Dear Author,

Here are the proofs of your article.

- You can submit your corrections **online**, via **e-mail** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- You can also insert your corrections in the proof PDF and **email** the annotated PDF.
- For fax submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the **journal title**, **article number**, and **your name** when sending your response via e-mail or fax.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during copy editing and insert your answers/ corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the *Edited manuscript*.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style.
  Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.
- Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI. **Further changes are, therefore, not possible.**
- The **printed version** will follow in a forthcoming issue.

Please note

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL: http://dx.doi.org/[DOI].
If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information go to: http://www.link.springer.com.

Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us if you would like to have these documents returned.

Metadata of the article that will be visualized in OnlineFirst

ArticleTitle	A measure of interrater absolute agreement for ordinal categorical data	
Article Sub-Title		
Article CopyRight	Springer-Verlag GmbH Germany, part of Springer Nature (This will be the copyright line in the final PDF)	
Journal Name	Statistical Methods & Applications	

Corresponding Author	Family Name	**Bove**
	Particle	
	Given Name	**Giuseppe**
	Suffix	
	Division	Dipartimento di Scienze della Formazione
	Organization	Università "Roma Tre"
	Address	via del Castro Pretorio 20, 00185, Roma, Italy
	Phone	
	Fax	
	Email	giuseppe.bove@uniroma3.it
	URL	
	ORCID	http://orcid.org/0000-0002-2736-5697
Author	Family Name	**Conti**
	Particle	
	Given Name	**Pier Luigi**
	Suffix	
	Division	Dipartimento di Scienze Statistiche
	Organization	Università 'La Sapienza"
	Address	P.le A. Moro, 5, 00185, Rome, Italy
	Phone	
	Fax	
	Email	pierluigi.conti@uniroma1.it
	URL	
	ORCID	
Author	Family Name	**Marella**
	Particle	
	Given Name	**Daniela**
	Suffix	
	Division	Dipartimento di Scienze della Formazione
	Organization	Università "Roma Tre"
	Address	via del Castro Pretorio 20, 00185, Roma, Italy
	Phone	
	Fax	
	Email	daniela.marella@uniroma3.it
	URL	

ORCID

Schedule	Received	
	Revised	
	Accepted	8 November 2020

Abstract

A measure of interrater absolute agreement for ordinal scales is proposed capitalizing on the dispersion index for ordinal variables proposed by Giuseppe Leti. The procedure allows to overcome the limits affecting traditional measures of interrater agreement in different fields of application. An unbiased estimator of the proposed measure is introduced and its sampling properties are investigated. In order to construct confidence intervals for interrater absolute agreement both asymptotic results and bootstrapping methods are used and their performance is evaluated. Simulated data are employed to demonstrate the accuracy and practical utility of the new procedure for assessing agreement. Finally, an application to a real case is provided.

Keywords (separated by '-') Ordinal data - Interrater agreement - Resampling

Footnote Information

Statistical Methods & Applications
https://doi.org/10.1007/s10260-020-00551-5

ORIGINAL PAPER

A measure of interrater absolute agreement for ordinal categorical data

Giuseppe Bove[2] · Pier Luigi Conti[1] · Daniela Marella[2]

Accepted: 8 November 2020

Abstract
A measure of interrater absolute agreement for ordinal scales is proposed capitalizing on the dispersion index for ordinal variables proposed by Giuseppe Leti. The procedure allows to overcome the limits affecting traditional measures of interrater agreement in different fields of application. An unbiased estimator of the proposed measure is introduced and its sampling properties are investigated. In order to construct confidence intervals for interrater absolute agreement both asymptotic results and bootstrapping methods are used and their performance is evaluated. Simulated data are employed to demonstrate the accuracy and practical utility of the new procedure for assessing agreement. Finally, an application to a real case is provided.

Keywords Ordinal data · Interrater agreement · Resampling

1 Introduction

Ordinal rating scales are frequently developed in study designs where several raters (or judges) evaluate a group of targets. For instance, in language studies new rating scales before their routine application are tested out by a group of raters, who assess the language proficiency of a corpus of argumentative (written or oral) texts

✉ Giuseppe Bove
giuseppe.bove@uniroma3.it

Pier Luigi Conti
pierluigi.conti@uniroma1.it

Daniela Marella
daniela.marella@uniroma3.it

[1] Dipartimento di Scienze Statistiche, Università 'La Sapienza", P.le A. Moro, 5, 00185 Rome, Italy

