AStA Advances in Statistical Analysis
https://doi.org/10.1007/s10182-023-00483-x

ORIGINAL PAPER

Measures of interrater agreement for quantitative data

Daniela Marella[1] · Giuseppe Bove[2]

Received: 11 February 2023 / Accepted: 14 September 2023

Abstract
In this paper measures of interrater absolute agreement for quantitative measurements based on the standard deviation are proposed. Such indices allow (i) to overcome the limits affecting the intraclass correlation index; (ii) to measure the interrater agreement on single targets. Estimators of the proposed measures are introduced and their sampling properties are investigated for normal and non-normal data. Simulated data are employed to demonstrate the accuracy and practical utility of the new indices for assessing agreement. Finally, an application to assess the consistency of measurements performed by radiologists evaluating tumor size of lung cancer is presented.

Keywords Interrater agreement · Intraclass correlation · One-way ANOVA · Resampling

1 Introduction

The agreement between ratings or measurements given by two or more raters (humans or devices) on a group of targets (subjects or objects) have been considered in applications regarding biomedical sciences, education, psychometrics and other disciplines (for a review see, for example, Shoukri 2011; Broemeling 2009 and von Eye and Mun 2005). For instance, the agreement among clinical diagnoses provided by more physicians on a nominal scale is analyzed for identifying the best treatment

Daniela Marella and Giuseppe Bove have contributed equally to this work.

✉ Daniela Marella
daniela.marella@uniroma1.it

Giuseppe Bove
giuseppe.bove@uniroma3.it

[1] Dipartimento di Scienze Sociali ed Economiche, Università “La Sapienza”, Piazzale Aldo Moro 5, 00185 Rome, Italy

[2] Dipartimento di Scienze della Formazione, Università Roma TRE, Via del Castro Pretorio 20, 00185 Rome, Italy

Published online: 10 October 2023

for the patient, or the agreement among ratings of educators who assess on a new ordinal rating scale the language proficiency of a corpus of argumentative (written or oral) texts is considered to test reliability of the new scale.

In this paper we focus on the analysis of the agreement among quantitative (discrete or continuous) measurements, like, for instance, those provided by radiologists measuring the tumor size of lung cancer patients who could be considered in a clinical trial (this example is presented in the application of Sect. 5). The main interest is to measure by an index the extent raters assign the same (or very similar) values (absolute agreement) to the targets evaluated, because only in this case the scale can be used with confidence. For quantitative discrete scales with a limited number of levels, extensions of the Cohen's weighted Kappa index (e.g., Gwet 2014; Mitani et al. 2017) are available, and interesting inequalities relationships are established among some of them (Warrens 2010). These extensions cannot be used for quantitative discrete scales with a large number of levels or for continuous scale and have some drawbacks: (1) Indices are based on agreement expected by chance, that depends on the observed proportions of subjects allocated to the categories of the scale by each rater, and this implies that the measure of agreement depends on the marginal distributions of the categories of the scale observed for each rater; (2) indices are formulated in terms of agreement statistics based on all pairs of raters, but some authors argue that simultaneous agreement among three or more raters can be alternatively considered (e.g., see Warrens 2012); (3) indices cannot be computed for a single-target (target-specific measure of agreement), because in that case the agreement expected by chance is not defined or statistically not relevant (e.g., see Bove et al. 2021 for a proposal of a single-target measure of interrater absolute agreement for ordinal scales); (4) indices cannot evaluate agreement in a group of targets where each target is evaluated by a different group of raters (e.g., when each teacher is evaluated by pupils in a different class).

For quantitative discrete scales with any number of levels or for continuous scales, the intraclass correlation coefficient (ICC) is the traditional approach. The main interpretation of the ICC is as a measure of the proportion of variance (variously defined) that is attributable to the objects of measurement, see Shrout and Fleiss (1979). Several versions of the ICC have been proposed, each form is appropriate for specific situations defined by the experimental design as discussed in Shrout and Fleiss (1979) and McGraw and Wong (1996). Intraclass correlation coefficients are affected by the following limitations: (1) the restriction of variance problem, that consists in an attenuation of estimates of rating similarity caused by an artifact reduction of the between-targets variance in ratings; (2) estimation and hypothesis testing procedures for intraclass correlation coefficients are, in general, sensitive to the assumption of normality and are subject to unstable variance; (3) cannot measure single-target interrater absolute agreement. Such single-target evaluations are particularly useful both in situations where the rating scale is being tested and when the agreement on single cases is poor and a specific comparison between raters is requested. The restriction of variance problem of the intraclass correlation coefficients and the other two limitations can be overcome defining target-specific measures of interrater agreement that work separately with each target in the corresponding row of ratings in the targets × raters data matrix.

In the next sections, indices measuring the interrater agreement for quantitative measurements on a single-target based on the standard deviation, that are not affected by the previous three limitations of the intraclass correlation coefficients, are proposed. Furthermore, a global measure of agreement obtained averaging the single-target agreement measures is considered.

The paper is organized as follows. In Sect. 2, we provide a brief background about the one-way random effects model and define the particular ICC of interest. In Sect. 3, we propose alternative measures of interrater agreement based on standard deviation whose sampling properties are analyzed in Sects. 3.1 and 3.2, for normal and non-normal data, respectively. Finally, a simulation study illustrating the theoretical results is performed in Sect. 4 and an application to a real dataset concerning the agreement of radiologists measuring tumor size is described in Sect. 5.

2 ICC in the one-way random effects ANOVA model

In our framework we assume a one-way random effects model, then the n_T targets being rated are randomly drawn from the population of targets. Each target is rated by a set of n_R raters (not necessarily the same raters in each set) randomly drawn from the population of raters. In the one-way random model the only random effect is due to the target since the effects due to raters and due to interaction cannot be separated from random error. See McGraw and Wong (1996) and Elfving et al. (1999) for examples of this setting. More specifically, in McGraw and Wong (1996) behavioral genetics data are used to assess familial resemblance. In Elfving et al. (1999) a reliability study of a method using electromyography on back muscles is described.

Denote by x_{ij} the measurement made on the ith target by the jth rater, for $i = 1, \dots, n_T$ and $j = 1, \dots, n_R$. In the one-way ANOVA model it is specifically assumed that each experimental value x_{ij} may be regarded as the sum of three contributions,

$$x_{ij} = \mu + a_i + \epsilon_{ij} \tag{1}$$

where μ is the grand mean of all measurements, a_i is the target effect and ϵ_{ij} is the random error. The target effect a_i and the random error ϵ_{ij} are assumed to be independent and normally distributed with mean 0 and variances σ_T^2 and σ_ϵ^2, respectively. Notice that ϵ_{ij} is a residual component equal to the sum of inseparable effects of the rater, the rater-and-target interaction and the error term. The intraclass correlation ρ in a one-way ANOVA model is given by,

