

REVIEW ARTICLE OPEN ACCESS

Confidence Intervals and Sample Size for the ICC in Two-Way ANOVA Models

Dipro Mondal¹  | Math J J M Candel¹  | Alberto Cassese²  | Sophie Vanbelle¹ 

¹Department of Methodology and Statistics, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, the Netherlands |

²Department of Statistics, Computer Science, Applications “Giuseppe Parenti” The University of Florence, Firenze, Italy

Correspondence: Dipro Mondal (d.mondal@maastrichtuniversity.nl)

Received: 3 December 2024 | **Revised:** 23 March 2025 | **Accepted:** 19 April 2025

Funding: This work was supported by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Aspasia grant 015.015.064).

Keywords: interrater reliability | intrarater reliability | measurement errors | observer variation | reproducibility of results

ABSTRACT

The reliability of measurement instruments is vital in fields like medicine and psychology, where these tools are often used for diagnostic purposes. In reliability studies where participants are assessed by the same set of raters, the data can be modeled using a two-way ANOVA, with the intraclass correlation coefficient (ICC) as a key metric. This paper focuses on the ICC for agreement, which is crucial when the measurement values themselves, rather than just their rank ordering, are of interest. However, selecting appropriate confidence interval methods and determining adequate sample sizes for the ICC for agreement remains challenging. This work advances the understanding of confidence interval methods for the ICC for agreement, provides practical tools, and offers recommendations for selecting confidence interval methods and sample size procedures for planning reliability studies. In particular, we provide a comprehensive review and simulation-based comparison of six classes of confidence interval methods for the ICC for agreement identified in the literature. Our analysis includes a method based on the F-distribution, previously omitted, which demonstrates the best statistical properties in some cases. Then, in conjunction with the best-performing methods, we evaluate three sample size determination procedures based on the expected width of confidence intervals that we identified in the literature. To address the lack of accessible tools, we further developed an interactive R Shiny app, freely available to researchers, to compute confidence intervals and sample sizes. The utility of these methods is illustrated by a study on fetal heart rates.

1 | Introduction

The reliability of measurement instruments plays an important role in numerous scientific disciplines, such as medicine and psychology, where these instruments, amongst others, are used for diagnostic purposes. A reliability study is generally performed to assess the reliability of these instruments (i.e., how well the instrument is able to distinguish among participants in a population), which comprises participants measured repeatedly under similar conditions by the same instrument or rater (intrarater

reliability), or by different instruments or raters (interrater reliability).

A widely used metric to evaluate reliability when the measurements are quantitative is the intraclass correlation coefficient (ICC), which can take different forms depending on the study design [1, 2]. This paper focuses on the ICC for a reliability study where some participants are measured by the same set of raters. A suitable model for this data is the two-way analysis of variance (ANOVA) model. When different sets of raters rate each partici-

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Statistics in Medicine* published by John Wiley & Sons Ltd.

pant, or when repeated measurements are made simultaneously on each participant by the same rater, and these measurements for a participant are exchangeable, a one-way ANOVA model can be used. For a review in this context, see Mondal et al. [3]. Under a two-way ANOVA model, two different ICC coefficients can be defined based on whether the systematic differences between the instruments/raters are relevant or not: The ICC for agreement and the ICC for consistency, respectively. The ICC for consistency is concerned with the rank ordering of the participants. In that case, whether a particular rater systematically assigns higher or lower scores relative to the other raters is irrelevant, as only the ranking is important [1]. The ICC for agreement, on the other hand, is concerned with the actual values of the measurements and takes into account the systematic differences between the raters as well [1]. In disciplines such as medicine, where measurements are used for diagnosing participants, the focus extends beyond merely rank ordering individuals to the actual values of the measurements. In such cases, the intraclass correlation coefficient (ICC) for agreement becomes crucial, which is the primary focus of this paper.

When planning a reliability study, researchers often face the challenge of determining an adequate sample size for the study. The procedures for determining the minimum number of participants and raters for the reliability study are based on the expected width of the confidence interval around a planned value for the ICC. In this procedure, the aim is to determine the minimum number of raters/participants (or a combination thereof) such that the expected width of the confidence interval is less than a pre-specified value. Therefore, the procedure relies on a confidence interval for the ICC.

A search through existing literature on the confidence interval for the ICC for agreement reveals mainly six different classes of confidence interval methods to obtain a confidence interval, namely the moment approximation method [4, 5], the modified profile likelihood method [6], the delta method [7, 8], the modified large sample method [9–11], the generalized confidence interval method [12] and the variance partitioning method [13]. Most of these methods also include some variations in how they were developed by different authors. Comparative studies indicated that the two-moment approximation method exhibited worse statistical properties [14, 15], while the generalized confidence interval showed better statistical properties compared to the others [11, 16]. However, these results cannot be considered final, as no study has simultaneously compared all the available confidence interval methods. For instance, methods based on F- and beta-distributions [13] have been proposed but excluded from prior comparative studies. Additionally, differences in simulation settings across studies pose challenges in drawing definitive conclusions. These gaps underscore the need for a comprehensive evaluation of all available methods for constructing confidence intervals for the ICC for agreement.

Several approaches to determine the minimum number of raters and/or participants were developed to obtain a certain expected width of the confidence interval. These procedures differ based on whether the number of participants or raters is fixed [7, 17, 18]. However, their practical application is hindered by the complexity of implementation and the unavailability of user-friendly software. These challenges frequently require advanced proficiency

in programming and statistical methods, posing significant obstacles for practical use.

Our contribution directly addresses these gaps. First, we review the six classes of confidence interval methods available in the literature and complete their comparison, based on their statistical properties under some realistic scenarios, by means of a simulation study. Specifically, we make a comprehensive comparison of the available confidence interval methods, as well as those previously overlooked, and evaluate their performance in terms of their coverage and average confidence interval width. Second, for the methods with the best coverage properties, we review the different procedures to obtain a sample size for a reliability study under the two-way ANOVA model. Third, we develop an R-Shiny application capable of providing end-users with an interactive way of obtaining sample sizes, using the methods described in the paper. Our work not only advances the understanding of confidence interval methods, but also equips researchers with practical tools and guidance for choosing an appropriate confidence interval method and a sample size determination procedure to plan a reliability study.

The review is organized as follows. Section 2 introduces the ICC for agreement and consistency in the context of a reliability study where each of a group of raters assesses participants once, and the data is modeled using a two-way ANOVA model. Section 3 provides the different methods to construct a confidence interval for the ICC for agreement. Section 4 describes a Monte-Carlo simulation setup to evaluate the different confidence interval methods and presents the result of the comparisons. Section 5 provides different sample size determination procedures. Section 6 provides an example of how the sample size determination procedures can be used in a reliability study. Finally, Section 7 provides a discussion on the various methods and procedures described in this review.

2 | Definition and Estimation of the Intraclass Correlation Coefficients

Consider a study in which each of n participants is measured once by the same k raters. If these measurements are quantitative, they can be modeled by a two-way ANOVA model. The outcome y_{ij} for participant i ($i = 1, 2, \dots, n$) and rater j ($j = 1, 2, \dots, k$), y_{ij} , can then be written as [1],

$$y_{ij} = \mu + s_i + r_j + e_{ij}, \quad (1)$$

where μ is the grand mean, s_i is the participant random effect ($s_i \sim \mathcal{N}(0, \sigma_s^2)$), r_j is the rater effect, and e_{ij} is the error ($e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$). The rater effect can be considered as random or fixed. With the intention of generalizing the reliability to any set of raters from the rater population, we consider the rater effects as random ($r_j \sim \mathcal{N}(0, \sigma_r^2)$) as it is customary in inter-rater reliability studies. Note that in Equation (1), no interaction term is included as it is not possible to estimate the interaction effect between participants and raters when there is only one observation for each participant-rater combination.

Table 1 gives the variance decomposition of the ANOVA model with the mean squares denoted by MSS (=

TABLE 1 | Variance decomposition for the ANOVA model Equation (1).

Sources of Variation	Degrees of Freedom	Mean Squares	Expected Mean Squares
Between participants	$n - 1$	MSS	$k\sigma_s^2 + \sigma_e^2$
Between raters	$k - 1$	MSR	$n\sigma_r^2 + \sigma_e^2$
Errors	$(n - 1)(k - 1)$	MSE	σ_e^2

$\frac{k}{n-1} \sum_{i=1}^n (\bar{y}_{i.} - \bar{y}_{..})^2$, $MSR = \frac{n}{k-1} \sum_{j=1}^k (\bar{y}_{.j} - \bar{y}_{..})^2$ and $MSE = \frac{1}{(n-1)(k-1)} \sum_{i=1}^n \sum_{j=1}^k (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$, where $\bar{y}_{i.} = \frac{1}{k} \sum_{j=1}^k y_{ij}$, $\bar{y}_{.j} = \frac{1}{n} \sum_{i=1}^n y_{ij}$ and $\bar{y}_{..} = \frac{1}{nk} \sum_{i=1}^n \sum_{j=1}^k y_{ij}$. The assumptions of the ANOVA model we have adopted imply that the outcomes are conditionally normal, and the mean squares follow independent scaled χ^2 -distributions. The χ^2 -distributions corresponding to the mean squares are given in Section 3.1.1. Using the variance decomposition, the ICC for agreement for the two-way ANOVA model Equation (1) is defined as,

$$\rho_a = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_e^2}, \quad 0 \leq \rho_a \leq 1 \quad (2)$$

and the ICC for consistency for the same model is defined as,

$$\rho_c = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}, \quad 0 \leq \rho_c \leq 1. \quad (3)$$

The coefficient ρ_a is concerned with absolute measurements rather than just the rank ordering of the participants as given by ρ_c . The difference between the two coefficients, ρ_a and ρ_c , lies in the denominator. Specifically, for ρ_a the denominator includes the systematic differences between the raters, as expressed by σ_r^2 , thereby accounting for the variation from all sources (see Equation 1), unlike ρ_c . As a consequence, ρ_a yields lower values than ρ_c .

Under the ANOVA model given by Equation (1), the aforementioned ICC for agreement is estimated as [1, 19],

$$\hat{\rho}_a = \frac{MSS - MSE}{MSS + (k - 1)MSE + \frac{k}{n}(MSR - MSE)} \quad (4)$$

and that for consistency is estimated as [1],

$$\hat{\rho}_c = \frac{MSS - MSE}{MSS + (k - 1)MSE}. \quad (5)$$

Note that Almehrzi and Emam [8] provided an alternative expression for $\hat{\rho}_a$ and for $\hat{\rho}_c$ using a matrix formulation. The expression for ρ_a is provided in Equation (8). Readers interested in the estimation of ICC for other cases, such as when the rater effects are considered fixed, can consult McGraw and Wong [1]. In the following sections, we focus only on the ICC for agreement. Henceforth we denote $\rho_a = \rho$ and its estimator by $\hat{\rho}_a = \hat{\rho}$.

3 | Confidence Intervals for the ICC for Agreement

In the literature, there are six classes of confidence interval methods to construct the lower (L) and upper (U) limits of the confidence interval for ρ . These are the moment approximation method [4, 5], the modified profile likelihood method [6], the delta method [7, 8], the modified large sample method [9–11], the generalized confidence interval method [12], and the variance partitioning method [13]. In this study, we do not evaluate the moment approximation and modified profile likelihood methods and focus on the latter four methods. The details of these two methods are provided in Supporting Information Data S1 (Section A), and the rationale for their exclusion is discussed in Section 4. Below, we summarize the most important features of the four classes of confidence interval methods we will evaluate.

3.1 | Delta Method

The delta method allows us to derive analytical approximations for the variance of $\hat{\rho}$ (or a transformation of it). This method was used by Saito et al. [7] to derive an approximation for the asymptotic variance of the log-transformed $\hat{\rho}$, $V(\log(\hat{\rho}))$, and by Almehrzi and Emam [8] to derive an expression for the approximate variance of $\hat{\rho}$ using matrix algebra.

3.1.1 | Delta Method Applied to Log-Transformation (W_{log})

Saito et al. [7] derived an approximate expression of the asymptotic variance, $V(\hat{\rho})$, based on the delta method. The variance on the log-transformed scale, $\log(\hat{\rho})$ is,

$$\begin{aligned} V(\log(\hat{\rho})) &= \frac{V(MSS) + V(MSE)}{[E(MSS) - E(MSE)]^2} \\ &+ \frac{V(MSS) + (k/n)^2 V(MSR) + (k - 1 - k/n)^2 V(MSE)}{[E(MSS) + (k/n)E(MSR) + (k - 1 - k/n)E(MSE)]^2} \\ &- 2 \frac{V(MSS) - (k - 1 - k/n)V(MSE)}{[E(MSS) - E(MSE)][E(MSS) + (k/n)E(MSR) + (k - 1 - k/n)E(MSE)]}. \end{aligned} \quad (6)$$

Since following from the assumptions of the ANOVA model in Equation (1), $MSS \sim \chi_{n-1}^2 \frac{k\sigma_s^2 + \sigma_e^2}{n-1}$, $MSR \sim \chi_{k-1}^2 \frac{n\sigma_r^2 + \sigma_e^2}{k-1}$ and $MSE \sim \chi_{(n-1)(k-1)}^2 \frac{\sigma_e^2}{(k-1)(n-1)}$ where χ_{n-1}^2 , χ_{k-1}^2 and $\chi_{(n-1)(k-1)}^2$ are a set of jointly independent χ^2 random variables, their expected values and variances can be obtained as,

$$\begin{aligned} E(MSS) &= k\sigma_s^2 + \sigma_e^2, \quad V(MSS) = \frac{2(k\sigma_s^2 + \sigma_e^2)^2}{n-1}, \\ E(MSR) &= n\sigma_r^2 + \sigma_e^2, \quad V(MSR) = \frac{2(n\sigma_r^2 + \sigma_e^2)^2}{k-1}, \\ E(MSE) &= \sigma_e^2, \quad V(MSE) = \frac{2\sigma_e^4}{(k-1)(n-1)}. \end{aligned}$$

Substituting these into Equation (6) we obtain the final expression for $V(\log(\hat{\rho}))$. The estimate of $V(\log(\hat{\rho}))$, $\hat{V}(\log(\hat{\rho}))$ is obtained by plugging in the estimates of σ_s^2 , σ_r^2 and σ_e^2 in Equation (6). The $(1 - \alpha) \times 100\%$ confidence interval for $\log(\rho)$ can be constructed using the Wald confidence interval method as,

$$L_{\log}, U_{\log} = \log(\hat{\rho}) \pm Z(1 - \alpha/2) \sqrt{\hat{V}(\log(\hat{\rho}))}, \quad (7)$$

where the expression of $\hat{V}(\log(\hat{\rho}))$ is computed as in Equation (6), and $Z(1 - \frac{\alpha}{2})$ denotes the $(1 - \frac{\alpha}{2}) \times 100$ percentile of a standard normal distribution. The confidence interval for ρ is obtained by exponentiating $L_{\log}$ and $U_{\log}$.

