Behav Res (2014) 46:808–822
DOI 10.3758/s13428-013-0415-1

Sample size requirements for the design of reliability studies: precision consideration

Gwowen Shieh

Published online: 12 December 2013

Abstract In multilevel modeling, the intraclass correlation coefficient based on the one-way random-effects model is routinely employed to measure the reliability or degree of resemblance among group members. To facilitate the advocated practice of reporting confidence intervals in future reliability studies, this article presents exact sample size procedures for precise interval estimation of the intraclass correlation coefficient under various allocation and cost structures. Although the suggested approaches do not admit explicit sample size formulas and require special algorithms for carrying out iterative computations, they are more accurate than the closed-form formulas constructed from large-sample approximations with respect to the expected width and assurance probability criteria. This investigation notes the deficiency of existing methods and expands the sample size methodology for the design of reliability studies that have not previously been discussed in the literature.

Keywords Intraclass correlation coefficient · Optimal design · Precision · Sample size

In view of the recommendations of 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), interval estimation is more informative about the magnitude of a targeted parameter than is null hypothesis significance testing, and it should be the best reporting practice in all empirical studies. Accordingly, the editorial policies and statistical guidelines of several prominent educational and psychological journals have called for greater use of confidence intervals for principal effect sizes. In addition, numerous practical principles and suggestions for selecting, calculating, and interpreting effect size indices of various types of statistical analyses have been addressed in the literature. Readers interested in learning more about the conceptual implications of point and interval estimates of effect sizes are directed to articles by Cumming (2012), Dunst and Hamby (2012), Ferguson (2009), Fritz, Morris, and Richler (2012), Grissom and Kim (2012), Odgaard and Fowler (2010), Robey (2004), Sun, Pan, and Wang (2010), Thompson (2007), and the references therein. Consequently, it has become a general consensus across many scientific disciplines to include appropriate effect size measures and associated confidence intervals when documenting the results of research studies.

There is considerable recent literature pertaining to both the theoretical and practical problems of investigating the hierarchical nature of individual and group influences in multilevel research. The statistical and methodological issues associated with hierarchical linear models can be found in Goldstein (2002), Hofmann (2002), Raudenbush and Bryk (2002), and Snijders and Bosker (2012). Within the context of the multilevel framework, measurements on individuals (e.g., employee, student, patient) within the same group (e.g., organization, classroom, clinic) are presumably more similar than measurements on individuals in different groups. The correlation among the measurements on individuals within the same group must be appropriately accounted for in a clustering study. Accordingly, the intraclass correlation coefficient (ICC) has been extensively used to measure reliability or degree of

Electronic supplementary material The online version of this article (doi:10.3758/s13428-013-0415-1) contains supplementary material, which is available to authorized users.

G. Shieh
Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan 30050, Republic of China

G. Shieh (✉)
Department of Management Science, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan 30010
e-mail: gwshieh@mail.nctu.edu.tw

resemblance among cluster members. Comprehensive reviews related to the ICC as an interrater reliability measure were presented in Bartko (1976), McGraw and Wong (1996), and Shrout and Fleiss (1979). Specifically, McGraw and Wong identified distinct definitions and relative merits of various ICC indices and also emphasized the advantage of the exact confidence interval procedure in coverage performance over the approximate method.

To facilitate the advocated statistical practice of presenting confidence intervals and to further improve the applicability of multilevel modeling, the focus of this article is on the interval estimation of the ICC effect size in a one-way random-effects model. The ICC(1) index, introduced by Fisher (1938), is the most frequently adopted measure of ICC. Hence, a convenient approach is to apply the large-sample theory to approximate the distribution of ICC(1). The numerical results of Donner and Wells (1986) showed that the resulting interval procedures provide consistently good coverage at all values of a population ICC and are competitive, in terms of the mean width, with the ANOVA F method. In contrast, Ukoumunne (2002) concluded that the methods based upon the variance ratio F statistic give greater coverage levels than those based upon the large-sample normality of the ICC(1) estimator, especially when the data have relatively few large groups and low ICC values. To some extent, the validity of the simulation results in Donner and Wells and in Ukoumunne is conditional on the assumed underlying model structures. Hence, further assessments with different model configurations would be helpful in understanding the intrinsic behavior of the competing interval procedures.

The stress on reporting effect sizes and confidence intervals in empirical investigation suggests that researchers should plan studies not only to select practically meaningful effect size indices but also to have sufficiently accurate interval estimates of effect sizes. Thus, it is prudent to aid this research practice by determining the sample sizes that are necessary to satisfy the desired precision of interval estimation in the planning stage of reliability studies. Due to the entrenched use of the ICC(1) index and the appealing simplicity of its asymptotic property, Bonett (2002) and Giraudeau and Mary (2001) adopted Fisher's (1938) approximate variance estimator of ICC(1) and presented a closed-form formula for determining the required sample size so that the confidence interval will have the desired expected width. Similarly, Zou (2012), using the same large-sample approximation, suggested an explicit equation for calculating the sample size that is needed to ensure that the interval half-width is within a designated value with the prespecified assurance probability. These simplified sample size methods are straightforward to apply and do not require an iterative solution. Moreover, numerical evidence was presented in Bonett, in Giraudeau and Mary, and in Zou (2012) to demonstrate the accuracy of these approximate procedures.

However, a detailed inspection reveals that Bonett (2002) provided only a limited number of comparisons between the exact and approximate sample sizes. Because no explicit computational algorithm was presented, the exact sample sizes reported in Bonett may actually be approximate values of a different method. Arguably, the selected evaluations between the supposedly exact and approximate sample sizes do not justify the accuracy of the suggested approach. On the other hand, the numerical investigation of Giraudeau and Mary (2001) involved several model configurations and performance measures. But their assessments were not conducted in an organized fashion by fixing all but one of the key factors and varying a single factor to help clarify the accuracy of the presented technique. Moreover, Zou (2012) considered only moderate and substantial levels of reliability, which may not represent settings likely to be encountered with real data. Thus, his evaluation did not cover a wide variety of parameter settings and was not thorough enough to elucidate the potential deficiency of the approximate sample size method. Consequently, a comprehensive investigation is required to complement the existing appraisals and demonstrations in Bonett, in Giraudeau and Mary, and in Zou. It is important to ensure that the effects of vital factors on sample size calculations are well understood before the technique can be recognized as a general tool for optimal design of reliability studies.

The simplicity of an approximate methodology may be appealing for inducing computational shortcuts, but it does not retain all of the essential features in the model formulation, and thus, the resulting sample size techniques tend to be restrictive and problematic. Specifically, it was noted in Donner and Koval (1983) that the accuracy of the asymptotic normality with Fisher's (1938) simple variance formulation depends on having moderately large numbers of groups—say, at least 30 groups. Hence, the corresponding sample size formula is vulnerable to model characteristics, and this restriction may impede its applicability. With advanced computer technology and prevalent statistical software, computational simplicity is no longer a major concern in sample size planning. For exact interval estimation constructed with the variance ratio F statistic, the sample size issues have important implications for conducting and interpreting reliability research but have received relatively little attention in the literature. Although Bonett (2002) and Donner (1999) have attempted some sample size calculations under the expected width principle, their approaches are approximate in nature and have not been fully evaluated empirically. Also, their approaches do not generalize to the assurance probability principle in a straightforward manner. It appears that no exact sample size procedures have been proposed for the standard confidence intervals based on the ANOVA F statistic.

Toward the goal of choosing the most appropriate methodology for reliability studies with potentially diverse model configurations, the present article describes exact sample size determinations for precise interval estimation of the ICC. It is

of practical interest and theoretical importance to reinforce the exact confidence intervals recommended by McGraw and Wong (1996) by developing the associated sample size procedures for various design schemes. Under both the expected width and assurance probability criteria, this study first examines research designs with the subject allocation constraint that the number of subjects per group is fixed or the number of groups is given. The prescribed studies of Bonett (2002), Giraudeau and Mary (2001), and Zou (2012) concentrated only on the first situation, and it is the only case in which closed-form formulas can be obtained from the large-sample approximation. Comprehensive appraisals were performed to demonstrate the advantages of the suggested exact sample size approaches over the approximate formulas under a wide range of parameter configurations and sample size structures.

Then the study extends the design strategies to accommodate both budgetary constraints and precision assessments. The cost implications suggest optimally assigning subjects to satisfy a designated precision level for the least cost or to attain maximum precision performance for a fixed cost. Accordingly, exact sample size procedures are proposed to obtain optimal solutions under both precision principles of expected width and assurance probability. The related cost issues in the design of reliability studies can be found in Flynn, Whitley, and Peters (2002), Shoukri, Asyali, and Donner (2004), Shoukri, Asyali, and Walter (2003), and the references therein. Moreover, Giraudeau and Mary (2001) demonstrated the approximate method and graphical displays for the practical problem of determining the optimal sample size combination in order to attain the narrowest expected width when the total number of subjects is fixed in advance. Since the formulation of the general cost function includes the total number of subjects as a special case, the corresponding illustration and algorithm provide an exact and more efficient approach to computing the optimal sample sizes. Essentially, this investigation updates and expands the current work for precise interval estimation of ICC by noting the fundamental deficiency of existing approximate formulas and demonstrating the usefulness of exact sample size procedures for various allocation and cost plans. Corresponding SAS/IML (SAS Institute, 2012) and R (R Development Core Team, 2013) computer programs are also developed to help researchers perform the recommended sample size procedures in the research design of reliability studies.