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_\epsilon^2}, \tag{2}$$

defined as the proportion of between-target variation relative to the total variation, for details see Shoukri et al. (2016) and reference therein. From (2), ρ varies between 0 and 1. More specifically, $\rho \leq 0.5$ denotes poor reliability, $0.5 < \rho \leq 0.75$ denotes good reliability, $\rho > 0.75$ excellent reliability, as suggested in Koo Terry and Li Mae (2016). It can be shown that ρ given by (2) is the correlation between two measurements on the same group (target) i. Thus, larger values of ρ indicate higher coherence among measurements on the same target by different raters. Let S_T and S_ϵ be the between-targets mean square and the residual mean square error, respectively, defined as,

$$S_T = n_R \sum_{i=1}^{n_T} (\bar{x}_{i.} - \bar{x}_{..})^2/(n_T - 1),$$

$$S_\epsilon = \sum_{i=1}^{n_T} \sum_{j=1}^{n_R} (x_{ij} - \bar{x}_{i.})^2/[n_T(n_R - 1)],$$

where $\bar{x}_{..} = \sum_{i=1}^{n_T} \sum_{j=1}^{n_R} x_{ij}/n_T n_R$ is the overall mean of $\{x_{ij}\}$ and $\bar{x}_{i.} = \sum_{j=1}^{n_R} x_{ij}/n_R$ is the mean of the measurements provided by the n_R raters on ith target. Since $E[S_T] = n_R\sigma_T^2 + \sigma_\epsilon^2$ and $E[S_\epsilon] = \sigma_\epsilon^2$, the most commonly used estimator for ρ is given by

$$\hat{\rho} = \frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2 + \hat{\sigma}_\epsilon^2}, \tag{3}$$

where $\hat{\sigma}_T^2 = (S_T - S_\epsilon)/n_R$, $\hat{\sigma}_\epsilon^2 = S_\epsilon$, see Liljequist et al. (2019). Notice that the expression for $\hat{\rho}$ in terms of the mean squares S_T and S_ϵ may become negative. This may occur by chance, especially if the sample size n_T is small. Finally, it should be borne in mind that while $\hat{\rho}$ is a consistent estimator of ρ, it is biased. Atenafu et al. (2012) investigated the issues related to bias correction of the ANOVA estimator of ICC from the one-way layout and the effect of non-normality through Monte-Carlo simulations by generating data from known skewed distributions. In Shoukri et al. (2016) the first order approximation for the bias and the variance of the ICC for a one-way random model are computed.

3 Single-target and global interrater agreement measures for a quantitative scale

A high intraclass correlation means that the points will be spread out along the line of equality in a n_R-dimensional space. The dispersion of a quantitative continuous variable assuming n_R values $(x_1, \dots, x_{n_R})$ can be measured computing its distance from the straight line $X_1 = X_2 = \cdots = X_{n_R}$, given by,

$$l = \sqrt{\left(\sum_{j=1}^{n_R} x_j^2\right) - \frac{\left(\sum_{j=1}^{n_R} x_j\right)^2}{n_R}} = \sqrt{n_R \sigma_\epsilon^2}, \tag{4}$$

where σ_ϵ^2 is the variance of the scores $(x_1, x_2, \dots, x_{n_R})$. Let m and M be the minimum and the maximum for the quantitative scale X, respectively, then

$$l_{\max} = \max\left(\sqrt{n_R\sigma_\epsilon^2}\right) \leq \sqrt{\frac{n_R(M-m)^2}{4}} = \frac{(M-m)\sqrt{n_R}}{2}. \tag{5}$$

Hence, it is possible to define a measure of dispersion normalized in the interval [0, 1] as follows,

$$g = \frac{l}{l_{\max}} = \frac{\sqrt{n_R\sigma_\epsilon^2}}{l_{\max}} = \frac{2\sigma_\epsilon}{M-m}. \tag{6}$$

Notice that $g = 1$ for maximum disagreement and $g = 0$ for perfect agreement. Maximum disagreement occurs when half of the scores are equal to M and half of the scores are equal to m. When the minimum and the maximum of X are unknown, a relative measure of agreement can be obtained by the coefficient of variation defined as,

$$\mathrm{CV} = \frac{\sigma_\epsilon}{\mu}, \tag{7}$$

where μ is the overall mean. Notice that high values of CV indicate disagreement. In Sects. 3.1 and 3.2 estimators of g and CV indices are proposed and their sampling properties are discussed both for the normal and non-normal case.

3.1 Sampling properties of *g* index

As previously stressed, the dispersion of a quantitative variable can be measured by the index (4). With regard to the ith target, the standard deviation σ_ϵ can be estimated by the sample standard deviation s_i defined as,

$$s_i = \sqrt{\frac{1}{n_R-1}\sum_{j=1}^{n_R}(x_{ij} - \bar{x}_{i.})^2}. \tag{8}$$

Note that even though the sample variance s_i^2 is an unbiased estimator of the variance σ_ϵ^2, that is $E(s_i^2) = \sigma_\epsilon^2$, the standard deviation s_i is a biased estimator of the standard deviation σ_ϵ. By Jensen's inequality, since the square root is a concave function, we obtain $E(s_i) = E\left(\sqrt{s_i^2}\right) \leq \sqrt{E(s_i^2)} = \sigma_\epsilon$ and the sample standard deviation s_i tends to underestimate σ_ϵ. Fortunately, the bias is typically minor if the sample size is reasonably large.

Lemma 1 *Under the normality assumption, it can be proved that*

$$E(s_i) = \sigma_\epsilon \frac{\sqrt{2}\Gamma(n_R/2)}{\sqrt{n_R-1}\Gamma((n_R-1)/2)} = \sigma_\epsilon A(n_R), \tag{9}$$

$$V(s_i) = \sigma_\epsilon^2\left(1 - \frac{2\Gamma(n_R/2)^2}{(n_R-1)\Gamma((n_R-1)/2)^2}\right) = \sigma_\epsilon^2(1 - A(n_R)^2), \tag{10}$$

where $A(n_R) = \frac{\sqrt{2}\Gamma(n_R/2)}{\sqrt{n_R-1}\Gamma((n_R-1)/2)} < 1$ and $\Gamma(.)$ *is the gamma function.*

Proof of Lemma 1 See Appendix.

Then, for each target i (for $i = 1, \ldots, n_T$) the following estimator of the g index (6) can be defined,

$$\hat{g}_i = \frac{\hat{l}_i}{l_{\max}} = \frac{\sqrt{n_R s_i^2}}{l_{\max}} = \frac{2s_i}{M-m}, \tag{11}$$

which measures the interrater agreement on scores concerning the ith target. The bias and the variance of $\hat{g}_i$ are computed in Lemma 2. □

Lemma 2 *Under the normality assumption, the bias and the variance of* $\hat{g}_i$ *are given by,*

$$B(\hat{g}_i) = E(\hat{g}_i) - g = \frac{2\sigma_\epsilon}{M-m}(A(n_R) - 1) = g(A(n_R) - 1) < 0, \tag{12}$$

$$V(\hat{g}_i) = \frac{4}{(M-m)^2}V(s_i) = \frac{4\sigma_\epsilon^2}{(M-m)^2}\left(1 - A(n_R)^2\right) = g^2\left(1 - A(n_R)^2\right). \tag{13}$$

Proof of Lemma 2 Immediate consequence of Lemma 1.