3.1.2 | Delta Method Applied Directly With Matrix Formulation (W_{mat})

Almehrizi and Emam [8] obtained the expression for $\hat{\rho}$ Equation (4) in matrix formulation as,

$$\hat{\rho} = \frac{\mathbf{1}'\mathbf{S}\mathbf{1} - tr\mathbf{S}}{\frac{1}{n}\mathbf{1}'\mathbf{S}\mathbf{1} + \frac{nk-n-k}{n}tr\mathbf{S} + k\mathbf{y}'_d\mathbf{y}_d}, \quad (8)$$

where $\mathbf{1}$ is a row vector of 1's ($k \times 1$), $\mathbf{S}$ is the estimated sample variance-covariance matrix of k measurements ($k \times k$), and $\mathbf{y}_d = (\bar{y}_{.1} - \bar{y}_{..}, \dots, \bar{y}_{.k} - \bar{y}_{..})'$ is a ($k \times 1$) vector of deviations of average measurements of each rater j ($\bar{y}_{.j}$) from the grand average ($\bar{y}_{..}$). Note that, $\mathbf{1}'\mathbf{S}\mathbf{1} = kMSS = \frac{k^2}{n-1} \sum_{i=1}^n (\bar{y}_{i.} - \bar{y}_{..})^2$ is the sample variance of total measurements of participants, $tr\mathbf{S} = MSS + (k-1)MSE = \sum_{j=1}^k \left(\frac{1}{n-1} \sum_{i=1}^n (y_{ij} - \bar{y}_{.j})^2 \right)$ is the sum of within rater sample variances, and, $\mathbf{y}'_d\mathbf{y}_d = \frac{k-1}{n}MSR = \sum_{j=1}^k (\bar{y}_{.j} - \bar{y}_{..})^2$ is the sum of squared deviations of average measurements of each rater j from the grand average. The three main terms $\mathbf{1}'\mathbf{S}\mathbf{1}$, $tr\mathbf{S}$, and $\mathbf{y}'_d\mathbf{y}_d$ are non-linear functions of the elements of the data matrix. Therefore, by the delta method, the estimated variance-covariance matrix of the three terms $\frac{n-1}{n}\mathbf{1}'\mathbf{S}\mathbf{1}$, $\frac{n-1}{n}tr\mathbf{S}$, and $\mathbf{y}'_d\mathbf{y}_d$ for the given model Equation (1) can be obtained as [8],

$$\hat{V}_y = \begin{pmatrix} \frac{2(\mathbf{1}'\mathbf{S}\mathbf{1})^2}{n+1} & \frac{2(\mathbf{1}'\mathbf{S}\mathbf{1})}{n+1} & 0 \\ \frac{2(\mathbf{1}'\mathbf{S}\mathbf{1})}{n+1} & \frac{2(tr\mathbf{S})^2}{n+1} & 0 \\ 0 & 0 & \frac{2(\mathbf{y}'_d\mathbf{y}_d)^2}{k+1} \end{pmatrix}.$$

Then the estimated variance of the estimator of ρ is,

$$\begin{aligned} \hat{V}(\hat{\rho}) &= \frac{2\hat{\rho}^2 \left[\left(1 + \frac{nk-n-k}{n}\hat{\rho} \right)^2 tr\mathbf{S}^2 + \left(1 - \frac{\hat{\rho}}{n} \right)^2 (\mathbf{1}'\mathbf{S}\mathbf{1})^2 - 2 \left(1 + \frac{nk-n-k}{n}\hat{\rho} \right) \left(1 - \frac{\hat{\rho}}{n} \right) \mathbf{1}'\mathbf{S}\mathbf{1} \right]}{(n+1)[\mathbf{1}'\mathbf{S}\mathbf{1} - tr\mathbf{S}]^2} \\ &\quad + \frac{8k\hat{\rho}^4\mathbf{y}'_d\mathbf{y}_d}{[\mathbf{1}'\mathbf{S}\mathbf{1} - tr\mathbf{S}]^2} \end{aligned} \quad (9)$$

The confidence interval for ρ is then obtained as [8],

$$L, U = \hat{\rho} \pm T_{n+1} \left(1 - \frac{\alpha}{2} \right) \sqrt{\hat{V}(\hat{\rho})},$$

where $\hat{V}(\hat{\rho})$ is replaced by Equation (9), and $T_{n+1}(1 - \frac{\alpha}{2})$ is the $(1 - \frac{\alpha}{2}) \times 100$ percentile of the t -distribution with $n + 1$ degrees of

freedom. Note that Almehrizi and Emam [8] use a t -distribution instead of the generally used normal distribution.

3.2 | Modified Large Sample Method

The modified large sample method introduced by Graybill and Wang [20] aims to find a refined approximation of the confidence interval for ρ starting with an approximate expression of a large sample confidence interval involving mean squares. The initial approximation is improved under limiting conditions of each of the variance components involved, thereby obtaining a modified large sample approximation of the confidence interval. Note that the final expression of the confidence limits may be different depending on the conditions for exactness that were defined during its derivation (see Burdick et al. [21] for further details).

Arteaga et al. [9] and Gui et al. [10] derived the confidence interval for ρ using the modified large sample method, each starting from different assumptions. The corresponding two-sided confidence intervals are different. Note that Shoukri [22] incorrectly refers to the hybrid moment method described in Supporting Information Data S1 (Section A.1.2) as the modified large sample method.

3.2.1 | Modified Large Sample Confidence Interval for ρ by Arteaga Et Al. [9] (MLS_A)

Arteaga et al. [9] derived the confidence limits for the transformation $\frac{k\rho}{n(1-\rho)}$ as,

$$\begin{aligned} L_A^* &= \frac{-1 + F_{n-1,\infty}^{-1}(1 - \frac{\alpha}{2})F_1 + (1 - F_{n-1,\infty}^{-1}(1 - \frac{\alpha}{2})F_5)F_3F_1^{-1}}{n - 1 + F_{n-1,\infty}^{-1}(1 - \frac{\alpha}{2})F_3F_2} \text{ and} \\ U_A^* &= \frac{-1 + F_{n-1,\infty}^{-1}(\frac{\alpha}{2})F_1 + (1 - F_{n-1,\infty}^{-1}(\frac{\alpha}{2})F_6)F_6F_1^{-1}}{n - 1 + F_{n-1,\infty}^{-1}(\frac{\alpha}{2})F_4F_2} \end{aligned} \quad (10)$$

where $F_1 = \frac{MSS}{MSE}$, $F_2 = \frac{MSR}{MSE}$, $F_3 = F_{n-1,k-1} \left(1 - \frac{\alpha}{2} \right)$, $F_4 = F_{n-1,k-1} \left(\frac{\alpha}{2} \right)$, $F_5 = F_{n-1,(n-1)(k-1)} \left(1 - \frac{\alpha}{2} \right)$ and $F_6 = F_{n-1,(n-1)(k-1)} \left(\frac{\alpha}{2} \right)$. Then, the limits of the confidence interval for ρ are,

$$L = \frac{n \max(0, L_A^*)}{k + n \max(0, L_A^*)}, \text{ and } U = \frac{n \max(0, U_A^*)}{k + n \max(0, U_A^*)},$$

where L_A^* and U_A^* are given in Equation (10). Henceforth, we refer to this method as MLS_A . Note that the expressions of L_A^* and U_A^* in Equation 2.9 from Arteaga et al. [9] is incorrect (see Appendix C for the correction of the formula in the paper of Arteaga et al. [9]).

3.2.2 | Modified Large Sample Confidence Interval for ρ by Gui Et Al. [10] (MLS_G)

Gui et al. [10] proposed an expression of the confidence interval according to the modified large sample method based on a general form of a correlation coefficient which was adapted by

Cappelleri and Ting [15] for ρ . The final expressions of the limits of the confidence interval are quite elaborate, so we refer to Appendix A of Cappelleri and Ting [15] for the exact expressions. Henceforth, we refer to this method as MLS_G . Note that, the expression of the limits of the confidence interval presented in Appendix B of Ionan et al. [11] corresponds to expressions for MLS_A , although it is referred to as the method by Cappelleri and Ting [15]. However, their evaluation of the accuracy of the confidence interval methods appears to use the correct expression for MLS_G .

3.3 | Generalized Confidence Interval Method (GCI)

Weerahandi [12] provided a general method for obtaining a confidence interval based on a generalized pivotal quantity. The generalized pivotal quantity is defined as a special pivotal quantity [23], essentially having two properties. First, the probability distribution of the pivotal quantity is free of any unknown parameters, and second, the observed pivot is also independent of unknown parameters. The generalized pivot for ρ is given by [18],

$$g(\mathbf{t}) = \frac{\max \left[0, \frac{(n-1)MSS}{k\chi^2_{n-1}} - \frac{(n-1)(k-1)MSE}{k\chi^2_{(n-1)(k-1)}} \right]}{\frac{(k-1)MSR}{n\chi^2_{k-1}} + \frac{(n-1)MSS}{k\chi^2_{n-1}} + \frac{(nk-n-k+1)(nk-n-k)MSE}{nk\chi^2_{(n-1)(k-1)}}}, \quad (11)$$

where $\mathbf{t}$ denotes a vector of mean squares. The $(1 - \alpha) \times 100\%$ confidence interval for ρ is given by the quantile function $\hat{G}$ as [11, 16, 18]

$$L, U = \hat{G}\left(\frac{\alpha}{2}\right), \hat{G}\left(1 - \frac{\alpha}{2}\right)$$

We denote this method henceforth as GCI . Unlike the other confidence interval methods, the generalized confidence interval method relies on empirical quantiles of the pivotal quantity that are obtained through sampling methods. Specifically, this is achieved by sampling from the χ^2 -distributions in Equation (11) and plugging in the values obtained [11, 16]. This is then repeated a large number of times, implying that the accuracy of this method relies on the number of simulations. Dobbin and Ionan [18] state that if the confidence interval is obtained properly (i.e., using the recommended 100 000 simulations or more [24]), then the difference between $\hat{G}$ and G is negligible. Note that the third term of the denominator of $g(\mathbf{t})$ is incorrect in Ionan et al. (see Appendix C for the correction of the formula in the paper of Ionan et al. [11]).

3.4 | Variance Partitioning Method (VP_F , VP_B)

Demetrashvili et al. [13] proposed two generic approaches to construct a confidence interval for an ICC coefficient for a general ANOVA model with any number of variance components. This can be applied to the ICC for agreement for the two-way ANOVA model. These approaches rely on the partitioning of the relevant variance components into two parts. Consider that there are a total of P variance components ($\sigma_1^2, \sigma_2^2, \dots, \sigma_Q^2, \sigma_{Q+1}^2, \dots, \sigma_P^2$) which can be divided into two partitions such that,

1. the first partition $q = 1, \dots, Q$ of the variance components $\sigma_1^2, \dots, \sigma_Q^2$ is unrelated to the rater process, the sum of which is denoted as σ_G^2 , and
2. the second partition $p = Q + 1, \dots, P$ of the variance components $\sigma_{Q+1}^2, \dots, \sigma_P^2$ represents part of the rater variability, the sum of which is denoted as σ_E^2 .

Then ρ Equation (2) can be defined in a general form, $\rho_G = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}$, where $\sigma_G^2 = \sum_{q=1}^Q \sigma_q^2$ and $\sigma_E^2 = \sum_{p=Q+1}^P \sigma_p^2$. For the two-way ANOVA Equation (1), $P = 3$, $Q = 1$, $\sigma_G^2 = \sigma_s^2$ and $\sigma_E^2 = \sigma_e^2 + \sigma_r^2$. Denoting the total variance $\sigma_T^2 = \sigma_s^2 + \sigma_E^2$, the estimated variance of the estimator of ρ is [13]

$$\hat{V}(\hat{\rho}) \approx \frac{\hat{\sigma}_E^4}{(\hat{\sigma}_T^2)^4} \hat{V}(\hat{\sigma}_s^2) + \frac{\hat{\sigma}_s^4}{(\hat{\sigma}_T^2)^4} \hat{V}(\hat{\sigma}_E^2) - \frac{2\hat{\sigma}_s^2 \hat{\sigma}_E^2}{(\hat{\sigma}_T^2)^4} \widehat{Cov}(\hat{\sigma}_s^2, \hat{\sigma}_E^2), \quad (12)$$

where $\hat{V}(\hat{\sigma}_E^2) = \hat{V}(\hat{\sigma}_r^2) + \hat{V}(\hat{\sigma}_e^2) + 2\widehat{Cov}(\hat{\sigma}_r^2, \hat{\sigma}_e^2)$ and $\widehat{Cov}(\hat{\sigma}_i^2, \hat{\sigma}_j^2)$ is the covariance of the estimators $\hat{\sigma}_i^2$ and $\hat{\sigma}_j^2$. Unlike the other confidence interval methods, the covariance between the variance estimators is required to obtain $\hat{V}(\hat{\rho})$. The variance-covariance estimates of the variance components are then obtained from the inverse of the Fisher information matrix, as provided in Searle et al. [25] (Chapter 4, page 155). One can use Equation (12) to construct a confidence interval for ρ using the Wald confidence interval relying on a normal approximation. However, the distribution of $\hat{\rho}$ is generally skewed. Therefore, instead of a normal distribution, Demetrashvili et al. [13] proposed using an F- or beta-distribution to approximate the distribution of $\hat{\rho}$, which we denote by the F- and Beta-methods, respectively.

3.4.1 | F-Method (VP_F)

The F-method is based on Satterthwaite's approximation of the variance components σ_s^2 and σ_E^2 , where the estimate of the two components follow independent χ^2 -distributions with the degrees of freedom df_1 and df_2 estimated as $d\hat{f}_1 = \frac{2\hat{\sigma}_s^4}{\hat{V}(\hat{\sigma}_s^2)}$ and $d\hat{f}_2 = \frac{2\hat{\sigma}_E^4}{\hat{V}(\hat{\sigma}_E^2)}$, respectively.

