Confidence Intervals for Intraclass Correlation in Inter-Rater Reliability

Rousson, V ; Gasser, Theo ; Seifert, Burkhardt

DOI: https://doi.org/10.1111/1467-9469.00353

 Vol 30: 617–624, 2003

Confidence Intervals for Intraclass Correlation in Inter-Rater Reliability

VALENTIN ROUSSON, THEO GASSER and BURKHARDT SEIFERT
University of Zürich

ABSTRACT. Calculation of a confidence interval for intraclass correlation to assess inter-rater reliability is problematic when the number of raters is small and the rater effect is not negligible. Intervals produced by existing methods are uninformative: the lower bound is often close to zero, even in cases where the reliability is good and the sample size is large. In this paper, we show that this problem is unavoidable without extra assumptions and we propose two new approaches. The first approach assumes that the raters are sufficiently trained and is related to a sensitivity analysis. The second approach is based on a model with fixed rater effect. Using either approach, we obtain conservative and informative confidence intervals even from samples with only two raters. We illustrate our point with data on the development of neuromotor functions in children and adolescents.

Key words: inter-rater data, intraclass correlation, lower bound of confidence interval, rater effect, sensitivity analysis

1. Introduction

Intraclass correlation is a widely used concept to assess inter-rater reliability (when several raters perform a single measurement on a group of subjects). A low reliability may indicate that the raters are not well trained or that the variable to be measured is not well defined or standardized. Hence, the reliability issue is of great importance in many fields.

We consider a random sample of n subjects for which a continuous variable Y is measured independently by d raters randomly selected from a population of raters. Denote by Y_{ij} the measurement made on the ith subject by the jth rater (for $i = 1,\ldots, n$ and $j = 1,\ldots, d$). Let us assume the model

$$Y_{ij} = \mu + s_i + r_j + e_{ij}, \tag{1}$$

where μ is fixed, and where s_i, r_j and e_{ij} are independent random effects which are normally distributed with mean 0 and variances σ_s^2, σ_r^2 and σ_e^2, respectively. The term s_i is the subject effect, whereas r_j is the rater effect – indicating rater bias – and e_{ij} is a measurement error. Intraclass correlation is defined as the (unconditional) correlation between two measurements Y_{ij_1} and Y_{ij_2} on the same subject i by two different raters j_1 and j_2, which is equal to

$$\rho = \frac{\mathrm{Cov}(Y_{ij_1}, Y_{ij_2})}{\mathrm{Var}(Y_{ij_1})^{1/2}\mathrm{Var}(Y_{ij_2})^{1/2}} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_e^2}.$$

This coefficient takes values between 0 and 1 such that the larger this quantity, the better the reliability.

In the present paper, we focus on calculating a $(1 - \alpha)$-confidence interval for ρ ($0 \leq \alpha \leq 1$). Thus, we wish to determine a lower bound L and an upper bound U such that the probabilities for ρ, to be smaller than L or larger than U, are both equal to $\alpha/2$. When negative values arise for L, they will be put to zero throughout the paper. In practice, the reliability is considered to be good if L is at least 0.75 (see e.g. Lee *et al.*, 1989).

Existing methods are reviewed in section 2 along with their drawbacks. In section 3, we show that any valid method for constructing a confidence interval for ρ will often be uninformative (i.e. the lower bound will often be close to zero) when the number of raters d is small and when the rater effect σ_r^2 is not negligible. We, therefore, investigated other approaches to define confidence intervals for intraclass correlation, to be presented in sections 4 and 5.

Throughout the paper, a data set assessing inter-rater reliability of a test battery of neuromotor functions in children and adolescents in used for illustration (see Largo *et al.*, 2001). This test battery consists of various fine motor tasks. Time needed to accomplish these tasks is recorded. We restrict our attention to four of these tasks, namely repetitive and alternating foot movements, repetitive finger movements and a pegboard task. In our sample, we have $n = 30$ children and $d = 2$ raters.

