Statistics
in Medicine

Research Article

Received 19 November 2011, Accepted 15 March 2012 Published online 4 July 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5466

Sample size formulas for estimating intraclass correlation coefficients with precision and assurance

G. Y. Zou[*†]

The number of subjects required to estimate the intraclass correlation coefficient in a reliability study has usually been determined on the basis of the expected width of a confidence interval. However, this approach fails to explicitly consider the probability of achieving the desired interval width and may thus provide sample sizes that are too small to have adequate chance of achieving the desired precision. In this paper, we present a method that explicitly incorporates a prespecified probability of achieving the prespecified width or lower limit of a confidence interval. The resultant closed-form formulas are shown to be very accurate. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: agreement; confidence interval; interrater; intrarater; measurements; reliability; reproducibility

1. Introduction

Unreliable measurements can have serious consequences in medical research. Reliability studies are therefore often conducted to assess the metric properties of observers or test methods in terms of reproducibility and level of agreement. For quantitative observations, the intraclass correlation coefficient (ICC), defined as the proportion of all variation that is not due to measurement error, is a widely used index for quantifying the extent of the reliability [1–3]. The interpretation of the ICC, denoted here as ρ, is often facilitated using the Landis and Koch [4] benchmarks that $\rho < 0$ reflects 'poor' reliability, 0 to 0.20 'slight', 0.21 to 0.4 'fair', 0.41 to 0.60 'moderate', 0.61 to 0.8 'substantial', and above 0.81 'almost perfect' reliability.

As with any scientific investigation, an essential step in conducting a reliability study is the calculation of a minimum sample size that will meet the given objective of the study. A study that is too large may waste resources, whereas a study that is too small will have little chance of meeting the study objective. Two general approaches have been adopted to estimate the sample size for reliability studies. From a hypothesis testing perspective, Donner and Eliasziw [5] have provided power graphs for selected numbers of raters and subjects. To broaden the application and simplify calculations involving numerical integration of incomplete beta functions, Walter *et al.* [6] provided a closed-form approximation to these results on the basis of the Fisher transformation.

The second approach to sample size estimation focuses on the precision of estimating the ICC, as quantified by the expected width of the confidence interval [7, 8]. This approach seems to be consistent with the current trend towards using confidence intervals to present study results. However, it is well known that methods based on expected interval width can underestimate the sample size needed to have an adequate chance of achieving the desired precision [9–12]. Sample size estimation procedures that incorporate assurance probability for estimating normal means have been proposed [11] and implemented in standard statistical procedures such as SAS 9.2 proc power (SAS Institute, Cary, NC, USA). Additional examples include sample sizes for odds ratios [13, 14], multiple comparisons of normal

Department of Epidemiology and Biostatistics and Robarts Clinical Trials of Robarts Research Institute, Schulich School of Medicine & Dentistry, University of Western Ontario, London, ON, Canada
**Correspondence to: G. Y. Zou, Department of Epidemiology and Biostatistics, Schulich School of Medicine & Dentistry, University of Western Ontario, London, ON, Canada N6A 5C1.*
†E-mail: gzou@robarts.ca

means [15], and single proportions [16]. The lack of a closed-form expression is a common feature of these procedures.

The purpose of this paper is to derive and evaluate simple sample size formulas for planning reliability studies focusing on the estimation of the ICC. In contrast to approaches currently available in the literature [7, 8], we explicitly incorporate a prespecified assurance probability of achieving the desired precision. It is shown that our results reduce to those obtained by Bonett [8] when the chance of achieving the interval width is set to 50%. We also show that the sample size formula based on the lower confidence limit is identical to that obtained using the approach of Walter *et al.* [6] when power is regarded as the assurance probability.

Section 2 first describes the model for point and confidence interval estimation of the ICC and then derives sample size formulas on the basis of the desired width and lower limit of the confidence interval for ρ, respectively. Section 3 reports on a simulation study evaluating the performance of the formulas. In particular, we first use the derived formulas to obtain sample sizes needed to achieve prespecified precision and prespecified assurance and then use a simulation study to evaluate the performance of the exact confidence interval procedure in terms of empirical coverage and the percentage of times the prespecified precision has been achieved. In Section 4, we illustrate the calculations involved by using published examples [8]. The paper closes with a discussion.

