneity of variance formulation provides a convenient and useful setup, it is not unusual for the homoscedasticity assumption to be violated in actual applications. For example, Grissom (2000) emphasized that there are theoretical reasons to expect and empirical results to document the existence of heteroscedasticity in clinical data. Moreover, Grissom and Kim (2005, pp. 10–14) provided additional explanations for the intrinsic causes of variance heterogeneity in real data. Notably, Grissom recommended employing suitable techniques that are superior to the traditional inferential methods under various conditions of heteroscedasticity.

For comparing the difference between two normal means that may have unequal population variance, the scenario is the well-known Behrens–Fisher problem (Kim & Cohen, 1998). Accordingly, Welch's (1938) approximate *t* procedure has been recognized as a satisfactory and robust solution over the two-sample *t* of the Behrens–Fisher problem. The same notion was independently suggested by Smith (1936) and Satterthwaite (1946); hence, the technique is sometimes referred to as the Smith–Welch–Satterthwaite procedure. The method not only is covered in

Electronic supplementary material The online version of this article (doi:10.3758/s13428-011-0139-z) 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

S.-L. Jan
Department of Applied Mathematics,
Chung Yuan Christian University,
Chungli, Taiwan 32023, Republic of China
e-mail: sljan@math.cycu.edu.tw

introductory textbooks of statistics and quantitative methods but also is available in several commonly used statistics packages—for example, Excel, Minitab, SAS, and SPSS. However, most research in this area is concerned with the null hypothesis significance tests for detecting mean differences—for example, Best and Rayner (1987) and Wang (1971). This dominance of hypothesis testing for making statistical inferences does not occur exclusively in the Behrens–Fisher problem. It more broadly reflects the longstanding and prevalent practice of significance tests in applied research across many scientific fields. As a compelling alternative, there has been a growing awareness in the use of confidence intervals instead of hypothesis tests for inference-making purposes, such as Hahn and Meeker (1991), Harlow, Mulaik, and Steiger (1997), Kline (2004), and Smithson (2003). But from both practical and scientific standpoints, it may be more informative to provide a reliable estimate of the magnitude of the examined effect, rather than simply to decide whether or not a finding is statistically significant. Accordingly, Wilkinson and the American Psychological Association's Task Force on Statistical Inference (1999) and the sixth edition of the *Publication Manual of the American Psychological Association* (APA, 2010) called for the greater use of confidence intervals. However, the interval estimation procedures are intrinsically stochastic in nature. From a study-planning point of view, researchers may wish to credibly address specific research questions and confirm meaningful treatment differences, so that the resulting confidence interval will meet the designated precision requirements. Hence, it is of practical interest and methodological importance to develop sample size procedures for precise interval estimation in the context of the Behrens–Fisher problem.

To ensure precision of the resulting confidence intervals, the notion of expected half-width for sample size calculations is frequently introduced in standard texts. However, considerable attention has focused on the criterion of tolerance probability of interval half-width within a given value. For example, see Beal (1989), Kelley, Maxwell, and Rausch (2003), Kupper and Hafner (1989), and Liu (2009) for related discussion in the context of estimating the mean difference between two normal populations with homoscedasticity. The empirical illustration in Kupper and Hafner shows that it typically requires a larger sample size to meet the necessary assurance of tolerance probability than the control of a designated expected half-width. Therefore, the sample sizes computed by the expected half-width approach tend to be inadequate to guarantee the desired tolerance level of interval half-width. Consequently, the assurance probability approach is recommended over the expected width criterion for sample size determination. However, it is noteworthy that the two principles of expected width and assurance probability are closely related to the two standard criteria of unbiasedness and consistency in statistical point estimation, respectively. In other words, these two measures impose unique and distinct aspects of precision characteristics on the resulting confidence intervals, and each principle has conceptual and empirical implications in its own right.

Within the framework of the Behrens–Fisher problem, Wang and Kupper (1997) derived a formula to compute the necessary sample size for a selected tolerance probability when the sample size ratio is given. Although the suggested sample size technique accommodates the more realistic situation of variance heterogeneity, three essential caveats of the results in Wang and Kupper should be pointed out. First, their theoretical presentations and algebraic expressions are noticeably awkward. The formulation is complicated in form, and the complexity requires intensive cumbersome evaluations. Furthermore, to our knowledge, there is no computer algorithm available for performing the necessary computation. Therefore, their result is of less practical value in application. Second, they suggest fixing the proportion of standard deviations as the allocation ratio to determine the optimal sample sizes for a designated tolerance level so that the total sample size is minimized. But the simplified algorithm employed by Wang and Kupper fails to take into account the underlying metric of integer sample sizes and often leads to suboptimal results. It is shown below in our numerical investigation that their procedure is not guaranteed to give the correct optimal sample sizes. Third, although there are mixed opinions on the effectiveness of expected width, they did not address the issue of how to perform the sample size calculations so that the expected confidence interval half-width will attain the planned precision. Thus, the results in Wang and Kupper should be clarified and extended with more transparent explications and exact computations. Note that the assurance probability for achieving a desired interval width can be further modified as a conditional probability that the confidence interval includes the true parameter. As was reported in Beal (1989), corresponding sample sizes computed with the conditional consideration are almost identical to or at most only slightly larger than those calculated with the aforementioned unconditional or tolerance probability approach. Nonetheless, our calculations also confirm that this phenomenon continues to exist in the Behrens–Fisher problem. Hence, the conditional criterion presented in Wang and Kupper will not be considered further in this article.