In the study of van den Heuvel [26], it was found that the F-method works better when the variance estimates do not include covariance terms. Then $\hat{\rho}$ is distributed according to,

$$\frac{\hat{\sigma}_s^2 f_{d\hat{f}_1, d\hat{f}_2}}{\hat{\sigma}_s^2 f_{d\hat{f}_1, d\hat{f}_2} + \hat{\sigma}_E^2},$$

where f_{v_1, v_2} denotes the density of an F-distribution with v_1 and v_2 degrees of freedom in the numerator and denominator, respectively. The $(1 - \alpha) \times 100\%$ confidence interval for ρ is given by,

$$L = \frac{\hat{\sigma}_s^2 F_{d\hat{f}_1, d\hat{f}_2}(\frac{\alpha}{2})}{\hat{\sigma}_s^2 F_{d\hat{f}_1, d\hat{f}_2}(\frac{\alpha}{2}) + \hat{\sigma}_E^2} \quad \text{and} \\ U = \frac{\hat{\sigma}_s^2 F_{d\hat{f}_1, d\hat{f}_2}(1 - \frac{\alpha}{2})}{\hat{\sigma}_s^2 F_{d\hat{f}_1, d\hat{f}_2}(1 - \frac{\alpha}{2}) + \hat{\sigma}_E^2}.$$

Note that, Demetrashvilli et al. [13] suggests to use $d\hat{f}_1 = \max\left(1, \frac{2\hat{\sigma}_1^4}{\hat{V}(\hat{\sigma}_1^2)}\right)$ to avoid small values of $d\hat{f}_1$ which might be computationally complex.

3.4.2 | Beta-Method (VP_B)

The Beta-method [13] aims to approximate the distribution of $\hat{\rho}$ with a beta-distribution with parameters $a > 0$ and $b > 0$. The parameters of the beta-distribution are obtained by matching the theoretical mean and variance of the beta-distribution to the mean and variance of $\hat{\rho}$ to obtain,

$$\begin{aligned}\hat{a} &= \frac{\hat{\rho}[\hat{\rho}(1 - \hat{\rho}) - \hat{V}(\hat{\rho})]}{\hat{V}(\hat{\rho})} \text{ and} \\ \hat{b} &= \frac{(1 - \hat{\rho})[\hat{\rho}(1 - \hat{\rho}) - \hat{V}(\hat{\rho})]}{\hat{V}(\hat{\rho})}\end{aligned}\quad (13)$$

Note that the estimation of these parameters a and b may lead to several complications, and Demetrashvilli et al. [13] suggest the following two corrections to avoid these problems.

1. A beta-distribution is U-shaped when both a and b are less than 1 ($a, b < 1$). This shape is not appropriate to approximate the distribution of $\hat{\rho}$. Therefore, to avoid parameter estimates $\hat{a}, \hat{b} < 1$, two fixes are depending on whether $\hat{\rho}$ is greater than or less than 0.5.
 - a. When $0 < \hat{a}, \hat{b} < 1$ and $\hat{\rho} = \frac{\hat{a}}{\hat{a} + \hat{b}} < 0.5$, the U-shaped beta-distribution is converted to a J-shaped beta-distribution, by increasing the mass closer to 0. This is done by fixing $\hat{b} = 1$ and leaving the other parameter $\hat{a} < 1$ (estimated by Equation 13).
 - b. When $0 < \hat{a}, \hat{b} < 1$ and $\hat{\rho} = \frac{\hat{a}}{\hat{a} + \hat{b}} > 0.5$, the U-shaped beta-distribution is converted to a J-shaped beta-distribution by increasing the mass closer to 1. This is done by fixing $\hat{a} = 1$ and leaving the other parameter $\hat{b} < 1$ (estimated by Equation 13).
2. The estimates for a and b from Equation (13) may be negative when the variance of $\hat{\rho}$ Equation (12) is larger than $\hat{\rho}(1 - \hat{\rho})$. Since a beta-distribution is defined for positive parameters, two fixes are proposed in this case, depending on whether $\hat{\rho}$ is greater or lesser than 0.5, for the same reasons mentioned above.
 - a. When $\hat{\rho} \leq 0.5$, $\hat{b}$ is fixed at 1 ($\hat{b} = 1$) and $\hat{a}$ is obtained as $\hat{a} = \hat{\rho}/(1 - \hat{\rho})$.
 - b. When $\hat{\rho} > 0.5$, $\hat{a}$ is fixed at 1 ($\hat{a} = 1$) and $\hat{b}$ is obtained as $\hat{b} = \hat{\rho}/(1 - \hat{\rho})$.

Based on this approximate distribution, the $(1 - \alpha) \times 100\%$ confidence interval for ρ is given by,

$$L = B_{\hat{a}, \hat{b}}\left(\frac{\alpha}{2}\right), \text{ and } U = B_{\hat{a}, \hat{b}}\left(1 - \frac{\alpha}{2}\right),$$

where $B_{\hat{a}, \hat{b}}(\frac{\alpha}{2})$ and $B_{\hat{a}, \hat{b}}(1 - \frac{\alpha}{2})$ are the $(\frac{\alpha}{2}) \times 100$ and $(1 - \frac{\alpha}{2}) \times 100$ percentiles of a beta-distribution with parameters $\hat{a}$ and $\hat{b}$.

4 | Comparison of Confidence Interval Methods

Several comparative studies [5, 8, 11, 14–16] investigated the statistical properties of the confidence interval methods reviewed in this paper. These studies are simulation based and compare the coverage probability and the average width of the different confidence interval methods for ρ . Coverage probability is defined as the proportion of confidence intervals from the simulations that contain the true value of ρ . The average width of the confidence interval is the difference between the upper and lower limits of the confidence interval, averaged over the simulations. A confidence interval method with coverage probability close to the nominal level and having the smallest average width is considered to be the best.

The choice of the confidence interval methods for our comparison is motivated by the results from the previous studies. Specifically, Almehrzi and Emam [8] demonstrated that their matrix formulation method outperformed the two-moment approximation method. Zou and McDermott [5] compared all the moment approximation methods, but could not recommend one method over the others. Rousson et al. [14] and Cappelleri and Ting [15] compared some of the better moment approximation methods with the modified large sample methods, leading to the conclusion that the latter methods are better. Ionan et al. [11] and Tian and Cappelleri [16] compared the modified large sample method proposed by Gui et al. [10] with the generalized confidence interval method, but suggested that neither of these methods is better than the others in all scenarios (i.e., combinations of ρ , R , k , and n). Xiao and Liu [6] compared their modified profile likelihood method with the generalized confidence interval method and showed that their method produces over-coverage in most cases. These comparisons are limited as, first, they do not encompass all the available confidence interval methods, and second, the scenarios in which these comparisons are made vastly differ across papers. For example, the matrix formulation was only compared to the two-moment approximation method, whereas the moment approximation methods have been shown to be inferior to several other confidence interval methods, like the modified large sample methods and the generalized confidence interval method. Therefore, we have included these methods (matrix formulation of $\hat{\rho}$, modified large sample methods, generalized confidence interval method) in our comparison. Further, the variance partitioning method, not included in previous comparisons, is added to our comparative study. The modified profile likelihood method is excluded from our study mainly for two reasons. First, the modified profile likelihood relies on a specification of a plausible range for ρ and R , which is generally unavailable in a reliability study. Second, this method uses two consecutive grid searches combined with Monte-Carlo simulations, implying a high computational cost of the method.

In summary, we compare the coverage probability and the average width of the delta method applied to the log-transformation and the matrix formulation of $\hat{\rho}$ (W_{log} and W_{mat} , respectively), the modified large sample methods by Arteaga et al. [9] and Gui et al. [10] (MLS_A and MLS_G , respectively), the general-

ized confidence interval method (*GCI*) and the variance partitioning methods using the *F*- and Beta-methods (VP_F and VP_B , respectively).

4.1 | Monte Carlo Simulation

We set up a Monte Carlo simulation study to evaluate the statistical properties of the seven confidence interval methods. Following the ANOVA model defined in Equation (1), for each replication, $N = n \times k$ errors are drawn from a normal distribution centered at zero with variance $\sigma_e^2 = 1$. Then, for a given value of R , k rater effects (r_j) are drawn for n participants following a normal distribution centered in zero with variance $\sigma_r^2 = R \times \sigma_e^2 = R$ (since $\sigma_e^2 = 1$). Next, for a given value of ρ , n participant effects (s_i) are drawn for k raters from a normal distribution centered in zero with variance $\sigma_s^2 = \frac{\rho(R+1)}{1-\rho}$. This process is replicated 10^5 times. For each replication, the confidence interval using the above-mentioned seven confidence interval methods is obtained. We study the properties of these confidence interval methods for various scenarios. Specifically, we choose $n \in \{20, 40, 60, 100\}$, which covers a wide range of the number of participants; $\rho \in \{0.7, 0.8, 0.9\}$, since a high ρ is generally of interest in a reliability study; and, $R \in \{0.1, 1, 2\}$ based on empirical evidence from several reliability studies [18, 27–35] which indicate that in most cases $R \leq 1$. As for the number of raters, we choose $k \in \{2, 3, 5, 8\}$. Two-sided confidence intervals are commonly used for sample size determination procedures (see Section 5). Therefore, we compare the obtained two-sided confidence intervals based on the coverage probability as well as the average width of the confidence interval. Confidence interval methods providing a coverage probability closer to the nominal coverage are more accurate, thus, we examine how far the coverage probability is from the nominal coverage. For a 95% confidence interval, a distance smaller than 0.01 (i.e., the coverage probability is between 0.94 and 0.96) is considered to be the best, whereas a distance larger than 0.05 (i.e., the coverage probability is below 0.90) is considered to be the worst. As for the width of the confidence interval, the average difference between the limits of the confidence interval is obtained. Since a shorter width of the confidence interval is desirable, among methods that have a coverage probability close to the nominal coverage probability (e.g., a distance less than 0.01 from 0.95 for a 95% confidence interval), a method with a smaller average width is preferred.

4.2 | Results of the Coverage Probability

We start by presenting the coverage probability results of the Monte Carlo simulation study. Since the coverage probability generally improves as ρ increases, Figure 1 summarizes the results for a two-sided 95% confidence interval with $\rho = 0.7$, which is the lowest value of ρ in our study. Complete results can be found in Supporting Information Table S1. The coverage probability of the confidence intervals is color-graded based on how far it is from 0.95. To indicate when the coverage probability of the confidence interval is greater than 0.95, we have marked it with a plus sign. The figure shows that the confidence interval method *GCI* provides a coverage probability between 0.94 and 0.96 in all scenarios. This is closely followed by

MLS_G , which provides coverage probability within the interval [0.94, 0.96] in many cases. In other cases, for example, when $R = 0.1$ and $k \leq 5$, MLS_G provides over-coverage with a coverage probability within the interval (0.96, 0.98]. Conclusions on under- or over-coverage can be drawn from the complete results in Supporting Information Table S1. The coverage probability of the other methods varies depending on the scenario. The general trends are discussed below on a case-by-case basis.

When $k = 2$ and $R = 0.1$, the confidence interval method MLS_A and VP_F provide coverage within the interval [0.94, 0.96] for $n \geq 20$, while VP_B shows under-coverage [0.92, 0.94] for $n \geq 20$. W_{mat} performs badly for small values of n and provides a coverage within the interval [0.94, 0.96] when $n \geq 60$. MLS_A , VP_F , VP_B and W_{mat} also provide similar coverage for $0 < R < 0.1$ (results not shown). As R increases, the performance of the methods generally worsens. Specifically, W_{log} , W_{mat} , VP_F , and VP_B show significant under-coverage with a coverage below 0.70 (see Supporting Information Table S1). When $R = 1$, MLS_A , however, shows over-coverage with a coverage within (0.96, 0.98], and its performance improves as R increases (for details, see Supporting Information Table S1).

Summarizing the results for $k > 2$, MLS_A provides a coverage within the interval (0.96, 0.98] for $R \leq 2$ and $n \geq 20$, and improves to a coverage within [0.94, 0.96] as R increases. As for the remaining methods, when $R = 0.1$, W_{log} and VP_B provide a coverage within [0.94, 0.96] for $n \geq 20$. For the same R , VP_F exhibits a coverage within [0.94, 0.96] when $n \geq 40$. W_{mat} provides a coverage within [0.94, 0.96] when $n \leq 40$ and over-coverage with a coverage within (0.96, 0.98] when $n \geq 60$. Increasing R , however, has a detrimental effect on the coverage probabilities of W_{log} , W_{mat} , VP_F , and VP_B . For example, when $k = 8$ and $n = 100$, the coverage probability of W_{log} decreases from within [0.94, 0.96] for $R = 0.1$ to less than 0.88 for $R = 2$ (see Supporting Information Table S1).