2. Methods

2.1. The model and the confidence interval estimators

Suppose that k observations are made on each of N subjects randomly drawn from some population of interest. Suppose further that the jth ($j = 1, 2, \ldots, k$) observation, Y_{ij}, for subject i ($i = 1, 2, \ldots, N$) can be modeled as

$$Y_{ij} = \mu + a_i + e_{ij},$$

where μ is the overall mean, a_i are random effects, which are assumed to follow a normal distribution with mean 0 and variance σ_a^2, and e_{ij} are measurement errors, which are assumed independent of a_i and normally distributed with mean 0 and variance σ^2. The ICC (ρ) is defined as the ratio of the variability between subjects to the total variability, that is,

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2}.$$

Note that this model can also be applied when each subject is observed repeatedly on each of several occasions for test–retest reliability. Throughout this paper, we focus on using the analysis of variance approach to estimation ρ, given by

$$\widehat{\rho} = \frac{\text{MSB} - \text{MSE}}{\text{MSB} + (k-1)\text{MSE}},$$

where MSB and MSE are the between-subject and within-subject mean squared errors, respectively. The large sample variance for $\widehat{\rho}$ is given [17] by

$$\text{var}(\widehat{\rho}) = \frac{2(1-\rho)^2[1+(k-1)\rho]^2}{k(k-1)(N-1)},$$

which can be consistently estimated by substituting $\widehat{\rho}$ for ρ. Thus, a large sample $(1-\alpha)100\%$ two-sided confidence interval is given by

$$\widehat{\rho} \mp z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{\rho})}, \tag{1}$$

where and throughout the paper $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. This procedure is usually referred to as the Wald method. Alternatively, one can obtain an exact confidence interval by noting that

$$F(\widehat{\rho}) = \frac{\text{MSB}}{\text{MSE}} = \frac{1+(k-1)\widehat{\rho}}{1-\widehat{\rho}}$$

is distributed as an F-distribution with degrees of freedom of $N-1$ and $N(k-1)$. Specifically, an exact $(1-\alpha)100\%$ two-sided confidence interval (l,u) is given by

$$l,u=\frac{F/F_u-1}{F/F_u+k-1},\quad \frac{F/F_l-1}{F/F_l+k-1}, \tag{2}$$

where F_l and F_u are the $\alpha/2$ and the $1-\alpha/2$th quantiles of the F-distribution with degrees of freedom $N-1$ and $N(k-1)$, respectively. Note that because there is no approximation involved, this procedure may be regarded as the standard procedure under the assumptions described previously for the one-way random effects model. This is also different from 'exact' procedures for discrete data such as the Clopper–Pearson procedure for a single proportion [18], which has been shown to be conservative because of discreteness [19].

One can also apply the Fisher transformation that $0.5\ln F(\widehat{\rho})$ follows approximately a normal distribution with mean $0.5\ln F(\rho)$ and variance σ_z^2, where

$$\sigma_z^2\approx\frac{1}{2(k-1)(N-1)}.$$

The confidence limits for ρ are obtained by first setting confidence limits for $0.5\ln F$, that is,

$$a,b=0.5\ln\frac{1+(k-1)\widehat{\rho}}{1-\widehat{\rho}}\mp\frac{z_{\alpha/2}}{\sqrt{2(k-1)(N-1)}},$$

which yields confidence limits for ρ given by

$$l=\frac{\exp(2a)-1}{\exp(2a)+k-1},\quad u=\frac{\exp(2b)-1}{\exp(2b)+k-1}. \tag{3}$$

Note that among the three procedures presented previously, the procedure in Equation (1) is the simplest but may not perform well in small samples, primarily because $\widehat{\rho}$ has a left-skewed sampling distribution. Thus, either the exact procedure in Equation (2) or the Fisher transformation-based interval in Equation (3) is commonly recommended for data analysis.