Interval estimation procedures

In reliability studies, a frequently adopted design is the one-way random-effects model

$$Y_{ij} = \mu + \gamma_i + \varepsilon_{ij}, i = 1, \ldots, G; j = 1, \ldots, N, \quad (1)$$

where Y_{ij} is the jth individual measurement within group i, μ is the grand mean, and γ_i and ε_{ij} are independent random variables with $\gamma_i \sim N(0, \sigma_\gamma^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$. The variance of Y_{ij} is then given by $\sigma_\gamma^2+\sigma_\varepsilon^2$, where σ_γ^2 represents the between-group variance and σ_ε^2 is the within-group variance. Accordingly, the ICC ρ is defined as

$$\rho = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\varepsilon^2}, \quad (2)$$

which is directly interpretable as the proportion of the total variance of the response that is accounted for by the clustering or group cohesion. To assess the magnitude of ρ, the well-known ICC(1) index is denoted by

$$\widehat{\rho} = \frac{MSB - MSW}{MSB + (N-1)MSW} = \frac{F^* - 1}{F^* + N - 1}, \quad (3)$$

where MSB is the between-group mean square, MSW is the within-group mean square, and $F^* = MSB/MSW$. Under the model assumption defined in Eq. 1, the ANOVA F test statistic F^* has the distribution

$$F^* \sim \tau F(G-1, G(N-1)), \quad (4)$$

where $\tau = 1 + N\rho/(1 - \rho)$ and $F(G - 1, G(N - 1))$ is the F distribution with $G - 1$ and $G(N - 1)$ degrees of freedom. For the purpose of the interval estimation of ρ, an exact 100(1 – α)% two-sided confidence interval $\{\widehat{\rho}_{EL}, \widehat{\rho}_{EU}\}$ of ρ can be constructed from Eq. 4 as

$$\left\{\widehat{\rho}_{EL}, \widehat{\rho}_{EU}\right\} = \left\{\frac{F^*/F_{\alpha/2} - 1}{F^*/F_{\alpha/2} + N - 1}, \frac{F^*/F_{(1-\alpha/2)} - 1}{F^*/F_{(1-\alpha/2)} + N - 1}\right\} \quad (5)$$

where $F_{\alpha/2}$ and $F_{(1-\alpha/2)}$ are the upper and lower 100(α/2)th percentiles of the F distribution $F(G - 1, G(N - 1))$, respectively. Accordingly, the upper and lower 100(1 − α)% one-sided confidence intervals of ρ are of the form $\{\widehat{\rho}_{EL}, 1\}$ and $\{0, \widehat{\rho}_{EU}\}$, respectively, where $\widehat{\rho}_{EL} = (F^*/F\alpha - 1)/(F^*/F\alpha + N - 1)$ and $\widehat{\rho}_{EU} = \left(F^*/F_{(1-\alpha)} - 1\right)/\left(F^*/F_{(1-\alpha)} + N - 1\right)$.

It was derived in Fisher (1938) that the large-sample variance of $\widehat{\rho}$ can be approximated by

$$\nu^2 = \frac{2(1-\rho)^2[1 + (N-1)\rho]^2}{N(N-1)(G-1)}.$$

This simplification gives a convenient approximation to the underlying distribution of $\widehat{\rho}$ with an asymptotic normal distribution

$$\widehat{\rho} \sim N\left(\rho, \nu^2\right). \quad (6)$$

Thus, an approximate 100(1 – α)% two-sided confidence interval $\{\widehat{\rho}_{AL},\ \widehat{\rho}_{AU}\}$ of ρ can be readily obtained as

$$\left\{\widehat{\rho}_{AL},\widehat{\rho}_{AU}\right\}=\left\{\widehat{\rho}-z_{\alpha/2}\widehat{\nu},\widehat{\rho}+z_{\alpha/2}\widehat{\nu}\right\}, \tag{7}$$

where $\nu=\left(\widehat{\nu}^2\right)^{1/2}$, $\widehat{\nu}^2=2(1-\widehat{\rho})^2[1+(N-1)\widehat{\rho}]^2/[N(N-1)(G-1)]$ is the estimated variance of $\widehat{\rho}$ and $z_{\alpha/2}$ is the upper 100(α/2)th percentile of the standard normal distribution. Also, the upper and lower 100(1 – α)% one-sided confidence intervals of ρ can be expressed as $\{\widehat{\rho}_{AL},1\}$ and $\{0,\widehat{\rho}_{AU}\}$, respectively, where $\widehat{\rho}_{AL}=\widehat{\rho}-z_{\alpha}\widehat{\nu}$ and $\widehat{\rho}_{AU}=\widehat{\rho}+z_{\alpha}\widehat{\nu}$. Although several comparable approximate interval procedures have been proposed as documented in Donner and Wells (1986) and Ukoumunne (2002), at present there is very little information on the development of supporting sample size methodology in the literature. The attention is restricted to the simple method with Fisher's (1938) asymptotic variance estimator because it is the only case where the accompanying sample size techniques for precise interval estimation are already available. The approximate formulas suggested in Bonett (2002), Giraudeau and Mary (2001), and Zou (2012) are utilized as a benchmark to contrast the performance of the exact approaches presented in the subsequent explication.

For ease of illustration, the widths of the 100(1 – α)% two-sided confidence intervals $\{\widehat{\rho}_{EL},\ \widehat{\rho}_{EU}\}$ and $\{\widehat{\rho}_{AL},\widehat{\rho}_{AU}\}$ given in Eqs. 5 and 7 are denoted by

$$W_{\mathrm{E}}=\widehat{\rho}_{\mathrm{EU}}-\widehat{\rho}_{\mathrm{EL}}\ \text{ and }\ W_{\mathrm{A}}=\widehat{\rho}_{\mathrm{AU}}-\widehat{\rho}_{\mathrm{AL}}=2z_{\alpha/2}\widehat{\nu}, \tag{8}$$

respectively. The subscripts E and A of W_{E} and W_{A} emphasize the dependence on the corresponding exact and approximate procedures. It is clear that the actual widths W_{E} and W_{A} depend on the confidence coefficient 1 – α, the sample size in terms of the number of groups and the number of subjects per group (G, N), and the observed statistic F^* or $\widehat{\rho}$. Since the ANOVA F^* and the $\widehat{\rho}$ index are random variables, their statistical properties determine the resulting confidence interval width. When planning a study for ensuring that the confidence interval is narrow enough to produce meaningful findings, researchers must consider the stochastic nature of interval widths W_{E} and W_{A}. Certainly, the sample size needed for precise interval estimation is affected in important and distinctive ways by the actual formulations of interval procedures and associated distributional properties. It will be shown later that the simple approximations may fail to properly account for the embedded features and may lead to a poor choice of sample size.

From an advance study design viewpoint, it is desirable to determine the optimal sample sizes so that the resulting confidence interval will meet the designated precision requirements. Two useful principles concern the control of the expected width and the assurance probability of the width within a designated value. Specifically, it is necessary to calculate the required sample size such that the expected width of a 100(1 – α)% two-sided confidence interval is within the given bound

$$E[W]\leq\delta,$$

where W is the interval width and δ (> 0) is a constant. On the other hand, one may compute the sample size needed to guarantee, with a given assurance probability, that the width of a 100(1 – α)% two-sided confidence interval will not exceed the planned value

$$P(W\leq\omega)\geq 1-\gamma,$$

where (1 – γ) is the specified assurance level and ω (>0) is a constant. Since there may be several possible choices of sample size combinations (G, N) that satisfy the chosen precision criterion in the process of sample size calculations, it is constructive to consider design schemes with subject and budget constraints that lead to a unique and optimal result. In the next two sections, the ideas of Shieh and Jan (2012) are applied to develop exact sample size procedures of precise interval estimation of ICC with four different allocation and cost settings under the expected width and assurance probability criteria, respectively.

Expected width criterion

An exact 100(1 – α)% two-sided confidence interval $\{\widehat{\rho}_{\mathrm{EL}},\widehat{\rho}_{\mathrm{EU}}\}$ can be readily obtained with the ANOVA statistic F^* and the designated percentiles $F_{\alpha/2}$ and $F_{(1-\alpha/2)}$; however, the evaluation of expected width $E[W_{\mathrm{E}}]$ does not permit an explicit expression. Note that the interval width W_{E} is a function of the F^* statistic, which has an F distribution given in Eq. 4. Since an F random variable is a one-to-one function of a beta random variable and, unlike the range of an F distribution, which is between 0 and infinity, a beta distribution is bounded between the values of 0 and 1, it is computationally simple and relatively stable to perform the numerical integration of $E[W_{\mathrm{E}}]$ with respect to a beta distribution, instead of an F distribution. Accordingly, the utility of the expected interval width is exploited to construct the optimal sample size procedures for exact confidence interval estimation in the following four design structures.