[2] Dipartimento di Scienze della Formazione, Università "Roma Tre", via del Castro Pretorio 20, 00185 Roma, Italy

Journal : Small-ext 10260 | Dispatch : 19-11-2020 | Pages : 19
Article No. : 551 | □ LE | □ TYPESET
MS Code : SMAP-D-19-00262R2 | ☑ CP | ☑ DISK

produced by a group of writers. Similar situations can be found in organizational, educational, biomedical, social, and behavioural research areas, where raters can be counsellors, teachers, clinicians, evaluators, or consumers and targets can be organization members, students, patients, subjects, or objects. When each rater evaluates each target, the raters provide comparable categorizations of the targets. The more the raters categorizations coincide, the more the rating scale can be used with confidence without worrying about which raters produced those categorizations. Hence, the main interest here consists in analysing the extent that raters assign the same (or very similar) values on the rating scale (interrater absolute agreement), that is to establish to what extent raters evaluations are close to an equality relationship (*e.g.*, in the case of only two raters, if the two sets of ratings are represented by x and y the relation of interest is $x = y$). Measures of interrater absolute agreement, as Cohen's Kappa [and extensions to take into account three or more raters, e.g., von Eye and Mun (2005)] and intraclass correlations (ICC) [(Shrout and Fleiss 1979; McGraw and Wong 1996)] are usually applied when dealing with rating performed by ordinal scales. A first problem of these procedures is that they are not originally defined for ordinal scales, and so they have to be adapted. For instance, the application of indices based on Cohen's Kappa need to assign numerical values to the ordinal level of the scale; intraclass correlation indices are based on ANOVA for repeated measures approach for interval data. Another limitation of the above mentioned measures is that they are affected by the restriction of variance problem [e.g., LeBreton et al. (2003)], that consists in an attenuation of estimates of rating similarity caused by an artefact reduction of the between-subjects variance in ratings. For instance, this happens in language studies when the same task is defined for native (L1) and non-native (L2) writers, and the analysis compare rater agreement in the two groups separately. Even in the presence of a very good absolute agreement, Cohen's Kappa coefficient and intraclass correlations can take low values, especially for L1 group, because the range of ratings provided by the raters are concentrated on one or two very high levels of the scale (a range restriction that determines a between-target variance restriction).

In order to overcome the restriction of variance problem, measure for absolute agreement (or consensus) have been proposed, see (LeBreton and Senter 2008) for a review. The main underlying idea is to measure the within-target variance of ratings (i.e., the between-rater variance) separately for each target, and summarize the results in a final average index (usually normalized in the interval [0, 1]). In this approach, the influence of the low level of the between-target variance is removed by separate analysis of the ratings of each target. One of the most popular index in this group, denoted by r_{WG}, was proposed by James et al. (1984), (1993). Let X be an ordinal categorical variable with K categories (*e.g.* a Likert scale), the index r_{WG} can be expressed as

$$r_{WG} = 1 - \frac{s_X^2}{\sigma_E^2} \tag{1}$$

where s_X^2 is the observed between-rater variance of the ratings and σ_E^2 is the between-rater variance obtained from a theoretical null distribution representing a

Journal : Small-ext 10260	Dispatch : 19-11-2020	Pages : 19
Article No. : 551	□ LE	□ TYPESET
MS Code : SMAP-D-19-00262R2	☑ CP	☑ DISK

complete lack of agreement among raters. Roughly speaking, the null distribution conceptually represents no agreement, which means that to calculate r_{WG}, one makes a direct comparison between the observed variance in raters' ratings with the variance one would expect if there was no agreement among raters. Higher numbers indicate a greater agreement.

For raters in perfect agreement we have $s_X^2 = 0$, with a corresponding value $r_{WG} = 1$. In applications, r_{WG} values greater than 0.7 (possibly 0.8) are considered associated with high level of interrater absolute agreement [see (LeBreton and Senter 2008), p. 836 Table 2]. Often researchers define the no agreement, or the null distribution, in terms of a uniform distribution. When the null distribution is assumed as uniform, the equation for the corresponding variance is

$$\sigma_E^2 = \frac{K^2 - 1}{12} \tag{2}$$

where K refers to the total number of levels of the scale X.