Finally, a referee has kindly suggested two new approximate confidence interval procedures. The first is analogous to the Wilson score interval for a proportion, comprising all values of ρ satisfying

$$|\widehat{\rho}-\rho|\leqslant z_{\alpha/2}\sqrt{\frac{2(1-\rho)^2[1+(k-1)\rho]^2}{k(k-1)(N-1)}}.$$

The second procedure entails obtaining confidence interval for ρ through σ_a^2/σ^2 by applying the method of variances estimates recovery [20,21] to a ratio [22]. Because neither method has been fully evaluated, we leave these methods for future research.

2.2. *Sample size estimation based on the width of the Wald interval*

It is generally desirable to derive the sample size precisely on the basis of procedures that would be used for data analysis. However, in the present context, the interval width from the exact procedure and the Fisher z-transformation cannot be expressed as a simple function of the sample size N. Therefore, we proceed with the large sample confidence interval procedure for simplicity. This is similar to the case of the difference between two binomial proportions where sample size estimation is usually based on the Wald procedure, but method 10 of Newcombe [23] is recommended for data analysis.

Our goal here is to have N large enough to ensure that the half width of a $(1-\alpha)100\%$ two-sided confidence interval is no larger than a prespecified width ω with a prespecified assurance probability $1-\beta$. Specifically, we need

$$1-\beta=\Pr\left(z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\rho})}\leqslant\omega\right). \tag{4}$$

Let

$$f(\rho)=\frac{2(1-\rho)^2[1+(k-1)\rho]^2}{k(k-1)}.$$

Equation (4) implies that

$$1-\beta = \Pr\left(z_{\alpha/2}\frac{\sqrt{f(\widehat{\rho})}}{\sqrt{N-1}} \leqslant \omega\right)$$
$$= \Pr\left(\sqrt{f(\widehat{\rho})} \leqslant \frac{\omega}{z_{\alpha/2}}\sqrt{N-1}\right).$$

To obtain N, an asymptotic distribution for $\sqrt{f(\widehat{\rho})}$ is needed. By the delta method, we have

$$\text{var}\left(\sqrt{f(\widehat{\rho})}\right) = \frac{[f'(\rho)]^2}{2(N-1)},$$

where $f'(\rho)$ is the first-order derivative of $f(\rho)$ with respect to ρ, given by

$$f'(\rho) = \frac{4(1-\rho)[1+(k-1)\rho]|k-2+2\rho-2k\rho|}{k(k-1)}.$$

By the central limit theorem, we have

$$\sqrt{f(\widehat{\rho})} \sim N\left[\sqrt{f(\rho)}, \frac{[f'(\rho)]^2}{2(N-1)}\right].$$

Thus, we have, asymptotically,

$$1-\beta = \Pr\left(Z \leqslant \frac{\omega/z_{\alpha/2}\sqrt{N-1}-\sqrt{f(\rho)}}{|f'(\rho)|/\sqrt{2(N-1)}}\right),$$

which gives

$$\frac{z_\beta|f'(\rho)|}{\sqrt{2(N-1)}} = \frac{\omega\sqrt{N-1}}{z_{\alpha/2}} - \sqrt{f(\rho)},$$

where z_β is the upper β quantile of the standard normal distribution. Solving for $\sqrt{N-1}$ and squaring the admissible solution, we obtain

$$N = 1 + \left[\frac{\sqrt{f(\rho)}+\sqrt{f(\rho)+2z_\beta|f'(\rho)|\omega/z_{\alpha/2}}}{2\omega/z_{\alpha/2}}\right]^2$$
$$= 1 + \left[\frac{Az_{\alpha/2}+\sqrt{A^2z_{\alpha/2}^2+4\omega z_{\alpha/2}z_\beta A|B|}}{\omega\sqrt{2k(k-1)}}\right]^2, \quad (5)$$

where $|B|$ is the absolute value of B,

$$A = (1-\rho)[1+(k-1)\rho] \quad \text{and} \quad B = k-2+2\rho-2k\rho.$$

Note that, for $\beta = 0.50$ so that $z_\beta = 0$, the formula reduces to

$$N = 1 + \frac{2z_{\alpha/2}^2(1-\rho)[1+(k-1)\rho]^2}{k(k-1)\omega^2}, \quad (6)$$

which can be recognized as that obtained by Bonett [8] with ω being the half-interval width. Note also that for given values of ρ, ω, and N, the assurance probability $1-\beta$ can be obtained using the expression

$$z_\beta = \frac{\sqrt{2}\omega(N-1)/z_{\alpha/2}-2(1-\rho)[1+(k-1)\rho]\sqrt{N-1}}{|k-2+2\rho-2k\rho|}.$$