2. Existing methods

For introduction of notation, let us define the well-known quantities $\mathrm{SS}_s = d\sum_{i=1}^{n}(\bar{Y}_{i.} - \bar{Y}_{..})^2$, $\mathrm{SS}_r = n\sum_{j=1}^{d}(\bar{Y}_{.j} - \bar{Y}_{..})^2$ and $\mathrm{SS}_e = \sum_{i=1}^{n}\sum_{j=1}^{d}(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$, where $\bar{Y}_{i.} = \sum_{j=1}^{d} Y_{ij}/d$, $\bar{Y}_{.j} = \sum_{i=1}^{n} Y_{ij}/n$ and $\bar{Y}_{..} = \sum_{i=1}^{n}\sum_{j=1}^{d} Y_{ij}/(nd)$. Under model (1), quantities $\mathrm{MS_s} = \mathrm{SS_s}/(n-1)$, $\mathrm{MS}_r = \mathrm{SS}_r/(d-1)$ and $\mathrm{MS}_e = \mathrm{SS}_e/((n-1)(d-1))$ have expectations $d\sigma_s^2 + \sigma_e^2$, $n\sigma_r^2 + \sigma_e^2$ and σ_e^2, respectively. Thus we obtain unbiased estimates of σ_s^2, σ_r^2 and σ_e^2 as $\hat{\sigma}_s^2 = (\mathrm{MS}_s - \mathrm{MS}_e)/d$, $\hat{\sigma}_r^2 = (\mathrm{MS}_r - \mathrm{MS}_e)/n$ and $\hat{\sigma}_e^2 = \mathrm{MS}_e$. In the remainder of the paper, we shall often consider the ratio $\psi_{r/e} = \sigma_r^2/\sigma_e^2$. An estimate of $\psi_{r/e}$ is then given by $\hat{\psi}_{r/e} = \hat{\sigma}_r^2/\hat{\sigma}_e^2$, while an estimate of intraclass correlation (see Bartko, 1966) is given by

$$\hat{\rho} = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_r^2 + \hat{\sigma}_e^2} = \frac{\mathrm{MS}_s - \mathrm{MS}_e}{\mathrm{MS}_s + (d-1)\mathrm{MS}_e + (d/n)(\mathrm{MS}_r - \mathrm{MS}_e)}.$$

Fleiss & Shrout (1978) considered an appropriate linear combination W of $\mathrm{MS_r}$ and $\mathrm{MS_e}$ and they determined a parameter ν_{W} such that the first two moments of $\nu_{\mathrm{W}} W/E[W]$ coincide with those of a chi-square distribution with ν_{W} degrees of freedom. The lower bound of an approximate $(1-\alpha)$-confidence interval for ρ was then given by

$$L = \frac{n(\mathrm{MS}_s - F_{1-\alpha/2,n-1,\nu_W}\mathrm{MS}_e)}{n\mathrm{MS}_s + F_{1-\alpha/2,n-1,\nu_W}(d\mathrm{MS}_r + (dn-d-n)\mathrm{MS}_e)},$$

where F_{α,ν_1,ν_2} denotes the α-quantile of an F distribution with ν_1 and ν_2 degrees of freedom, while the upper bound may be obtained from the same formula replacing $(1-\alpha/2)$ by $\alpha/2$. This method is asymptotically valid if both n and d tend to infinity. If only n tends to infinity, however, the chi-square approximation is not good enough and the method tends to be much too liberal; see our simulations below.

More recently, Zou & McDermott (1999) proposed similar methods based on three- or four-moment approximations to determine ν_{W}. Previously, Arteaga *et al.* (1982) derived an explicit formula for a confidence interval for intraclass correlation which is asymptotically exact when n or d tends to infinity. In Gui *et al.* (1995), a general approach for obtaining a confidence interval for a ratio of expected mean squares is proposed, which is applicable to intraclass correlation.