In view of the potential variance heterogeneity one might encounter in applied work, the present article contributes to the applications of Welch's (1938) procedure by providing feasible sample size methodology for constructing precise confidence intervals under two distinct perspectives. One method gives the minimum sample size such that the

expected confidence interval half-width is within the designated bound. The other approach provides the sample size needed to guarantee, with a given tolerance probability, that the half-width of a confidence interval will not exceed the planned value. Furthermore, conventional sample size calculations do not consider allocation schemes with participant constraints or cost implications. However, researchers have explored design strategies that take into account the impact of different constraints of the sample scheme and project funding while maintaining adequate power (Allison, Allison, Faith, Paultre, & Pi-Sunyer, 1997, and references therein). Jan and Shieh (2011) considered the problem of determining optimal sample sizes to meet a designated power for Welch's test under various allocation and cost considerations that call for independent random samples from two normal populations with possibly unequal variances. The same principles would apply for a study seeking a precise estimate of the mean difference between two treatments. It is well known that there exists a direct connection between hypothesis testing and interval estimation, although the two procedures are philosophically different in the power and precision viewpoints. Not surprisingly, the sample size required to test a hypothesis regarding the specific value of a parameter with desired power can be markedly different from the sample size needed to obtain adequate precision of interval estimation in the same study. Since there are crucial and useful tactics for study design other than the minimization of total sample size, it is prudent to present a comprehensive account of design configurations in terms of various participant and budget constraints. In this article, exact methods are presented to give proper sample sizes when either the ratio of group sizes is fixed in advance or one sample size is fixed. In addition, detailed procedures are provided to determine the optimal sample sizes to maximize the precision for a given total cost and to minimize the cost for a specified precision. Finally, corresponding SAS computer codes are developed to facilitate computations of the exact necessary sample size in actual applications.

Precise interval estimation

In line with the advocated practice of greater use of confidence intervals, we attempt to develop the sample size methodology under precision consideration for Welch's (1938) approximate t procedure in the context of the Behrens–Fisher problem. Consider independent random samples from two normal populations with the following formulations:

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

where μ_1, μ_2, σ_1^2, σ_2^2 are unknown parameters, $j = 1, \ldots, N_i$, and $i = 1$ and 2. To detect the difference between two group means, the well-known Welch's pivotal quantity is of the form

$$V = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\left(S_1^2/N_1 + S_2^2/N_2\right)^{1/2}},$$

where $\overline{X}_1 = \sum_{j=1}^{N_1} X_{1j}/N_1$, $\overline{X}_2 = \sum_{j=1}^{N_2} X_{2j}/N_2, S_1^2 = \sum_{j=1}^{N_1} (X_{1j} - \overline{X}_1)^2/(N_1 - 1)$ and $S_2^2 = \sum_{j=1}^{N_2} \left(X_{2j} - \overline{X}_2\right)^2/(N_2 - 1)$. Accordingly, Welch proposed the approximate distribution for V:

$$V \dot{\sim} t\left(\hat{\nu}\right), \tag{1}$$

where $t\left(\hat{\nu}\right)$ is the t distribution with degrees of freedom $\hat{\nu}$ and $\hat{\nu} = \hat{\nu}\left(N_1, N_2, S_1^2, S_2^2\right)$ with

$$1/\hat{\nu} = \frac{1}{N_1 - 1}\left\{\frac{S_1^2/N_1}{S_1^2/N_1 + S_2^2/N_2}\right\}^2 + \frac{1}{N_2 - 1} \times \left\{\frac{S_2^2/N2}{S_1^2/N_1 + S_2^2/N_2}\right\}^2.$$

Thus, an approximate $100(1 - \alpha)\%$ two-sided confidence interval of mean difference $(\mu_1 - \mu_2)$ is of the form (L, U), where $L = \left(\overline{X}_1 - \overline{X}_2\right) - t_{\hat{\nu},\alpha/2}\left(S_1^2/N_1 + S_2^2/N_2\right)^{1/2}$, $U = \left(\overline{X}_1 - \overline{X}_2\right) + t_{\hat{\nu},\alpha/2}\left(S_1^2/N_1 + S_2^2/N_2\right)^{1/2}$, and $t_{\hat{\nu},\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the t distribution $t\left(\hat{\nu}\right)$ with degrees of freedom $\hat{\nu}$. For ease of presentation, the half-width of the $100(1 - \alpha)\%$ two-sided confidence interval is denoted by

$$H = t_{\hat{\nu},\alpha/2}\left(S_1^2/N_1 + S_2^2/N_2\right)^{1/2} \tag{2}$$

It is clear that the actual half-width H depends on the sample sizes N_1 and N_2, the confidence coefficient $1 - \alpha$, as well as on variance estimates S_1^2 and S_2^2. More importantly, both S_1^2 and S_2^2 are scaled chi-square random variables with degrees of freedom $(N_1 - 1)$ and $(N_2 - 1)$, respectively, and thus jointly determine the distributional feature of the half-width H of a confidence interval. When planning a study for ensuring that the confidence interval is narrow enough to produce meaningful findings, researchers must consider the stochastic nature of sample variances.

For the purpose of advanced research design, it is desirable to determine the sample sizes required to achieve the designated precision properties of a confidence interval. Two useful principles concern the control of the expected half-width and the tolerance probability of the half-width within a preassigned value. Specifically, it is necessary to determine the required sample size such that the expected

half-width of a 100(1 − α)% confidence interval is within the given bound

$$E[H] = \delta, \tag{3}$$

where the expectation $E[H]$ is taken with respect to the joint distribution of S_1^2 and S_2^2, and δ (> 0) is a constant. On the other hand, one may compute the sample size needed to guarantee, with a given tolerance probability, that the half-width of a 100(1 − α)% confidence interval will not exceed the planned value

$$P\{H < \omega\} = 1 - \gamma, \tag{4}$$

where (1 − γ) is the specified tolerance level, and ω (> 0) is a constant.