In order to obtain an agreement estimate on the whole group of targets, the following estimator of g index can be considered,

$$\bar{\hat{g}} = \frac{1}{n_T}\sum_{i=1}^{n_T}\hat{g}_i = \frac{1}{n_T}\sum_{i=1}^{n_T}\frac{2s_i}{M-m}. \tag{14}$$

More specifically, $\bar{\hat{g}}$ is an estimator of g obtained averaging the n_T estimates $\hat{g}_1, \ldots, \hat{g}_{n_T}$. In Proposition 1 both the sampling properties and the asymptotic distribution of $\bar{\hat{g}}$ are analyzed under the normality assumption. □

Proposition 1 *Under the normality assumption, the bias and the variance of* $\bar{\hat{g}}$ *estimator are given by*,

$$B(\bar{\hat{g}}) = E(\bar{\hat{g}}) - g = g(A(n_R) - 1), \tag{15}$$

$$V(\bar{\hat{g}}) = \frac{g^2}{n_T}(1 - A(n_R)^2). \tag{16}$$

Furthermore, $\overline{\widehat{g}}$ *has a gamma distribution with shape parameter* $\tau = k(n_T(n_R - 1))/2$ *and scale parameter* $\theta = 2/n_T$*, where* $k = 2\sigma_\epsilon/((M-m)\sqrt{n_R - 1})$*. For large* τ (e.g., as n_T *goes to infinity) the gamma distribution can be approximated by a normal distribution with mean* $\tau\theta$ *and variance* $\tau\theta^2$*.*

Proof of Proposition 1 See Appendix. □

Remark 1 From Proposition 1, an unbiased estimator of g can be defined as $\overline{\widehat{g}}^{*} = \overline{\widehat{g}}/A(n_R)$.

In Proposition 2 both the sampling properties and the asymptotic distribution of $\overline{\widehat{g}}$ are analyzed for large n_T ($e.g$, $n_T > 30$) and moderate n_R ($e.g$, $n_R = 7 - 10$) when the normality assumption is not satisfied.

Proposition 2 *The estimator* $\overline{\widehat{g}}$ *is a biased estimator of g with expectation and variance given by*

$$E(\overline{\widehat{g}}) = \frac{1}{n_T}\sum_{i=1}^{n_T}\frac{2E(s_i)}{M-m} \leq g, \tag{17}$$

$$V(\overline{\widehat{g}}) = \frac{1}{n_T^2}\sum_{i=1}^{n_T}\frac{4V(s_i)}{(M-m)^2}. \tag{18}$$

Furthermore, since $\widehat{g}_1, \ldots, \widehat{g}_{n_T}$ *are i.i.d., for the central limit theorem,* as n_T *goes to infinity the random variable* $\overline{\widehat{g}}$ *tends to a normal distribution with mean and variance given by* (17) *and* (18), *respectively.*

The results in Propositions 1 and 2 are useful to construct point and interval estimates for g. They are also useful for testing null hypotheses such as $H_0 : g \leq g_0$, where g_0 be a real number in [0, 1]. Consider the hypothesis problem,

$$\begin{cases} H_0 : g \leq g_0 \\ H_1 : g > g_0 \end{cases} \tag{19}$$

As a consequence of Propositions 1 and 2, a test with an asymptotic significance level α consists in rejecting H_0 whenever

$$\overline{\widehat{g}} > g_0 + z_{1-\alpha}\sqrt{\widehat{V}(\overline{\widehat{g}})}, \tag{20}$$

where $z_{1-\alpha}$ is the $(1-\alpha)$-th quantile of the standard normal distribution and $\widehat{V}(\overline{\widehat{g}})$ is the estimate of variance of $\overline{\widehat{g}}$.

Finally, when the normality assumption is not satisfied and no exact expressions for (17) and (18) are available, the magnitude of the bias of $\overline{\widehat{g}}$ as well as its standard error can be evaluated by bootstrap method (see Efron 1979; Mashreghi et al. 2016 and Conti et al. 2020) according to the following steps:

Step 1: Generate B simple random samples with replacement of n_T targets from the original sample (bootstrap samples).
Step 2: For each bootstrap sample b (for $b = 1, \dots, B$) the estimate of g index (14) is computed obtaining $\overline{\widehat{g}}_1, \dots, \overline{\widehat{g}}_B$.
Step 3: Compute the mean and the variance of B bootstrap estimates $\overline{\widehat{g}}_1, \dots, \overline{\widehat{g}}_B$. Formally,

$$\overline{\widehat{g}}^* = \frac{1}{B}\sum_{b=1}^{B}\overline{\widehat{g}}_b, \quad s^{2*} = \frac{1}{B-1}\sum_{b=1}^{B}(\overline{\widehat{g}}_b - \overline{\widehat{g}}^*)^2. \tag{21}$$

Then, the bootstrap estimate of bias is given by $\widehat{B}(\overline{\widehat{g}}) = \overline{\widehat{g}}^* - t(\widehat{F})$ where $t(\widehat{F})$ is the plug-in estimator of the parameter g. An unbiased estimator of g can then be defined by subtracting the bias from the original estimate (bias correction).

Remark 2 In order to homogenize the values assumed by g and ρ, the index $1 - g$ can be considered.

3.2 Sampling properties of CV index

If the minimum (m) and the maximum (M) of X are unknown, an alternative measure of agreement may be the coefficient of variation defined as $\mathrm{CV} = \sigma_\epsilon/\mu$. For each target i, an estimator of CV can be defined as $\widehat{\mathrm{CV}}_i = s_i/\overline{x}_{..}$ where $\overline{x}_{..} = \sum_{i=1}^{n_T}\sum_{j=1}^{n_R} x_{ij}/n_R n_T$.

In order to analyze the properties of $\widehat{\mathrm{CV}}_i$ we use the Taylor linearization technique (or delta method) approximating the nonlinear estimator $\widehat{\mathrm{CV}}_i$ by a pseudo-estimator, which is a linear function of s_i and $\overline{x}$, thus easy to handle. The technique for finding such a pseudo-estimator consists of the first Taylor approximation of $\widehat{\mathrm{CV}}_i$, expanding around the point $\theta = (\mu, \sigma_\epsilon)$, and neglecting the remainder term. Formally,

$$\widehat{\mathrm{CV}}_i = \frac{\sigma_\epsilon}{\mu} + \frac{1}{\mu}(s_i - \sigma_\epsilon) - \frac{\sigma_\epsilon}{\mu^2}(\overline{x}_{..} - \mu) + R, \tag{22}$$

where R is a remainder of smaller order than the terms in the equation.

Lemma 3 *Under the normality assumption, the bias and the variance of* CV_i *are given by*

$$B(\widehat{\mathrm{CV}}_i) \approx \mathrm{CV}(A(n_R) - 1), \tag{23}$$

$$V(\widehat{\mathrm{CV}}_i) \approx (\mathrm{CV})^2 \left(1 - A(n_R)^2 + \frac{\mathrm{CV}^2}{n_T n_R} \right). \tag{24}$$

Proof of Lemma 3 See Appendix. □

For each target i, $\widehat{\mathrm{CV}}_i$ measures the interrater agreement on measures concerning the ith target. In order to obtain a global interrater agreement estimate, the following estimator of CV index is considered,

$$\overline{\widehat{\mathrm{CV}}} = \frac{1}{n_T} \sum_{i=1}^{n_T} \widehat{\mathrm{CV}}_i. \tag{25}$$

More specifically, $\overline{\widehat{\mathrm{CV}}}$ is an estimator of CV obtained averaging the n_T estimates $\widehat{\mathrm{CV}}_1, \dots, \widehat{\mathrm{CV}}_{n_T}$ obtained from the n_T sample targets.

Lemma 4 *Under the normality assumption, the bias and the variance of* $\overline{\widehat{\mathrm{CV}}}$ *estimator are given by*

$$B(\overline{\widehat{\mathrm{CV}}}) \approx \mathrm{CV}(A(n_R) - 1), \tag{26}$$

$$V(\overline{\widehat{\mathrm{CV}}}) \approx \frac{\mathrm{CV}^2}{n_T} \left(1 - A(n_R)^2 + \frac{\mathrm{CV}^2}{n_T n_R} \right). \tag{27}$$

Proof of Lemma 4 Immediate consequence of Lemma 3.