2.3. Sample size for achieving a prespecified lower limit of asymmetrical interval procedure

Rather than focusing on the width of confidence interval, an investigator may be more concerned that the reliability coefficient is not less than a prespecified value, ρ_0. This is reasonable if the primary objective of the study is to determine whether the ICC is of acceptable magnitude for the purpose at hand. Although it may not be easy to specify ρ_0 in general, the benchmarks provided by Landis and Koch [4] can be adopted here. Thus, we wish to derive a solution to

$$1-\beta = \Pr(\rho_L \geqslant \rho_0).$$

We now apply the approach based on the Fisher transformation, because this procedure performs better than the Wald method presented in Equation (1). Thus, we have

$$\begin{aligned} 1-\beta &= \Pr\left[0.5\ln F(\rho_L) \geqslant 0.5\ln F(\rho_0)\right] \\ &= \Pr\left[0.5\ln F(\widehat{\rho}) - z_\alpha\sqrt{\sigma_z^2} \geqslant 0.5\ln F(\rho_0)\right] \\ &= \Pr\left[\frac{0.5\ln F(\widehat{\rho}) - 0.5\ln F(\rho)}{\sqrt{\sigma_z^2}} \leqslant \frac{0.5\ln\frac{F(\rho)}{F(\rho_0)} - z_\alpha\sqrt{\frac{k}{2(k-1)(N-1)}}}{\sqrt{\frac{k}{2(k-1)(N-1)}}}\right]. \end{aligned}$$

Thus, we have

$$z_\beta\sqrt{\frac{k}{2(k-1)(N-1)}} = \frac{1}{2}\ln\frac{F(\rho)}{F(\rho_0)} - z_\alpha\sqrt{\frac{k}{2(k-1)(N-1)}}.$$

Solving for N yields

$$N = 1 + \frac{2(z_\alpha + z_\beta)^2 k}{\{\ln[F(\rho)/F(\rho_0)]\}^2(k-1)}, \tag{7}$$

which is equivalent to that given by Walter *et al.* [6] who derived the formula from a hypothesis testing perspective. Previous authors have not recognized this equivalence [7, 8]. Again given values of ρ, ρ_0, and N, the assurance probability is readily available, because

$$z_\beta = \sqrt{\frac{(k-1)(N-1)}{2k}}\ln\frac{F(\rho)}{F(\rho_0)} - z_\alpha.$$

In addition, it is straightforward to derive the achievable lower limit for given values of N, ρ, k, α, and β. Let

$$L_0 = \ln F(\rho) - \frac{z_\alpha + z_\beta}{\sqrt{\frac{(k-1)(N-1)}{2k}}}.$$

The achievable lower limit ρ_0 is given by

$$\rho_0 = \frac{\exp(L_0) - 1}{\exp(L_0) + k - 1}. \tag{8}$$

3. Evaluation

Because the formulas in Equations (5) and (7) rely on asymptotic theory for their validity, we evaluated their accuracy in finite samples with the use of simulations, where, without loss of generality, we only consider the parameter value $\alpha = 0.05$. We performed the simulation study with the use of `proc iml` in SAS Version 9.2 software.

To evaluate the formula for given confidence interval width, we first used Equation (5) to obtain the required sample size N for a prespecified combination of ρ, k, ω, and β and then randomly generated 10,000 datasets of size N from a multivariate normal distribution (With 10,000 simulation runs, we consider empirical coverage percentage for a 95% confidence interval falling into the range between 94.6%

and 95.4% to be acceptable.) We then calculated the two-sided 95% confidence limits for each dataset with the use of Equation (2). Finally, we estimated the empirical coverage as the proportion of the 10,000 confidence intervals that contained the parameter value of ρ and calculated the empirical assurance probability (EAP) as the proportion of the 10,000 interval half widths that were less than ω. We considered $\rho = 0.6, 0.7$, and 0.8, $k = 2, 3, 5$, and 10, $\omega = 0.1$ and 0.15, and $\beta = 0.5, 0.2$, and 0.1.