The index r_{WG} and other indices reviewed in LeBreton and Senter (2008) (*e.g.*, standard and average deviation indices) allow to avoid the problem of variance restriction, but as traditional measures of interrater agreement they are defined only for interval data. Besides, the accuracy of r_{WG} depends strongly on the specification of the null distribution. One disadvantage of r_{WG} is the ambiguity in choosing the reference distribution. Although (James et al. 1984) recommended using the uniform distribution, Lindell and Brandt (1997) recommended using maximum dissensus. Burke et al. (1999), however, cautioned against the use of maximum dissensus because its use may lead to the overestimation of interrater agreement. Finally, depending on the choice of the null distribution, negative values could be obtained for r_{WG}. For these reasons, in this contribution we propose a new procedure to measure absolute agreement for ordinal rating scales by using the dispersion index proposed by Leti (1983) (pp. 290–297) for ordinal variables. In this way, we take into consideration the ordinal level of the measurement scales. The new measure is not affected by restriction of variance problems and does not depend on the choice of a particular null distribution. In this paper we assume a two-way random sampling design, where the sampling design involves a sample of raters as well as a sample of targets, all of which are rated by each sampled rater.

The paper is organized as follows. In Sect. 2 the dispersion index proposed by Leti (1983) (pp. 290–297) for ordinal variables is introduced and its sampling properties are analyzed in Sect. 3. Such results allow to construct confidence interval without resorting to bootstrap method, as generally happened for inference on measure of interrater absolute agreement, see (Cohen et al. 2001) and reference therein. Section 4 contains results of a simulation experiment used to illustrate both the performance of the proposed interrater agreement index and to compare it with the bootstrap method in constructing confidence intervals. Finally, in Sect. 5 an application to real data is performed.

Journal : Small-ext 10260	Dispatch : 19-11-2020	Pages : 19
Article No. : 551	□ LE	□ TYPESET
MS Code : SMAP-D-19-00262R2	☑ CP	☑ DISK

2 Leti index as a measure of interrater absolute agreement for ordinal scales

The dispersion of an ordinal categorical variable can be measured by the index proposed in Leti (1983) (pp. 290–297), which is given by

$$D = 2\sum_{k=1}^{K-1} F_k(1 - F_k) \tag{3}$$

where K is the number of categories of the variable X and F_k is the cumulative proportion associated to category k, for $k = 1, \ldots, K$. It is interesting to notice that D has properties of within and between dispersion decomposition analogous to the well-known variance decomposition (Grilli and Rampichini 2002). Index (3) is nonnegative and it is easy to prove that $D = 0$ if and only if all observed categories are equal (absence of dispersion). The maximum value of the index (D_{max}) is obtained when all observations are concentrated in the two extreme categories of the variable (maximum dispersion), and it is

$$D_{max} = \frac{K-1}{2} \tag{4}$$

as N is even,

$$D_{max} = \frac{K-1}{2}\left(1 - \frac{1}{N^2}\right) \tag{5}$$

as N is odd, N being the total number of observations. For N moderately large, the maximum of the index can be assumed equal to $(K-1)/2$. Hence, it is possible to define a measure of dispersion normalized in the interval [0, 1] given by

$$d = \frac{D}{D_{max}} = \frac{2}{K-1}D. \tag{6}$$

The lower the value of d the higher the raters agreement. Note that, when $d = 0$ (maximum agreement between raters) $r_{WG} = 1$. When $d = 1$

$$r_{WG} = 1 - \frac{(K-1)^2}{4}\frac{1}{\sigma_E^2} \tag{7}$$

and if the uniform distribution is assumed as null distribution, (7) becomes

$$r_{WG} = \frac{4-2K}{K+1} \tag{8}$$

taking value lower than zero when $K > 2$. In accordance with (LeBreton et al. 2005) out-of-bounds values ($r_{WG} < 0$ or $r_{WG} > 1$) are generally setted to zero. Unlike r_{WG}, d can never be out of the range [0, 1].

Advantages of our proposal respect to measures of absolute agreement like r_{WG} are:

Journal : Small-ext 10260 | Dispatch : 19-11-2020 | Pages : 19
Article No. : 551 | □ LE | □ TYPESET
MS Code : SMAP-D-19-00262R2 | ☑ CP | ☑ DISK

(i) d takes into consideration the ordinal level of the measurement scales;
(ii) d allows to avoid the problem of restriction of variance;
(iii) d does not depend by the formulation of a null distribution for normalization;
(iv) the sampling proprieties of d are known, as showed in Sect. 3.

Remark 1 In order to homogenize the values assumed by d and r_{WG}, the index $1 - d$ can be considered.

Suggestions for interpreting the value of $1 - d$ appropriately are in Table 1, where a comparison between r_{WG} and d $(1 - d)$ is reported. More specifically, datasets with different level of raters agreement have been generated and the indices r_{WG}, d and $1 - d$ have been computed.