4.3 | Results of the Average Width

As for the width of the confidence interval, similar to the coverage probability, we chose to present the results for $\rho = 0.7$, as it represents the worst case scenario. The width of the confidence intervals becomes smaller as ρ increases (see Supporting Information Table S1). Figure 2 summarizes the results of the average width from the Monte Carlo simulation study for the 95% two-sided confidence interval for $\rho = 0.7$. Generally, keeping n , k , and ρ constant, as R increases, the average width of the confidence interval increases, which was also observed in previous studies [11, 15]. Furthermore, as expected, increasing k while keeping n , R , and ρ constant, the average width of the confidence interval decreases. From Figure 2 it can be seen that VP_F provides a shorter average width of the confidence interval in most cases. It is important to note that although *GCI* provides a coverage within [0.94, 0.96] in all scenarios studied (see Figure 1), the width of the confidence interval obtained using this method is quite large (see Figure 2). For example, when $R = 0.1$, $k = 2$ and $n = 40$, the width of the confidence interval for *GCI* is about twice as large as that of VP_F (see Supporting Information Table S1). However, VP_F provides

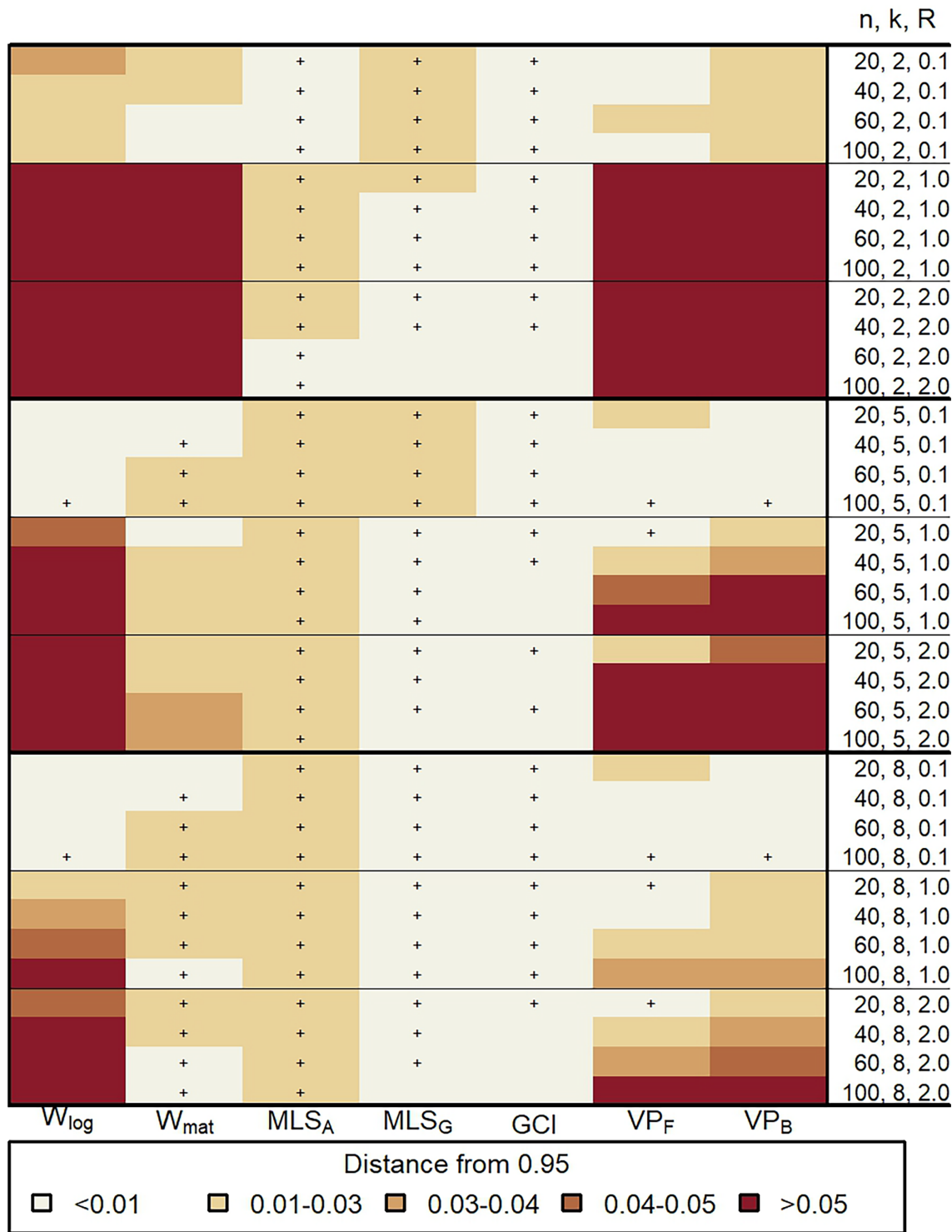


FIGURE 1 | Heatmap depicting the empirical coverage probability from the Monte Carlo simulation study for the 95% two-sided confidence interval for $\rho = 0.7$. The distances from 0.95 are color-graded with a "+" sign denoting over-coverage. A distance ≤ 0.01 indicates that the coverage probability is within the interval $[0.94, 0.96]$ with a "+" sign denoting coverage probability within $[0.95, 0.96]$; a distance $(0.01, 0.03]$ indicates that the coverage probability is within $[0.92, 0.94]$ or $[0.96, 0.98]$ ("+" sign); a distance $(0.03, 0.04]$ indicates that the coverage probability is within $[0.91, 0.92]$ or $[0.98, 0.99]$ ("+" sign); a distance $(0.04, 0.05]$ indicates that the coverage probability is within $[0.90, 0.91]$ or $[0.99, 1.00]$ ("+" sign) and a distance > 0.05 indicates that the coverage probability is < 0.90 .

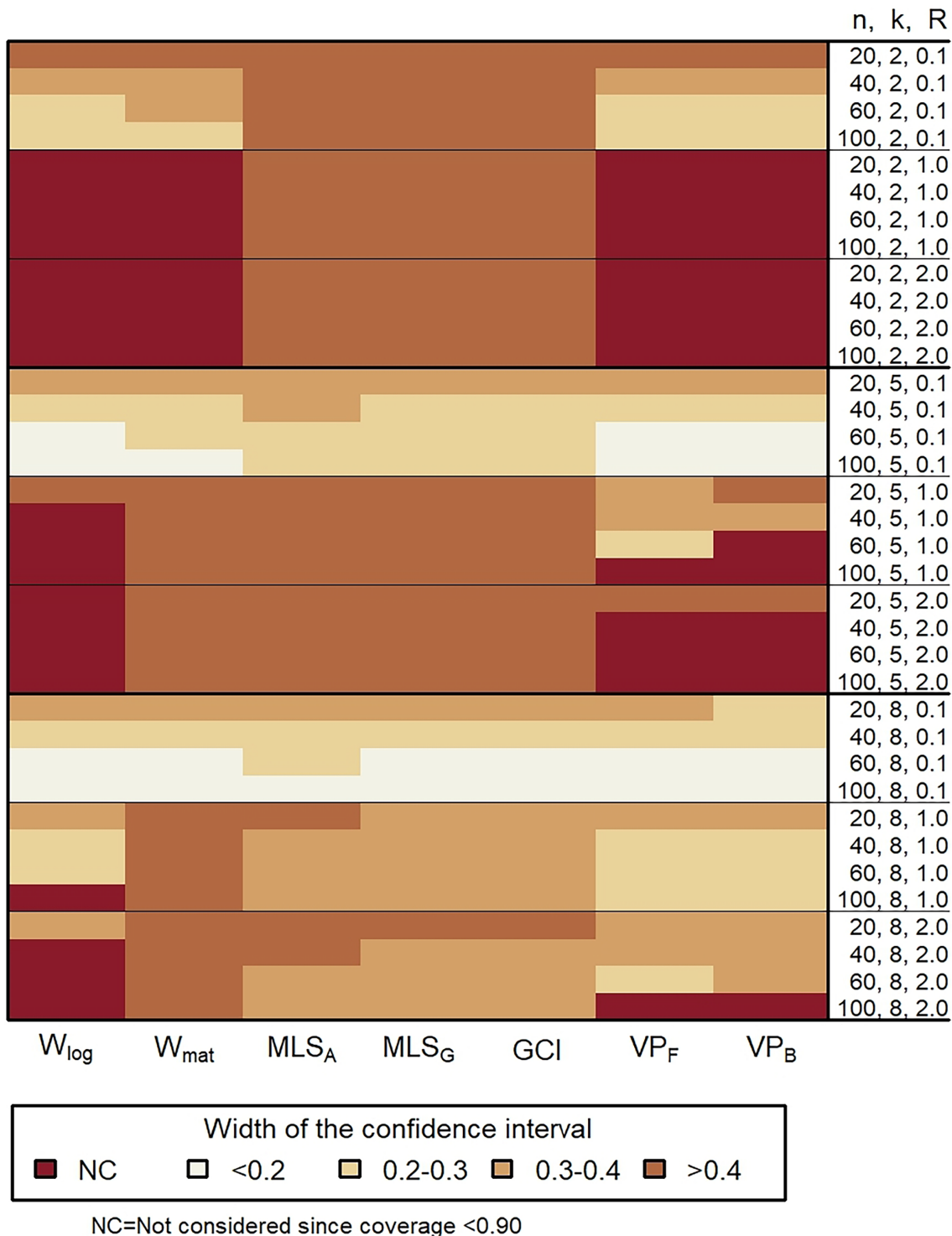


FIGURE 2 | Heatmap depicting the results of the average width of the confidence interval in the Monte Carlo simulation study for the 95% two-sided confidence interval for $\rho = 0.7$. The average width of the confidence interval is not considered (NC) when the coverage probability was < 0.90 (see Figure 1).

a coverage within $[0.94, 0.96]$ almost only for $R \leq 0.1$. Nonetheless, as k increases, for example, when $k = 8$, the average width of the confidence interval for GCI and VP_F is comparable, especially for $n \geq 40$. Note that the average width of MLS_G and GCI were shown to be asymptotically equivalent [18], and this is also observed in Figure 2.

4.4 | Remarks and Guide to Choose a Confidence Interval Method

When selecting a method for constructing a confidence interval for ρ , it is essential to consider several factors. Figures 1 and 2 provide insights into the performance of the different methods

in terms of coverage probability and confidence interval width under various scenarios. The figures show that the number of raters, k , and the rater-to-error variance ratio, R , play a major role in the performance of the methods.

From Figures 1 and 2, it is evident that no method produces a confidence interval width smaller than 0.2 when $R \geq 1$. However, two critical points should be considered. First, we expect the values of R to be smaller than 1 in many applications [18, 27–35], which results in acceptable widths of the confidence intervals and acceptable coverage probabilities. Second, the results presented in this section are for $\rho = 0.7$, which corresponds to the worst-case scenario in our study. Reliability studies typically target a larger value of ρ ($\rho \geq 0.7$). As ρ increases, the width of the confidence interval decreases for all methods, and the coverage probability approaches the nominal coverage for most confidence interval methods, leading to acceptable performance (for details see Supporting Information Table S1).

Table 2 summarizes the results for $\rho \geq 0.7$. When the number of raters is large (e.g., $k \geq 8$), then we recommend the use of GCI or MLS_G . When the number of raters is small (e.g., $k < 8$), we recommend the use of VP_F when $R \leq 0.1$, and again of GCI or MLS_G when $R > 0.1$. Since during the planning stage of a reliability study R is commonly unknown, when one is uncertain about the value of R or when R is known and $R > 0.1$, MLS_G and GCI are the only reasonable options. Notably, no method appears to yield a width of the confidence interval less than 0.2 when $R \geq 1$.

TABLE 2 | Recommendations for confidence interval method combining the results of the coverage probability and average width of the different confidence interval methods.

	$R \leq 0.1$	$R > 0.1$
$k < 8$	VP_F	$MLS_G, GCI, (MLS_A$ when $R \geq 2)$
$k \geq 8$	MLS_G, GCI	MLS_G, GCI

5 | Sample Size Determination for an Expected Width of the Confidence Interval

The number of raters/participants can be determined based on the expected width of the confidence interval approach [36], which utilizes the different confidence interval construction methods discussed in Section 3. The approach searches for the minimum number of participants, n and/or raters, k , such that the expected width of a confidence interval (W) for a planned ICC value, ρ , is approximately equal to or less than a constant ω . We regard the resulting n and k as the optimal combination of n and k . Note that this approach depends on a planned ICC value, ρ , and the rater-to-error variance ratio $R (= \sigma_r^2/\sigma_e^2)$ and thereby yields valid results when the initial guess for ρ and R is accurate.

In the literature, three procedures to determine sample sizes based on the expected width of the confidence interval can be identified. These procedures determine the optimal combination of n and k by searching through possible nodes of the grid for n and k . Saito et al. [7] proposed a procedure that utilizes an analytical formula to search through a small number of nodes in the grid to identify the optimal combination. While using an analytical formula avoids searching through all possible nodes of the grid for n and k , the procedure has two significant drawbacks that limit its practical use. The procedure starts with an arbitrary value of $N = n \times k$ where the combination of n and k , for a given N , is determined using an analytical formula (see Equation 5 in Supporting Information Data S1), which may produce non-integer values. Since integer values for n and k are required, combinations of n and k near the calculated values are obtained. The width of the confidence interval for these combinations is then determined via simulations. If the confidence interval width is not sufficiently small or is much smaller than ω , the previous steps are repeated by increasing or decreasing N , respectively. However, when increasing N , the minimum width of the confidence interval for an integer combination of n and k may not necessarily decrease, as exemplified in Figure 3, where the width is determined using the MLS_G method. A plausible solution is to iteratively determine the confidence interval width by

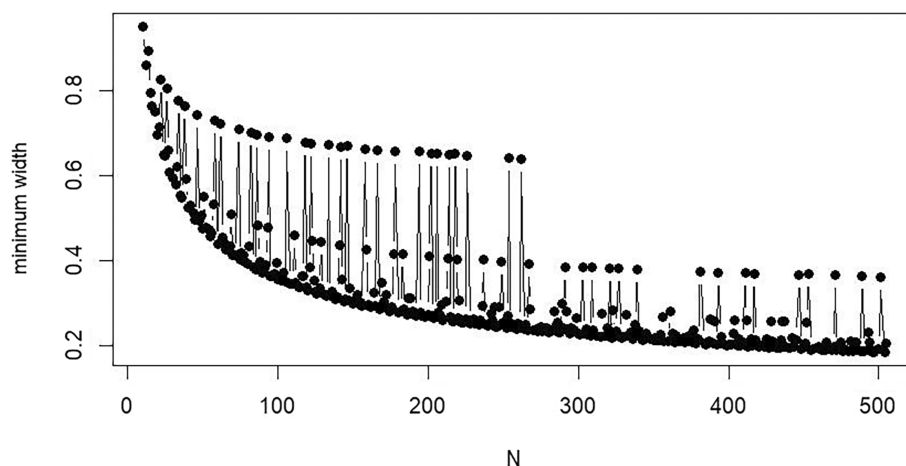


FIGURE 3 | Plot of the minimum width of the confidence interval of MLS_G obtained using the combinations of n and k according to the procedure of Saito et al. [7] for $\rho = 0.7$, $R = 0.1$, and $N \in \{10, 505\}$.

increasing N from the minimum feasible value and checking for each N whether the width is less than ω . However, this significantly increases the computation time required for the procedure and is infeasible if the range of N is not small enough (e.g., if it exceeds three or four values). Another drawback is that, since the procedure relies on specifying a range for the total number of observations N , it may result in a combination of n and k that is not feasible in practice. The details of the procedure are given in Supporting Information Data S1 (Section B), but this procedure is not considered further in our evaluation.

Doros and Lew [17] proposed a general procedure that is purely simulation-based and searches through all possible nodes of n and k to identify the optimal combination. This procedure also requires extensive simulations to evaluate the width of the confidence interval for each combination of n and k . Consequently, it can be rather time-consuming and inefficient. However, this procedure can be adapted by fixing k to a limited set of feasible values, thus reducing the computational burden. Finally, Dobbin and Ionan [18] developed a procedure that allows either n or k to be fixed. By using analytical formulas for the expected width of the confidence intervals for GCI and also MLS_A , this procedure has the advantage of having the shortest computation time. However, since the analytical results cannot be easily extended to other confidence interval methods, this procedure is also the least general. These two procedures are described in detail in the following subsections.

5.1 | Purely Simulation-Based Procedure

Doros and Lew [17] proposed a purely simulation-based procedure for finding the optimal combination of n and k based on the expected width of the confidence interval. The procedure entails:

1. Simulate y_{ij} according to Equation (1), and calculate the mean squares as defined in Table 1;
2. Estimate the expected width of the confidence interval by averaging over a large number of simulations for a variety of n and k values;
3. Return the combination of n and k which results in an expected width of the confidence interval smaller than or approximately equal to the predefined value (ω).