Design I: the number of subjects per group is fixed

Assuming that the number of subjects in each group is prespecified, the expected width $E[W_{\mathrm{E}}]$ is a monotone

function of the number of groups when all other factors remain constant. Thus, for a selected threshold δ, a simple incremental search can be employed to find the minimum sample size G such that

$$E[W_{\mathrm{E}}] \leq \delta \tag{9}$$

for the chosen size N, confidence level $(1 - \alpha)$, and target ICC value ρ. Essentially, it requires a special purpose algorithm to perform the involved numerical computation. In contrast, Bonett (2002) and Giraudeau and Mary (2001) used the approximation $E[W_{\mathrm{A}}] = E\left[2z_{\alpha/2}\widehat{\nu}\right] = 2z_{\alpha/2}\nu$ to show that the minimum sample size G such that $E[W_{\mathrm{A}}] \leq \delta$ is equivalent to the smallest integer G that satisfies the inequality

$$G \geq \frac{8z_{\alpha/2}^2(1-\rho)^2[1+(N-1)\rho]^2}{N(N-1)\delta^2} + 1. \tag{10}$$

Clearly, the suggested sample size formula for the approximate confidence interval $\{\widehat{\rho}_{\mathrm{AL}}, \widehat{\rho}_{\mathrm{AU}}\}$ is relatively easy to apply and does not involve any iterative computation. But the confidence intervals and sample size method have some undesirable properties, and they should not be used indiscriminately.

For the purposes of assessing the behavior of interval estimation and sample size procedures, an extensive numerical examination was performed for the model settings in Table 1 of Bonett (2002), and it was also extended to other configurations that were not considered there. To demonstrate a profound implication of the approximate and exact methodology, the empirical study was conducted in two steps. The first step involved sample size calculations for the expected width criterion across a wide range of model configurations. In the second step, a Monte Carlo simulation study was performed to provide insights into the precision behavior for the interval estimation and sample size calculations under the design characteristics specified in the first step.

First, a systematic numerical investigation of nested design was conducted by fixing the confidence level $(1 - \alpha) = 0.95$, and varying the other three factors of interval bound $\delta = 0.2$ and 0.3, the number of subjects per group $N = 2, 3, 5, 10, 20$, and the population ICC value $\rho = 0$–0.90 with an increment of 0.1. With these specifications, the required numbers of groups were computed for the two approaches described in Eqs. 9 and 10, respectively. To conserve space, only the computed sample sizes associated with $N = 5$ and 10 for $\delta = 0.2$ are presented in Tables 1 and 2, respectively, whereas the corresponding results for $\delta = 0.3$ are listed in Tables 3 and 4. Moreover, the estimated or achieved expected widths $E[W_{\mathrm{E}}]$ and $E[W_{\mathrm{A}}] = 2z_{\alpha/2}\nu$

Table 1 Sample size, coverage, and precision of the exact and approximate procedures for $1 - \alpha = 0.95$, $N = 5$, and expected width $\delta = 0.2$

ρ	G	Upper 97.5 %	Error	Lower 97.5 %	Error	Two-sided 95 %	Error	Simulated $E[W]$	Estimated $E[W]$	Error
Exact approach										
0	42	0.9776	0.0026	0.9742	−0.0008	0.9518	0.0018	0.1974	0.1975	−0.0001
0.1	63	0.9732	−0.0018	0.9725	−0.0025	0.9457	−0.0043	0.1984	0.1985	−0.0001
0.2	80	0.9731	−0.0019	0.9754	0.0004	0.9485	−0.0015	0.1995	0.1995	0.0000
0.3	91	0.9748	−0.0002	0.9771	0.0021	0.9519	0.0019	0.1993	0.1992	0.0001
0.4	93	0.9722	−0.0028	0.9746	−0.0004	0.9468	−0.0032	0.1992	0.1993	−0.0001
0.5	86	0.9723	−0.0027	0.9772	0.0022	0.9495	−0.0005	0.1991	0.1992	−0.0001
0.6	71	0.9745	−0.0005	0.9758	0.0008	0.9503	0.0003	0.1993	0.1993	0.0000
0.7	51	0.9754	0.0004	0.9754	0.0004	0.9508	0.0008	0.1983	0.1986	−0.0003
0.8	29	0.9748	−0.0002	0.9767	0.0017	0.9515	0.0015	0.1992	0.1991	0.0001
0.9	12	0.9763	0.0013	0.9760	0.0010	0.9523	0.0023	0.1901	0.1909	−0.0008
Approximate method										
0	40	0.9917	**0.0167**	0.9433	**−0.0317**	0.9350	**−0.0150**	0.1964	0.1985	−0.0021
0.1	62	0.9898	**0.0148**	0.9524	**−0.0226**	0.9422	**−0.0078**	0.1980	0.2000	−0.0020
0.2	81	0.9839	**0.0089**	0.9563	**−0.0187**	0.9402	**−0.0098**	0.1978	0.1996	−0.0018
0.3	93	0.9784	**0.0034**	0.9647	**−0.0103**	0.9431	**−0.0069**	0.1976	0.1990	−0.0015
0.4	95	0.9789	0.0039	0.9675	−0.0075	0.9464	−0.0036	0.1982	0.1995	−0.0013
0.5	88	0.9761	0.0011	0.9719	−0.0031	0.9480	−0.0020	0.1984	0.1993	−0.0010
0.6	73	0.9709	−0.0041	0.9762	0.0012	0.9471	−0.0029	0.1983	0.1987	−0.0004
0.7	51	0.9658	−0.0092	0.9821	0.0071	0.9479	−0.0021	0.2004	0.1998	0.0005
0.8	29	0.9599	−0.0151	0.9824	0.0074	0.9423	−0.0077	0.2014	0.1968	0.0046
0.9	10	0.9433	**−0.0317**	0.9918	**0.0168**	0.9351	**−0.0149**	0.2183	0.1901	0.0283

Table 2 Sample size, coverage, and precision of the exact and approximate procedures for 1 – α = 0.95, N = 10, and expected width δ = 0.2

ρ	G	Upper 97.5 %	Error	Lower 97.5 %	Error	Two-sided 95 %	Error	Simulated $E[W]$	Estimated $E[W]$	Error
Exact approach										
0	14	0.9748	−0.0002	0.9750	0.0000	0.9498	−0.0002	0.1904	0.1903	0.0001
0.1	28	0.9729	−0.0021	0.9747	−0.0003	0.9476	−0.0024	0.1999	0.1999	0.0000
0.2	45	0.9753	0.0003	0.9774	0.0024	0.9527	0.0027	0.1985	0.1983	0.0002
0.3	58	0.9731	−0.0019	0.9749	−0.0001	0.9480	−0.0020	0.1991	0.1991	0.0000
0.4	65	0.9718	−0.0032	0.9746	−0.0004	0.9464	−0.0036	0.1992	0.1992	0.0000
0.5	64	0.9742	−0.0008	0.9735	−0.0015	0.9477	−0.0023	0.1996	0.1996	0.0000
0.6	56	0.9765	0.0015	0.9758	0.0008	0.9523	0.0023	0.1989	0.1988	0.0001
0.7	42	0.9750	0.0000	0.9751	0.0001	0.9501	0.0001	0.1976	0.1977	−0.0001
0.8	25	0.9757	0.0007	0.9736	−0.0014	0.9493	−0.0007	0.1965	0.1964	0.0001
0.9	10	0.9751	0.0001	0.9724	−0.0026	0.9475	0.0025	0.1944	0.1948	−0.0004
Approximate method										
0	10	0.9996	**0.0246**	0.8740	**−0.1010**	0.8736	**−0.0764**	0.1900	0.1948	−0.0048
0.1	26	0.9942	**0.0192**	0.9295	**−0.0455**	0.9237	**−0.0263**	0.1965	0.1998	−0.0033
0.2	44	0.9879	**0.0129**	0.9465	**−0.0285**	0.9344	**−0.0156**	0.1970	0.1996	−0.0026
0.3	59	0.9854	**0.0104**	0.9528	**−0.0222**	0.9382	**−0.0118**	0.1963	0.1987	−0.0024
0.4	67	0.9816	0.0066	0.9601	−0.0149	0.9417	−0.0083	0.1967	0.1985	−0.0018
0.5	66	0.9794	0.0044	0.9648	−0.0102	0.9442	−0.0058	0.1980	0.1993	−0.0014
0.6	57	0.9739	−0.0011	0.9693	−0.0057	0.9432	−0.0068	0.1992	0.1999	−0.0007
0.7	43	0.9708	−0.0042	0.9729	−0.0021	0.9437	−0.0063	0.2003	0.1999	0.0005
0.8	24	0.9668	−0.0082	0.9788	0.0038	0.9456	−0.0044	0.2051	0.1998	0.0053
0.9	9	0.9534	**−0.0216**	0.9856	**0.0106**	0.9390	**−0.0110**	0.2198	0.1880	0.0318

are also summarized in the tables. Due to the underlying metric of integer sample sizes, the resulting expected widths are marginally smaller than the selected width, δ = 0.2 or 0.3.

It follows from the sample sizes in Tables 1 and 2 that the exact numbers G do not completely agree with those reported in Table 1 of Bonett (2002). Hence, this clarifies that the correct sample size, denoted by n_c, in Bonett, provides only an approximate number of groups needed for the exact confidence interval $\{\widehat{\rho}_{EL}, \widehat{\rho}_{EU}\}$ to have the desired expected width. The computations for N = 2 also reveal that the results in Table 5 of Donner (1999) are not exact solutions. On the other hand, the differences between the required numbers of groups of the exact and approximate methods in Tables 1, 2, 3, and 4 are 1 or 2 in most cases, with the worst case of 4 when δ = 0.2, N = 10, and ρ = 0. However, the exact and approximate confidence intervals are constructed with distinctive F and normal distributions, respectively. The resulting sample sizes needed to attain the designated precision are not necessarily equivalent. The small discrepancy between the sample size calculations may reveal an interesting phenomenon of the two techniques, but it does not serve to assure the actual performance of the approximate sample size formula. The relative performance of the exact and approximate approaches is further justified in the second stage of this empirical study.