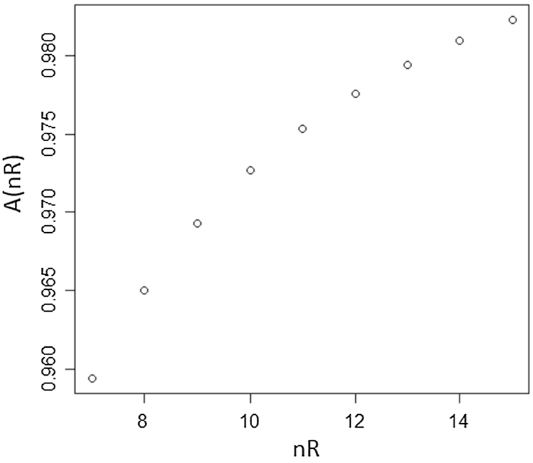


Fig. 1 Plot of $A(n_R)$ against the values of n_R, for $n_R = 7 - 15$

Notice that both the bias and the variance of $\overline{\widehat{g}}$ and $\overline{\widehat{\mathrm{CV}}}$ decrease as $A(n_R)$ increases. In Fig. 1 a plot of $A(n_R)$ against the values of n_R, for $n_R = 7 - 15$ is reported. □

Remark 3 From Lemma (4), an unbiased estimator of CV can be defined as $\overline{\widehat{\mathrm{CV}}}/A(n_R)$.

In Proposition 3 both the sampling properties and the asymptotic distribution of $\overline{\widehat{\mathrm{CV}}}$ are analyzed for large n_T ($e.g$, $n_T > 30$) and moderate n_R ($e.g$, $n_R = 7 - 10$).

Proposition 3 *The estimator* $\overline{\widehat{\mathrm{CV}}}$ *has expectation*

$$E(\overline{\widehat{\mathrm{CV}}}) = \frac{1}{n_T}\sum_{i=1}^{n_T} E\left(\widehat{\mathrm{CV}}_i\right) \leq \mathrm{CV}, \tag{28}$$

and variance

$$V(\overline{\widehat{\mathrm{CV}}}) = \frac{1}{n_T^2}\sum_{i=1}^{n_T} V(\widehat{\mathrm{CV}}_i). \tag{29}$$

Furthermore, since $\widehat{\mathrm{CV}}_1, \dots, \widehat{\mathrm{CV}}_{n_T}$ *are i.i.d., for the central limit theorem, as* n_T *goes to infinity the random variable* $\overline{\widehat{\mathrm{CV}}}$ *tends to a normal distribution with mean and variance given by* (28) *and* (29), *respectively.*

When the normality assumption is not satisfied, the magnitude of the bias of $\overline{\widehat{\mathrm{CV}}}$ as well as its variance can be evaluated by resampling methods, as discussed at the end of Sect. 3.1. Analogously to g, the results in Proposition 3 are useful to construct point and interval estimates of CV and to perform statistical tests.

4 Simulation study

In order to evaluate the performance of the indices discussed in Sects. 2 and 3, a simulation experiment with moderate $n_R (n_R = 7)$ is performed, since in the real applications the number of raters is generally limited. As stressed in Koo Terry and Li Mae (2016), as a rule of thumb, researchers should try to obtain at least 30 targets and involve at least 3 raters.

For normal outcome, data were simulated according to the framework of the one-way random effects model described in Sect. 2, results are reported in Sect. 4.1. For non-normal data, simulation study and its results are illustrated in Sect. 4.2.

We focus on confidence intervals for the aforementioned indices because confidence intervals indicate the range within which the population parameters g, CV and

ρ (the interrater agreement in the population) are likely to fall, as well as precision of these estimates (i.e., the size of the range). That is, confidence intervals show the range of plausible values for interrater agreement in the population. The simulation study was carried out by R (R Core Team 2022).

4.1 Simulation study for normal data

For normal data the simulation study consists of the following steps:

Step 1 Generate a sample s of $n_R = 7$ raters and $n_T = 50$ targets from a one-way random model (1) with parameters $\mu = 8$, $\sigma_T^2 = 1$ and σ_ϵ^2. Different values of σ_ϵ^2 are considered in order to obtain alternative values for ρ. More specifically, for $\sigma_\epsilon^2 = 2, 0.6, 0.2$ we obtain $\rho = 0.33, 0.63, 0.83$ corresponding to low, moderate and high agreement, respectively. Analogously, the indices g and CV are computed according to (6) and (7), respectively. Then, $g = 0.08, 0.12, 0.15$ and $\mathrm{CV} = 0.06, 0.10, 0.18$, for $\sigma_\epsilon^2 = 2, 0.6, 0.2$. Notice that, in the computation of g the minimum (m) and the maximum (M) in (6) are computed simulating 10,000,000 observations from the one-way random model for each value of σ_ϵ^2 with $\mu = 8$ and $\sigma_T^2 = 1$.

Suggestions for interpreting the value of g and CV are in Table 1, where a comparison between the indices ρ, g and CV is reported. More specifically, datasets with different level of raters agreement are generated according to the aforementioned one-way random model for different values of $\sigma_\epsilon \in [0, 3]$. As Table 1 shows, for $\rho \leq 0.5$ (low agreement) both g and CV are larger than 0.14 and 0.13, respectively. For moderate agreement $\rho \in (0.5{-}0.75]$, the index g is in (0.10, 0.14] and CV is in (0.07, 0.13]. For high agreement $\rho > 0.75$, g assumes values in [0, 0.10) and CV in [0, 0.07).

Step 2 Compute bias and variance of the estimators $\overline{\widehat{g}}$ and $\overline{\widehat{\mathrm{CV}}}$. Furthermore, confidence intervals for g ($[L_g^s, U_g^s]$) and CV ($[L_{\mathrm{CV}}^s, U_{\mathrm{CV}}^s]$) of level $1 - \alpha = 0.95$ based on the asymptotic normal approximation are computed, see Proposition 1 and Proposition 3.

Step 3 Compute the intraclass correlation coefficient ρ, its bias and variance. Furthermore, confidence intervals for ρ ($[L_\rho^s, U_\rho^s]$) are obtained as follows,

Table 1 Comparison between ρ, g and CV, for $\mu = 8$, $\sigma_T^2 = 1$ and $\sigma_\epsilon^2 \in [0, 3]$

ρ	g	CV
≤ 0.5	> 0.14	> 0.13
(0.5–0.75]	(0.10, 0.14]	(0.07, 0.13]
> 0.75	[0, 0.10)	[0, 0.07)

$$L^s_\rho = \frac{F_L - 1}{F_L + n_R - 1}, \quad U^s_\rho = \frac{F_U - 1}{F_U + n_R - 1} \tag{30}$$

where $F_L = F_O/F_{1-\alpha/2,V_2,V_1}$, $F_U = F_O F_{1-\alpha/2,V_1,V_2}$ and $F_O = S_T/S_\epsilon$. The degrees of freedom (dof, for short) are $V_1 = n_T(n_R - 1)$ and $V_2 = n_T - 1$, see for details (Shrout and Fleiss 1979).

Step 4 Steps 1–3 are repeated $S = 5000$ times.