To simplify presentation and computation, the following alternative formulation for H is derived:

$$H = t_{\hat{\nu},\alpha/2}(K \cdot G/\kappa)^{1/2} \tag{5}$$

where $\kappa = N_1 + N_2 - 2, K = (N_1 - 1)S_1^2/\sigma_1^2 + (N_2 - 1)S_2^2/\sigma_2^2 \sim \chi^2(\kappa), G = \left[\left(\sigma_1^2/N_1\right)\{B/p\} + \left(\sigma_2^2/N_2\right)\{(1-B)/(1-p)\}\right]$, $p = (N_1 - 1)/\kappa$, and $B = \left\{(N_1 - 1)S_1^2/\sigma_1^2\right\}/K \sim \text{Beta}\{(N_1 - 1)/2, (N_2 - 1)/2\}$. Note that the random variables K and B are independent. Also, it can be shown that

$$1/\hat{\nu} = B_1^2/(N_1 - 1) + B_2^2/(N_2 - 1) \tag{6}$$

where $B_2 = 1 - B_1$ and $B_1 = \left[\left(\sigma_1^2/N_1\right)\{B/p\}\right]/\left[\left(\sigma_1^2/N_1\right)\{B/p\} + \left(\sigma_2^2/N_2\right)\{(1-B)/(1-p)\}\right]$. Hence, both G and $\hat{\nu}$ are functions of the random variable B.

It is clear from the distinct formulations in Eqs. 2 and 5 that the underlying core distribution of H transforms from the joint distribution of two independent chi-square random variables to the joint distribution of a chi-square random variable K and a beta random variable B. The suggested transformation appears at first sight to be of not much use, but actually it greatly simplifies our analytical and computational illustrations. Note that the product form of a chi-square random variable K and other terms associated with a beta random variable B in Eq. 5 permit more transparent representations than those presented in Wang and Kupper (1997). Moreover, a beta distribution is bounded by 0 and 1, and requires less computational effort than a chi-square distribution. Therefore, the numerical computation of exact values of $E[H]$ and $P\{H < \omega\}$ can be conducted with the evaluations of both the one-dimensional integration with respect to a beta probability distribution function, and the cumulative distribution function of a chi-square random variable. Since all related functions are readily available in major statistical packages, the exact computations can be performed with current computing capabilities.

In order to permit a practical treatment of sample size planning, additional concerns are considered to accommodate the participant and cost constraints in practical situations. In the next two sections, we will synthesize the ideas of Jan and Shieh (2011) and Kupper and Hafner (1989) to develop exact procedures of precise interval estimation with four different design and budget settings under the expected width and tolerance probability considerations, respectively. All calculations are performed using programs written with SAS/IML (SAS Institute, 2008a), and they are available in the supplementary files.

Expected width consideration

With the distributional properties described in Eqs. 5 and 6, the assessment of expected half-width $E[H]$ in Eq. 3 can be simplified as

$$E[H] = E_K\left[K^{1/2}\right] \cdot E_B\left[t_{\hat{\nu},\alpha/2} \cdot G^{1/2}\right]/\kappa^{1/2}. \tag{7}$$

It follows from the standard result of a chi-square distribution with κ degrees of freedom that $E_K\left[K^{1/2}\right] = 2^{1/2} \cdot \Gamma\{(\kappa+1)/2\}/\Gamma\{\kappa/2\}$. Moreover, the expectation $E_B\left[t_{\hat{\nu},\alpha/2} \cdot G^{1/2}\right]$ is taken with respect to the distribution of B and does not permit a closed-form expression. Although the expected width can still be numerically evaluated for all proper model configurations, it is prudent to focus on those with significant implications. To simplify the exposition, the following two allocation constraints are considered because of their potential usefulness. First, the ratio $r = N_2/N_1$ between the two group sizes may be fixed in advance, so the goal is to find the minimum sample size N_1 ($N_2 = rN_1$) required to achieve the selected precision level. Second, one of the two sample sizes, say, N_2, may be determined in advance, so the smallest size N_1 required to satisfy the specified precision should be determined.

Sample size ratio is fixed

Consider that the sample size ratio $r = N_2/N_1$ is preassigned, and without loss of generality, the ratio is assumed as $r \geq 1$. Thus, for a specified precision δ, a simple incremental search can be conducted to find the minimum sample size N_1 such that $E[H] \leq \delta$ for the chosen confidence level (1 − α) and error variances (σ_1^2, σ_2^2). Note that the expected half-width is asymptotically equivalent to $E[H] \doteq z_{\alpha/2}\left(\sigma_1^2/N_1 + \sigma_2^2/N_2\right)^{1/2}$, where $z_{\alpha/2}$ is the upper 100(α/2)th percentile of the standard normal distribution. The particular result provides a convenient initial value for N_1. Accordingly, it is more efficient to start the computation process with the sample size N_{1Z}, which is the smallest integer that satisfies the inequality

$$N_{1z} \geq z_{\alpha/2}^2\left(\sigma_1^2 + \sigma_2^2/r\right)/\delta^2. \tag{8}$$

For demonstration, when $\delta = 0.5$ and $1 - \alpha = 0.95$, the sample sizes N_1 and $N_2 = r{\cdot}N_1$ are presented in Table 1 for selected values of $r = 1$, 2, and 3; $\sigma_1 = 1/3$, 1/2, 1, 2 and 3; and $\sigma_2 = 1$. The actual expected half-width $E[H]$ is also listed, and the values are slightly less than the nominal value of 0.5.

One sample size is fixed

Assume the sample size N_2 of the second group is held constant, and that it is desirable to find the proper sample size N_1 to achieve the selected precision in terms of expected half-width. Just as in the previous case, the minimum sample size N_1 needed to ensure confidence intervals with the specified expected half-width δ can be found by a simple iterative search for the chosen confidence level $(1 - \alpha)$ and parameter values (σ_1^2, σ_2^2). In this case, the starting sample size N_{1Z}, based on the asymptotic approximation, is the smallest integer that satisfies the inequality

$$N_{1z} \geq \sigma_1^2 / \left\{ \left(\delta / z_{\alpha/2}\right)^2 - \sigma_2^2 / N_2 \right\}. \tag{9}$$

Note that the chosen sample size N_2 should not be too small because it is problematic to consider a small $N_2 < \sigma_2^2 / \left(\delta / z_{\alpha/2}\right)^2$ since the initial value N_{1Z} and resulting N_1 may be negative. In addition, it should be noted the resulting N_{1Z} and N_1 values are unbounded and impractical if one considers a value of $N_2 \doteq \sigma_2^2 / \left(\delta / z_{\alpha/2}\right)^2$. Accordingly, Table 2 presents the computed sample size N_1 and the actual expected half-width with chosen value N_2 for the same settings with $\delta = 0.5$, $1 - \alpha = 0.95$, and the five standard deviation settings of σ_1 and σ_2 in Table 1.