With the given sample sizes and parameter configurations, estimates of the true coverage probability and expected width were computed through Monte Carlo simulation of 10,000 independent data sets. For each replicate, the confidence limits associated with a 95 % two-sided confidence interval were computed and were also employed to construct the upper and lower 97.5 % one-sided confidence intervals. Accordingly, a total of three different sets of confidence intervals were obtained. Thus, the simulations cover a much broader range of situations than those considered in Donner and Wells (1986), Giraudeau and Mary (2001), and Ukoumunne (2002), which 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 ICC ρ. The accuracy of the examined confidence interval procedure is determined by the difference between the simulated coverage probability and the designated coverage probability as error = simulated coverage probability – nominal coverage probability. In addition, the average interval width of ρ was also computed for the 10,000 replicated widths of 95 % two-sided confidence intervals. The adequacy of a sample size procedure for precise interval estimation is determined by the following formula: error = simulated expected width – estimated expected width. The simulated

Table 3 Sample size, coverage, and precision of the exact and approximate procedures for $1 - \alpha = 0.95$, $N = 5$, and expected width $\delta = 0.3$

ρ	G	Upper 97.5 %	Error	Lower 97.5 %	Error	Two-sided 95 %	Error	Simulated $E[W]$	Estimated $E[W]$	Error
Exact approach										
0	20	0.9734	−0.0016	0.9733	−0.0017	0.9467	−0.0033	0.2970	0.2959	0.0011
0.1	29	0.9751	0.0001	0.9724	−0.0026	0.9475	−0.0025	0.2948	0.2953	−0.0005
0.2	36	0.9761	0.0011	0.9749	−0.0001	0.9510	0.0010	0.2968	0.2971	−0.0003
0.3	40	0.9750	0.0000	0.9765	0.0015	0.9515	0.0015	0.2987	0.2987	0.0000
0.4	41	0.9742	−0.0008	0.9763	0.0013	0.9505	0.0005	0.2977	0.2978	−0.0001
0.5	38	0.9724	−0.0026	0.9760	0.0010	0.9484	−0.0016	0.2972	0.2974	−0.0002
0.6	31	0.9726	−0.0024	0.9760	0.0010	0.9486	−0.0014	0.2997	0.2999	−0.0002
0.7	23	0.9752	0.0002	0.9750	0.0000	0.9502	0.0002	0.2972	0.2966	0.0006
0.8	14	0.9754	0.0004	0.9744	−0.0006	0.9498	−0.0002	0.2953	0.2944	0.0009
0.9	7	0.9750	0.0000	0.9735	−0.0015	0.9485	−0.0015	0.2771	0.2745	0.0026
Approximate method										
0	19	0.9955	**0.0205**	0.9202	**−0.0548**	0.9157	**−0.0343**	0.2866	0.2922	−0.0056
0.1	29	0.9893	**0.0143**	0.9384	**−0.0366**	0.9277	**−0.0223**	0.2893	0.2952	−0.0059
0.2	37	0.9849	**0.0099**	0.9480	**−0.0270**	0.9329	**−0.0171**	0.2921	0.2975	−0.0054
0.3	42	0.9809	**0.0059**	0.9558	**−0.0192**	0.9367	**−0.0133**	0.2931	0.2981	−0.0050
0.4	43	0.9791	0.0041	0.9597	−0.0153	0.9388	−0.0112	0.2942	0.2984	−0.0042
0.5	40	0.9729	−0.0021	0.9667	−0.0083	0.9396	−0.0104	0.2947	0.2977	−0.0031
0.6	33	0.9684	−0.0066	0.9708	−0.0042	0.9392	−0.0108	0.2962	0.2980	−0.0018
0.7	24	0.9636	−0.0114	0.9759	0.0009	0.9395	−0.0105	0.2967	0.2947	0.0020
0.8	14	0.9587	−0.0163	0.9804	0.0054	0.9391	−0.0109	0.3046	0.2888	0.0158
0.9	5	0.9372	**−0.0378**	0.9803	**0.0053**	0.9175	**−0.0325**	0.3779	0.2851	0.0928

results of three types of coverage probabilities, average widths, and corresponding errors for the exact approach and the approximate method are also presented in Tables 1, 2, 3, and 4.

The numerical results indicate that the simulated coverage probabilities of the one- and two-sided exact confidence intervals closely agree with the nominal confidence levels for all 40 combined cases. In particular, the case of the 95 % two-sided confidence interval estimation with $\delta = 0.3$, $N = 10$, $G = 8$, and $\rho = 0$ yielded a simulated coverage probability 0.9450 and induced the largest absolute error 0.0050. Thus, the accurate performance of the exact confidence interval procedures is extremely stable for all model configurations including the settings with small numbers of groups and small ICC values. However, it is not the case for the approximate confidence intervals.

A closer look at the coverage behavior of the approximate confidence intervals shows that the discrepancy between simulated and nominal coverage probabilities tends to decrease for larger numbers of groups. Although this general pattern is consistent with the findings of Donner and Koval (1983) about the accuracy of the asymptotic normality with Fisher's variance estimator, the sizable errors of the one- and two-sided confidence intervals reveal that the approximation remains problematic even for some $G \geq 30$. For example, the errors associated with the upper 97.5 % confidence intervals in Table 1 are 0.0167, 0.0148, 0.0089, and 0.0034 for $\rho = 0$, 0.1, 0.2, and 0.3 with $G = 40$, 62, 81, and 93, respectively. For the lower 97.5 % confidence intervals, the resulting coverage differences are −0.0317, –0.0226, –0.0187, and −0.0103. Hence, the combined coverage errors of the 95 % two-sided confidence intervals are −0.0150, –0.0078, –0.0098, and −0.0069. Unfortunately, there are many similar and problematic cases in Tables 2, 3, and 4 as well. Although the simple guideline suggests that the asymptotic normality with Fisher's variance estimator is reasonable for $G \geq 30$, the approximation is less accurate when the population ICC is small. Moreover, the confidence limits of the 95 % two-sided confidence interval are the same as the respective lower and upper limits of the one-sided upper and lower 97.5 % confidence intervals. Thus, it is misleading to report that a two-sided interval procedure is accurate on the basis of a combination of some noticeable under- and overestimated one-sided coverage probabilities. Consequently, a mere coverage probability assessment of two-sided confidence intervals may obscure potential biases in confidence limits of the transformed equidistant confidence intervals based on the large-sample approximation. On the other hand, the simulated expected width of the approximate confidence intervals showed good agreement with the estimated interval width unless the number of groups is small. To

Table 4 Sample size, coverage, and precision of the exact and approximate procedures for $1 - \alpha = 0.95$, $N = 10$, and expected width $\delta = 0.3$

ρ	G	Upper 97.5 %	Error	Lower 97.5 %	Error	Two-sided 95 %	Error	Simulated $E[W]$	Estimated $E[W]$	Error
Exact approach										
0	8	0.9711	−0.0039	0.9739	−0.0011	0.9450	−0.0050	0.2912	0.2910	0.0002
0.1	14	0.9759	0.0009	0.9738	−0.0012	0.9497	−0.0003	0.2978	0.2982	−0.0004
0.2	21	0.9766	0.0016	0.0764	0.0014	0.9530	0.0030	0.2959	0.2954	0.0005
0.3	26	0.9734	−0.0016	0.9744	−0.0006	0.9478	−0.0022	0.2985	0.2981	0.0004
0.4	29	0.9773	0.0023	0.9723	−0.0027	0.9496	−0.0004	0.2968	0.2969	−0.0001
0.5	28	0.9722	−0.0028	0.9736	−0.0014	0.9458	−0.0042	0.2995	0.2995	0.0000
0.6	25	0.9745	−0.0005	0.9751	0.0001	0.9496	−0.0004	0.2958	0.2956	0.0002
0.7	19	0.9742	−0.0008	0.9735	−0.0015	0.9477	−0.0023	0.2941	0.2938	0.0003
0.8	12	0.9750	0.0000	0.9759	0.0009	0.9509	0.0009	0.2890	0.2898	−0.0008
0.9	6	0.9726	−0.0024	0.9761	0.0011	0.9487	−0.0013	0.2710	0.2741	−0.0031
Approximate method										
0	5	0.9998	**0.0248**	0.8144	**−0.1606**	0.8142	**−0.1358**	0.2752	0.2922	−0.0170
0.1	13	0.9973	**0.0223**	0.8998	**−0.0752**	0.8971	**−0.0529**	0.2777	0.2885	−0.0107
0.2	21	0.9921	**0.0171**	0.9244	**−0.0506**	0.9165	**−0.0335**	0.2832	0.2927	−0.0095
0.3	27	0.9882	**0.0132**	0.9392	**−0.0358**	0.9274	**−0.0226**	0.2892	0.2968	−0.0076
0.4	30	0.9831	0.0081	0.9501	−0.0249	0.9332	−0.0168	0.2932	0.2995	−0.0063
0.5	30	0.9807	0.0057	0.9545	−0.0205	0.9352	−0.0148	0.2937	0.2984	−0.0047
0.6	26	0.9745	−0.0005	0.9632	−0.0118	0.9377	−0.0123	0.2966	0.2992	−0.0026
0.7	20	0.9707	−0.0043	0.9701	−0.0049	0.9408	−0.0092	0.2960	0.2936	0.0024
0.8	12	0.9608	−0.0142	0.9724	−0.0026	0.9332	−0.0168	0.3042	0.2889	0.0153
0.9	5	0.9501	**−0.0249**	0.9739	**−0.0011**	0.9240	**−0.0260**	0.3476	0.2659	0.0817

enhance the illustration, the errors between the simulated expected width and estimated expected width of the exact approach and the approximate method in Table 4 are plotted in Fig. 1 as a supplement. Although the simple sample size formula was generally effective for attaining the expected width requirement, the overall coverage of the approximate confidence intervals may not be convincing enough for making sound applications. In short, these detailed appraisals confirmed that the exact interval estimation and sample size procedures are superior to the approximate techniques in terms of coverage probability performance and expected width precision for all situations even if the number of groups is substantially large.