After having computed the confidence intervals $[L^s_t, U^s_t]$ for $t = g, \mathrm{CV}, \rho$ for sample s $(s = 1, \dots, S = 5000)$, their accuracy has been evaluated by the following indicators.

(1) Estimated coverage probability, in per cent, for the interval,

$$\mathrm{ECP} = \frac{100}{S} \sum_{s=1}^{S} I(L^s_t \le t \le U^s_t). \tag{31}$$

(2) Estimated left-tail and right-tail errors (lower and upper error rates) in per cent,

$$\mathrm{LE} = \frac{100}{S} \sum_{s=1}^{S} I(L^s_t > t), \tag{32}$$

$$\mathrm{RE} = \frac{100}{S} \sum_{s=1}^{S} I(U^s_t < t). \tag{33}$$

(3) Estimated average length (AL) of all 5000 simulated intervals given by

$$\mathrm{AL} = \sum_{s=1}^{S} \frac{U^s_t - L^s_t}{S}, \tag{34}$$

where $I(a) = 1$ if a is true and $I(a) = 0$ elsewhere, and $t = g, \mathrm{CV}, \rho$.

In Table 2 the bias and the standard deviation of the estimates $\bar{\hat{g}}$, $\overline{\widehat{\mathrm{CV}}}$ and $\hat{\rho}$ over the $S = 5000$ samples are reported.

Table 2 Bias and standard deviation of the estimates ($\bar{\hat{g}}$, $\overline{\widehat{\mathrm{CV}}}$, $\hat{\rho}$) for normal data and ($n_R = 7, n_T = 50$), over the $S = 5{,}000$ samples for different (g, CV, ρ) coefficients. Results in the last six columns are multiplied by 100

σ^2_ϵ	g	CV	ρ	$B(\bar{\hat{g}})$	$B(\overline{\widehat{\mathrm{CV}}})$	$B(\hat{\rho})$	$\mathrm{Sd}(\bar{\hat{g}})$	$\mathrm{Sd}(\overline{\widehat{\mathrm{CV}}})$	$\mathrm{Sd}(\hat{\rho})$
2	0.15	0.18	0.33	−0.57	−0.71	−0.45	0.55	0.78	6.25
0.6	0.12	0.10	0.63	−0.47	−0.39	−0.64	0.46	0.42	5.72
0.2	0.08	0.06	0.83	−0.35	−0.23	−0.45	0.33	0.24	3.28

Table 3 Performance of confidence intervals for g, CV and ρ for normal data and ($n_R = 7$,$n_T = 50$)

$(\sigma_\epsilon^2, g, \mathrm{CV}, \rho)$	Indicators	g	CV	ρ
(2, 0.15, 0.18, 0.33)	CP	95.18	95.02	95.02
	LE	2.76	2.56	2.36
	RE	2.06	2.42	2.62
	AL	0.02	0.03	0.24
(0.6, 0.12, 0.10, 0.63)	CP	95.14	95.12	94.62
	LE	2.58	2.66	2.62
	RE	2.28	2.22	2.76
	AL	0.02	0.02	0.21
(0.2, 0.08, 0.06, 0.83)	CP	94.90	94.96	94.84
	LE	2.72	2.80	2.50
	RE	2.38	2.24	2.66
	AL	0.01	0.01	0.12

As results in Table 2 show, all the estimators $\bar{\hat{g}}$, $\overline{\widehat{\mathrm{CV}}}$ and $\hat{\rho}$ underestimate the corresponding parameters (g, CV, ρ). Note that as g, CV decrease (high agreement) the bias decreases from −0.57% to −0.35% for g and from −0.71% to −0.23% for CV, respectively. The same consideration holds for the standard error. Finally, with respect to $\bar{\hat{g}}, \overline{\widehat{\mathrm{CV}}}$ the intraclass correlation estimator $\hat{\rho}$ is characterized by larger standard errors.

Finally, the confidence intervals for g, CV, ρ are computed. Results are reported in Table 3. More specifically, Table 3 presents the estimated coverage probabilities of 95% confidence intervals (CP), the estimated left-tail (LE) and right-tail (RE) errors (nominal values is 2.5% for both) and the average length (AL) for the indices g, CV and ρ, when $(n_R = 7, n_T = 50)$.

As reported in Table 3, the confidence intervals for g and CV obtained with the normal approximation perform very well. Coverage probabilities are approximately equal to 95% nominal value for g and CV indices, respectively, with an average length of 0.01 for $(g = 0.08, \mathrm{CV} = 0.06)$, of 0.02 for $(g = 0.12, \mathrm{CV} = 0.10)$ and of about 0.02 for $(g = 0.15, \mathrm{CV} = 0.18)$. Confidence interval for ρ performs as well as the confidence intervals for g and CV in terms of coverage probability but the interval average length is wider. Analogous results are obtained when $n_R = 11$.

4.2 Simulation study for non-normal data

In this section the robustness of the estimators $\bar{\hat{g}}, \overline{\widehat{\mathrm{CV}}}$ and $\hat{\rho}$ to deviations from the normality is evaluated. According to the framework of the one-way random effects model, the simulation study consists of the following steps:

Step 1: generate ϵ_{ij} from a normal distribution with mean 0 and variance σ_ϵ^2. As in Sect. 4.1, different values for σ_ϵ^2 $(\sigma_\epsilon^2 = 2, 0.6, 0.2)$ are considered so to distinguish between low, moderate and high value for ρ $(\rho = 0.33, 0.63, 0.83)$.

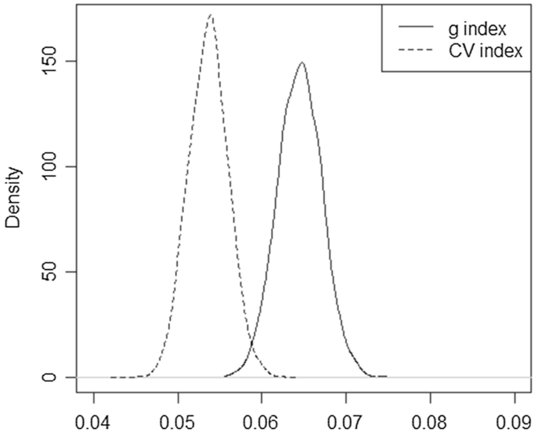


Fig. 2 Kernel density estimate of g and CV indices from the $S = 5000$ original samples, true values are $g = 0.07$, CV $= 0.06$

Table 4 Bias and standard deviation of the estimates $(\overline{\widehat{g}}, \overline{\widehat{\mathrm{CV}}}, \widehat{\rho})$ for non-normal data and $(n_R = 7, n_T = 50)$, over the $S = 5000$ samples for different (g, CV, ρ) coefficients. Results in the last six columns are multiplied by 100

σ^2_ϵ	g	CV	ρ	$B(\overline{\widehat{g}})$	$B(\overline{\widehat{\mathrm{CV}}})$	$B(\widehat{\rho})$	$\mathrm{Sd}(\overline{\widehat{g}})$	$\mathrm{Sd}(\overline{\widehat{\mathrm{CV}}})$	$\mathrm{Sd}(\widehat{\rho})$
2	0.13	0.18	0.33	−0.53	−0.72	−2.06	0.51	0.77	11.21
0.6	0.09	0.10	0.63	−0.37	−0.39	−3.40	0.35	0.42	12.01
0.2	0.07	0.06	0.83	−0.28	−0.23	−3.02	0.23	0.24	7.90

Step 2: Generate a_i from a gamma distribution with shape parameter $\alpha = 1/2$ and scale parameter $\theta = \sqrt{2}$. The mean and variance are $E(a_i) = \alpha\theta = \sqrt{2}/2$ and $V(a_i) = \alpha\theta^2 = 1$, respectively. The skewness of the distribution is $2/\sqrt{\alpha} = 2.8$ and the kurtosis coefficient is $6/\alpha = 12$. Recall that, the skewness and kurtosis coefficients for a normally distributed random variable are 0 and 3, respectively.