We used similar steps to evaluate formula (7) for estimating ρ with a prespecified lower limit. We considered similar parameter combinations, with the exception of defining ρ_0 to be smaller than the selected values for ρ by 0.10 and 0.15.

Results in Table I show that the empirical coverage percentages (ECP) are very close to the nominal level of 95%. This is because the exact confidence interval for ρ based on the F-distribution does not involve any approximation. As expected, the sample sizes based on Equation (6) can only provide around a 50% assurance probability of achieving the desired precision. We can also observe that the EAPs are in reasonable agreement with the nominal levels, with EAP slightly lower than the nominal when the number of raters is 2 or 3. The suboptimal EAP is due largely to the use of the Wald confidence interval procedure in the derivation. However, we should not use the same restrictive criteria to interpret empirical coverage probability and EAP, because of the approximation nature in sample size planning.

Results in Table II show that the formula in Equation (7) for prespecified lower limits is very accurate in achieving the nominal assurance probability. This is likely because the formula is derived on the basis of the Fisher transformation for constructing a confidence interval for ρ, which is known to perform well.

Table I. Performance of the sample size formula (5) for number of subjects N required to ensure that the half width of a 95% two-sided confidence interval for ρ is no greater than ω with 50%, 80%, and 90% assurance probabilities.

			50% Assurance			80% Assurance			90% Assurance		
ρ	ω	k	N	ECP[†]	EAP[‡]	N	ECP	EAP	N	ECP	EAP
0.60	0.10	2	159	95.17	49.87	183	95.17	79.77	196	95.28	90.84
		3	101	94.69	53.42	114	95.18	83.64	120	94.95	92.02
		5	73	94.94	58.86	80	95.06	84.92	84	94.78	94.54
		10	57	95.00	55.96	63	94.88	91.37	65	94.96	97.68
	0.15	2	71	94.85	49.24	87	95.06	80.34	95	95.50	90.26
		3	46	95.09	56.52	54	95.05	84.07	58	95.16	93.55
		5	33	95.30	58.64	38	94.91	88.88	41	94.83	98.64
		10	26	95.24	60.35	30	95.16	100.0	31	94.93	100.0
0.70	0.10	2	101	94.90	47.19	124	94.62	78.72	135	95.06	89.30
		3	68	95.05	51.81	81	95.01	80.21	88	94.63	90.54
		5	51	94.96	53.50	61	95.17	84.01	65	94.74	91.29
		10	42	95.07	54.04	50	95.21	86.55	53	94.82	93.33
	0.15	2	46	95.35	47.20	61	95.29	79.70	68	95.29	89.66
		3	31	95.42	52.03	40	95.24	83.14	44	95.00	90.97
		5	24	94.94	57.54	30	95.21	86.14	33	94.96	94.28
		10	20	94.58	60.36	24	95.43	86.40	27	94.93	96.52
0.80	0.10	2	51	94.97	44.96	69	94.84	75.96	77	94.97	86.91
		3	36	94.77	49.82	48	94.60	79.22	54	94.90	89.27
		5	29	94.69	52.93	37	95.00	79.35	42	94.83	90.44
		10	24	94.94	52.20	32	95.14	83.73	35	95.43	90.62
	0.15	2	24	94.85	45.02	35	95.01	74.84	41	95.29	87.00
		3	17	95.23	48.86	25	94.75	80.67	28	95.13	88.38
		5	14	94.92	56.55	19	94.83	80.63	22	94.84	90.71
		10	12	95.39	58.42	16	95.24	81.47	19	95.15	92.72

[†]Empirical coverage percentage (ECP) based on 10,000 simulation runs.
[‡]Empirical assurance (EAP) defined as percentage of times that the half width of the two-sided 95% confidence interval is no greater than ω.