In addition to the prescribed allocation constraints of participants, it is often sensible to consider cost and effectiveness issues when research funding is limited. Moreover, the costs of obtaining subjects may differ across the two groups. Suppose c_1 and c_2 are the costs per subject in the first and second groups, respectively; then, the total cost of the study is $C = c_1 N_1 + c_2 N_2$. Thus, the following two questions arise naturally in choosing the optimal sample sizes. First, how can the maximum precision be achieved in a study with a limited budget? Second, what is the least cost for an investigation to maintain its desired level of precision? In general, balanced group sizes do not necessarily yield the optimal solution in the aforementioned two scenarios. This assertion can be easily justified from the simplified asymptotic approximation of $E[H] \doteq z_{\alpha/2}\left(\sigma_1^2/N_1 + \sigma_2^2/N_2\right)^{1/2}$, that the optimal sample size allocation ratio for the appraisals of cost and precision is

$$\frac{N_2}{N_1} = \theta, \tag{10}$$

where $\theta = \sigma_2 c_1^{1/2} / \left(\sigma_1 c_2^{1/2}\right)$. Although this identity reveals the obvious disadvantage of a naive, balanced design, it has its own weakness as a rule of thumb. It is readily seen from Eq. 7 that the exact properties of the expected half-width depend on the joint distribution of a chi-square random variable K and a beta random variable B. The resulting behavior of $E[H]$ for finite sample sizes can be notably different from that of asymptotic theory. Hence, the simple guideline of Eq. 10 does not guarantee an optimal result when the sample sizes are small. Instead, the identity is employed as a benchmark in the following detailed and systematic presentation of optimal sample size allocation.

Total cost is fixed and expected width needs to be minimized

It can be shown under a fixed value of total cost $C = c_1 N_{1Z} + c_2 N_{2Z}$ and $N_{2Z}/N_{1Z} = \theta$ that the resulting sample sizes are

$$N_{1Z} = \frac{C\left(\sigma_1 c_2^{1/2}\right)}{c_1\left(\sigma_1 c_2^{1/2}\right) + c_2\left(\sigma_2 c_1^{1/2}\right)} \text{ and} \tag{11}$$

$$N_{2z} = \frac{C\left(\sigma_2 c_1^{1/2}\right)}{c_1\left(\sigma_1 c_2^{1/2}\right) + c_2\left(\sigma_2 c_1^{1/2}\right)}.$$

Table 1 Computed sample sizes (N_1, N_2) and expected half-width $E[H]$ when sample size ratio $r = N_2/N_1$ is fixed with $\delta = 0.5$ and $1 - \alpha = 0.95$

$\sigma_1{:}\sigma_2$															
r	1/3:1			1/2:1			1:1			2:1			3:1		
	N_1	N_2	$E[H]$	N_1	N_2	$E[H]$	N_1	N_2	$E[H]$	N_1	N_2	$E[H]$	N_1	N_2	$E[H]$
1	19	19	.4959	21	21	.4947	32	32	.4980	79	79	.4973	156	156	.4988
2	11	22	.4788	13	26	.4843	25	50	.4901	71	142	.4989	148	296	.4995
3	8	24	.4788	10	30	.4897	22	66	.4958	69	207	.4972	146	438	.4986

Table 2 Computed sample sizes (N_1, N_2) and expected half-width $E[H]$ when sample size N_2 is fixed with $\delta = 0.5$ and $1 - \alpha = 0.95$

σ_1:σ_2															
1/3:1			1/2:1			1:1			2:1			3:1			
N_1	N_2	$E[H]$	N_1	N_2	$E[H]$	N_1	N_2	$E[H]$	N_1	N_2	$E[H]$	N_1	N_2	$E[H]$	
7	24	.4888	12	25	.4982	27	40	.4970	78	80	.4993	166	100	.4989	
6	27	.4831	10	30	.4897	23	60	.4927	71	140	.4993	152	200	.4994	
5	30	.4910	9	35	.4843	21	80	.4958	69	200	.4978	148	300	.4993	

As was described previously, although this sample size combination minimizes the magnitude $\left(\sigma_1^2/N_1 + \sigma_2^2/N_2\right)^{1/2}$ or asymptotic expected half-width $z_{a/2}\left(\sigma_1^2/N_1 + \sigma_2^2/N_2\right)^{1/2}$, it may be suboptimal with respect to the actual precision level $E[H]$. In practice, the sample sizes need to be integers, and it is unlikely that the values of N_{1Z} and N_{2Z} in Eq. 11 are actually whole numbers. Consequently, any sample size adjustment or rounded numbers made on N_{1Z} and N_{2Z} will introduce further inexactness into the optimization analysis. To find the exact solution, a detailed precision calculation and comparison is performed for the sample size combinations with N_1 from N_{1min} to N_{1max} and $N_2 = Floor\{(C - c_1N_1)/c_2\}$, where $N_{1min} = Max\{Floor(N_{1Z}) - 10, 5\}$, $N_{1max} = Ceil\{(C - c_2N_{2min})/c_1\}$, $N_{2min} = Max\{Floor(N_{2Z}) - 10, 5\}$, the function $Floor(a)$ returns the largest integer that is less than or equal to a, and $Ceil(a)$ returns the smallest integer that is greater than or equal to a. Note that the constants of 10 and 5 are chosen to prevent computation error and to ensure that an optimal solution is covered. Thus, the optimal sample size allocation is the one giving the maximum precision or minimum expected half-width. For illustration, numerical results are presented in Table 3 for $(c_1, c_2) = (1, 1)$, (1, 2), and (1, 3), and fixed total cost C = 30, 40, 60, 150, and 240 in accordance with the standard deviation combinations reported in the previous two tables. The results in Table 3 reveal that the actual expected half-width for a given total cost increases considerably as the unit cost c_2 increases from 1 to 3. Furthermore, the simplified allocation scheme does not yield the optimal sample sizes in several cases. For example, the optimal sample sizes are $N_1 = 24$ and $N_2 = 18$ for $(\sigma_1, \sigma_2) = (1, 1)$ and $(c_1, c_2) = (1, 2)$, in contrast with the result of $N_{1Z} = 24.8528$ and $N_{2Z} = 17.5736$ computed by Eq. 11. Correspondingly, the optimal ratio $N_2/N_1 = 18/24 = 0.7500$ is slightly greater than the ratio computed with the simple formula presented in Eq. 10: $\theta = (1 \cdot 1^{1/2})/(1 \cdot 2^{1/2}) = 0.7071$.