Step 3: Steps 1–2 are repeated $S = 5000$ times.

As in Sect. 4.1, bias and variance of the estimators $\overline{\widehat{g}}, \overline{\widehat{\mathrm{CV}}}$ and $\widehat{\rho}$ as well as confidence intervals for g, CV and ρ are computed. Notice that, according to the model (1) the measurements x_{ij} have mean equal to 8 and variance $\sigma^2_T + \sigma^2_\epsilon$ with $\sigma^2_T = \alpha\theta^2 = 1$.

Figure 2 shows the kernel density of g and CV indices estimated from the $S = 5000$ original samples for $n_T = 50, n_R = 7$ and $\sigma^2_\epsilon = 0.2$. The true values for g and CV indices are 0.07 and 0.06, respectively. The bandwidth selection rule is as proposed by Sheather and Jones (1991). Notice that both the estimators follow a normal distribution.

In Table 4 the bias and the standard deviation of the estimates $\overline{\widehat{g}}$, $\overline{\widehat{\mathrm{CV}}}$ and $\widehat{\rho}$ over the $S = 5000$ samples are reported.

Table 5 Performance of confidence intervals for g, CV and ρ for non-normal data and ($n_R = 7$,$n_T = 50$)

$(\sigma_\epsilon^2, g, \mathrm{CV}, \rho)$	Indicators	g	CV	ρ
(2,0.13,0.18,0.33)	CP	94.96	95.16	69.64
	LE	2.74	2.58	12.18
	RE	2.30	2.26	18.18
	AL	0.02	0.03	0.23
(0.6,0.09,0.10,0.63)	CP	95.30	94.92	62.58
	LE	2.64	2.66	14.82
	RE	2.06	2.42	22.60
	AL	0.01	0.02	0.21
(0.2,0.07,0.06,0.83)	CP	94.88	95.04	61.48
	LE	2.48	2.56	14.66
	RE	2.64	2.40	23.86
	AL	0.01	0.01	0.13

The conclusions of Table 4 are similar to those drawn from Table 2 for $\overline{\widehat{g}}$ and $\overline{\widehat{\mathrm{CV}}}$ both in terms of bias and standard error. As expected, the worst performance is shown by $\widehat{\rho}$ with larger bias and standard error. Finally, the confidence intervals for g, CV and ρ are computed. Results are reported in Table 5. The confidence intervals for g and CV are robust to deviations from the normality assumption with coverage probability of about 95%. The same result does not hold for ρ with a coverage probability approximately equal to 69.64% for $\rho = 0.33$, 62.58% for $\rho = 0.63$ and 61.48% for $\rho = 0.83$. Furthermore, the interval average length for ρ is wider than g and CV. Notice that the average length of the confidence intervals for ρ is approximately the same as in the case of normal data. Analogous results are obtained when $n_R = 11$.

The same simulation has been performed assuming a marked deviation from normality, that is $\alpha = 1/9$ and $\theta = 3$. The results for $\widehat{g}$ are approximately the same. With regard to CV the coverage probability shows a slight decrease to 93%. Same consideration holds for ρ which coverage probability decreases to 43% for $\rho = 0.33$ and 35% for $\rho = 0.63, 0.83$, respectively.

Table 6 Descriptive statistics of tumor measurements (centimeters) of five radiologists on 40 lung lesions

Radiologist	Mean	Median	Range	Sd
1	3.92	3.80	1.5–8.0	1.57
2	3.71	3.80	1.2–7.8	1.48
3	4.42	4.20	1.5–9.0	1.52
4	4.37	4.10	1.5–9.0	1.58
5	4.14	3.95	1.7–9.0	1.52

5 An application to tumor size of lung cancer

In Erasmus et al. (2003) a study to assess the agreement between radiologists evaluating lung tumors is considered. Notice that, this is a critical component of many cancer trials because measurements can be used to justify additional testing of an agent or to decide whether or not to continue the therapy.

Patients were selected with non-small-cell lung cancer and with 40 lung lesions whose size exceeded at least 1.5 cm. Measurements were performed independently by five thoracic radiologists using printed film by computed tomography. Each radiologist reads each of 40 images performing unidimensional and bidimensional measures. More specifically, a) the longest diameter and b) the longest diameter and the perpendicular longest diameter of each lesion.

Measurements were repeated after 5–7 days, then each radiologist looked at the same image twice. Table 6.18 in Broemeling (2009) contains the data of the two replications of the unidimensional measurements. In order to ascertain how to improve measurement consistency, in Erasmus et al. (2003) variations between and within the two replications of the five radiologists are estimated by statistical modeling.

We proceed to analyze agreement computing the proposed indices g and CV to the unidimensional measurements of the five radiologists. With this regard, some descriptive statistics regarding the first replication of the unidimensional measurement are provided in Table 6. The similarity of the means in Table 6 reflects a pretty good level of agreement, with radiologist 2 reporting the smallest mean tumor size and the smallest standard deviation.

Table 7 Descriptive statistics of the $\widehat{\mathrm{CV}}_i$ and $\widehat{g}_i$

$\widehat{\mathrm{CV}}_{\min}$	$\widehat{\mathrm{CV}}_{\max}$	$\widehat{\mathrm{CV}}_i \leq 0.15$	$0.15 < \widehat{\mathrm{CV}}_i \leq 0.30$	$\widehat{\mathrm{CV}}_i > 0.30$
0.01	0.42	28 (70%)	10 (25%)	2 (5%)
$g_{\min}$	$g_{\max}$	$g_i \leq 0.15$	$0.15 < g_i \leq 0.30$	$g_i > 0.30$
0.01	0.44	25 (62.5%)	13 (32.5%)	2 (5%)

Table 8 Comparison between ρ, g and CV, for $\mu = 4.11$, $\sigma_T^2 = 2.12$ and $\sigma_\epsilon^2 \in [0, 3]$

ρ	g	CV
≤ 0.5	> 0.38	> 0.35
(0.5–0.75]	(0.22, 0.38]	(0.20, 0.35]
> 0.75	[0, 0.22)	[0, 0.21)

Table 9 Bias and standard deviation of $\overline{\widehat{g}}$ and $\overline{\widehat{\mathrm{CV}}}$ indices. Bias values are multiplied by 100

$\overline{\widehat{g}}$	$\overline{\widehat{\mathrm{CV}}}$	Bias($\overline{\widehat{g}}$)	Bias($\overline{\widehat{\mathrm{CV}}}$)	Sd($\overline{\widehat{g}}$)	Sd($\overline{\widehat{\mathrm{CV}}}$)
0.14	0.13	−0.03	−0.04	0.01	0.01

In Table 7 some descriptive statistics regarding $\widehat{\mathrm{CV}}_i$ and $\widehat{g}_i$ are reported. Percentages are reported in round brackets. In order to compute g the minimum and the maximum value of the measurements in the dataset are considered.