Table II. Performance of the sample size formula (7) for number of subjects N required to ensure that the lower limit of a 95% one-sided confidence limit for ρ is no less than ρ_0 with 50%, 80%, and 90% assurance probabilities.

			50% Assurance			80% Assurance			90% Assurance		
ρ	ρ_0	k	N	ECP†	EAP‡	N	ECP	EAP	N	ECP	EAP
0.60	0.50	2	132	95.15	49.16	300	95.16	79.34	415	95.05	90.21
		3	82	94.89	51.67	184	95.15	80.28	255	95.01	90.36
		5	57	94.78	51.89	129	95.38	80.45	178	95.16	89.87
		10	44	95.01	53.06	99	94.99	81.11	137	94.79	90.23
	0.45	2	64	94.96	50.20	144	94.78	80.82	199	95.22	90.19
		3	39	95.27	52.26	87	94.81	81.07	120	94.65	90.26
		5	27	94.87	53.77	60	95.46	80.93	83	94.91	90.13
		10	21	94.96	54.09	46	95.27	81.42	63	94.48	89.86
0.70	0.60	2	91	95.10	49.47	205	95.29	80.11	284	94.89	90.28
		3	59	94.85	52.18	134	94.78	80.21	184	95.51	90.40
		5	44	94.92	53.01	99	94.91	81.48	136	94.97	90.34
		10	36	94.96	54.33	80	95.14	81.05	110	94.88	90.18
	0.55	2	45	94.89	49.63	101	95.38	79.77	140	94.65	89.74
		3	29	95.18	51.97	65	94.86	80.79	90	95.14	90.70
		5	22	94.57	53.92	48	95.15	81.65	66	94.74	90.21
		10	18	95.05	55.78	38	94.99	81.25	53	95.20	90.88
0.80	0.70	2	52	94.96	49.45	117	95.33	79.45	162	95.04	89.72
		3	36	95.00	51.22	80	95.27	80.59	110	95.06	90.27
		5	28	94.79	54.46	62	95.14	81.23	85	94.86	90.63
		10	24	94.91	55.38	52	94.98	81.37	71	94.83	90.23
	0.65	2	27	94.74	50.58	61	94.90	80.35	83	94.94	89.32
		3	19	94.88	52.97	41	95.07	80.91	57	95.01	90.78
		5	15	94.99	55.08	32	94.86	82.14	44	95.09	90.51
		10	12	95.32	54.58	27	95.27	82.15	36	94.62	90.45

†Empirical coverage percentage (ECP) based on 10,000 simulation runs.
‡Empirical assurance (EAP) defined as percentage of times that the lower limit of one-sided confidence interval is no less than ρ_0.

4. Worked examples

As an illustration, we now reconsider the examples given by Bonett [8] who intended to highlight the advantage that his sample size formula is more robust to different values of ρ than the formula by Walter *et al.* [6]. Bonett [8, p. 1335] stated that

> A planning value of ρ is needed for sample size determination in both hypothesis testing and interval estimation application but the effect of an inaccurate planning value is more serious in hypothesis applications. For instance, to test $H_0 : \rho_0 = 0.7$ at $\alpha = \beta = 0.05$ with $k = 3$, the required sample size is about 3376, 786, 167 for $\tilde{\rho}_1 = 0.725, 0.75$ and 0.8, respectively. In comparison, the sample size required to estimate ρ_1 with a 95% confidence interval width of 0.2 is 60, 52 and 37 for $\tilde{\rho}_1 = 0.725, 0.75$ and 0.8, respectively.

Even though it has found its way into textbooks (e.g., [24]), this comparison is misleading for several reasons. First, in light of our results, the former set of sample sizes provide an assurance probability of 95%, whereas the latter sizes only have 50% assurance probability. Second, all the lower limits for the former cases were set at 0.7, whereas for the latter, the sample sizes are for lower limits of 0.625, 0.65, and 0.7, respectively. Third, the former sample size estimation takes into account the skewness of the sampling distribution of the ICC, whereas the latter method treats it as symmetric.