Target expected width is fixed and total cost needs to be minimized

In this case, the large sample approximation shows that in order to ensure the nominal expected half-width $\delta = z_{\alpha/2}\left(\sigma_1^2/N_{1z} + \sigma_2^2/N_{2z}\right)^{1/2}$ while minimizing total cost $C = c_1N_{1Z} + c_2N_{2Z}$, the best sample size combination is

$$N_{1z} = \frac{\theta\sigma_1^2 + \sigma_2^2}{\theta\left(\delta/z_{\alpha/2}\right)^2} \text{ and } N_{2z} = \frac{\theta\sigma_1^2 + \sigma_2^2}{\left(\delta/z_{\alpha/2}\right)^2}, \tag{12}$$

where θ is the optimal ratio defined in Eq. 10. Similar to the usage of sample sizes in Eq. 11, the computed values of N_{1Z} and N_{2Z} in Eq. 12 are modified to expedite a screening of sample size combinations in order to find the optimal allocation that maintains the desired expected half-width with the least cost. Specifically, the exact precision computation and cost evaluation are conducted for sample size combinations with N_1, from N_{1min} to N_{1max} satisfying the required precision, where $N_{1min} = Max[Floor(N_{1Z}) - 10, Ceil\left\{\sigma_1^2/\left(\delta/z_{\alpha/2}\right)^2\right\}, 6]$, $N_{1max} = Ceil\left[\sigma_1^2/\{(\delta/z_{\alpha/2})^2 - \sigma_2^2/N_{2\min}\}\right] + 20$, $N_{2min} = Max[Floor(N_{2Z}) - 10, Ceil\{\sigma_2^2/$

Table 3 Computed sample sizes (N_1, N_2) and expected half-width $E[H]$ when the total cost is fixed with $1 - \alpha = 0.95$

σ_1:σ_2																			
c_1:c_2	1/3:1				1/2:1				1:1				2:1				3:1			
	Cost	N_1	N_2	$E[H]$	Cost	N_1	N_2	$E[H]$	Cost	N_1	N_2	$E[H]$	Cost	N_1	N_2	$E[H]$	Cost	N_1	N_2	$E[H]$
1:1	30	8	22	.4960	40	13	27	.4779	60	30	30	.5150	150	100	50	.4833	240	180	60	.5081
1:2	30	6	12	.6726	40	10	15	.6231	60	24	18	.6285	150	88	31	.5517	240	162	39	.5615
1:3	30	6	8	.8366	40	7	11	.7497	60	21	13	.7204	150	78	24	.6052	240	150	30	.6031

$\left(\delta/z_{\alpha/2}\right)^2\}$, 6]. The constants of 6, 20, and 10 are chosen to prevent computation error and to enhance the optimal search. For each fixed value of N_1, the matching sample size N_2 is calculated to satisfy the required expected half-width. Thus, the optimal sample size allocation is the one giving the smallest cost while maintaining the specified expected half-width value. In cases in which there is more than one combination yielding the same least cost, the one producing the maximum precision is reported. Table 4 provides the corresponding optimal sample size allocation, cost, and actual expected half-width for the configurations of (c_1, c_2) = (1, 1), (1, 2), and (1, 3), and the five standard deviation settings of σ_1 and σ_2. It is clear that the total cost for a required precision and for fixed standard deviations increases substantially as the unit cost c_2 changes from 1 to 3. The optimal allocations have the simple ratio θ for the three cases of (σ_1, σ_2) = (1, 1), (2, 1), and (3, 1) when (c_1, c_2) = (1, 1). However, most of the sample size ratios are close to, but different from, the ratio θ. The largest discrepancy occurs with the case N_2/N_1 = 22/8 = 2.7500 for (c_1, c_2) = (1, 1) and (σ_1, σ_2) = (1/3, 1), whereas the approximate ratio $\theta = \left(1 \cdot 1^{1/2}\right)/\left(1/3 \cdot 1^{1/2}\right) = 3$.

Tolerance probability consideration

Instead of the expected half-width criterion, an useful alternative approach for sample size determination is to ensure that the actual confidence interval half-width will not exceed the planned bound with a given tolerance probability. For analytic clarity and computational ease, the probability $P\{H < \omega\}$ given in Eq. 4 is expressed as

$$P\{H < \omega\} = E_B\left[F_K\left\{(\kappa/G)\left(\omega/t_{\hat{\nu},\alpha/2}\right)^2\right\}\right], \qquad (13)$$

where $F_K(\cdot)$ is the cumulative density function of $K \sim \chi^2(\kappa)$. Note that the expression in Eq. 13 provides a more clear and concise exposition of the assurance probability of precision than does Eq. 14 in Wang and Kupper (1997). The formulation also expedites the subsequent computational task for various participant and cost constraints. Since there may be several possible sample sizes N_1 and N_2 that meet the required tolerance level, it is worthwhile to consider the same practical circumstances as in the case of expected interval half-width. Accordingly, the examinations presented here simplify and expand the existing and limited results in Wang and Kupper.