As reported in LeBreton et al. (2003) values of r_{WG} greater than 0.7 (possibly 0.8) are considered associated with high level of interrater absolute agreement. As shown in Table 1 the same consideration holds for the $1 - d$ index.

Finally, in this paper a single item on Likert scale with K categories has been considered. For J items, the index r_{WG} (denoted by $r_{WG(J)}$) can been defined as shown in Cohen et al. (2001). Analogoulsy to $r_{WG(J)}$, extensions to J items for d index based on the average of J values d_j, each computed for each single item, can be considered.

3 Sampling properties of *d* index

A sample of n_R raters and a sample of n_T targets are drawn by simple random sampling without replacement from a finite population of targets and raters, respectively. Let us denote with X_{ij} the score given by the jth rater to the ith target on a K-point scale, for $i = 1, \ldots, n_T$ and $j = 1, \ldots, n_R$. Formally, X_{ij}s are independent categorical random variables having K categories with $p_k^{(ij)} = P(X_{ij} = k)$, for $i = 1, \ldots, n_T$, $j = 1, \ldots, n_R$ and $k = 1, \ldots, K$. In the sequel we assume that both the targets and the raters are homogeneus (*targets-raters homogeneity assumption*), which implies that the probability $p_k^{(ij)} = p_k$ does not depend on rater j or target i, for $i = 1, \ldots, n_T$, $j = 1, \ldots, n_R$, $k = 1, \ldots, K$. As a consequence of *homogeneity assumptions*, the variables X_{ij} are independent and identically distributed (*i.i.d.*).

Table 1 Comparison between r_{WG} and d $(1 - d)$

r_{WG}	d	$1 - d$
0.07	0.81	0.19
0.34	0.61	0.39
0.49	0.53	0.47
0.74	0.32	0.68
0.83	0.14	0.86

Journal : Small-ext 10260 | Dispatch : 19-11-2020 | Pages : 19
Article No. : 551 | □ LE | □ TYPESET
MS Code : SMAP-D-19-00262R2 | ☑ CP | ☑ DISK

Remark 2 With regard to the raters homogeneity, variability in scores provided by raters may depend on a number of raters characteristics such as their expertise, familiarity with the assessment process, or amount of training raters received prior to the rating task, etc. Cumming et al. (2002) showed that rating was positively influenced by earlier rating experience and by experience as an EFL/ESL or English L1 teacher. Thompson (1991) indicated that training in linguistics and knowledge of other languages may lead to higher degrees of interrater reliability. Roughly speaking, assuming raters homogeneity means to eliminate the effect of such characteristics on raters score.

Evaluations of interrater agreement can be applied to a number of different contexts and are frequently encountered in social, medicine, psychology and education. An application in medicine and in education are illustrated in Examples 1 and 2, respectively.

Example 1 Gleason grading is a used grading system for prostatic carcinoma. The Gleason Score is the grading system used to determine the aggressiveness of prostate cancer. This grading system can be used to choose appropriate treatment options. The Gleason Score ranges from 1 to 5 and describes how much the cancer from a biopsy looks like healthy tissue (lower score) or abnormal tissue (higher score). In Allsbrook et al. (2001) 46 needle biopsies containing prostatic carcinoma were assigned Gleason scores by 10 urologic pathologists. Clearly the urologic pathologists do not necessarily give the same grading for each patient. However, we would expect that they tend to agree with each other. The hypothesis that X_{ij} are *i.i.d* comes from the assumption of targets and raters homogeneity. With regard to Allsbrook et al. (2001) study: (i) the 10 urologic pathologists are homogeneous since they have the same background knowledge and familiarity with grading system; (ii) the 46 patients are homogeneous because affected by the same kind of prostatic carcinoma.

Example 2 A study of agreement among raters in educational research is in Kuiken and Vedder (2014), where raters' judgements of writing performance in L2 and L1 has been analyzed. More specifically, all texts in L2 and L1 were rated by expert raters on both communicative adequacy and linguistic complexity on a six-point Likert scale. All raters were experienced L2-teachers and native speakers of the target language. Furthermore, they are homogeneous with respect to the familiarity with the assessment process and the amount of training raters received prior to the rating task.

As previously stressed, the dispersion of an ordinal categorical variable can be measured by the index (3).