Suppose now we wish to ensure, with 80% probability, that the lower limit of the one-sided 95% confidence interval is no less that 0.7 when the anticipated value of ρ is 0.725. Then, by Equation (7), the number of subjects we need is

$$N = 1 + \frac{2(1.645 + 0.84)^2 \times 3}{\left[\ln \frac{1+(3-1)0.725}{1-0.725} / \frac{1+(3-1)0.7}{1-0.7}\right]^2 (3-1)} \approx 1601.$$

Similar steps result in sample sizes of 374 and 80 for $\rho = 0.75$ and 0.8, respectively. In the second half of the comment, Bonett [8] referred to the sample size requirements for lower limits of 0.725, 0.75, and 0.80. According to formula (7), the sample sizes are 120, 107, and 80, respectively. Note that these are not as dramatic as claimed [8], although they are materially different from the respective sample sizes of 60, 52, and 37, given by Bonett [8] who implicitly assumed 50% assurance probability. For 80% assurance probability, when $\rho = 0.725$, with $N = 60$ and $k = 3$, the achievable lower limit ρ_0 is, by Equation (8),

$$\rho_0 = \frac{\exp(L_0) - 1}{\exp(L_0) + k - 1} = 0.577,$$

where

$$L_0 = \ln \frac{1 + (3-1)0.725}{1 - 0.725} - \frac{1.645 + 0.84}{\sqrt{(3-1)(60-1)/6}} \approx 1.63.$$

With Equation (5), we know that, with 50% probability, 'the sample size required to estimate ρ_1 with a 95% confidence interval width of 0.2 is 60, 52, and 37 for $\tilde{\rho}_1 = 0.725, 0.75$, and 0.8, respectively.' Suppose we wish to increase the assurance probability from 50% to 80%. With Equation (5) applied, the required sample sizes are 73, 65, and 48, respectively.

5. Discussion

It has now been suggested that reliability studies be reported following a set of guidelines [25]. Among them are the justification of sample size and the presentation of results using confidence intervals. A variety of confidence interval procedures for ICCs are now available [2, 3], but there is a paucity of sample size formulas. In the study design stage, one could use the formula derived from the hypothesis testing prospective by Walter *et al.* [6], but it may give an impression that using the formula is somewhat inconsistent with confidence interval estimation. The popular alternative would then be the formula suggested by Bonett [8], but as we have shown in this paper (Table I), this practice results in a chance of about 50% of achieving the desired precision. This is analogous to designing a study with 50% power. The problem of planning sample sizes on the basis of expected confidence interval width has long been recognized [9, 10]. Until now, there has been little work carried out for reliability studies to avoid this problem.

This paper has presented simple formulas for the calculation of the required number of subjects for a given number of observations per subject that is needed for estimating the ICC with desired precision based either on interval width or lower limit. In contrast to previous formulas that ignore the stochastic nature of confidence limits [7, 8], our approach treats interval width and lower limit as random variables and thus explicitly incorporates a desired assurance probability. We have also shown that a previous sample size formula derived from a hypothesis testing perspective can be used for estimating the lower limit, especially when the Landis–Koch benchmarks are adopted for specifying acceptable levels of reliability.

The sample sizes based on our formulas can be substantially larger than those without explicitly considering the assurance probability [8]. However, we believe that our approach is more appropriate, because in the planning of a study, it is important to assure that there is a high chance of achieving the desired precision. In the case of a sample size that is too large from a practical standpoint, one may consider the possibility of reducing confidence level or precision but not the assurance probability, otherwise it would result in a false sense of assurance.

We derived the formulas by assuming a fixed number of raters for two reasons. First, the number of subjects plays a much higher role in determining the precision of the estimation. Second, the same principle used here can be followed if a formula for the number of raters is deemed necessary.

As discussed previously in the literature [5, 6], the method used here is simplified by adopting a one-way random effects model in which the subject is the only factor considered. In some situations, it might be more appropriate to allow for both subject and rater effects with a two-way random effects model. Application of our formulas in such situations would, in general, result in an overestimation of sample size. More accurate formulas for the two-way random effects model may be derived by following the same principles here but requiring prespecification of the magnitude of the rater effect. Because a conservative sample size may not necessarily be a serious drawback in the design stage of a study, we concur with Walter *et al.* [6] that sample size formulas based on a one-way model can be useful even in the context of a two-way model, especially when the rater effect is not expected to be very large.