Sample size ratio is fixed

With the allocation ratio $r = N_2/N_1 > 1$, specified width ω, tolerance probability $(1 - \gamma)$, confidence coefficient $(1 - \alpha)$, and error variances (σ_1^2, σ_2^2), a straightforward iterative process is performed to find the minimum sample size N_1, such that $P\{H < \omega\} \leq 1 - \gamma$. To simplify the incremental search, the initial value of N_1 in the algorithm is based on Eq. 8 with $\delta = \omega$, because the optimal solutions here for large level of $(1 - \gamma)$ are greater than those of the expected interval width approach with the same interval bound. This situation is similar to those noted in Kupper and Hafner (1989) for the traditional two-sample problem. More concrete examples are presented in Table 5 for $(1 - \gamma)$ = 0.90 and ω = 0.5. For ease of comparison, the other parameter configurations of $(1 - \alpha)$, (σ_1^2, σ_2^2) and r are identical to those in Table 1. In addition to its complex formulation, the numerical calculation of Wang and Kupper (1997) is also questionable. Specifically, for the settings of ω = 0.3, $(1 - \alpha)$ = 0.95, (σ_1^2, σ_2^2) = (1, 2), and r = 1, our computations yield the optimal sample sizes $N_1 = N_2$ = 139 and $N_1 = N_2$ = 149 for $(1 - \gamma)$ = 0.80 and 0.95, respectively. The corresponding results reported in Table 1 of Wang and Kupper are $N_1 = N_2$ = 138 and $N_1 = N_2$ = 144. Note that SAS procedure PROC POWER (SAS Institute, 2008b) provides the useful feature of finding the optimal sample sizes $N_1 = N_2$ (r = 1) for the desired tolerance probability with confidence intervals of mean difference under homogeneous variances assumption. However, it does not consider the corresponding sample size calculations for the Behrens–Fisher problem with arbitrary sample size ratio $r \geq 1$, as is illustrated here.

Table 4 Computed sample sizes (N_1, N_2), cost, and expected half-width $E[H]$ when the total cost needs to be minimized with target expected half-width δ = 0.5 and 1 – α = 0.95

σ_1:σ_2																			
c_1:c_2	1/3:1				1/2:1				1:1				2:1				3:1			
	Cost	N_1	N_2	$E[H]$	Cost	N_1	N_2	$E[H]$	Cost	N_1	N_2	$E[H]$	Cost	N_1	N_2	$E[H]$	Cost	N_1	N_2	$E[H]$
1:1	30	8	22	.4960	37	12	25	.4982	64	32	32	.4980	141	94	47	.4987	248	186	62	.4998
1:2	51	9	21	.4987	60	16	22	.4998	93	37	28	.4995	182	106	38	.4999	303	205	49	.4992
1:3	72	12	20	.4966	82	19	21	.4996	120	42	26	.4984	218	116	34	.4998	348	219	43	.4996

Table 5 Computed sample sizes (N_1, N_2) and tolerance probability $P\{H < \omega\}$ when sample size ratio $r = N_2/N_1$ is fixed with $\omega = 0.5$, $1 - \gamma = 0.90$, and $1 - \alpha = 0.95$

σ_1:σ_2															
r	1/3:1			1/2:1			1:1			2:1			3:1		
	N_1	N_2	$P\{H < \omega\}$	N_1	N_2	$P\{H < \omega\}$	N_1	N_2	$P\{H < \omega\}$	N_1	N_2	$P\{H < \omega\}$	N_1	N_2	$P\{H < \omega\}$
1	26	26	.9285	27	27	.9058	39	39	.9137	91	91	.9017	176	176	.9098
2	14	28	.9406	16	32	.9246	31	62	.9310	84	168	.9048	168	336	.9009
3	10	30	.9348	13	39	.9357	28	84	.9086	82	246	.9094	166	498	.9048

One sample size is fixed

A different restriction of the design setting is to find the minimum sample size, say, N_1, that ensures a required tolerance probability when the other sample size, N_2, is fixed in advance. With the substitution of $\delta = \omega$ in Eq. 9, the resulting sample size is utilized as the starting value for the incremental search of optimal solution. The corresponding results with $(1 - \gamma) = 0.90$ and $\omega = 0.5$ are listed in Table 6 for the same configurations of $(1 - \alpha) = 0.90$, (σ_1^2, σ_2^2), and N_2 in Table 2. It is clear that the computed sample size N_1 in Table 6 is larger than that for the same setting in Table 2. Since there is no explicit low bound of N_2, it is possible that the specified N_2 is too small, and the matching N_1 may be unbounded. Thus, the iterative search of optimal N_1 is programmed to terminate when N_1 reaches the value 1,001, because the resulting sample size combination appears to be impractical or unusual.

In the following section, we will turn our attention to the budget issue with varying unit cost per subject in each group.

Total cost is fixed and tolerance probability needs to be maximized

The notion of maximizing the tolerance level with a fixed value of total cost $C = c_1N_1 + c_2N_2$ is considered, where c_1 and c_2 are the known costs for each participant of the two groups. To find the best sample size allocation, the prescribed logic and algorithm under the expected width criterion is applied to the optimization of cost and tolerance probability with the substitution of precision criterion $P\{H < \omega\}$ for $E[H]$. With a selective set of designated total cost C = 50, 60, 80, 180, and 300, and heterogeneity levels, the optimal sample sizes are summarized in Table 7 for $\omega = 0.5$, $1 - \alpha = 0.95$, and three unit cost settings. As was described earlier for the expected half-width consideration in Table 3, the results in Table 7 also have the same behavior, in that the actual tolerance probability for a given total cost deceases substantially as the unit cost c_2 increases from 1 to 3. Therefore, researchers should be cautious about the prominent impact of heterogeneity on precision performance when the sources are limited.