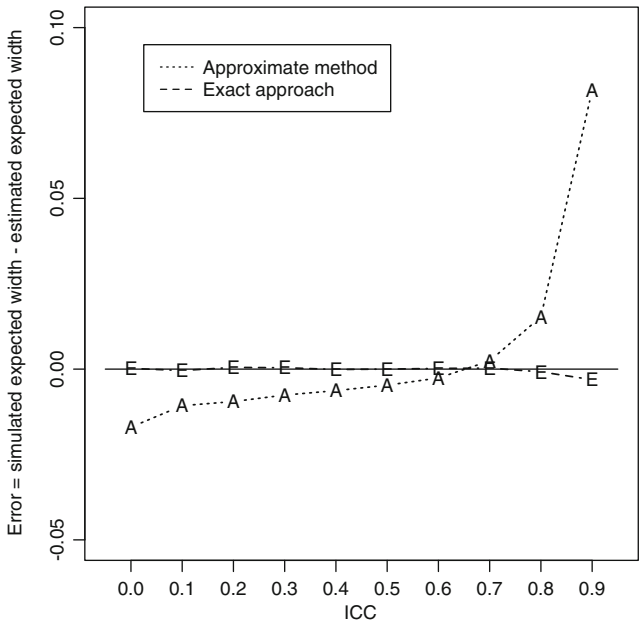

Fig. 1 The performance of expected width with $N = 10$ and $\delta = 0.3$

Table 5 Number of subjects per group of the exact approach for $1 - \alpha = 0.95$, expected width $\delta = 0.2$, and the number of groups $G = 20$, 40, 60, 80, and 100

G	20	40	60	80	100
ρ	N	N	N	N	N
0	8	6	5	4	4
0.1	16	8	6	5	4
0.2	>2,000	12	7	5	5
0.3	>2,000	41	10	6	5
0.4	>2,000	>2,000	13	7	5
0.5	>2,000	>2,000	13	6	4
0.6	>2,000	>2,000	8	4	3
0.7	>2,000	12	4	3	3
0.8	>2,000	3	2	2	2
0.9	3	2	2	2	2

Design II: the number of groups is specified

An alternative setting is to find the optimal sample size when the number of groups G is fixed in advance. Hence, it reduces the problem to the determination of the required number of subjects N per group to achieve the desired expected width. Basically, the prescribed exact and iterative algorithm can be modified to compute the minimum sample size N such that the expected width of the exact confidence interval $E[W_E] \leq \delta$ for the desired bound δ, selected size G, confidence level $(1 - \alpha)$, and target ICC value ρ. In this case, for the approximate confidence intervals, the particular form of Fisher's asymptotic variance estimator ν^2 does not permit a closed-form formula for calculating the necessary number of subjects N in each group. It appears that this setup has not been considered in the literature, including Bonett (2002) and Giraudeau and Mary (2001).

For demonstration, the optimal sample sizes N in each group are listed in Table 5 for $1 - \alpha = 0.95$, $\delta = 0.2$, $G = 20$, 40, 60, 80, and 100, and $\rho = 0$–0.9 with an increment of 0.1. Since the number of groups G plays an essential role in the evaluation of expected width, the search of the optimal value N does not always give practically useful results when G is relatively small. Accordingly, the computation was terminated when $N > 2{,}000$, as noted in Table 5. Despite these incomplete results, it is seen that the computed sample size N decreases with an increasing value of G and is a concave function of ρ when all other factors are fixed. As a cautionary note, the findings suggest that it requires a careful examination of the design structure when the number of groups is less than 40 and the population ρ is in the neighborhood of 0.5. Without a detailed appraisal, one may unknowingly conduct a study with an underestimated sample size, which leads to the undesirable consequences of inadequate precision performance and an unsatisfactory research outcome.

Design III: total cost is fixed and the expected width needs to be minimized

To assess the cost of a reliability study, Eliasziw and Donner (1987) considered the following linear cost function:

$$C = C_O + C_G G + C_N N + C_{GN} GN, \tag{11}$$

where C_O is the overhead cost for the study, C_G reflects costs proportional to the number of groups, C_N denotes costs proportional to the number of subjects per group, and C_{GN} stands for costs proportional to both the number of groups and the number of subjects per group. Their focus, however, was on the cost and power issues of a hypothesis-testing procedure. Instead, the precision of confidence intervals is examined here under the cost setup. Accordingly, for a fixed total cost, a problem of practical interest is to decide the optimal design in terms of (G, N) in order to have the narrowest expected width of a confidence interval.

Table 6 Optimal sample sizes (G, N) and estimated expected widths of the exact procedure when the maximum number of subjects is $C = 300$ for $1 - \alpha = 0.95$

ρ	G	N	GN	Estimated Expected Width
0	10	30	300	0.0894
0.1	30	10	300	0.1924
0.2	50	6	300	0.2299
0.3	75	4	300	0.2437
0.4	75	4	300	0.2416
0.5	100	3	300	0.2252
0.6	100	3	300	0.1991
0.7	100	3	300	0.1640
0.8	150	2	300	0.1175
0.9	150	2	300	0.0625

In view of the discrete nature of sample sizes, the optimal solution can be found through a screening of a finite number of (G, N) combinations that attain the minimum expected width subject to the cost constraint. First, when $N = 2$, the maximum number of groups G_{max} is computed by $G_{max} = Floor\{(C - C_O - 2C_N)/(C_G + 2C_{GN})\}$ for the specified total cost C and cost coefficients (C_O, C_G, C_N, C_{GN}), where the function $Floor(a)$ returns the largest integer that is less than or equal to a. Then detailed expected width calculations and comparisons are performed for the sample size combinations (G, N) with $N = Floor\{(C - C_O - C_G G)/(C_N + C_{GN} G)\}$ for $G = 2$ to G_{max}. Ultimately, the optimal sample size allocation is the one giving the least expected width. It is noteworthy that the cost function in Eq. 11 reduces to $C = GN$ with $C_O = C_G = C_N = 0$ and $C_{GN} = 1$. Therefore, the consideration of a fixed total number of subjects in Giraudeau and Mary (2001) is a special case of the fixed total cost framework. Conceivably,

Table 7 Optimal sample sizes (G, N), and total number of subjects of the exact procedure when the total number of subjects needs to be minimized for $1 - \alpha = 0.95$ and expected width $\delta = 0.2$

ρ	G	N	GN	Estimated Expected Width
0	9	14	126	0.1976
0.1	31	9	279	0.1993
0.2	66	6	396	0.1999
0.3	112	4	448	0.2000
0.4	111	4	444	0.1992
0.5	128	3	384	0.1992
0.6	100	3	300	0.1991
0.7	68	3	204	0.1994
0.8	56	2	112	0.1983
0.9	21	2	42	0.1962

Table 8 Sample size, coverage, and precision of the exact and approximate procedures for $1 - \alpha = 0.95$, $N = 5$, $P(W \leq \omega) = 0.9$, and $\omega = 0.2$

ρ	G	Upper 97.5 %	Error	Lower 97.5 %	Error	Two-sided 95 %	Error	Simulated $E[W]$	Estimated $E[W]$	Error
Exact approach										
0	55	0.9767	0.0017	0.9723	−0.0027	0.9490	−0.0010	0.9123	0.9157	−0.0034
0.1	75	0.9771	0.0021	0.9733	−0.0017	0.9504	0.0004	0.9152	0.9125	0.0027
0.2	89	0.9724	−0.0026	0.9757	0.0007	0.9481	−0.0019	0.9160	0.9203	−0.0043
0.3	94	0.9786	0.0036	0.9765	0.0015	0.9551	0.0051	1.0000	1.0000	0.0000
0.4	94	0.9742	−0.0008	0.9754	0.0004	0.9496	−0.0004	1.0000	1.0000	0.0000
0.5	93	0.9772	0.0022	0.9758	0.0008	0.9530	0.0030	0.9476	0.9492	−0.0016
0.6	82	0.9771	0.0021	0.9741	−0.0009	0.9512	0.0012	0.9017	0.9036	−0.0019
0.7	65	0.9740	−0.0010	0.9738	−0.0012	0.9478	−0.0022	0.9122	0.9159	−0.0037
0.8	42	0.9758	0.0008	0.9720	−0.0030	0.9478	−0.0022	0.8989	0.9046	−0.0057
0.9	20	0.9746	−0.0004	0.9741	−0.0009	0.9487	−0.0013	0.9133	0.9128	0.0005
Approximate method										
0	54	0.9898	0.0148	0.9440	−0.0310	0.9338	−0.0162	0.9200	0.9099	0.0101
0.1	76	0.9884	0.0134	0.9561	−0.0189	0.9445	−0.0055	0.9416	0.9125	0.0291
0.2	91	0.9850	0.0100	0.9609	−0.0141	0.9459	−0.0041	0.9685	0.9113	0.0572
0.3	97	0.9769	0.0019	0.9659	−0.0091	0.9428	−0.0072	1.0000	0.9146	0.0854
0.4	97	0.9763	0.0013	0.9658	−0.0092	0.9421	−0.0079	1.0000	0.9807	0.0193
0.5	95	0.9748	−0.0002	0.9671	−0.0079	0.9419	−0.0081	1.0000	0.9056	**0.0944**
0.6	84	0.9708	−0.0042	0.9747	−0.0003	0.9455	−0.0045	0.9202	0.9021	0.0181
0.7	65	0.9684	−0.0066	0.9780	0.0030	0.9464	−0.0036	0.8968	0.9009	−0.0041
0.8	42	0.9633	−0.0117	0.9816	0.0066	0.9449	−0.0051	0.8906	0.9145	−0.0239
0.9	18	0.9515	−0.0235	0.9924	0.0174	0.9439	−0.0061	0.8567	0.9170	−0.0603