Doros and Lew [17] proposed this sample size determination procedure for the two-moment approximation method, but it may be combined with the other confidence interval methods. To reduce the computational cost of this procedure, and because the search can result in a combination of n and k not feasible in practice, the grid search can be restricted by fixing either n or k . Considering a limited range of values for k , and utilizing the confidence interval methods VP_F and MLS_G , we implemented an algorithm to obtain the minimum number of participants, n^* ensuring that the expected width of the confidence interval remains below a pre-specified value. For each k , the corresponding number of participants n^* is determined using the bisection algorithm, and the steps of the algorithm are provided in Figure A1 in Appendix A. Note that this procedure is very time-consuming when combined with the GCI method. When the confidence interval width is

obtained according to GCI , the procedure proposed by Dobbin and Ionan [18] is preferable (see Section 5.2).

Table 3 shows sample sizes obtained for the confidence interval method VP_F and MLS_G using the procedure proposed by Doros and Lew [17] by fixing $k \in \{5, 10, 15\}$, $\rho \in \{0.7, 0.8, 0.9\}$, and $R \in \{0.1, 2\}$ such that the expected width of the confidence interval is less than ω , with $\omega \in \{0.2, 0.3\}$. The width of the confidence interval was obtained from 10^4 Monte Carlo simulations using a two-sided 95% confidence interval. Sample sizes for VP_F were not calculated for $R = 2$ since when $R = 2$, we found the coverage probability for VP_F to be unacceptable (see Section 4.2). As for MLS_G , it can be further observed from Table 3 that for some scenarios the sample size determination procedure did not return a finite value ($n \leq 500$), denoted by the blanks in the table. This probably occurs since the asymptotic width of MLS_G for the given R , k , and ρ is greater than ω . Comparing the sample sizes for VP_F and MLS_G , it can be observed that sample sizes for MLS_G are rather close to those of VP_F when $R = 0.1$ and $k \geq 10$. It can be further seen from Table 3 that as k or ρ increase, a smaller n is required to achieve the same ω . However, the decrease in n with an increase in k is not as large when k increases from 10 to 15, compared to when k increases from 5 to 10.

TABLE 3 | Sample sizes obtained with the procedure of Doros and Lew [17] using the expected width of the two-sided 95% confidence interval for the method VP_F and MLS_G , for $\rho \in \{0.7, 0.8, 0.9\}$, $R \in \{0.1, 2\}$, $k \in \{5, 10, 15\}$, and $\omega \in \{0.2, 0.3\}$. The width of the confidence interval is based on 10^4 Monte Carlo simulations. No sample size was obtained for VP_F for $R = 2$, since when $R = 2$, we found the coverage probability for VP_F to be unacceptable (see Section 4.2). The blanks for the sample size for MLS_G indicate that a width of the confidence interval less than ω could not be obtained with $n \leq 500$.

R	k	ρ	$\omega = 0.2$		$\omega = 0.3$	
			VP_F	MLS_G	VP_F	MLS_G
0.1	5	0.7	53	85	28	33
		0.8	34	42	16	20
		0.9	15	15	8	7
	10	0.7	45	45	20	18
		0.8	27	27	14	14
		0.9	14	11	8	7
	15	0.7	40	39	19	18
		0.8	25	22	13	11
		0.9	12	11	8	5
2	5	0.7				
		0.8				
		0.9				
	10	0.7				114
		0.8				19
		0.9		19		7
	15	0.7				32
		0.8		89		17
		0.9		17		5

5.2 | Sample Size Determination Procedure According to Dobbin and Ionan [18]

The sample size calculation proposed by Dobbin and Ionan [18] is based on the average width of the confidence interval according to MLS_A and GCI . Note that Dobbin and Ionan [18] refer through their paper to the paper of Cappelleri and Ting [15], who only discussed MLS_G , which could be misleading. Denoting the width of the confidence intervals obtained with the MLS_A and GCI methods as $\hat{v}_{MLS_A}(\mathbf{t})$ and $\hat{v}_{GCI}(\mathbf{t})$, the average widths are,

$$\begin{aligned}\mu_{MLS_A}(q) &= E[\hat{v}_{MLS_A}(\mathbf{t})|q] \text{ and} \\ \mu_{GCI}(q) &= E[\hat{v}_{GCI}(\mathbf{t})|q]\end{aligned}$$

where $q = (\sigma_s^2, \sigma_r^2, \sigma_e^2, n, k) = (\rho, R, n, k)$, that is, the unknown quantities. The sample size determination procedure using the generalized confidence interval method (Section 3.3) aims to find n when k is fixed, by finding the minimum value of n with an expected width smaller than ω , or, similarly, the minimum value of k , when n is fixed. When the modified large sample confidence interval is used, μ_{GCI} is replaced by μ_{MLS_A} in the above criterion. Dobbin and Ionan [18] further derived the asymptotics (see Appendix B), which enables one to check whether a width of the confidence interval is achievable under the parameters specified. When the width is achievable, Dobbin and Ionan [18] showed that the functions μ_{MLS_A} , and μ_{GCI} can be obtained numerically via Monte Carlo simulations.

5.2.1 | Estimating $\mu_{MLS_A}(q)$

A novel procedure referred to as “dependent conditioning” was developed by Dobbin and Ionan [18] to obtain μ_{MLS_A} which can be summarized in the following four steps.

1. The limits of the confidence interval U , L and their difference $U - L$ are written in terms of two F -distributed random variables, $X_{SE} \left(= \frac{MSS/(k\sigma_s^2 + \sigma_e^2)}{MSE/\sigma_e^2} \right)$ and $X_{RE} \left(= \frac{MSR/(n\sigma_s^2 + \sigma_e^2)}{MSE/\sigma_e^2} \right)$.
2. The bivariate joint density of X_{SE} and X_{RE} is obtained, after which the conditional density, $f(x_{RE}|x_{SE})$ is calculated.
3. The variable X_{RE} is integrated out of the conditional density by expressing analytically the integral using a hypergeometric function [18].
4. The final width estimator is obtained using Monte Carlo simulations of a single variable following $F_{n-1, (n-1)(k-1)}$. The width estimator is given as

$$\hat{v}_{MLS_A} = \frac{1}{M} \sum_{m=1}^M (E[U|(x_{SE})_m] - E[L|(x_{SE})_m]),$$

where M is the total number of Monte Carlo simulations and $(x_{SE})_m$ are independent and identically distributed draws from $F_{n-1, (n-1)(k-1)}$. The analytic expressions for $E[U|(x_{SE})_m]$ and $E[L|(x_{SE})_m]$ are obtained using hypergeometric functions, or can be obtained numerically using Monte Carlo simulations [18].

5.2.2 | Estimating $\mu_{GCI}(q)$

The process of obtaining μ_{GCI} requires obtaining $\hat{v}_{GCI}(\mathbf{t})$. Unlike $\hat{v}_{MLS_A}(\mathbf{t})$, obtaining $\hat{v}_{GCI}(\mathbf{t})$ is computationally intensive as it requires estimating quantiles of the pivotal quantity, $G(\mathbf{t})$ (see Equation 11) which is obtained through extensive (e.g., 100 000) [24] Monte Carlo simulations of three χ^2 distributions. A less computationally intensive method to obtain $\hat{v}_{GCI}(\mathbf{t})$ was suggested by Dobbin and Ionan [18] using two variance reduction techniques [37]. First, a modified version of a variance reduction technique called Rao-Blackwellization [38] is used, termed the ‘inverse Rao-Blackwellization estimator’,

$$\begin{aligned}\hat{G}_{RB}(\alpha) \\ = \inf \left\{ g : \alpha \leq \frac{1}{M} \sum_{m=1}^M P[g(\mathbf{t}) \leq g | (\chi_{n-1}^2)_m, (\chi_{(n-1)(k-1)}^2)_m] \right\},\end{aligned}\quad (14)$$

where $(\chi_{n-1}^2)_m, (\chi_{(n-1)(k-1)}^2)_m$ are m^{th} Monte Carlo observations of χ^2 variables to estimate MSS and MSE . Using this estimator, one can obtain the conditional probability in Equation (14) directly from the percentiles of an inverted Gamma-distribution, thus reducing the computational complexity. Second, a variance reduction technique using control variates [39] was applied [18]. Identifying an approximate linear relationship between $v(t) = \hat{G}(1 - \alpha/2) - \hat{G}(\alpha/2)$ and the control variate $\frac{MSR}{MSS}$ when $\rho > 0.7$, the lower and upper limits of the $(1 - \alpha) \times 100\%$ confidence interval are obtained as,

$$\begin{aligned}\hat{U}_{GCI}^{ctrl} &= \hat{U}_{GCI} - \frac{cov(\hat{U}_{GCI}, \frac{MSR}{MSS})}{var(\frac{MSR}{MSS})} \times \left(\frac{MSR}{MSS} - E \left[\frac{MSR}{MSS} \right] \right), \text{ and} \\ \hat{L}_{GCI}^{ctrl} &= \hat{L}_{GCI} - \frac{cov(\hat{L}_{GCI}, \frac{MSR}{MSS})}{var(\frac{MSR}{MSS})} \times \left(\frac{MSR}{MSS} - E \left[\frac{MSR}{MSS} \right] \right),\end{aligned}\quad (15)$$

where $\hat{U}_{GCI} = \hat{G}_{RB}(1 - \alpha/2)$ and $\hat{L}_{GCI} = \hat{G}_{RB}(\alpha/2)$ are the upper and lower limits based on the inverse Rao-Blackwellization estimator, $E[MSR/MSS] = \frac{n\sigma_s^2 + \sigma_e^2}{k\sigma_s^2 + \sigma_e^2} \times \frac{n-1}{n-3}$ and $var(MSR/MSS) = \frac{n\sigma_s^2 + \sigma_e^2}{k\sigma_s^2 + \sigma_e^2} \times \frac{2(n-1)^2(n+k-4)}{(k-1)(n-3)^2(n-5)}$. The remaining terms related to the mean squares in Equation (15) are obtained from the Monte Carlo simulation output. The width estimator is then obtained as,

$$\hat{v}_{GCI}^{ctrl} = \hat{U}_{GCI}^{ctrl} - \hat{L}_{GCI}^{ctrl}.\quad (16)$$

When $\omega > \mu_{MLS_A}$ or $\omega > \mu_{GCI}$, Dobbin and Ionan [18] proposed an algorithm to obtain the sample size n when the other parameters in q are given. We implemented a slightly modified version of the algorithm (see Figure A1 in Appendix A), where $\hat{v}$ is replaced by $\hat{v}_{MLS_A}$ in case of MLS_A , or by $\hat{v}_{GCI}$ in case of GCI . The difference in our implemented algorithm and the one by Dobbin and Ionan concerns the steps of the bisection algorithm. Specifically, where Dobbin and Ionan [18] proposed to obtain a sample size by applying a bisection algorithm in several intervals within the range of plausible values of n , we perform the bisection algorithm over the entire region (n_{min}, n_{max}) in a single step. Both procedures lead to the same solution (n^*).

The same procedure can be applied by fixing n and considering $g(k) = \omega - \hat{v}_{MLS_A}(\sigma_s^2, \sigma_r^2, \sigma_e^2, n, k)$. Dobbin and Ionan [18] proposed numerical methods reducing the computational complexity of the MLS_A and GCI methods, leading to feasible sample size determination procedures. Note that, unlike the previous procedures, this cannot be easily extended to other confidence interval methods. However, the simulation results showed that, although the width of the confidence interval is quite large, the GCI method provides adequate coverage in almost all scenarios. The MLS_A method, however, yields over-coverage in all scenarios that we studied.

Table 4 shows the summary of the sample sizes obtained by the procedure by Dobbin and Ionan [18] using the expected width of the 95% two-sided confidence interval for the methods MLS_A and GCI . The values for the number of participants for MLS_A and GCI indicate the average number of participants obtained using the procedure with 100 replications, while the values in the brackets indicate their standard deviation. The sample size was determined for $\rho \in \{0.7, 0.8, 0.9\}$, $R \in \{0.1, 2\}$, $k \in \{5, 10\}$ such that the expected width of the confidence interval is less than ω , $\omega \in \{0.2, 0.3\}$. Comparing the sample sizes for $R = 0.1$ to the sample sizes obtained in Table 3 for VP_F , we see that VP_F requires a smaller number of participants for the same ω and k . This is to be expected since the average width of the confidence interval for VP_F is smaller than that of MLS_A or GCI (see Supporting Information Table S1).

Table 4 shows that in some scenarios the sample size determination procedure did not return a finite value, denoted by the blanks in the table. This occurs when the asymptotic width of the confidence interval cannot be achieved for the given R , k , and ρ (see Appendix B for further details). Comparing the scenarios with blanks in Table 4 to the average widths of the confidence

interval in Table B2 (Appendix B), it can be seen that the average widths in those scenarios are larger than ω . For example, when $\rho = 0.7$, $k = 5$ and $R = 2$, the asymptotic widths of the confidence interval are 0.46 and 0.42 for MLS_A and GCI , respectively, which are larger than 0.3, the maximum ω in Table 4. This implies that when $n \rightarrow \infty$, the minimum width of the confidence intervals for given ρ , k , and ω is greater than 0.3, and therefore a sample size with $\omega = 0.3$ cannot be obtained. In some scenarios however, for example for $\rho = 0.9$, $k = 5$ and $R = 2$, the asymptotic average width of MLS_A is 0.298 which is rather close to $\omega = 0.3$ and then, a large standard deviation for n was observed in Table 4. We, therefore, recommend that researchers check the asymptotic width of the confidence interval for their choices of ρ , R , and k before proceeding with sample size determination via this procedure.

Furthermore, we also observe blanks (i.e., $n > 500$) for sample sizes determined using MLS_G in Table 3, implying that the asymptotic width of MLS_G is larger than ω for those scenarios. Comparing the blanks in Tables 3 and 4 for MLS_G and MLS_A , respectively, we notice that a finite sample size ($n \leq 500$) could also not be obtained for MLS_A for similar combinations of R , k , ρ and ω . Since an expression for the asymptotic width of the confidence interval for MLS_G is too complicated, this suggests that one can use the asymptotic width of the confidence interval for MLS_A instead.