Target tolerance probability is fixed and total cost needs to be minimized

In contrast with the previous case in which the total costs were fixed, the cost and precision assessment can be conversely performed by finding the optimal sample sizes to minimize cost when the target tolerance level is given. The utility of this procedure for the evaluation of expected half-width is extended to accommodate the precision criterion of assurance probability that the interval half-width is enclosed in the desirable range. To demonstrate the

Table 6 Computed sample sizes (N_1, N_2) and tolerance probability $P\{H < \omega\}$ when sample size N_2 is fixed with $\omega = 0.5$, $1 - \gamma = 0.90$, and $1 - \alpha = 0.95$

σ_1:σ_2															
1/3:1				1/2:1			1:1			2:1			3:1		
N_1	N_2	$P\{H < \omega\}$		N_1	N_2	$P\{H < \omega\}$	N_1	N_2	$P\{H < \omega\}$	N_1	N_2	$P\{H < \omega\}$	N_1	N_2	$P\{H < \omega\}$
199	24	.9000		60	25	.9001	38	40	.9126	94	80	.9076	189	100	.9057
13	27	.9075		18	30	.9156	31	60	.9239	86	140	.9115	174	200	.9086
9	30	.9084		14	35	.9247	28	80	.9009	83	200	.9076	169	300	.9020

Table 7 Computed sample sizes (N_1, N_2) and tolerance probability $P\{H < \omega\}$ when the total cost is fixed with half-width $\omega = 0.5$, and $1 - \alpha = 0.95$

σ_1:σ_2																			
c_1:c_2	1/3:1				1/2:1				1:1				2:1				3:1			
	Cost	N_1	N_2	$P\{H < \omega\}$	Cost	N_1	N_2	$P\{H < \omega\}$	Cost	N_1	N_2	$P\{H < \omega\}$	Cost	N_1	N_2	$P\{H < \omega\}$	Cost	N_1	N_2	$P\{H < \omega\}$
1:1	50	13	37	.9988	60	20	40	.9988	80	40	40	.9402	180	120	60	.9937	300	225	75	.9925
1:2	50	10	20	.4885	60	16	22	.5128	80	34	23	.2394	180	106	37	.4723	300	204	48	.4765
1:3	50	11	13	.1546	60	15	15	.1615	80	38	14	.0679	180	105	25	.1127	300	201	33	.0986

interrelation of the parameter configurations, numerical results are presented in Table 8 for the target tolerance probability $1 - \gamma = 0.90$, $\omega = 0.5$, and $1 - \alpha = 0.95$, along with several combinations of unit costs (c_1, c_2) and standard deviations (σ_1, σ_2). Similar to the expected width situation, the resulting total cost for fixed values of tolerance probability and standard deviations is drastically increasing as the unit cost c_2 changes from 1 to 3. It is suggested in Wang and Kupper (1997, p. 735) that the optimal sample sizes ratio is $N_2/N_1 = \sigma_2/\sigma_1$ for the problem of minimizing the total number of sample sizes. However, none of the optimal allocation ratios in their Table 5 agrees with this guideline. Essentially, a systematic search and detailed inspection of sample size combinations is required to find the optimal allocation that attains the desired precision while giving the least total sample size. This extra procedure and resulting merit in sample size determination is not addressed in Wang and Kupper (1997). In contrast, all of the issues are considered in our suggested procedure and the developed program.

Numerical example

To illustrate the usefulness and discrepancy of the proposed sample size procedures under different various situations of precision criteria and design schemes, we extend the numerical demonstration in Jan and Shieh (2011) from hypothesis testing to interval estimation for the difference of ability tests administered online and in the laboratory. Since the demographical structure of online samples can differ from that of offline samples acquired in traditional laboratory settings (Ihme, Lemke, Lieder, Martin, Muller & Schmidt, 2009), the planning parameter values are chosen as $\mu_{Lab} = 11$, $\mu_{Online} = 10$, $\sigma_{Lab} = 2.3$, and $\sigma_{Online} = 2.7$ to reflect the underlying treatment effect and heteroscedasticity. Moreover, online testing has the advantages of ease of obtaining a large sample and low cost. It would seem sensible that more samples could be obtained online rather than offline. The determination of actual sample sizes depends on the precision properties that the research wants to ensure for the resulting confidence intervals as well as other essential design features. First, it is intuitively reasonable to consider the expected width criterion. Suppose that the sample ratio is $N_{Online}/N_{Lab} = 4$. It follows that the sample sizes $N_{Lab} = 110$ and $N_{Online} = 440$ are required for the 95% confidence intervals of mean differences to have the expected interval half-width $\delta \leq 0.5$. On the other hand, if the sample size for the online sample is fixed at $N_{Online} = 400$, then it would need $N_{Lab} = 115$ to meet the same precision. To account for a budgetary concern where the total cost is $C = 200$ and the respective unit costs per subject are $c_{Lab} = 1$ and $c_{Online} = 0.2$, the optimal allocation of sample sizes is $N_{Lab} = 132$ and $N_{Online} = 340$,

Table 8 Computed sample sizes (N_1, N_2), cost, and tolerance probability $P\{H < \omega\}$ when the total cost needs to be minimized with target tolerance probability = 0.90, $\omega = 0.5$, and $1 - \alpha = 0.95$

σ_1:σ_2																			
c_1:c_2	1/3:1				1/2:1				1:1				2:1				3:1			
	Cost	N_1	N_2	$P\{H < \omega\}$	Cost	N_1	N_2	$P\{H < \omega\}$	Cost	N_1	N_2	$P\{H < \omega\}$	Cost	N_1	N_2	$P\{H < \omega\}$	Cost	N_1	N_2	$P\{H < \omega\}$
1:1	39	10	29	.9141	47	16	31	.9042	78	39	39	.9137	161	107	54	.9002	276	207	69	.9032
1:2	67	11	28	.9091	77	19	29	.9036	114	44	35	.9072	210	120	45	.9017	338	226	56	.9057
1:3	94	13	27	.9075	106	22	28	.9055	147	48	33	.9015	254	128	42	.9060	391	238	51	.9042