As results in Table 7 show, the 70% and 62.5% of the forty lung lesions show $\widehat{\mathrm{CV}}_i$ and $\widehat{g}_i$ values less than or equal to 0.15. On the other hand, some high $\widehat{\mathrm{CV}}_i$ and $\widehat{g}_i$ values are present (e.g., images 16, 37), these images could be selected for a comparison between radiologists and to detect particular types of lesions (irregular edge and/or irregular contour) difficult to measure. More specifically, for the image 16 ($\widehat{g}_i = 0.38, \widehat{\mathrm{CV}}_i = 0.36$), for the image 37 ($\widehat{g}_i = 0.44, \widehat{\mathrm{CV}}_i = 0.42$).

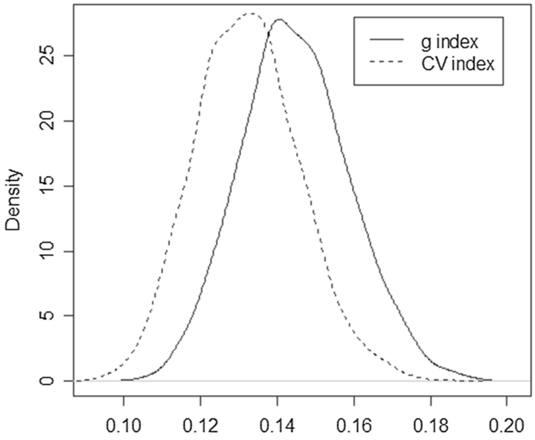
The dataset was previously analyzed in Bove (2022), showing an high level of agreement with an intraclass correlation coefficient equal to 0.83. The normality assumption tested by the Shapiro–Wilk test is not rejected at 1% level of significance. However, as shown in the simulation study of Sect. 4.2 the measures g and CVare both robust to violation of normality assumption.

Notice that the estimates of the one-way random model parameters are $\widehat{\mu} = 4.11$, $\widehat{\sigma}^2_T = 2.12$ and $\widehat{\sigma}^2_\epsilon = 0.42$, respectively. In order to interpret the values of g and CV, datasets with 10,000,000 observations are generated from the estimated one-way random model for different values of $\sigma^2_\epsilon \in [0, 3]$. Results are reported in Table 8.

Finally, in Table 9 the values of $\overline{\widehat{g}}$ and $\overline{\widehat{\mathrm{CV}}}$ given by (6) and (7) respectively, their bias and standard deviation are reported. The values of $\overline{\widehat{g}}$ and $\overline{\widehat{\mathrm{CV}}}$ are 0.14 and 0.13 showing as $\rho = 0.83$ an high agreement between measurements. The magnitude of the bias and standard deviation (Sd) of $\overline{\widehat{g}}$ and $\overline{\widehat{\mathrm{CV}}}$ are evaluated by bootstrap method, $B = 5000$ bootstrap samples are drawn from the initial sample. Results are reported in Table 9.

Figure 3 shows the kernel density of the g and CV indices estimated from the $B = 5000$ bootstrap samples. The bandwidth selection rule is as proposed by Sheather and Jones (1991).

Fig. 3 Kernel density estimate of g and CV indices from the 5000 bootstrap samples



The $(1-\alpha)=0.95$ confidence intervals using the normal approximation are [0.12, 0.16] and [0.11, 0.15] for g and CV, respectively, and the error is at most 0.02.

6 Concluding remarks

In order to analyze the agreement between quantitative measurements provided by a set of raters for a group of targets several versions of the intraclass correlation coefficient have been proposed. Such versions are affected by the restriction of variance problem, cannot measure target-specific agreement and are sensitive to the assumption of normality. In this paper, indices that allow to evaluate the agreement between two or more raters for each single-target have been proposed, and a global measure of agreement obtained averaging the single-target agreement measures is considered. Sampling properties for the global measures were analyzed both under normal and non-normal data. A quite extensive simulation study and an application to a real data set illustrated the good performance of the proposed indices and their robustness to deviations from normality assumptions.

Appendix A: Appendix

Proof Lemma 1 The expectation of s_i can be written as follows,

$$E(s_i)=\sqrt{\frac{\sigma_\epsilon^2}{n_R-1}}E\left(\sqrt{\frac{(n_R-1)s_i^2}{\sigma_\epsilon^2}}\right). \tag{A1}$$

Under the normality assumption $(n_R-1)s_i^2/\sigma_\epsilon^2$ follows a Chi-square distribution with n_R-1 dof. Then, the expectation in (A1) regards the square root of a Chi-square distributed variable. Thus,

$$\begin{aligned}
E(s_i)&=\sqrt{\frac{\sigma_\epsilon^2}{n_R-1}}\int_0^\infty \sqrt{x}\frac{(1/2)^{(n_R-1)/2}x^{((n_R-1)/2)-1}\exp\{-x/2\}}{\Gamma((n_R-1)/2)}dx\\
&=\sqrt{\frac{2\sigma_\epsilon^2}{n_R-1}}\frac{\Gamma(n_R/2)}{\Gamma((n_R-1)/2)}\int_0^\infty\frac{(1/2)^{(n_R)/2}x^{(n_R/2-1)}\exp\{-x/2\}}{\Gamma(n_R/2)}dx\\
&=\sqrt{\frac{2\sigma_\epsilon^2}{n_R-1}}\frac{\Gamma(n_R/2)}{\Gamma((n_R-1)/2)}\\
&=\sigma_\epsilon\frac{\sqrt{2}\Gamma(n_R/2)}{\sqrt{n_R-1}\Gamma((n_R-1)/2)}=\sigma_\epsilon A(n_R),
\end{aligned} \tag{A2}$$

where $A(n_R) = \frac{\sqrt{2}\Gamma(n_R/2)}{\sqrt{n_R-1}\Gamma((n_R-1)/2)} < 1$ and $\Gamma(.)$ is the gamma function. Notice that the integral in the second equality regards the density of a Chi-square distribution with n_R dof.

With regard to the variance of s_i, we obtain,

$$\begin{aligned} V(s_i) = E(s_i^2) - [E(s_i)]^2 &= \sigma_\epsilon^2\left(1 - \frac{2\Gamma(n_R/2)^2}{(n_R-1)\Gamma((n_R-1)/2)^2}\right) \\ &= \sigma_\epsilon^2(1 - A(n_R)^2) \end{aligned} \tag{A3}$$

Proof of Proposition 1 The bias and the variance of $\bar{\hat{g}}$ follow from Lemma 2. Formally,

$$B(\bar{\hat{g}}) = E(\bar{\hat{g}}) - g = \frac{1}{n_T}\sum_{i=1}^{n_T} E(\hat{g}_i) - g = g(A(n_R) - 1), \tag{A4}$$

$$V(\bar{\hat{g}}) = \frac{1}{n_T^2}\sum_{i=1}^{n_T} V(\hat{g}_i) = \frac{g^2}{n_T}(1 - A(n_R)^2). \tag{A5}$$

Finally,

$$\begin{aligned} \bar{\hat{g}} &= \frac{1}{n_T}\sum_{i=1}^{n_T} \hat{g}_i \\ &= \frac{2\sigma_\epsilon}{(M-m)\sqrt{n_R-1}}\frac{1}{n_T}\sum_{i=1}^{n_T}\sqrt{\frac{(n_R-1)s_i^2}{\sigma_\epsilon^2}} \\ &= kG \end{aligned} \tag{A6}$$

where $k = \frac{2\sigma_\epsilon}{(M-m)\sqrt{n_R-1}}$ and $G = \frac{1}{n_T}\sum_{i=1}^{n_T}\sqrt{\frac{(n_R-1)s_i^2}{\sigma_\epsilon^2}}$. Notice that,

1. G is the sample mean of n_T independent and identically distributed Chi-squared variables with $(n_R - 1)$ dof, then G is distributed as a gamma distribution with shape parameter $n_T(n_R - 1)/2$ and scale parameter $2/n_T$.
2. The gamma distribution has the scaling property. That is, if G follows a gamma distribution of parameters $(n_T(n_R - 1)/2, 2/n_T)$ then $Y = kG$ also has a gamma distribution with parameters (τ, θ) where $\tau = kn_T(n_R - 1)/2$ and $\theta = 2/n_T$.