To check how these methods work for small values of d, we simulated 5000 replications of size $d = 2$ and $n = 30$, 100, 500, according to model (1) with $\rho = 0.8$ and $\psi_{r/e} = 1$ or $1/3$. The confidence level $(1-\alpha) = 0.95$ was used throughout (see Table 1 for a summary of the results). Methods proposed by Fleiss & Shrout (1978) and Zou & McDermott (1999) did not

Table 1. *Comparison of performance of five methods for constructing a 95% confidence interval for intra-class correlation based on 5000 replications, where $d = 2$ and $\rho = 0.8$ for different values of n and $\psi_{r/e}$. Methods are abbreviated by FS: Fleiss & Shrout; ZM3: Zou & McDermott (three-moment approximation); ZM4: Zou & McDermott (four-moment approximation); A: Arteaga* et al.*; G: Gui* et al. *PrI, PrL and PrU are approximations of the probability (in percentage) for ρ to be inside, on the left side and on the right side of the interval, respectively. Ideally, they should be equal to 95.0, 2.5 and 2.5. medL and medU are the median left bound and the median right bound of the 5000 intervals*

		$\psi_{r/e} = 1$					$\psi_{r/e} = 1/3$				
Method	n	PrI	PrL	PrU	medL	medU	PrI	PrL	PrU	medL	medU
FS	30	87.8	11.2	1.0	0.52	0.93	95.3	3.6	1.1	0.59	0.91
ZM3	30	91.4	8.2	0.4	0.00	0.95	96.9	2.9	0.2	0.40	0.93
ZM4	30	91.8	7.8	0.4	0.00	0.95	97.1	2.7	0.2	0.31	0.93
A	30	97.0	1.7	1.3	0.01	0.92	97.4	0.6	2.0	0.03	0.90
G	30	96.0	1.3	2.7	0.00	0.92	96.7	0.3	3.0	0.00	0.90
FS	100	72.8	26.4	0.8	0.62	0.92	92.0	7.9	0.1	0.70	0.89
ZM3	100	83.3	16.5	0.2	0.00	0.95	94.0	6.0	0.0	0.00	0.93
ZM4	100	84.8	15.0	0.2	0.00	0.95	94.3	5.7	0.0	0.18	0.93
A	100	96.5	1.8	1.7	0.02	0.90	97.6	1.0	1.4	0.03	0.87
G	100	95.1	1.8	3.1	0.00	0.90	96.1	0.6	3.3	0.00	0.87
FS	500	67.0	32.5	0.5	0.64	0.91	73.7	26.3	0.0	0.74	0.87
ZM3	500	86.4	13.4	0.2	0.00	0.96	84.5	15.5	0.0	0.00	0.94
ZM4	500	88.0	11.8	0.2	0.00	0.96	86.2	13.8	0.0	0.00	0.94
A	500	96.1	2.1	1.8	0.02	0.89	97.1	1.6	1.3	0.03	0.85
G	500	95.1	2.2	2.7	0.00	0.88	95.3	1.7	3.0	0.00	0.85

perform well, especially when $\psi_{r/e} = 1$. Moreover, their performances deteriorated for large values of n. The method of Fleiss & Shrout (1978) was, in general, much too liberal. For large rater effect or large n, the intervals proposed by Zou & McDermott (1999) were not informative as the median left bound was equal to zero (which is a trivial bound). On the other hand, the approximations proposed by Arteaga *et al*. (1982) and Gui *et al*. (1995) appeared to be reasonably accurate and did not deteriorate for large values of n. Unfortunately, these intervals were very large and hence not informative. Simulations with larger values of ρ and those with $d = 3$ led to similar conclusions.

The application of these methods to neuromotor data showed that problems also arise for real data. Lower bounds of the various 95 per cent confidence intervals are given in Table 2, together with estimates of ρ and $\psi_{r/e}$. For repetitive finger movements, the estimated inter-rater reliability $\hat{\rho}$ was small and a low lower bound of the confidence interval seems appropriate. For alternative foot movements and pegboard, however, the last four methods produced lower bounds close to zero, although $\hat{\rho}$ was pretty large. In fact, according to these methods, only repetitive foot movements achieved a good inter-rater reliability.