Table 9 Sample size, coverage, and precision of the exact and approximate procedures for $1 - \alpha = 0.95$, $N = 10$, $P(W \leq \omega) = 0.9$, and $\omega = 0.2$

ρ	G	Upper 97.5 %	Error	Lower 97.5 %	Error	Two-sided 95 %	Error	Simulated $E[W]$	Estimated $E[W]$	Error
Exact approach										
0	20	0.9739	−0.0011	0.9755	0.0005	0.9494	−0.0006	0.8950	0.9009	−0.0059
0.1	38	0.9743	−0.0007	0.9717	−0.0033	0.9460	−0.0040	0.8984	0.9019	−0.0035
0.2	54	0.9737	−0.0013	0.9752	0.0002	0.9489	−0.0011	0.9045	0.9068	−0.0023
0.3	64	0.9727	−0.0023	0.9748	−0.0002	0.9475	−0.0025	0.9055	0.9101	−0.0046
0.4	67	0.9769	0.0019	0.9774	0.0024	0.9543	0.0043	1.0000	1.0000	0.0000
0.5	67	0.9743	−0.0007	0.9747	−0.0003	0.9490	−0.0010	1.0000	1.0000	0.0000
0.6	63	0.9759	0.0009	0.9733	−0.0017	0.9492	−0.0008	0.9141	0.9140	0.0001
0.7	52	0.9725	−0.0025	0.9734	−0.0016	0.9459	−0.0041	0.9098	0.9132	−0.0034
0.8	35	0.9746	−0.0004	0.9741	−0.0009	0.9487	−0.0013	0.8996	0.9037	−0.0041
0.9	17	0.9750	0.0000	0.9736	−0.0014	0.9486	−0.0014	0.9088	0.9058	0.0030
Approximate method										
0	18	0.9989	0.0239	0.9048	−0.0702	0.9037	−0.0463	0.9227	0.9224	0.0003
0.1	37	0.9945	**0.0195**	0.9416	**−0.0334**	0.9361	−0.0139	0.9245	0.9039	0.0206
0.2	55	0.9877	0.0127	0.9482	−0.0268	0.9359	−0.0141	0.9476	0.9156	0.0320
0.3	66	0.9843	0.0093	0.9553	−0.0197	0.9396	−0.0104	0.9835	0.9132	0.0703
0.4	69	0.9816	0.0066	0.9609	−0.0141	0.9425	−0.0075	1.0000	0.9409	0.0591
0.5	69	0.9781	0.0031	0.9652	−0.0098	0.9433	−0.0067	1.0000	0.9272	**0.0728**
0.6	65	0.9730	−0.0020	0.9704	−0.0046	0.9434	−0.0066	0.9538	0.9088	0.0450
0.7	53	0.9723	−0.0027	0.9724	−0.0026	0.9447	−0.0053	0.9128	0.9091	0.0037
0.8	35	0.9682	−0.0068	0.9798	0.0048	0.9480	−0.0020	0.8815	0.9075	−0.0260
0.9	16	0.9592	−0.0158	0.9888	0.0138	0.9480	−0.0020	0.8616	0.9260	−0.0644

the proposed algorithm is more efficient than their graphical displays for determining the correct optimal sample size. Also, it is important not to overlook the fundamental distinction between the suggested exact methods and their approximate procedures. To illustrate the suggested algorithm for optimal sample size determination when the total number of subjects is fixed or a cost function with $C_O = C_G = C_N = 0$ and $C_{GN} = 1$, the optimal combination (G, N) and achieved expected width $E[W_E]$ are given in Table 6 for $1 - \alpha = 0.95$, $C = 300$, and $\rho =$ 0–0.9 with an increment of 0.1. It is interesting to see that the number of groups G and the number of subjects per group N of the best allocation set increases and decreases with an increasing value of ρ, respectively. Whereas the estimated expected width is a concave function of ρ with a maximum 0.2437 around $\rho = 0.3$.

Design IV: target expected width is fixed and the total cost needs to be minimized

Another cost-related issue is to find the optimal sample size combination to meet a specified expected width requirement for the least cost. This problem is actually more involved than the previous one and was not considered in Giraudeau and Mary (2001). The search of the optimal result needs to synthesize the two procedures of determining the ideal G and N when the other term is given as described in the design settings I and II, respectively. The suggested procedure is conducted in three steps. First, the previous algorithm is applied to find the optimal number of groups G_{max} needed to achieve the desired precision with expected width δ for the specified coverage probability $1 - \alpha$, parameter value ρ, and the number of subjects per group $N = 2$. Then, a sequence of sample size calculations are performed to determine the optimal number of subjects per group, denoted by ($N_2, \ldots, N_{G\max - 1}$), required to meet the target expected width δ for the specified coverage probability $1 - \alpha$, parameter value ρ, and $G = 2$ to ($G_{max} - 1$). This is a direct application of the exact approach presented in the second design setting. In the third and last stage, the optimal solution (G^*, N^*) is the pair of values (G, N) giving the smallest cost for all combinations $(G, N) = \{(2, N_2), (3, N_3), \ldots, (G_{max} - 1, NG_{max - 1}), (G_{max}, 2)\}$. However, there may be more than one combination giving the same amount of least cost. A further screening and selection process is conducted to find the one producing the narrowest expected width. Clearly, a computer program is more efficient than a graphical chart for determining the necessary outcome.

The corresponding optimal evaluations are exemplified in Table 7 when the total number of subjects GN needs to be minimized with $1 - \alpha = 0.95$, $\delta = 0.20$, and $\rho = 0$–0.9 with an

Table 10 Sample size, coverage, and precision of the exact and approximate procedures for $1 - \alpha = 0.95$, $N = 5$, $P(W \le \omega) = 0.9$, and $\omega = 0.3$

ρ	G	Upper 97.5 %	Error	Lower 97.5 %	Error	Two-sided 95 %	Error	Simulated $E[W]$	Estimated $E[W]$	Error
Exact approach										
0	28	0.9755	0.0005	0.9765	0.0015	0.9520	0.0020	0.9106	0.9131	−0.0025
0.1	36	0.9765	0.0015	0.9762	0.0012	0.9527	0.0027	0.9173	0.9103	0.0070
0.2	41	0.9753	0.0003	0.9750	0.0000	0.9503	0.0003	0.9283	0.9273	0.0010
0.3	42	0.9749	−0.0001	0.9747	−0.0003	0.9496	−0.0004	1.0000	1.0000	0.0000
0.4	42	0.9774	0.0024	0.9748	−0.0002	0.9522	0.0022	1.0000	1.0000	0.0000
0.5	42	0.9745	−0.0005	0.9754	0.0004	0.9499	−0.0001	1.0000	1.0000	0.0000
0.6	39	0.9731	−0.0019	0.9732	−0.0018	0.9463	−0.0037	0.9326	0.9302	0.0024
0.7	32	0.9767	0.0017	0.9738	−0.0012	0.9505	0.0005	0.9196	0.9193	0.0003
0.8	22	0.9736	−0.0014	0.9750	0.0000	0.9486	−0.0014	0.9065	0.9057	0.0008
0.9	12	0.9764	0.0014	0.9747	−0.0003	0.9511	0.0011	0.9140	0.9151	−0.0011
Approximate method										
0	28	0.9938	0.0188	0.9305	−0.0445	0.9243	−0.0257	0.9398	0.9208	0.0190
0.1	37	0.9889	0.0139	0.9409	−0.0341	0.9298	−0.0202	0.9487	0.9057	0.0430
0.2	43	0.9835	0.0085	0.9563	−0.0187	0.9398	−0.0102	1.0000	0.9042	**0.0958**
0.3	45	0.9795	0.0045	0.9579	−0.0171	0.9374	−0.0126	1.0000	0.9310	0.0690
0.4	44	0.9760	0.0010	0.9625	−0.0125	0.9385	−0.0115	1.0000	0.9636	0.0364
0.5	45	0.9730	−0.0020	0.9661	−0.0089	0.9391	−0.0109	1.0000	0.9297	0.0703
0.6	41	0.9678	−0.0072	0.9717	−0.0033	0.9395	−0.0105	0.9586	0.9184	0.0402
0.7	33	0.9656	−0.0094	0.9792	0.0042	0.9448	−0.0052	0.9157	0.9166	−0.0009
0.8	22	0.9587	−0.0163	0.9847	0.0097	0.9434	−0.0066	0.8800	0.9139	−0.0339
0.9	10	0.9473	−0.0277	0.9923	0.0173	0.9396	−0.0104	0.8198	0.9043	−0.0845

increment of 0.1. In this case, the number of subjects per group of the optimal combination decreases with an increasing value of ρ. On the other hand, the number of groups and the total number of subjects are concave functions of ρ. Specifically, the required total number of subjects $GN = 448$ for $\rho = 0.3$ confirms that the specified magnitude $C = GN = 300$ presented in the preceding empirical illustration in Table 6 is not large enough to ensure the designated precision level with $\delta = 0.20$.