Table 9 Computed sample sizes (N_1, N_2) for precise interval estimation under various participant and cost constraints when $\sigma_1 = 2.3$, $\sigma_2 = 2.7$, $1 - \alpha = 0.95$, $\delta = 0.5$, $1 - \gamma = 0.90$, $\omega = 0.5$, $c_1 = 1$, and $c_2 = 0.2$

	Expected Width		Tolerance Probability	
	(N_1, N_2)	$E[H]$	(N_1, N_2)	$P\{H < \omega\}$
1. Fixed allocation ratio: $r = N_2/N_1 = 4$	(110, 440)	0.4986	(125, 500)	0.9084
II. One sample size is fixed: $N_2 = 400$	(115, 400)	0.4990	(134, 400)	0.9068
III. Fixed cost: $C = 200$	(132, 340)	0.4878	(133, 335)	0.7253
IV. Fixed target precision: $\delta = 0.5$, $\omega = 0.5$, and $1 - \gamma = 0.90$	(125, 328)	0.4998	(143, 340)	0.9004
		Minimum cost $C = 190.6$		Minimum cost $C = 211$

thus producing the maximum precision within the cost constraint. Conversely, the sample size combination $N_{\text{Lab}} = 125$ and $N_{\text{Online}} = 328$ induces the lowest cost $C = 190.6$, while ensuring the expected interval half-width $E[H] \leq 0.5$. The computed sample sizes and the corresponding actual values of expected interval half-width are summarized in Table 9 for ease of discussion.

Alternatively, it may be necessary for the assurance level of confidence interval half-widths to be enclosed by a designated bound. Assume that the tolerance probability $1 - \gamma = 0.90$, and 95% confidence interval half-width $\omega = 0.5$. A study with the sample ratio $r = N_{\text{Online}}/N_{\text{Lab}} = 4$ must have the sample sizes $N_{\text{Lab}} = 125$ and $N_{\text{Online}} = 500$ to meet the precision specification. When the online sample is predetermined at $N_{\text{Online}} = 400$, the computation shows that the laboratory group must at least have the sample size $N_{\text{Lab}} = 134$ in order to satisfy the designated precision. In the case of limited total cost $C = 200$, with $c_{\text{Lab}} = 1$ and $c_{\text{Online}} = 0.2$, the best set of sample sizes is $N_{\text{Lab}} = 133$ and $N_{\text{Online}} = 335$, and the resulting tolerance level is the highest for all sample sizes N_{Lab} and N_{Online}, with $N_{\text{Lab}} + (0.2)N_{\text{Online}} \leq 200$. However, for the tolerance probability $1 - \gamma = 0.90$ and 95% confidence interval half-width $\omega = 0.5$, the minimum cost is $C = 211$ for the optimal sample sizes $N_{\text{Lab}} = 143$ and $N_{\text{Online}} = 340$. These results and associated tolerance probabilities are also presented in Table 9. It is noteworthy that the computed sample sizes under the expected width consideration are smaller than those of the tolerance probability criterion. The only exception is the third case, with fixed total cost $C = 200$. Accordingly, the optimal sample sizes $N_{\text{Lab}} = 132$ and $N_{\text{Online}} = 340$ yield the expected half-width 0.4878, whereas the best sample size combination $N_{\text{Lab}} = 133$ and $N_{\text{Online}} = 335$ gives a tolerance level of merely $0.7253 < 1 - \gamma = 0.90$. These contrasting behaviors may be useful for researchers to justify their design strategy and financial support. The reader is referred to Ihme et al. (2009) for further details about the comparison of ability tests administered online and in the laboratory.

Conclusions

In order to enhance the applicability of confidence intervals and the fundamental usefulness of Welch's (1938) procedure, in the present article, we present the corresponding sample size techniques under various precision principles and design schemes. The precision criteria consist of the control of the expected width and the assurance of tolerance probability of confidence intervals. The design perspective includes four different allocation constraints and cost considerations. Detailed sample size tables are provided to help researchers have a better understanding of the intrinsic relationships that exist between the optimal sample sizes and the associated model, precision, and design configurations. Since existing software packages do not accommodate sample size calculations with the same degree of generality as is illustrated in this article, computer programs are developed to facilitate the use of the suggested procedures. The proposed sample size methodology should be useful for behavioral and other areas of social sciences to plan two-group comparison studies in which variances differ across groups.

Author Note The authors thank the editor, Gregory Francis, for enhancing the clarity of the article's presentation, Professor Chao-Ying Joanne Peng of Indiana University, and an anonymous referee, whose suggestions extended and strengthened its content immensely.

References

Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods, 2,* 20–33.

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

Beal, S. L. (1989). Sample size determination for confidence intervals on the population mean and on the difference between two population means. *Biometrics, 45,* 969–977.

Best, D. J., & Rayner, J. C. W. (1987). Welch's approximate solution for the Behrens–Fisher problem. *Technometrics, 29,* 205–210.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68,* 155–165.

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

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

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah: Erlbaum.

Ihme, J. M., Lemke, F., Lieder, K., Martin, F., Muller, J. C., & Schmidt, S. (2009). Comparison of ability tests administered online and in the laboratory. *Behavior Research Methods, 41,* 1183–1189.

Jan, S. L., & Shieh, G. (2011). Optimal sample sizes for Welch's test under various allocation and cost considerations. *Behavior Research Methods*. doi:10.3758/s13428-011-0095-7.

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

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

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

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

Liu, X. S. (2009). Sample size and the width of the confidence interval for mean difference. *British Journal of Mathematical and Statistical Psychology, 62,* 201–215.

SAS Institute. (2008a). *SAS/IML User's Guide, Version 9.2*. Cary: SAS Institute Inc.

SAS Institute. (2008b). *SAS/STAT User's Guide, Version 9.2*. Cary: SAS Institute Inc.

Satterthwaite, F. E. (1946). An approximate distribution of estimate of variance components. *Biometrics Bulletin, 2,* 110–114.

Smith, H. F. (1936). The problem of comparing the results of two experiments with unequal errors. *Journal of the Council for Scientific and Industrial Research, 9,* 211–212.

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

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