Table 2. *Lower bounds of 95% confidence intervals derived by seven methods for inter-rater reliability of four neuromotor tasks. Estimates of ρ and $\psi_{r/e}$ are also provided. The last two columns contain estimates obtained using the methods proposed in sections 4 and 5*

			Lower bounds of 95% CI for ρ						
Task	$\hat{\rho}$	$\hat{\psi}_{r/e}$	FS	ZM3	ZM4	A	G	$L(1)$	$\tilde{L}$
Repetitive foot movements	0.98	0.00	0.96	0.96	0.96	0.75	0.76	0.91	0.94
Alternating foot movements	0.92	0.68	0.63	0.00	0.00	0.02	0.00	0.81	0.73
Repetitive finger movements	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pegboard	0.96	1.12	0.66	0.00	0.00	0.04	0.00	0.93	0.86

3. Asymptotically exact confidence intervals

In this section, we investigate in more detail the problem of obtaining a low lower bound of a confidence interval for intraclass correlation. We shall see that this problem is not specific to the methods of Arteaga *et al.* (1982) or Gui *et al.* (1995), but arises for any method which produces an exact confidence interval when d is fixed and n tends to infinity.

In such a case, the lower bound L_n should satisfy $\lim_{n\to\infty} \Pr\{\rho \leq L_n\} = \alpha/2$. As MS_s and MS_e are consistent estimates of $d\sigma_s^2 + \sigma_e^2$ and σ_e^2, respectively, we have for any positive random variable X

$$\lim_{n\to\infty} \Pr\left\{\sigma_r^2 \geq \frac{(1-X)(MS_s - MS_e) - dX MS_e}{dX}\right\} = \Pr\{\rho \leq X\}.$$

Thus, an asympotic formula for L_n is given by solving the equation

$$\frac{(1-L_n)(MS_s - MS_e) - dL_n MS_e}{dL_n} = A_n,$$

where A_n is the upper bound of a $(1-\alpha)$-confidence interval for σ_r^2 which is exact when n tends to infinity. Note that $(d-1)MS_r/(n\sigma_r^2)$ is asymptotically distributed as χ^2_{d-1}. As a consequence, we may define $A_n = (d-1)MS_r/(n\chi^2_{\alpha/2,d-1})$, where $\chi^2_{\alpha,\nu}$ denotes the α-quantile of a chi-square distribution with ν degrees of freedom. Therefore, any confidence interval for ρ which is exact when n tends to infinity will have a lower bound which is asymptotically equal to

$$L_n = \frac{MS_s - MS_e}{MS_s + d(d-1)MS_r/(n\chi^2_{\alpha/2,d-1}) + (d-1)MS_e}.$$

We may now calculate the probability that L_n is larger than a given value ρ_0 (e.g. $\rho_0 = 0.75$) when n tends to infinity. By considering $\rho_0 \leq \rho$, we have

$$\lim_{n\to\infty} \Pr\{L_n \geq \rho_0\} = \Pr\left\{\frac{X^2}{\chi^2_{\alpha/2,d-1}} \leq \frac{\rho}{1-\rho} \cdot \left(1 + \frac{\sigma_e^2}{\sigma_r^2}\right) \cdot \left(\frac{1}{\rho_0} - \frac{1}{\rho}\right) + 1\right\},$$

where X^2 is distributed as χ^2_{d-1}. This probability increases with $1/\psi_{r/e}$, ρ, $1/\rho_0$ and d (see Table 3 for some examples with $\alpha = 0.05$ and $\rho_0 = 0.75$).

We see that these probabilities are low for small values of d and large values of $\psi_{r/e}$, even when ρ is as high as 0.95. This means that an asymptotically exact confidence interval for ρ will be useless in these cases because its lower bound will rarely be larger than 0.75. In other words, even in cases where the true intraclass correlation ρ is very high, we will not be able to prove it with few raters and $\psi_{r/e}$ not negligible.