Table 4 further shows the sample sizes obtained with this procedure, with MLS_A generally larger than that with GCI . Furthermore, the standard deviation of the sample sizes obtained by repeating the procedure with both confidence interval methods was quite large, especially when $R = 2$. Therefore, when $R > 1$, to obtain a sample size with a small mean and standard deviation using the procedure by Dobbin and Ionan [18], we recommend

TABLE 4 | Mean and standard deviation of the number of participants obtained by repeating the sample size determination procedure by Dobbin and Ionan [18] 100 times. The sample size is obtained each time using the expected width of the 95% two-sided confidence interval for the methods MLS_A and GCI , for $\rho \in \{0.7, 0.8, 0.9\}$, $R \in \{0.1, 2\}$, $k \in \{5, 10\}$, and $\omega \in \{0.2, 0.3\}$ utilizing 10^3 Monte Carlo simulations for GCI and 10^5 for MLS_A . The values in brackets are the standard deviation (s.d.) computed on the 100 replications. The blanks in the sample size indicate that a width of the confidence interval less than ω could not be obtained with $n \leq 10^7$.

<i>R</i>	<i>k</i>	ρ	$\omega = 0.2$		$\omega = 0.3$	
			<i>MLS_A</i>	<i>GCI</i>	<i>MLS_A</i>	<i>GCI</i>
			mean(<i>n</i>), (s.d.(<i>n</i>))	mean(<i>n</i>), (s.d.(<i>n</i>))	mean(<i>n</i>), (s.d.(<i>n</i>))	mean(<i>n</i>), (s.d.(<i>n</i>))
0.1	5	0.7	185.3 (0.95)	96.2 (15.62)	41.1 (0.27)	29.6 (5.03)
		0.8	73.5 (0.50)	46 (7.19)	22 (0)	17.7 (2.73)
		0.9	18.7 (0.47)	15.2 (1.95)	8.3 (0.44)	7.8 (1.06)
	10	0.7	58 (0)	42.4 (7.67)	22 (0)	18.3 (3.22)
		0.8	31 (0)	24.9 (4.28)	13 (0)	11.7 (1.9)
		0.9	11 (0)	10.4 (1.46)	6 (0)	5.7 (0.67)
2	5	0.7				
		0.8				
		0.9			7706.7 (4301.11)	287.9 (46.26)
	10	0.7			861.4 (24.85)	150.7 (20.72)
		0.8			41.9 (0.39)	25.8 (3.39)
		0.9	30 (0.17)	21.2 (2.52)	8 (0)	6.7 (0.74)

choosing a large value for k if possible (e.g., $k > 10$). Furthermore, we recommend that researchers repeat the procedure multiple times to decide on an adequate sample size.

5.3 | Guide to Choose a Sample Size Determination Procedure

Three procedures are available in the literature for determining the sample size that ensures that the expected width of the confidence interval for ρ remains below a pre-specified value. These procedures differ in their computational burden and flexibility in selecting confidence interval methods for implementation.

The choice of a confidence interval method directly influences the selection of an appropriate sample size procedure. Table 5 summarizes our recommendations for choosing a confidence interval method and a corresponding sample size procedure. Following our recommendations in Section 4.4, when $R \leq 0.1$ and $k < 8$, selecting the VP_F method requires using the procedure by Doros and Lew. In other cases, GCI or MLS_G can be selected as the confidence interval method. While both methods can be implemented using the procedure by Doros and Lew, its high computational burden makes the procedure by Dobbin and Ionan a more practical choice. However, this limits the choice of the confidence interval to GCI . Note that in this case, the procedure must be repeated several times (e.g., 100 times), and the sample size can be determined based on a summary measure (e.g., the mean) from those repetitions.

5.4 | Software for Sample Size Calculation

Currently, to the best of our knowledge, there is no available software to obtain sample sizes for the ICC for agreement when the data can be modeled using the two-way ANOVA model given in Equation (1). Therefore, a shiny app was developed and made available in Github, enabling researchers to obtain sample sizes of participants and/or raters using the procedures described in Section 5. Additionally, the app permits the calculation of the asymptotic width for MLS_A , GCI and VP_F , as well as determining confidence intervals for empirical data using the methods MLS_A , MLS_G , GCI and VP_F .

TABLE 5 | Recommendations for selecting a sample size determination procedure and a confidence interval method, based on the combined results of coverage probability, average confidence interval width, and the evaluation of sample size determination procedures.

	$R \leq 0.1$	$R > 0.1$
$k < 8$	Purely simulation-based procedure (VP_F)	Procedure according to Dobbin and Ionan (GCI)
$k \geq 8$	Procedure according to Dobbin and Ionan (GCI)	Procedure according to Dobbin and Ionan (GCI)

6 | Empirical Illustration

In this section, we illustrate how the confidence interval methods discussed in Sections 3 and 4, and the sample size determination approaches introduced in Section 5 can be used within the context of a reliability study. The study by Engelhart et al. [40] is used for the purpose of this illustration. This is an obstetric study measuring fetal heart rate (FHR). During intermittent auscultation, the heartbeat of the fetus is measured by midwives using a hand-held Doppler device. If an abnormality (acceleration or deceleration from the baseline heart rate) is detected, then a treatment decision is made. Inter-rater agreement and reliability are important in this context as disagreements may lead to different interventions for the mother and her baby. The primary aim of the study by Engelhart et al. [40] was to evaluate the inter-rater reliability and agreement of fetal heart rate measured using a handheld Doppler device (beats per minute, bpm). The FHR was determined on the same $n = 154$ women undergoing labor by $k = 16$ midwives. After removing missing values, data from 145 FHRs by 16 midwives remained.

The data from this study can be suitably modeled using the two-way ANOVA model (see Equation 1). The ANOVA model makes the assumptions of normality of the FHR measurements by each midwife and that the variance of the measurements by different midwives is homoscedastic. Exploratory analysis revealed that the absolute value of the skewness is less than 1 and the value of the excess kurtosis (averaged across midwives) is less than 2, implying that the midwife measurements can be considered normal [41, 42]. Mild heteroscedasticity is observed. The estimated ICC for agreement, $\hat{\rho} = 0.735$, and the estimated ratio of rater-to-error variance, $\hat{R} = 0.09$. The value of the ICC for agreement indicates that 73.5% of the FHR variability can be attributed to variability between the fetuses. The confidence interval for the ICC for agreement obtained using the seven confidence interval methods is given in Table 6. The estimated ICC for agreement, $\hat{\rho} = 0.735$ (0.682, 0.781), was considered to be sufficient for using the measurements in FHR monitoring.

From Table 6 it can be seen that the limits of the confidence intervals are rather close across the different methods. The width of the confidence interval for the different methods is approximately 0.1. Considering that the value for the ICC for agreement for the population of fetuses and midwives is around $\rho = 0.7$ and the corresponding rater-to-error variance ratio is $R \leq 0.1$, from Figure 1, we can deduce that all the confidence interval methods provide adequate coverage for $n = 145$ and $k = 16$.

We now illustrate how the different sample size determination procedures can be used to plan a reliability study. A researcher

TABLE 6 | Confidence intervals for the ICC for agreement using the seven confidence interval methods.

Limit	W_{log}	W_{mat}	MLS_A	MLS_G	GCI	VP_F	VP_B
Lower	0.688	0.683	0.672	0.684	0.682	0.681	0.685
Upper	0.786	0.787	0.786	0.783	0.781	0.779	0.782

may be interested in planning a reliability study to measure the FHR (bpm) using a handheld Doppler device. Suppose that the researchers expect the ICC for agreement to be 0.7, and want to obtain sample sizes such that the width of the confidence interval is less than 0.1. The researchers further assume the rater-to-error variance ratio to be 0.1 or less. Then, if the number of raters should be between 10 and 15, following our recommendations in Section 4.4 and Section 5.3, the researchers can obtain sample sizes using the procedure of Dobbin and Ionan [18] described in Section 5.2 using the confidence interval method *GCI*. The procedure is repeated 100 times and the mean (and standard deviation) of the calculated number of participants indicates that the researchers can opt for 225 (± 41) participants when a maximum of 10 raters are available, or 167 (± 31) participants when a maximum of 15 raters are available for the reliability study.

7 | Discussion

During the planning stage of a reliability study, it is essential to determine the number of participants and/or raters to limit the uncertainty associated with the sampling process. When the measurements from the reliability study can be modeled using a two-way ANOVA model as in Equation (1), two different ICC coefficients can be obtained, namely the ICC for agreement and the ICC for consistency. In this paper, we provide an overview of the different procedures that were developed to determine the number of participants and/or the number of raters for a reliability study when using the ICC for agreement. These procedures are based on the expected width of the confidence interval.

Before examining sample size determination procedures, we evaluated the statistical properties of seven confidence interval methods for the ICC for agreement, completing the comparison of all available methods. Our simulation study highlighted the critical role of the rater-to-error variance ratio, R , in determining whether a method achieves coverage probabilities close to the nominal level. The results showed that *GCI* consistently provides the closest coverage probability to the nominal level across scenarios. *MLS_G* performed similarly to *GCI* with respect to its coverage probability under most scenarios, with comparable interval widths. These findings align with Ionan et al. [11] earlier study. The estimator of the ICC is a ratio with both numerator and denominator as linear combinations of scaled χ^2 random variables. *GCI* does not rely on approximations but leverages a large number of simulations utilizing samples from scaled χ^2 -distributions (see Equation 11). *MLS_G* also avoids standard approximation techniques, having its lower and upper limits derived separately using the Fieller-type confidence interval for ratios [43]. As a result, *GCI* and *MLS_G* are expected to perform well. However, note that, when $k \leq 8$, the confidence intervals according to *MLS_G* and *GCI* are relatively wide. VP_F , a method not included in prior comparisons, performed well for $R \leq 0.1$, offering narrower confidence intervals while maintaining a coverage probability that was the closest to the nominal level. For $k \geq 8$, *GCI* and *MLS_G* intervals approach the width of VP_F . Higher values of ρ (e.g., $\rho = 0.8, 0.9$; see Supporting Information Table S1) improved coverage probabilities and provided shorter interval widths. Based on these results, we recommend using

GCI or *MLS_G* for $k \geq 8$. For $k < 8$, VP_F is preferable if $R \leq 0.1$; otherwise, *MLS_G* or *GCI* is recommended.

Three procedures to determine the optimal combinations of n and k for a reliability study were identified in the literature. These procedures aim to identify the combination of n and k by exploring a grid of their feasible values, such that the width of the confidence interval is less than a pre-defined constant ω when a planning value for ρ and R is given. The first procedure, proposed by Saito et al. [7] uses an analytical formula to search within a small grid of feasible n and k values. However, this method requires specifying a range for the total number of observations $N = n \times k$, which may result in combinations of n and k that are not feasible in practice. Additionally, if the specified range of N is too broad (e.g., the range of N exceeds three or four values), the procedure can become rather time-intensive. For these reasons, this procedure was not investigated further. The second procedure, proposed by Doros and Lew [17], involves exploring a grid of all feasible values of n and k and uses simulations to calculate the expected confidence interval width for each combination. Although general in its applicability, this approach is also computationally intensive. We adapted this procedure by fixing k and exploring a grid of plausible n , making it computationally less intensive. The third procedure, introduced by Dobbin and Ionan [18] is based on analytical formulas for the confidence interval widths of *MLS_A* and *GCI*. While computationally the most efficient, this procedure is limited to these two confidence interval methods and may not be easily adapted to others. Further, it was observed that for some combinations of ρ and R , no finite sample sizes could be obtained for a given value of k . In such cases, the asymptotic width of the confidence interval methods *MLS_A* and *GCI* derived by Dobbin and Ionan [18] allows to calculate, via simulation, the minimum width the confidence interval one can achieve under the given scenarios. We recommend that researchers examine the asymptotic width of the confidence interval for their chosen ρ , R , and k before proceeding to sample size determination. Furthermore, repeating the sample size procedure by Dobbin and Ionan [18] multiple times, we found that the resulting sample sizes could be rather different. Therefore, we urge researchers using this procedure to repeat the procedure several times and obtain the final sample size using some summary measure, such as taking the mean of sample sizes obtained by repeating this procedure 100 times.

Our findings demonstrate that the choice of sample size determination procedure and confidence interval method significantly impacts the final combination of raters, k , and participants, n , with results highly dependent on the planning values for ρ and R . We therefore advise researchers to carefully consider the requirements of the reliability study as a guide to choose the appropriate procedure. As for the choice of the values of ρ and R to determine a sample size, when one is uncertain about their values, we recommend choosing ρ as small as plausible and R as large as plausible, since this would lead to a conservative sample size. As for the choice of the sample size determination procedure, one may use the procedure of Doros and Lew [17] (adapted by fixing k) for *MLS_G* and VP_F , or of Dobbin and Ionan [18] for *MLS_A* and *GCI*. The use of these procedures was illustrated using FHR measurements by Engelhart et al. [40]. In addition, these

procedures were all implemented in an R Shiny app, which enables users to obtain sample sizes via an interactive and easy-to-use graphical interface.

Our work improves the understanding of the methods to construct a confidence interval for the ICC for agreement by making a comparison of all available methods. Additionally, we evaluate procedures for determining sample sizes for reliability studies, ensuring a precision around the ICC estimate, and offer tools and recommendations for designing a reliability study. However, our study is not without limitations. First, the estimator for the ICC for agreement and its corresponding confidence interval rely on the assumptions of normality of the outcomes as well as homoscedasticity of the variances, in line with the two-way ANOVA model mentioned in Equation 1. The statistical properties of these confidence intervals are impacted by violations of these conditions. We therefore advise researchers to carefully check the assumptions of the ANOVA model before applying the methods mentioned in this paper. Ionan et al. [11] reported that the methods showing good coverages, that is, GCI and MLS_G , are not robust to violations of normality. In such a case, one may use the non-parametric estimation of the ICC introduced by Lu et al. [44], which does not make the assumption of normality. However, a procedure for sample size determination is not yet available in this case. Second, in this study, we consider a complete design, that is, each rater rates every participant. This may not be the case when, for example, some measurements are missing by design. The estimation of the ICC for agreement in such a case has been discussed by ten Hove [45], however, no sample size determination has yet been developed under this scenario. Third, the ANOVA model considered in this paper does not include an interaction term. However, when raters rate each participant more than once, an interaction effect between the raters and participants can be included in the model and estimated from the data. Whether the same methods discussed in this paper can be extended to such a design requires further investigation. Fourth, in this paper, we reviewed the sample size determination procedures under the width of the confidence interval approach. In another approach, researchers may be interested in finding sample sizes that ensure a high probability of the width of the confidence interval being less than a pre-specified value. This is known as the assurance probability approach [3, 46]. How to extend the methods studied in this manuscript, analytically and/or computationally, to accommodate the assurance probability approach is left for future work. Fifth, a researcher may, however, be interested in determining sample sizes for a study where the aim is to statistically test whether the ICC for agreement is larger than a pre-defined value. No work has yet been done in this regard, and hence it is a topic for future research.