From 1 and 2 follows the result. Clearly, for large τ the gamma distribution can be approximated by a normal distribution with mean $\tau\theta$ and variance $\tau\theta^2$. □

Proof of Lemma 3 The bias and the variance follow from Lemma 1. Formally, from the first Taylor approximation of CV_i around the point $\theta = (\mu, \sigma_\epsilon)$ we obtain,

$$\widehat{\mathrm{CV}}_i = \frac{\sigma_\epsilon}{\mu} + \frac{1}{\mu}(s_i - \sigma_\epsilon) - \frac{\sigma_\epsilon}{\mu^2}(\bar{x}_{..} - \mu) + R, \tag{A7}$$

where R is a remainder of smaller order than the terms in the equation. Then, neglecting R the bias is

$$\begin{aligned} B(\widehat{\mathrm{CV}}_i) = E(\widehat{\mathrm{CV}}_i) - \mathrm{CV} &\approx \frac{1}{\mu}(E(s_i) - \sigma_\epsilon) - \frac{\sigma_\epsilon}{\mu^2}(E(\bar{x}) - \mu) \\ &= \mathrm{CV}(A(n_R) - 1) \end{aligned} \tag{A8}$$

where $E(s_i) = \sigma_\epsilon A(n_R)$ and $E(\bar{x}) = \mu$. The variance is

$$\begin{aligned} V(\widehat{\mathrm{CV}}_i) &\approx \frac{1}{\mu^2}V(s_i) + \frac{\sigma_\epsilon^2}{\mu^4}V(\bar{x}) - 2\frac{\sigma_\epsilon}{\mu^3}Cov(s_i, \bar{x}) \\ &= \frac{\sigma_\epsilon^2}{\mu^2}\left(\frac{V(s_i)}{\sigma_\epsilon^2} + \frac{V(\bar{x})}{\mu^2} - 2\frac{\mathrm{Cov}(s_i, \bar{x})}{\mu\sigma_\epsilon}\right) \\ &= \mathrm{CV}^2\left(1 - A(n_R)^2 + \frac{\mathrm{CV}^2}{n_T n_R}\right) \end{aligned} \tag{A9}$$

where $V(s_i) = \sigma_\epsilon^2(1 - A(n_R)^2$, $V(\bar{x}) = \frac{\sigma_\epsilon^2}{n_T n_R}$ and $Cov(s_i, \bar{x}) = 0$ since $Cov(s_i, \bar{x}_i) = 0$.

□

Acknowledgements The dataset in the application of Sect. 5 was also presented in Bove (2022).

Funding Open access funding provided by Università degli Studi di Roma La Sapienza within the CRUI-CARE Agreement.

Declarations

Conflict of interest All authors declare that they have no conflicts of interest.

References

Atenafu, E.G., Hamid, J.S., To, T., Willan, A., Feldman, B., Beyene, J.: Bias-corrected estimator for the intraclass correlation coefficient in the balanced one-way random effects model. BMC Med. Res. Methodol. **12**(126), 110 (2012)

Bove, G., Conti, P.L., Marella, D.: A measure of interrater absolute agreement for ordinal categorical data. Stat. Methods Appl. **30**, 927–945 (2021)

Bove, G.: Measures of interrater agreement based on the standard deviation. In: Balzanella, A., Bini, M., Cavicchia, C., Verde, R. (eds.) 51st Scientific Meeting of the Italian Statistical Society (SIS), Book of short papers, pp. 1644–1649. Pearson, Milano. ISBN 9788891932310 (2022)

Broemeling, L.D.: Bayesian Methods for Measures of Agreement. Chapman & Hall/CRC, London (2009)

Conti, P.L., Marella, D., Mecatti, F., Andreis, F.: A unified principled framework for resampling based on pseudo-populations: asymptotic theory. Bernoulli **26**(2), 1044–1069 (2020)

Efron, B.: Bootstrap methods: another look at the jackknife. Ann. Stat. **7**(1), 1–26 (1979)

Elfving, B., Nemeth, G., Arvidsson, I., Lamontagne, M.: Reliability of EMG spectral parameters in repeated measurements of back muscle fatigue. J Electromyogr Kinesiol **9**(4), 235–243 (1999)

Erasmus, J.J., Gladish, G.W., Broemeling, L., Sabloff, B.S., Truong, M.T., Herbst, R.S., Munden, R.F.: Interobserver and intraobserver variability in measurement of non- small-cell carcinoma lung lesions: Implications for assessment of tumor response. J. Clin. Oncol. **21**(13), 2574–2582 (2003)

Gwet, K.L.: Handbook of Inter-Rater Reliability, 4th edn. Advanced Analytics LLC, Gaithersburg MD (2014)

Liljequist, D., Elfving, B., Skavberg Roaldsen, K.: Intraclass correlation—a discussion and demonstration of basic features. PLoS ONE **14**(7), 10 (2019). https://doi.org/10.1371/journal.pone.0219854

Mashreghi, Z., Haziza, D., Léger, C.: A survey of bootstrap methods in finite population sampling. Stat. Surv. **10**, 1–52 (2016)

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

Mitani, A.A., Freer, P.E., Nelson, K.P.: Summary measures of agreement and association between many raters' ordinal classifications. Ann. Epidemiol. **27**(10), 677–685 (2017)

R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2022). https://www.R-project.org/

Sheather, S.J., Jones, M.C.: A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. B **53**(3), 683–690 (1991)

Shoukri, M.M., Al-Hassan, T., DeNiro, M., El Dali, A., Al-Mohanna, F.: Bias and mean square error of reliability estimators under the one and two random effects models: the effect of non-normality. Open J. Stat. **6**(2), 254–273 (2016)

Shoukri, M.M.: Measures of Interobserver Agreement and Reliability. Taylor and Francis Group, Boca Raton (2011)

Shrout, P.E., Fleiss, J.L.: Intraclass correlations: use in assessing rater reliability. Psychol. Bull. **86**(2), 420–428 (1979)

Koo Terry, K., Li Mae, Y.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. **15**(2), 155–163 (2016)

von Eye, A., Mun, E.Y.: Analyzing rater agreement. Manifest variable methods. Lawrence Erlbaum Associates, Mahwah (2005)

Warrens, M.J.: Equivalences of weighted kappas for multiple raters. Stat. Methodol. **9**(3), 407–422 (2012)

Warrens, M.J.: Inequalities between multi-rater kappas. Adv. Data Anal. Class. **4**, 271–286 (2010). https://doi.org/10.1007/s11634-010-0073-4