Table 3. *Probabilities that the lower bound of an asymptotically ($n \to \infty$) exact 95% confidence interval for ρ is larger than 0.75 for different values of d, $\psi_{r/e}$ and ρ*

	$\psi_{r/e} = 3$			$\psi_{r/e} = 1$			$\psi_{r/e} = 1/3$		
d	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
2	0.03	0.05	0.07	0.03	0.06	0.09	0.04	0.07	0.12
3	0.04	0.09	0.19	0.04	0.12	0.26	0.06	0.20	0.43
4	0.04	0.15	0.37	0.05	0.22	0.53	0.08	0.42	0.81
5	0.05	0.22	0.58	0.06	0.34	0.77	0.11	0.64	0.97
10	0.08	0.64	0.99	0.12	0.86	1.00	0.29	1.00	1.00
20	0.15	0.97	1.00	0.27	1.00	1.00	0.65	1.00	1.00
30	0.23	1.00	1.00	0.41	1.00	1.00	0.86	1.00	1.00
50	0.39	1.00	1.00	0.66	1.00	1.00	0.99	1.00	1.00
100	0.70	1.00	1.00	0.94	1.00	1.00	1.00	1.00	1.00

4. Population of trained raters

We have shown that it may be impossible to obtain a valid and informative confidence interval in realistic situations. The reason for this is that the rater effect σ_r^2 cannot be accurately estimated from a sample with few raters. This does not mean though that no conclusions at all can be obtained. We need, however, to make some extra assumptions about σ_r^2, or more precisely about the ratio $\psi_{r/e}$. This ratio compares two kinds of errors. A value larger than 1 would indicate that the raters have difficulties to agree on the general level of rating, compared with assessing individual deviations from this level. As an appropriate training of raters and standardization of the task allows to reduce the rater effect, $\psi_{r/e} \leq 1$ appears to be a reasonable assumption in many cases. Given such an upper bound, we may define a conservative and informative confidence interval for ρ, as shown in this section.

Let $\psi_{s/e} = \sigma_s^2/\sigma_e^2$. Under model (1), quantities $(n-1)\mathrm{MS}_s/(d\sigma_s^2 + \sigma_e^2)$ and $(n-1)(d-1)\mathrm{MS}_e/\sigma_e^2$ are independent and distributed as χ^2_{n-1} and $\chi^2_{(n-1)(d-1)}$, respectively. As a consequence, $\mathrm{MS}_s/\mathrm{MS}_e$ is distributed as $(d\psi_{s/e} + 1)F_{n-1,(n-1)(d-1)}$. Thus, an exact $(1-\alpha)$-confidence interval for $\psi_{s/e}$ is obtained as $[A(\alpha/2);\ A(1-\alpha/2)]$ where

$$A(\alpha) = \frac{F_{\alpha,(n-1)(d-1),n-1}\mathrm{MS}_s - \mathrm{MS}_e}{d\mathrm{MS}_e}.$$

From this, we may derive an exact confidence interval for a quantity which depends monotonically on $\psi_{s/e}$. Observe that we may write $\rho = \psi_{s/e}/(\psi_{s/e} + \psi_{r/e} + 1)$. Thus, for any fixed value of $\psi_{r/e}$, $\rho = \rho(\psi_{s/e})$ is increasing in $\psi_{s/e}$. As a consequence, an exact $(1-\alpha)$-confidence interval for ρ given the value of $\psi_{r/e}$ is obtained as $[L(\psi_{r/e});\ U(\psi_{r/e})]$, where

$$L(\psi_{r/e}) = \frac{A(\alpha/2)}{A(\alpha/2) + \psi_{r/e} + 1}$$

and $U(\psi_{r/e})$ is obtained by replacing $\alpha/2$ by $(1-\alpha/2)$.

Thus, if we knew $\psi_{r/e}$, we would have an exact confidence interval for ρ. In practice, we may often assume that $\psi_{r/e}$ lies in an interval $[\psi_0;\ \psi_1]$, for example in the interval $[0;\ 1]$ as discussed above. In such a case, the confidence interval $[L(\psi_1);\ U(\psi_0)]$ for ρ is conservative. Most importantly, for any fixed value of d, the length of this interval tends to zero when n tends to ∞.