Acknowledgments

We want to thank Dr. K. Dobbin for providing the R codes for their sample size determination procedure, and Dr. N. Demetrashvili for providing the codes for their implementation of the variance partitioning confidence interval. Their contributions have significantly helped in this study. We also thank Dr. C. Engelhart and Dr. E. Blix for providing the data.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that supports the findings of this study are available in the Supporting Information of this article.

References

1. K. McGraw and S. Wong, "Forming Inferences About Some Intraclass Correlation Coefficients," *Psychological Methods* 1 (1996): 30–46.
2. P. Shrout and J. Fleiss, "Intraclass Correlations: Uses in Assessing Rater Reliability," *Psychological Bulletin* 86, no. 2 (1979): 420–428.
3. D. Mondal, S. Vanbelle, A. Cassese, and M. Candel, "Review of Sample Size Determination Methods for the Intraclass Correlation Coefficient in the One-way Analysis of Variance Model," *Statistical Methods in Medical Research* 33 (2024): 532–553, <https://doi.org/10.1177/09622802231224657>.
4. J. Fleiss and P. Shrout, "Approximate Interval Estimation for a Certain Intraclass Correlation Coefficient," *Psychometrika* 43 (1978): 259–262, <https://doi.org/10.1007/BF02293867>.
5. K. Zou and M. McDermott, "Higher-Moment Approaches to Approximate Interval Estimation for a Certain Intraclass Correlation Coefficient," *Statistics in Medicine* 18 (1999): 2051–2061.
6. Y. Xiao and H. Liu, "Modified Profile Likelihood Approach for Certain Intraclass Correlation Coefficients," *Computational Statistics* 28 (2013): 2241–2265, <https://doi.org/10.1007/s00180-013-0405-x>.
7. Y. Saito, T. Sozu, C. Hamada, and I. Yoshimura, "Effective Number of Subjects and Number of Raters for Inter-Rater Reliability Studies," *Statistics in Medicine* 25 (2006): 1547–1560.
8. R. Almehrzi and M. Emam, "Asymptotic Standard Errors of Intraclass Correlation Coefficients for Two-Way Model," *Communications in Statistics: Simulation and Computation* 52, no. 5 (2021): 2073–2092.
9. C. Arteaga, S. Jeyaratnam, and F. Graybill, "Confidence Intervals for Proportions of Total Variance in the Two-Way Cross Component of Variance Model," *Communications in Statistics - Theory and Methods* 11 (1982): 1643–1658.
10. R. Gui, F. Graybill, R. Burdick, and N. Ting, "Confidence-Intervals on Ratios of Linear-Combinations for Non-Disjoint Sets of Expected Mean Squares," *Journal of Statistical Planning and Inference* 48 (1995): 215–227.
11. A. Ionan, M. Polley, L. McShane, and K. Dobbin, "Comparison of Confidence Interval Methods for an Intra-Class Correlation Coefficient (ICC)," *BMC Medical Research Methodology* 14 (2014): 121.
12. S. Weerahandi, "Generalized Confidence-Intervals," *Journal of the American Statistical Association* 88 (1993): 899–905.
13. N. Demetrashvili, E. Wit, and E. Heuvel, "Confidence Intervals for Intraclass Correlation Coefficients in Variance Components Models," *Statistical Methods in Medical Research* 25 (2016): 2359–2376.
14. V. Rousson, T. Gasser, and B. Seifert, "Confidence Intervals for Intraclass Correlation in Inter-Rater Reliability," *Scandinavian Journal of Statistics* 30 (2003): 617–624.
15. J. Cappelleri and N. Ting, "A Modified Large-Sample Approach to Approximate Interval Estimation for a Particular Intraclass Correlation Coefficient," *Statistics in Medicine* 22 (2003): 1861–1877.
16. L. Tian and J. Cappelleri, "A New Approach for Interval Estimation and Hypothesis Testing of a Certain Intraclass Correlation Coefficient: The Generalized Variable Method," *Statistics in Medicine* 23 (2004): 2125–2135.
17. G. Doros and R. Lew, "Design Based on Intra-Class Correlation Coefficients," *American Journal of Biostatistics* 1 (2010): 1–8.
18. K. Dobbin and A. Ionan, "Sample Size Methods for Constructing Confidence Intervals for the Intra-Class Correlation Coefficient," *Computational Statistics & Data Analysis* 85 (2015): 67–83.

19. J. Bartko, "The Intraclass Correlation Coefficient as a Measure of Reliability," *Psychological Reports* 19 (1966): 3–11, <https://journals.sagepub.com/doi/abs/10.2466/pr0.1966.19.1.3>.
20. F. Graybill and C. Wang, "Confidence Intervals on Nonnegative Linear Combinations of Variances," *Journal of the American Statistical Association* 75 (1980): 869–873, <http://www.jstor.org/stable/2287174>.
21. R. Burdick, C. Borror, and D. Montgomery, *Design and Analysis of Gauge R&R Studies* (Society for Industrial, 2005), <https://epubs.siam.org/doi/abs/10.1137/1.978089871837910.1137/1.9780898718379>.
22. M. Shoukri and ProQuest, *Measures of Interobserver Agreement and Reliability* (CRC Press, 2011).
23. J. Hannig, H. Iyer, and P. Patterson, "Fiducial Generalized Confidence Intervals," *Journal of the American Statistical Association* 101 (2006): 254–269, <https://doi.org/10.1198/016214505000000736>.
24. R. Burdick, C. Borror, and D. Montgomery, *Design and Analysis of Gauge R & R Studies: Making Decisions With Confidence Intervals in Random and Mixed Anova Models* (Asa-Siam Series on Statistics and Applied Probability 17) (Society for Industrial, 2005).
25. S. Searle, G. Casella, and C. McCulloch, *Variance Components* (Wiley, 1992).
26. E. Heuvel, "A Comparison of Estimation Methods on the Coverage Probability of Satterthwaite Confidence Intervals for Assay Precision With Unbalanced Data," *Communications in Statistics-Simulation and Computation* 39 (2010): 777–794.
27. A. Kassambara, "Datarium: Data Bank for Statistical Analysis and Visualization," 2019, <https://CRAN.R-project.org/package=datarium>.
28. N. Ting, J. Cappelleri, and A. Bushmakina, "Two Confidence Interval Approaches on the Dependability Coefficient in a Two-Factor Crossed Design," *Journal of Biopharmaceutical Statistics* 19 (2009): 610–624.
29. D. Streiner and G. Norman, *Health Measurement Scales: A Practical Guide to Their Development and Use* (Oxford University Press, 1995).
30. A. Cutti, A. Ferrari, P. Garofalo, M. Raggi, A. Cappello, and A. Ferrari, "Outwalk: A Protocol for Clinical Gait Analysis Based on Inertial and Magnetic Sensors," *Medical & Biological Engineering & Computing* 48 (2010): 17–25.
31. D. Hart, "Equivalence Reliability and Convergent Validity of Percent Body Fat Prediction Equations," *American Journal of Sports Science and Medicine* 7 (2019): 45–50, <http://pubs.sciepub.com/>.
32. J. Bland and D. Altman, "Measuring Agreement in Method Comparison Studies," *Statistical Methods in Medical Research* 8 (1999): 135–160, <https://doi.org/10.1177/096228029900800204>.
33. B. Carstensen, L. Gurrin, C. Ekström, and M. Figurski, "MethComp: Functions for Analysis of Agreement in Method Comparison Studies," 2013.R Package Version 122.
34. J. Carrasco and L. Jover, "Estimating the Generalized Concordance Correlation Coefficient Through Variance Components," *Biometrics* 59 (2003): 849–858.
35. J. L. Fleiss, *Repeated Measurements Studies in the Design and Analysis of Clinical Experiments* (Wiley Online Books, 1999), 220–240, <https://doi.org/10.1002/9781118032923.ch8>.
36. D. Bonett, "Sample Size Requirements for Estimating Intraclass Correlations With Desired Precision," *Statistics in Medicine* 21 (2002): 1331–1335, <https://www.ncbi.nlm.nih.gov/pubmed/12111881>.
37. W. L. Dunn and J. K. Shultis, "Chapter 5–Variance Reduction Techniques," in *Exploring Monte Carlo Methods*, 2nd ed., ed. W. L. Dunn, J. K. Shultis, W. L. Dunn, and J. K. Shultis (Elsevier, 2023), 151–187, <https://doi.org/10.1016/B978-0-12-819739-4.00013-5>.
38. C. Robert and G. Casella, *Monte Carlo Statistical Methods* (Springer New York, 1999), <https://doi.org/10.1007/978-1-4757-3071-5>.
39. P. Rothery, "The Use of Control Variates in Monte Carlo Estimation of Power," *Journal of the Royal Statistical Society. Series C* 31, no. 2 (1982): 125–129, <https://doi.org/10.2307/2347974>.
40. C. H. Engelhart, S. Vanbelle, P. Øian, A. S. D. Pay, A. Kaasen, and E. Blix, "How Well Can the Fetal Heart Rate Baseline be Assessed by Intrapartum Intermittent Auscultation? An Interrater Reliability and Agreement Study," *Birth* 51, no. 4 (2024): 835–842.
41. J. Hair, W. Black, B. Babin, and R. Anderson, *Multivariate Data Analysis* (Cengage Learning, 2019).
42. P. Curran, S. West, and J. Finch, "The Robustness of Test Statistics to Nonnormality and Specification Error in Confirmatory Factor Analysis," *Psychological Methods* 1 (1996): 16–29.
43. E. Fieller, "Some Problems in Interval Estimation," *Journal of the Royal Statistical Society: Series B: Methodological* 16 (1954): 175–185, <https://doi.org/10.1111/j.2517-6161.1954.tb00159.x>.
44. N. Lu, T. Chen, P. Wu, et al., "Functional Response Models for Intraclass Correlation Coefficients," *Journal of Applied Statistics* 41 (2014): 2539–2556.
45. D. Hove, T. Jorgensen, and L. Ark, "Updated Guidelines on Selecting an Intraclass Correlation Coefficient for Interrater Reliability, With Applications to Incomplete Observational Designs," *Psychological Methods* 29, no. 5 (2024): 967–979.
46. G. Zou, "Sample Size Formulas for Estimating Intraclass Correlation Coefficients With Precision and Assurance," *Statistics in Medicine* 31 (2012): 3972–3981, <https://www.ncbi.nlm.nih.gov/pubmed/22764084>.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1.** Appendices. **Table S1.** Coverage of the seven confidence interval methods for different combinations of n, k, R and rho. Average width of the seven confidence interval methods for different combinations of n, k, R and rho. <https://github.com/DiproMondal/sample-size-ICCagreement-2wayANOVA>.

Appendix A

Steps of the Sample Size Determination Algorithm

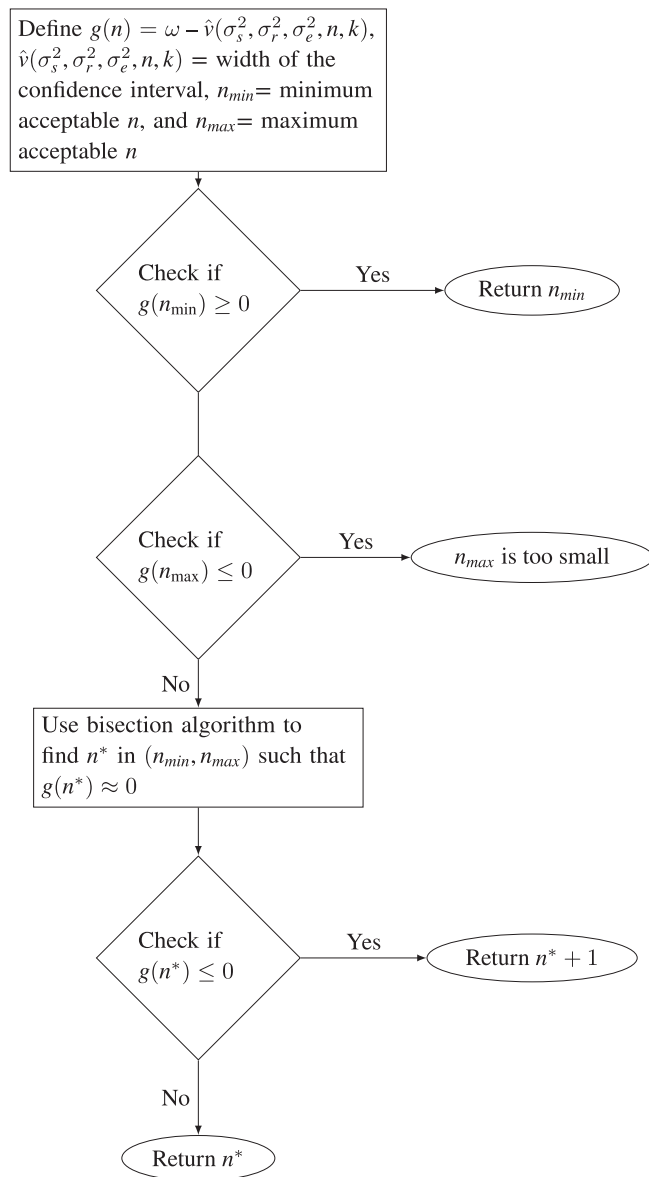


FIGURE A1 | The steps of the algorithm to obtain a sample size of participants, n for a pre-specified expected width of the confidence interval ω , a number of raters, k , a planned value for the ICC for agreement, ρ and the rater-to-error variance ratio, R .

Appendix B

Asymptotics

Asymptotics for the Generalized Confidence Interval

$n \rightarrow \infty$

Denoting $k_{lim}^2 = \lim_{n \rightarrow \infty} \frac{1}{k-1} \sum_{j=1}^{k-1} (\bar{y}_{.j} - \bar{y}_{..})^2$, under the assumptions of the ANOVA model (Equation 1) the pivotal quantity reduces to [18],

$$g(t) \xrightarrow{d} \max \left\{ 0, \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2 + (k_{lim}^2/[W_1/(k-1)])} \right\}$$

where $\xrightarrow{d}$ represent convergence in distribution and $W_1 \sim \chi_{k-1}^2$.