Acknowledgements

The comments from three anonymous reviewers and Dr Julia Taleban greatly improved the presentation. This work was partially supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada. Dr Zou is a recipient of the Early Researcher Award of Ontario Ministry of Research and Innovation, Canada.

References

1. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* 1966; **19**:3–11.
2. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 1979; **86**:420–428. DOI: 10.1037/0033-2909.86.2.420.
3. McGraw KO, Wong SP. Forming inference about some intraclass correlation coefficients. *Psychological Methods* 1996; **1**:30–46. DOI: 10.1037/1082-989X.1.1.30.
4. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**:159–174. DOI: 10.2307/2529310.
5. Donner A, Eliasziw M. Sample size requirements for reliability studies. *Statistics in Medicine* 1987; **6**:441–448. DOI: 10.1002/sim.4780060404.
6. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Statistics in Medicine* 1998; **17**:101–110. DOI: 10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-E.
7. Giraudeau B, Mary JY. 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* 2001; **20**:3205–3214. DOI: 10.1002/sim.935.abs.
8. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine* 2002; **21**:1331–1335. DOI: 10.1002/sim.1108.
9. Greenland S. On sample size and power calculations for studies using confidence intervals. *American Journal of Epidemiology* 1988; **128**:231–237.
10. Kupper LL, Hafner KB. How appropriate are popular sample size formulas? *The American Statistician* 1989; **43**:101–105. DOI: 10.2307/2684511.
11. Beal SL. Sample size determination for confidence intervals on the population mean and on the difference between two population means. *Biometrics* 1989; **45**:969–977. DOI: 10.2307/2531696.
12. Daly LE. Confidence intervals and sample sizes: don't throw out all your old sample size tables. *British Medical Journal* 1991; **302**:331–336.
13. Satten GA, Kupper LL. Sample size requirements for interval estimation of the odds ratio. *American Journal of Epidemiology* 1990; **131**:177–184.
14. Satten GA, Kupper LL. Sample size determination for pair-matched case-control studies where the goal is interval estimation of the odds ratio. *Journal of Clinical Epidemiology* 1990; **43**:55–59. DOI: 10.1016/0895-4356(90)90056-U.
15. Pan Z, Kupper LL. Sample size determination for multiple comparison studies treating confidence interval width as random. *Statistics in Medicine* 1999; **18**:1475–1488. DOI: 10.1002/(SICI)1097-0258(19990630)18:12<1475::AID-SIM144>3.3.CO;2-S.
16. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *Journal of Clinical Epidemiology* 2005; **58**:859–862. DOI: 10.1016/j.jclinepi.2004.12.009.
17. Fisher RA. *Statistical Methods for Research Workers*. Oliver and Boyd: Edinburgh, 1925.
18. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; **26**:404–413.
19. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 1998; **17**:857–872. DOI: 10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E.
20. Zou GY. On the estimation of additive interaction by use of the four-by-two table and beyond. *American Journal of Epidemiology* 2008; **168**:212–224. DOI: 10.1093/aje/kwn104.
21. Zou GY, Donner A. Construction of confidence limits about effect measures: a general approach. *Statistics in Medicine* 2008; **27**:1693–1702. DOI: 10.1002/sim.3095.
22. Li Y, Koval JJ, Donner A, Zou GY. Interval estimation for the area under the receiver operating characteristic curve when data are subject to error. *Statistics in Medicine* 2011; **29**:2521–2531. DOI: 10.1002/sim.4015.

23. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* 1998; **17**:873–890. DOI: 10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I.
24. Shoukri MM. *Measures of Interobserver Agreement and Reliability*, (2nd ed). Taylor & Francis Group: New York, 2011.
25. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, *et al.* Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International Journal of Nursing Studies* 2011; **48**:661–671. DOI: 10.1016/j.ijnurstu.2011.01.016.