Figure 1 plots such 95 per cent confidence intervals $[L(\psi_{r/e});\ U(\psi_{r/e})]$ for values of $\psi_{r/e} \leq 3$ and for the four neuromotor tasks. The estimates of $\psi_{r/e}$ and ρ given in Table 2 are also represented by a dotted vertical line and a dotted horizontal line, respectively. These confidence intervals provide useful information in the spirit of a sensitivity analysis.

Lower bounds $L(1)$ are given in Table 2. We may draw here the following conclusion: if the raters in the sample came from a population of trained raters (such that the rater effect is smaller than measurement error), intraclass correlation would be larger than $L(1)$ with a probability larger than 0.975. As $L(1)$ is larger than 0.75 for repetitive foot movements and alternating foot movements, we may have some confidence that the reliability is good for these tasks. For pegboard, $L(1)$ is also pretty high, but the estimate of $\psi_{r/e}$ is slightly larger than 1. Nevertheless, as $L(3)$ is still clearly larger than 0.75 (see Fig. 1), the reliability can be expected to be good for this task too. Recall that such conclusions were not possible using the existing methods.

5. Model with fixed rater effect

In this section, we consider the raters in the sample to be the whole population of raters. Thus, we consider model (1) with fixed rater effect r_j. We use a parametrization such that $\sum_{j=1}^{d} r_j = 0$, while quantifying the rater effect by $\tilde{\sigma}_r^2 = \sum_{j=1}^{d} r_j^2/(d-1)$. Intraclass

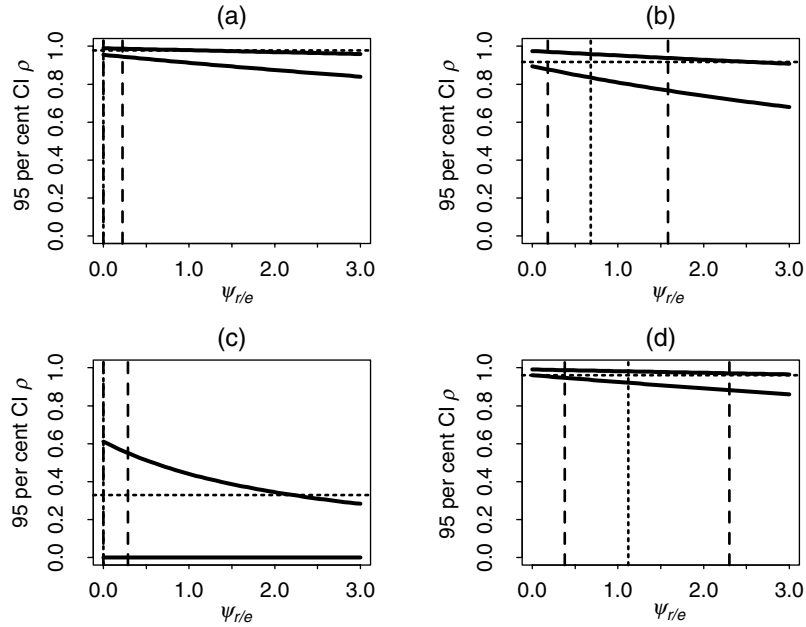

Fig. 1. Plot of 95 per cent confidence intervals $[L(\psi_{r/e}),\ U(\psi_{r/e})]$ for all values $\psi_{r/e}$ between 0 and 3, assessing inter-rater reliability of (a) repetitive foot movements, (b) alternating foot movements, (c) repetitive finger movements and (d) pegboard. Dotted vertical line refers to the sample estimate of $\psi_{r/e}$ (or $\tilde{\psi}_{r/e}$). Dashed vertical lines refer to the limits of a 95 per cent confidence interval of $\tilde{\psi}_{r/e}$. Dotted horizontal line refers to the sample estimate of ρ.

correlation is here defined as $\tilde{\rho} = \sigma_s^2/(\sigma_s^2 + \tilde{\sigma}_r^2 + \sigma_e^2)$. A similar approach has been considered by Dolezal *et al.* (1998).