$k \rightarrow \infty$

Denoting $n_{lim}^2 = \lim_{k \rightarrow \infty} \frac{1}{n-1} \sum_{i=1}^{k-1} (\bar{y}_{i.} - \bar{y}_{..})^2$, under the assumptions of the ANOVA model Equation (1) [18], the pivotal quantity reduces to,

$$g(t) \xrightarrow{d} \max \left\{ 0, \frac{n_{lim}^2/[W_2/(n-1)]}{\sigma_r^2 + \sigma_e^2 + (n_{lim}^2/[W_2/(n-1)])} \right\}$$

where $W_2 \sim \chi_{n-1}^2$.

Asymptotics for the Modified Large Sample Confidence Interval

The derivation of the asymptotics by Dobbin and Ionan [18] utilizes the expression of the limits of the confidence interval method MLS_A .

$n \rightarrow \infty$

Denoting $k_{lim}^2 = \lim_{n \rightarrow \infty} \frac{1}{k-1} \sum_{j=1}^k (\bar{y}_{.j} - \bar{y}_{..})^2$, the lower (L) and upper (U) limits of the confidence interval can be written as [18],

$$Ld \rightarrow \max \left\{ 0, \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2 + k_{lim}^2/F_{k-1,\infty}(\frac{\alpha}{2})} \right\}, \text{ and}$$

$$Ud \rightarrow \max \left\{ 0, \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2 + k_{lim}^2/F_{k-1,\infty}(1 - \frac{\alpha}{2})} \right\}$$

$k \rightarrow \infty$

Denoting $n_{lim}^2 = \lim_{k \rightarrow \infty} \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_{i.} - \bar{y}_{..})^2$, the lower (L) and upper (U) limits of the confidence interval can be written as [18],

$$Ld \rightarrow \max \left\{ 0, \frac{F_{\infty,n-1}(\frac{\alpha}{2})n_{lim}^2}{\sigma_r^2 + \sigma_e^2 + F_{\infty,n-1}(\frac{\alpha}{2})n_{lim}^2} \right\}, \text{ and}$$

$$Ud \rightarrow \max \left\{ 0, \frac{F_{\infty,n-1}(1 - \frac{\alpha}{2})n_{lim}^2}{\sigma_r^2 + \sigma_e^2 + F_{\infty,n-1}(1 - \frac{\alpha}{2})n_{lim}^2} \right\}$$

Asymptotics for the Variance Partitioning Confidence Interval

We derive the asymptotic expression for the variance partitioning confidence interval (VP_F) considering the case where the estimator of the variance components and the variance of the variance components are constructed via ANOVA estimation. Recall that the distribution of $\hat{\rho}$ when using the confidence interval method VP_F , is given by (see Section 3.4.1),

$$\frac{\hat{\sigma}_s^2 d\hat{f}_1 d\hat{f}_2}{\hat{\sigma}_s^2 d\hat{f}_1 d\hat{f}_2 + \hat{\sigma}_E^2}$$

where $d\hat{f}_1 = \frac{2\hat{\sigma}_s^4}{\hat{V}(\hat{\sigma}_s^2)}$ and $d\hat{f}_2 = \frac{2\hat{\sigma}_E^4}{\hat{V}(\hat{\sigma}_E^2)} = \frac{2(\hat{\sigma}_s^2 + \hat{\sigma}_r^2)^2}{\hat{V}(\hat{\sigma}_s^2) + \hat{V}(\hat{\sigma}_r^2)}$, respectively. To find the expression for the asymptotic width of the confidence interval, we need the asymptotic expressions for $\hat{\sigma}_s^2$, $\hat{\sigma}_r^2$, $\hat{\sigma}_e^2$, $\hat{V}(\hat{\sigma}_s^2)$, $\hat{V}(\hat{\sigma}_r^2)$ and $\hat{V}(\hat{\sigma}_e^2)$.

Following Dobbin and Ionan [18], $\hat{\sigma}_s^2$ converges in distribution to σ_s^2 , and $\hat{\sigma}_e^2$ converges in distribution to σ_e^2 , and $\hat{\sigma}_r^2$ converges in distribution to $k_{lim}^2 = \lim_{n \rightarrow \infty} \frac{1}{k-1} \sum_{j=1}^k (\bar{y}_{.j} - \bar{y}_{..})^2$ (see Appendix B of Dobbin and Ionan [18]).

We obtain the variance of the estimator of the variances from the variance of mean squares given in Section 3.1.1. The estimated variance of the estimator of the variance of the participants is obtained as,

$$\hat{V}(\hat{\sigma}_s^2) = \frac{1}{k^2} (V(MSS) + V(MSE))$$

$$= \frac{2}{k^2(n-1)} \left((k\hat{\sigma}_s^2 + \hat{\sigma}_e^2)^2 + \frac{(\hat{\sigma}_e^2)^2}{k-1} \right).$$

This converges in distribution to,

$$V(\hat{\sigma}_s^2) = \frac{2}{k^2(n-1)} \left((k\sigma_s^2 + \sigma_e^2)^2 + \frac{\sigma_e^4}{k-1} \right).$$

Therefore, under the limit $n \rightarrow \infty$, $V(\hat{\sigma}_s^2) \rightarrow 0$.

The estimated variance of the estimator of the variance of the raters is obtained as,

$$\begin{aligned} \hat{V}(\hat{\sigma}_r^2) &= \frac{1}{n^2} (V(MSR) + V(MSE)) \\ &= \frac{2}{n^2(k-1)} \left((n\hat{\sigma}_r^2 + \hat{\sigma}_e^2)^2 + \frac{(\hat{\sigma}_e^2)^2}{n-1} \right) \\ &= \frac{2}{(k-1)} \left(\left(\hat{\sigma}_r^2 + \frac{\hat{\sigma}_e^2}{n} \right)^2 + \frac{(\hat{\sigma}_e^2)^2}{n^2(n-1)} \right). \end{aligned}$$

This converges in distribution to,

$$V(\hat{\sigma}_r^2) = \frac{2}{(k-1)} \left(\left(k_{lim}^2 + \frac{\sigma_e^2}{n} \right)^2 + \frac{\sigma_e^4}{n^2(n-1)} \right).$$

Therefore, under the limit $n \rightarrow \infty$, $V(\hat{\sigma}_r^2) \rightarrow \frac{2(k_{lim}^2)^2}{k-1}$.

The estimated variance of the estimator of the variance of the errors is obtained as,

$$\hat{V}(\hat{\sigma}_e^2) = V(MSE) = \frac{2(\hat{\sigma}_e^2)^2}{(n-1)(k-1)}.$$

This converges in distribution to,

$$V(\hat{\sigma}_e^2) = \frac{2\sigma_e^4}{(n-1)(k-1)}.$$

Therefore, under the limit $n \rightarrow \infty$, $V(\hat{\sigma}_e^2) \rightarrow 0$.

Then, plugging $V(\hat{\sigma}_s^2)$ in the equation for $d\hat{f}_1$ gives, under the limit $n \rightarrow \infty$, $d\hat{f}_1 \rightarrow \infty$. Plugging $V(\hat{\sigma}_r^2)$ and $V(\hat{\sigma}_e^2)$ in the equation for $d\hat{f}_2$ gives, under the limit $n \rightarrow \infty$, $d\hat{f}_2 \rightarrow (k-1) \left(1 + \frac{\sigma_e^2}{k_{lim}^2} \right)^2$.

Using Slutsky's theorem, the limits of the confidence interval can now be written as,

$$\begin{aligned} L &\xrightarrow{d} \frac{\sigma_s^2 F_{\infty, (k-1) \left(1 + \frac{\sigma_e^2}{k_{lim}^2} \right)^2 \left(\frac{\alpha}{2} \right)}}{\sigma_s^2 F_{\infty, (k-1) \left(1 + \frac{\sigma_e^2}{k_{lim}^2} \right)^2 \left(\frac{\alpha}{2} \right) + k_{lim}^2 + \sigma_e^2} \text{ and} \\ U &\xrightarrow{d} \frac{\sigma_s^2 F_{\infty, (k-1) \left(1 + \frac{\sigma_e^2}{k_{lim}^2} \right)^2 \left(1 - \frac{\alpha}{2} \right)}}{\sigma_s^2 F_{\infty, (k-1) \left(1 + \frac{\sigma_e^2}{k_{lim}^2} \right)^2 \left(1 - \frac{\alpha}{2} \right) + k_{lim}^2 + \sigma_e^2}. \end{aligned}$$

When R is small however, for example $R \leq 0.1$, numerical results presented in Table B1 show that k_{lim}^2 can be replaced by σ_r^2 . Note that, in Section 3.4, the variance of the variance components is based on maximum-likelihood estimation. However, when $n \rightarrow \infty$ and $k \rightarrow \infty$, ANOVA and ML estimation lead to similar results (see Searle et al. [25]).

TABLE B1 | Comparison of the Asymptotic Width of the Confidence Interval With VP_F Under Different Scenarios, Using Different Specifications of df_2 . That Is, $df_2 = (k-1)(1 + \frac{\sigma_e^2}{k_{lim}^2})^2$ (*Model 1*), and $df_2 = (k-1)(1 + \frac{\sigma_e^2}{\sigma_r^2})^2$ (*Model 2*). the Results for Models 1 and 2 Are Obtained Using $n = 1\,000$ and Using 10^5 Simulations.

ρ	k	R	<i>Model 1</i>	<i>Model 2</i>
0.7	2	0.1	0.0942	0.1053
0.8	2	0.1	0.0735	0.0802
0.9	2	0.1	0.0429	0.0452
0.7	5	0.1	0.0517	0.0528
0.8	5	0.1	0.0397	0.0403
0.9	5	0.1	0.0225	0.0227
0.7	10	0.1	0.0352	0.0353
0.8	10	0.1	0.0269	0.0269
0.9	10	0.1	0.0152	0.0151

TABLE B2 | Width of the Confidence Interval With MLS_A , GCI , and VP_F for $k \in \{5, 10\}$, $R \in \{0.1, 2\}$, and $\rho \in \{0.7, 0.8, 0.9\}$. the Results for MLS_A and GCI Are Obtained Using $n = 1\,000$, and the Table Shows the Average Width of the Confidence Interval From 10^5 Simulations, While the Values in Brackets Are the Standard Deviation of the Widths. the Results for VP_F Are Obtained Analytically (According to *Model 2*).

k	R	ρ	<i>MLS_A</i>	<i>GCI</i>	<i>VP_F</i>
			$E[W], sd(W)$	$E[W], sd(W)$	$E[W]$
5	0.1	0.7	0.120 (0.065)	0.100 (0.061)	0.053
		0.8	0.098 (0.057)	0.080 (0.052)	0.040
		0.9	0.059 (0.038)	0.048 (0.034)	0.023
	2	0.7	0.458 (0.113)	0.421 (0.127)	
		0.8	0.412 (0.131)	0.369 (0.139)	
		0.9	0.298 (0.128)	0.256 (0.126)	
10	0.1	0.7	0.051 (0.021)	0.047 (0.021)	0.035
		0.8	0.040 (0.017)	0.036 (0.017)	0.027
		0.9	0.023 (0.01)	0.021 (0.01)	0.015
	2	0.7	0.287 (0.063)	0.272 (0.067)	
		0.8	0.237 (0.066)	0.221 (0.068)	
		0.9	0.148 (0.053)	0.136 (0.052)	

Results of the Asymptotic Width With the Generalized Confidence Interval, the Modified Large Sample Confidence Interval, and Variance Partitioning Method

In Table B2 the asymptotic width of the confidence interval with MLS_A and GCI are obtained using simulations, according to Dobbin and Ionan [18] (see Appendix B). As for the width of the confidence interval with VP_F , the asymptotic width of the confidence interval is obtained analytically following our derivation (see *Model 2* in Appendix B).

Appendix C

Revisions to Published Formulas

Modified Large Sample Confidence Interval for ρ (MLS_A)

The derivation of the upper and lower limits of the confidence interval for MLS_A does not correspond to the final equation for the limits of the con-

fidence interval for the transformation of ρ . The final expression provided in Arteaga et al. [9] (Equation 2.9) is,

$$L_A^* = \frac{-1 + F_{\alpha/2:n_1,\infty}^{-1} F_1 + (1 - F_{\alpha/2:n_1,\infty}^{-1} F_{\alpha/2:n_1,n_3}) F_{\alpha/2:n_1,n_3} F_1}{n_1 + F_{\alpha/2:n_1,\infty}^{-1} F_{\alpha/2:n_1,n_2} F_2},$$

$$U_A^* = \frac{-1 + F_{1-\alpha/2:n_1,\infty}^{-1} F_1 + (1 - F_{1-\alpha/2:n_1,\infty}^{-1} F_{1-\alpha/2:n_1,n_3}) F_{1-\alpha/2:n_1,n_3} F_1^{-1}}{n_1 + F_{1-\alpha/2:n_1,\infty}^{-1} F_{1-\alpha/2:n_1,n_2} F_2}.$$

The correct expression should be

$$L_A^* = \frac{-1 + F_{\alpha/2:n_1,\infty}^{-1} F_1 + (1 - F_{\alpha/2:n_1,\infty}^{-1} F_{\alpha/2:n_1,n_3}) F_{\alpha/2:n_1,n_3} F_1^{-1}}{n_1 + F_{\alpha/2:n_1,\infty}^{-1} F_{\alpha/2:n_1,n_2} F_2},$$

$$U_A^* = \frac{-1 + F_{1-\alpha/2:n_1,\infty}^{-1} F_1 + (1 - F_{1-\alpha/2:n_1,\infty}^{-1} F_{1-\alpha/2:n_1,n_3}) F_{1-\alpha/2:n_1,n_3} F_1^{-1}}{n_1 + F_{1-\alpha/2:n_1,\infty}^{-1} F_{1-\alpha/2:n_1,n_2} F_2}.$$

For a definition of all quantities in the above equation, consult Arteaga et al. [9].

Generalized Confidence Interval Method

The expression for the generalized pivotal quantity given by Ionan et al. [11] (in Appendix A, Page 10) contains a typo for c_4 . The expression is

$$G = \frac{\max\{0, c_2 s_2^2 / W_2 - c_3 s_2^2 / W_3\}}{c_1 s_1^2 / W_1 + c_2 s_2^2 / W_2 + c_4 s_2^2 / W_3},$$

where $c_2 = (b_0 - 1)/(l_0 r_0)$, $c_3 = (b_0 l_0 r_0 - b_0 - l_0 + 1)/(l_0 r_0)$, $c_1 = (l_0 - 1)/(b_0 r_0)$, and $c_4 = c_3(b_0 l_0 r_0 - b_0 - l_0 + 1)/b_0$.

The correct expression for c_4 should be $c_4 = c_3(b_0 l_0 r_0 - b_0 - l_0)/b_0$. For a definition of all quantities in the above equation, consult Ionan et al. [11].