Assurance probability criterion

In addition to the expected width criterion, a useful alternative approach for sample size determination is to ensure that the actual confidence interval width will not exceed the planned bound with a given assurance probability. As in the case of expected width, this precision principle is explicated in the following four design schemes with different allocation and cost concerns.

Design I: the number of subjects per group is fixed

With a preassigned number of subjects in each group, the assurance probability $P\{W_E \le \omega\}$ for the width of a 100(1 – α)% exact two-sided confidence interval to not exceed the planned value ω is a one-to-one function of the number of groups when the target ICC value ρ remains constant. The exact value $P\{W_E \le \omega\}$ can be computed by numerical integration with respect to a beta probability density function. Then it is straightforward to find the minimum sample size G such that

$$P(W_E \le \omega) \ge 1-\gamma \tag{12}$$

through an iterative search where $(1 - \gamma)$ is the specified assurance level and ω (> 0) is a width bound. Due to the involved computation of the exact assurance probability, a computer algorithm similar to the one in the expected width criterion was developed.

Along the same line of the approximation perspective of Bonett (2002) and Giraudeau and Mary (2001), Zou (2012) proposed that the minimum sample size G such that $P(W_A \le \omega) \ge 1 - \gamma$ is identical to the smallest integer G that satisfies the inequality

$$G \ge \frac{\left\{[f(\rho)]^{1/2} + \left[f(\rho) + 2z_\gamma \omega f'(\rho)/z_{\alpha/2}\right]^{1/2}\right\}^2}{4\omega^2/z^2_{\alpha/2}} + 1, \tag{13}$$

Table 11 Sample size, coverage, and precision of the exact and approximate procedures for $1 - \alpha = 0.95$, $N = 10$, $P(W \le \omega) = 0.9$, and $\omega = 0.3$

ρ	G	Upper 97.5 %	Error	Lower 97.5 %	Error	Two-sided 95 %	Error	Simulated $E[W]$	Estimated $E[W]$	Error
Exact approach										
0	12	0.9765	0.0015	0.9753	0.0003	0.9518	0.0018	0.9008	0.9013	−0.0005
0.1	20	0.9731	−0.0019	0.9757	0.0007	0.9488	−0.0012	0.8999	0.9050	−0.0051
0.2	27	0.9756	0.0006	0.9761	0.0011	0.9517	0.0017	0.9472	0.9445	0.0027
0.3	30	0.9774	0.0024	0.9749	−0.0001	0.9523	0.0023	1.0000	1.0000	0.0000
0.4	30	0.9766	0.0016	0.9744	−0.0006	0.9510	0.0010	1.0000	1.0000	0.0000
0.5	30	0.9744	−0.0006	0.9735	−0.0015	0.9479	−0.0021	1.0000	1.0000	0.0000
0.6	29	0.9773	0.0023	0.9711	−0.0039	0.9484	−0.0016	0.9116	0.9102	0.0014
0.7	25	0.9743	−0.0007	0.9757	0.0007	0.9500	0.0000	0.9071	0.9043	0.0028
0.8	19	0.9763	0.0013	0.9727	−0.0023	0.9490	−0.0010	0.9259	0.9273	−0.0014
0.9	10	0.9755	0.0005	0.9760	0.0010	0.9515	0.0015	0.9002	0.9019	−0.0017
Approximate method										
0	10	0.9996	0.0246	0.8741	−0.1009	0.8737	−0.0763	0.9196	0.9129	0.0067
0.1	20	0.9960	0.0210	0.9190	−0.0560	0.9150	−0.0350	0.9559	0.9273	0.0286
0.2	27	0.9910	0.0160	0.9373	−0.0377	0.9283	−0.0217	0.9556	0.9051	0.0505
0.3	32	0.9875	0.0125	0.9504	−0.0246	0.9379	−0.0121	1.0000	0.9317	0.0683
0.4	32	0.9825	0.0075	0.9523	−0.0227	0.9348	−0.0152	1.0000	0.9521	0.0479
0.5	32	0.9779	0.0029	0.9595	−0.0155	0.9374	−0.0126	1.0000	0.9296	0.0704
0.6	31	0.9725	−0.0025	0.9643	−0.0107	0.9368	−0.0132	1.0000	0.9017	**0.0983**
0.7	27	0.9702	−0.0048	0.9685	−0.0065	0.9387	−0.0113	0.9360	0.9268	0.0092
0.8	19	0.9634	−0.0116	0.9770	0.0020	0.9404	−0.0096	0.8962	0.9277	−0.0315
0.9	9	0.9528	−0.0222	0.9854	0.0104	0.9382	−0.0118	0.8271	0.9160	−0.0889

where $f(\rho)=2(1-\rho)^2[1+(N-1)\rho]^2/[N(N-1)]$ and $f'(\rho)=\{4(1-\rho)[1+(N-1)\rho]|N-2+2\rho-2N\rho|\}/[N(N-1)]$. It was noted in Zou that Eq. 13 reduces to Eq. 10 for $\gamma = 0.5$ and $\omega = \delta$. Hence, the smallest sample size that suffices to have an expected width $E[W_A] \le \omega$ is the same as that to obtain an assurance probability $P(W_A \le \omega) \ge 0.5$. However, the approximate formula in Eq. 13 may not provide satisfactory results even when the number of groups is larger than 30. For example, when $\rho = 0.6$, $\omega = 0.2$ and $1 - \gamma = 0.90$, the simulated assurance probabilities reported in Table 1 of Zou are 0.9084, 0.9202, 0.9454, and 0.9768 for $(G, N) = (196, 2)$, (120, 3), (84, 5), and (65, 10), respectively. It follows from the asymptotic normal approximation that the corresponding attained or estimated assurance probabilities are 0.9062, 0.9012, 0.9021, and 0.9088, with the absolute errors 0.0022, 0.0190, 0.0433, and 0.0680, respectively. Moreover, it is noteworthy that the differences between the simulated assurance probability and estimated assurance probability for $\omega = 0.3$ are more prominent because the approximation deteriorates for smaller values of G.

To further demonstrate the discrepancy between the two sample size procedures in Eqs. 12 and 13, extensive examinations were conducted for a wide range of model settings. Specifically, the model configurations were chosen with the assurance level $1 - \gamma = 0.90$, the interval bound $\omega = 0.2$ and 0.3, the number of subjects per group $N = 2, 3, 5, 10, 20$, and the population ICC value $\rho = 0$–0.90 with an increment of 0.1. Following the same notion in the study of expected width, the numerical investigations included sample size calculations, simulated one- and two-sided coverage probabilities, simulated assurance probabilities, estimated assurance probability, and associated errors. For brevity, Tables 8 and 9 present the computed sample sizes for $\omega = 0.2$ with $N = 5$ and 10, respectively. Also, the empirical results associated with $\omega = 0.3$ are illustrated in Tables 10 and 11 for $N = 5$ and 10, respectively.

According to the computed sample size G presented in Tables 1, 2, 3, 4, 8, 9, 10, and 11 for the same value of N, it often requires a larger sample size to meet the necessary precision of assurance probability than the control of a designated expected width. The pattern of results between the two precision principles is similar to those reported in Kupper and Hafner (1989) and Shieh and Jan (2012). More important, the coverage behavior of the exact confidence intervals and the assurance performance of the exact sample size procedure maintain their excellence for all the model configurations examined here. In contrast, the approximate and equidistant interval procedure still demonstrates the same disadvantage of over- and underestimated one-sided coverage probabilities for small ICC values. For example, in the case of $\rho = 0.1$, $G = 37$, $N = 10$ in Table 9, the resulting errors of the upper and lower 97.5 % confidence intervals of 0.0195 and −0.0334 suggest that the two confidence limits are both smaller than the respective exact value. Regarding the accuracy of the approximate sample size formula given in Eq. 13, the procedure does not always guarantee the precision performance for the examined cases with varied combinations of ρ, G, and N. Although there are some absolute errors within a small range of 0.01 or 0.02, the incurred largest absolute errors are 0.0944, 0.0728, 0.0958, and 0.0983 for $(\rho, G) = (0.5, 95)$, (0.5, 69), (0.2, 43), and (0.6, 31) in Tables 8, 9, 10, and 11, respectively. For illustration, the errors between the simulated assurance probability and estimated assurance probability of the exact approach and the approximate method in Table 11 are plotted in Fig. 2 as a supplement. Hence, the accuracy of the simple

Table 12 Number of subjects per group of the exact approach for $1 - \alpha = 0.95$, $P(W \le 0.2) = 0.9$, and the number of groups $G = 20$, 40, 60, 80, and 100

G	20	40	60	80	100
ρ	N	N	N	N	N
0	10	7	5	5	4
0.1	45	10	7	5	5
0.2	>2,000	22	9	6	5
0.3	>2,000	>2,000	12	7	5
0.4	>2,000	>2,000	15	7	5
0.5	>2,000	>2,000	15	7	5
0.6	>2,000	>2,000	12	6	4
0.7	>2,000	>2,000	6	4	3
0.8	>2,000	6	3	3	2
0.9	5	2	2	2	2