Using a model with fixed rater effect leaves estimation unchanged. The expectation of $\mathrm{MS_r}$ is still equal to $n\tilde{\sigma}_r^2 + \sigma_e^2$, so that $\hat{\sigma}_r^2$ and $\hat{\rho}$ (as defined in section 2) are still estimates of $\tilde{\sigma}_r^2$ and $\tilde{\rho}$. The distribution of these estimates is, however, different. In particular, $\hat{\sigma}_r^2$ and $\hat{\rho}$ are consistent for $\tilde{\sigma}_r^2$ and $\tilde{\rho}$ when d is fixed and n tends to infinity. As a consequence, it is possible to obtain a conservative confidence interval for $\tilde{\rho}$ whose length tends to zero. Note, however, that the conclusions drawn from such an interval do not necessarily extend to other raters.

Let $\tilde{\psi}_{r/e} = \tilde{\sigma}_r^2/\sigma_e^2$. The quantity $\mathrm{MS}_r/\mathrm{MS}_e$ is distributed as non-central F with $(d-1)$ and $(n-1)(d-1)$ degrees of freedom, and with non-centrality parameter $(d-1)n\tilde{\psi}_{r/e}$. We have hence

$$\Pr\left\{\frac{\mathrm{MS}_r}{\mathrm{MS}_e} \leq F_{\alpha,d-1,(n-1)(d-1),(d-1)n\tilde{\psi}_{r/e}}\right\} = \alpha.$$

Thus, an exact $(1-\alpha)$-confidence interval for $\tilde{\psi}_{r/e}$ is obtained as $[B(\alpha/2);\ B(1-\alpha/2)]$, where $B(\alpha) = \lambda(\alpha)/((d-1)n)$ and $\lambda(\alpha)$ satisfies

$$F_{1-\alpha,d-1,(n-1)(d-1),\lambda(\alpha)} = \frac{\mathrm{MS}_r}{\mathrm{MS}_e}.$$

As cumulative probabilities for non-central F-distributions are tabulated in many standard statistical packages (e.g. in S-Plus), and as the quantiles of a non-central F-distribution are decreasing in the non-centrality parameter, the values of $\lambda(\alpha)$ can be easily computed by

bisection. Corresponding 95 per cent confidence intervals for $\tilde{\psi}_{r/e}$ for the four neuromotor tasks are represented by dashed vertical lines in Fig. 1.

Note that $[L(\tilde{\psi}_{r/e}), U(\tilde{\psi}_{r/e})]$ (as defined in section 4) is an exact $(1 - \alpha)$-confidence interval for $\tilde{\rho}$ given the value of $\tilde{\psi}_{r/e}$. As a consequence, $\tilde{L} = L(B(1 - \alpha/2))$ is a conservative lower bound for $\tilde{\rho}$ at the level $(1-2\alpha)$. To see this, note that as $L(\tilde{\psi}_{r/e})$ is decreasing in $\tilde{\psi}_{r/e}$, we have

$$\Pr\left\{\tilde{\rho} \geq L(B(1-\alpha/2)) \middle| B(1-\alpha/2) \geq \tilde{\psi}_{r/e}\right\} \geq \Pr\left\{\tilde{\rho} \geq L(\tilde{\psi}_{r/e})\right\} = 1 - \alpha/2.$$

It follows:

$$\begin{aligned}\Pr\{\tilde{\rho} \geq L(B(1-\alpha/2))\} &\geq \Pr\left\{\tilde{\rho} \geq L(B(1-\alpha/2)) \text{ and } \tilde{\psi}_{r/e} \leq B(1-\alpha/2)\right\} \\ &\geq \Pr\left\{\tilde{\rho} \geq L(\tilde{\psi}_{r/e})\right\} \cdot \Pr\left\{\tilde{\psi}_{r/e} \leq B(1-\alpha/2)\right\} \\ &\geq 1 - \alpha.\end{aligned}$$