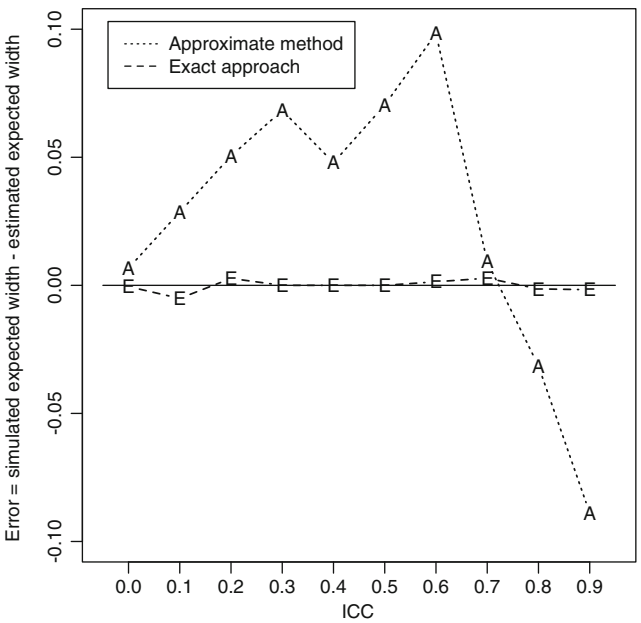

Fig. 2 The performance of assurance probability with $N = 10$, $1 - \gamma = 0.9$, and $\omega = 0.3$

formula is inconsistent. These findings demonstrate the shortcoming of the existing method of Zou (2012) and the adequacy of the presented exact sample size technique.

Design II: the number of groups is specified

The exact computation of assurance probability can be readily modified for determining the necessary number of subjects to attain the desired assurance probability when the number of groups is already decided. The exemplified sample size calculations of this design setting are demonstrated in Table 12 for $1 - \alpha = 0.95$, $P(W \leq 0.2) = 0.9$, the number of groups $G = 20$, 40, 60, 80, and 100, and $\rho = 0$–0.9 with an increment of 0.1. Essentially, the computed numbers of subjects per group are larger than those in Table 5, but they still possess the same relationship with ρ and G as in the expected width criterion.

Design III: total cost is fixed and the assurance probability needs to be maximized

To extend the applicability of the suggested exact confidence intervals and assurance probability principle, the general cost function defined in Eq. 11 can be employed to determine the optimal sample size (G, N) with fixed total cost. It is straightforward to adjust the proposed algorithm for the minimization of expected width to the maximization of assurance probability. Hence, the computations require only some minor changes of computer code and do not involve any extra complexity and effort. For ease of comparison with the outcomes of the expected width criterion, Table 13 shows the optimal sample sizes (G, N) and estimated assurance probability of the exact procedure when the maximum number of subjects is $C = 300$ for $1 - \alpha = 0.95$, interval bound $\omega = 0.2$, and $\rho = 0$–0.9 with an increment of 0.1. It is interesting to note that the optimal combinations (G, N) are not necessarily identical even when the total costs are fixed as 300 in both precision settings.

Table 13 Optimal sample sizes (G, N) and estimated assurance probability of the exact procedure when the maximum number of subjects is $C = 300$ for $1 - \alpha = 0.95$, and interval bound $\omega = 0.2$

ρ	G	N	GN	Estimated Assurance Probability
0	25	12	300	0.9994
0.1	37	8	296	0.5873
0.2	30	10	300	0.0656
0.3	3	100	300	0.0147
0.4	3	100	300	0.0083
0.5	100	3	300	0.0341
0.6	100	3	300	0.5131
0.7	100	3	300	0.9835
0.8	100	3	300	1.0000
0.9	148	2	296	1.0000

Table 14 Optimal sample sizes (G, N), and total number of subjects of the exact procedure when the total number of subjects needs to be minimized for $1 - \alpha = 0.95$ and assurance probability $P(W \leq 0.2) = 0.9$

ρ	G	N	GN	Estimated Assurance Probability
0	13	15	195	0.9055
0.1	52	7	364	0.9119
0.2	89	5	445	0.9208
0.3	114	4	456	1.0000
0.4	114	4	456	1.0000
0.5	143	3	429	0.9065
0.6	119	3	357	0.9092
0.7	88	3	264	0.9052
0.8	81	2	162	0.9021
0.9	35	2	70	0.9067

Design IV: target assurance probability is fixed and the total cost needs to be minimized

When the assurance performance is set at a given level, it is of practical value to adopt the optimal design with the least cost. The optimal (G, N) solution can be obtained from the prescribed three-step procedure by replacing the expected width calculation with the evaluation of assurance probability. The usefulness of this design scheme is explicated for the special concern that the overall cost is the total number of subjects. Accordingly, Table 14 contains the optimal sample sizes (G, N) and total number of subjects for the exact 95 % interval procedure to have the assurance probability $P(W \leq 0.2) \geq .9$ when the total number of subjects needs to be minimized. As was expected, the required total numbers of subjects or cost is substantially greater than those with similar configurations in Table 7 for the expected width principle.

Conclusions

For advance design of reliability studies, instead of conducting hypothesis testing with sufficient power, an alternative way to plan a study is to control the precision of the confidence interval. Within the context of a one-way random effects model, a variety of approximate confidence intervals of ICC have been proposed in the literature. Special attention has been focused on the confidence interval constructed with the asymptotic normality and Fisher's variance estimator of ICC(1). Although its closed-form expression is easy to apply and permits a simple derivation of explicit sample size formulas for some design structures, the desirable properties of coverage probability and interval width remain the major and decisive factors for selecting an interval

procedure. Detailed numerical examinations were made to evaluate the coverage behavior of the approximate interval method and the overall performance of the associated sample size techniques. The comprehensive results showed that the trade of high accuracy for computational simplicity may not always be a wise bargain. According to these findings, the lack of reliable and extended sample size methods for different design schemes impedes the approximate interval procedure as a well-founded method for practical applications. In order to facilitate the application of the exact confidence intervals, exact sample size procedures are developed for various allocation and cost design schemes under both the expected width and assurance probability criteria. In addition, computer programs are presented to aid the usefulness and implementation of the proposed techniques for the optimal design of reliability studies.

Acknowledgments The author wishes to express his gratitude to the editor, Gregory Francis, and the two anonymous reviewers for their constructive suggestions.

References

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.

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

Bartko, J. J. (1976). On various intraclass correlation reliability. *Psychological Bulletin, 83,* 762–765.

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine, 21,* 1331–1335.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge/Taylor & Francis Group.

Donner, A. D. (1999). Sample size requirements for interval estimation of the intraclass Kappa statistic. *Communications in Statistics-Simulation & Computation, 28,* 415–429.

Donner, A., & Koval, J. J. (1983). A note on the accuracy of Fisher's approximation to the large sample variance of an intraclass correlation. *Communications in Statistics-Simulation & Computation, 12,* 443–449.

Donner, A., & Wells, G. (1986). A comparison of confidence interval methods for the intraclass correlation coefficient. *Biometrics, 42,* 401–412.

Dunst, C. J., & Hamby, D. W. (2012). Guide for calculating and interpreting effect sizes and confidence intervals in intellectual and developmental disability research studies. *Journal of Intellectual & Developmental Disability, 37,* 89–99.

Eliasziw, M., & Donner, A. (1987). A cost-function approach to the design of reliability studies. *Statistics in Medicine, 6,* 647–655.

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

Fisher, R. A. (1938). *Statistical methods for research workers* (7th ed.). Edinburgh: Oliver and Boyd.

Flynn, T. N., Whitley, E., & Peters, T. J. (2002). Recruitment strategies in a cluster randomized trial: Cost implications. *Statistics in Medicine, 21,* 397–405.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141,* 2–18.

Giraudeau, B., & Mary, J. Y. (2001). Planning a reproducibility study: How many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Statistics in Medicine, 20,* 3205–3214.

Goldstein, H. (2002). *Multilevel statistical models* (3rd ed.). New York: Wiley.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York: Routledge/Taylor & Francis Group.

Hofmann, D. A. (2002). Issues in multilevel research: Theory development, measurement and analysis. In S. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 247–274). New York: Blackwell.

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

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46.

Odgaard, E. C., & Fowler, R. L. (2010). Confidence intervals for effect sizes: Compliance and clinical significance in the Journal of Consulting and Clinical Psychology. *Journal of Consulting and Clinical Psychology, 78,* 287–297.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage.

R Development Core Team. (2013). R: A language and environment for statistical computing [Computer software and manual]. Retrieved from http://www.r-project.org

Robey, R. R. (2004). Reporting point and interval estimates of effect-size foe planned contrasts: Fixed within effect analyses of variance. *Journal of Fluency Disorders, 29,* 307–341.

SAS Institute. (2012). *SAS/IML user's guide, version 9.3*. Cary: SAS Institute Inc.

Shieh, G., & Jan, S. L. (2012). Optimal sample sizes for precise interval estimation of Welch's procedure under various allocation and cost considerations. *Behavior Research Methods, 44,* 202–212.

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: Review and new results. *Statistical Methods in Medical Research, 13,* 251–271.

Shoukri, M. M., Asyali, M. H., & Walter, S. D. (2003). Issues of cost and efficiency in the design of reliability studies. *Biometrics, 59,* 1107–1112.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 102,* 989–1004.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44,* 423–432.

Ukoumunne, O. C. (2002). A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Statistics in Medicine, 21,* 3757–3774.

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

Zou, G. Y. (2012). Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in Medicine, 31,* 3972–3981.