Similarly, the quantity $\tilde{U} = U(B(\alpha/2))$ is a conservative upper bound for $\tilde{\rho}$ at the level $(1-2\alpha)$. For example, if $[B(\alpha/2); B(1-\alpha/2)]$ and $[L(\tilde{\psi}_{r/e}), U(\tilde{\psi}_{r/e})]$ are 95 per cent confidence intervals for $\tilde{\psi}_{r/e}$ and for $\tilde{\rho}$ given the value of $\tilde{\psi}_{r/e}$, respectively, then $[\tilde{L}, \tilde{U}]$ is a conservative 90 per cent confidence interval for $\tilde{\rho}$.

From plots as in Fig. 1, the lower bound $\tilde{L}$ of a 90 per cent confidence interval for $\tilde{\rho}$ may be obtained as the ordinate of the intersection of the lower curve with the right dashed vertical line. Similarly, the upper bound $\tilde{U}$ is obtained as the ordinate of the intersection of the upper curve with the left dashed vertical line.

Lower bounds $\tilde{L}$ of 95 percent confidence intervals for $\tilde{\rho}$ are given in Table 2. These lead to the following conclusion: if the raters in the sample were the whole population of raters, intraclass correlation would be larger than $\tilde{L}$ with a probability larger than 0.975. Here also, these values are close to or higher than 0.75 for all tasks, except for the repetitive finger movements. Again, we come to the conclusion that three tasks achieved a good reliability, whereas the fourth one had a clearly insufficient inter-rater reliability.

6. Conclusions

Intraclass correlation is a useful concept to assess inter-rater reliability. Unfortunately, any valid confidence interval for intraclass correlation will often be uninformative with few raters. In real life, however, we often have only two or three raters. In such a situation, one *has to* make some extra assumptions about the rater effect if one wishes to get relevant information about reliability. In this paper, we have proposed two realistic approaches which allow to get confidence that a reliability is good, even from samples with two raters, as illustrated by our example.

Acknowledgements

This research was supported by the Swiss National Science Foundation (Project no. 3200-045829.95/2).

References

Arteaga, C., Jeyaratnam, S. & Graybill, F. A. (1982). Confidence intervals for proportions of total variance in the two-way cross component of variance model. *Comm. Statist. Theory Methods* **11**, 1643–1982.

Bartko, J. J. (1966). Intraclass correlation coefficient as a measure of reliability. *Psychol. Rep.* **19**, 3–11.
Dolezal, K. K., Burdick, R. K. & Birch, N. J. (1998). Analysis of a two-factor R & R study with fixed operators. *J. Qual. Technol.* **30**, 163–170.
Fleiss, J. L. & Shrout, P. E. (1978). Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika* **43**, 259–262.
Gui, R., Graybill, F. A., Burdick, R. K. & Ting, N. (1995). Confidence intervals on ratios of linear combinations for non-disjoint sets of expected mean squares. *J. Statist. Plann. Inference* **48**, 215–227.
Largo, R., Caflisch, J., Hug, F., Muggli, K., Sheehy, A., Gasser, Th. & Molinari, L. (2001). Neuromotor development from 5 to 18 years. Part I: timed performance. *Develop. Med. Child Neurol.* **43**, 436–443.
Lee, J., Koh, D. & Ong, C. N. (1989). Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Comput. Biol. Med.* **19**, 61–70.
Zou, K. H. & McDermott, M. P. (1999). Higher-moment approaches to approximate interval estimation for a certain intraclass correlation coefficient. *Statist. Med.* **18**, 2051–2061.

Received April 2002, in final form December 2002

Valentin Rousson, Department of Biostatistics, University of Zürich, Sumatrastrasse 30, CH-8006 Zürich, Switzerland.
E-mail: rousson@ifspm.unizh.ch