With regard to ith target, let us denote with $\widehat{F}_k^{(i)}$ the empirical cumulative distribution function defined as

Journal : Small-ext 10260 Dispatch : 19-11-2020 Pages : 19
Article No. : 551 □ LE □ TYPESET
MS Code : SMAP-D-19-00262R2 ☑ CP ☑ DISK

$$\widehat{F}_k^{(i)} = \frac{1}{n_R} \sum_{j=1}^{n_R} I_{(X_{ij} \leq k)} \tag{9}$$

where the numerator represents the number of raters giving score less than or equal to k to the ith target. It is known that $E(\widehat{F}_k^{(i)}) = F_k^{(i)} = F_k$, where the last equality comes from the *targets homogeneity assumptions*. Furthermore, $V(\widehat{F}_k^{(i)}) = F_k(1 - F_k)$ and $Cov(\widehat{F}_k^{(i)}, \widehat{F}_l^{(i)}) = min(F_k, F_l) - F_k F_l$. In order to estimate (3), for each target i the following estimator can be defined

$$\widehat{D}_i = 2 \sum_{k=1}^{K-1} \widehat{F}_k^{(i)} (1 - \widehat{F}_k^{(i)}). \tag{10}$$

As stressed in Piccarreta (2001), (10) can be alternatively expressed as

$$\begin{aligned} \widehat{D}_i &= \sum_{k=1}^{K} \sum_{l=1}^{K} |k - l| \widehat{p}_k^{(i)} \widehat{p}_l^{(i)} \\ &= \frac{1}{n_R^2} \sum_{j=1}^{n_R} \sum_{j'=1}^{n_R} |X_{ij} - X_{ij'}| \end{aligned} \tag{11}$$

where

$$\widehat{p}_k^{(i)} = \frac{1}{n_R} \sum_{j=1}^{n_R} I_{(X_{ij} = k)} \tag{12}$$

is an unbiased estimator of p_k.

Proposition 1 *The random variable (r.v)* $n_R(\widehat{p}_1, \ldots, \widehat{p}_K)'$, *with* $\widehat{p}_k = \sum_{i=1}^{n_T} \widehat{p}_k^{(i)} / n_T$ *for* $k = 1, \ldots, K$, *follows a multinomial distribution with parameters* n_R *and* $(p_1, \ldots, p_K)$.

The expression (11) allows to compute easily the expectation and the variance of estimator (10) as shown in Proposition 2, see Lomnicki (1952) for details.

Proposition 2 *The estimator* $\widehat{D}_i$ *has expectation*

$$E(\widehat{D}_i) = \left(1 - \frac{1}{n_R}\right) D \tag{13}$$

and variance given by

$$Var(\widehat{D}_i) = \left(\frac{1}{n_R^2} - \frac{1}{n_R^3}\right)(4\sigma^2 + 4(n_R - 2)J - 2(2n_R - 3)D^2) = V \tag{14}$$

where

Journal : Small-ext 10260 | Dispatch : 19-11-2020 | Pages : 19
Article No. : 551 | □ LE | □ TYPESET
MS Code : SMAP-D-19-00262R2 | ☑ CP | ☑ DISK

$$\sigma^2 = Var(X_{ij}) = \sum_{k=1}^{K} k^2 p_k - \left(\sum_{k=1}^{K} k p_k\right)^2 \tag{15}$$

$$J = \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{l=1}^{K} |k-h||k-l| p_k p_h p_l. \tag{16}$$

Proof Both (13) and (14) come from the results in Lomnicki (1952). With regard to (13), we have

$$\begin{aligned} E(\widehat{D}_i) &= E\left(\frac{1}{n_R^2}\sum_{j=1}^{n_R}\sum_{j'=1}^{n_R}|X_{ij}-X_{ij'}|\right) \\ &= \frac{n_R(n_R-1)}{n_R^2} E\left(\frac{1}{n_R(n_R-1)}\sum_{j=1}^{n_R}\sum_{j'=1}^{n_R}|X_{ij}-X_{ij'}|\right) \\ &= \frac{n_R(n_R-1)}{n_R^2} 2\sum_{k=1}^{K-1} F_k(1-F_k) \\ &= \left(\frac{n_R-1}{n_R}\right) D. \end{aligned} \tag{17}$$

For the variance (14) we obtain

$$\begin{aligned} Var(\widehat{D}_i) &= Var\left(\frac{1}{n_R^2}\sum_{j=1}^{n_R}\sum_{j'=1}^{n_R}|X_{ij}-X_{ij'}|\right) \\ &= \left(\frac{n_R-1}{n_R}\right)^2 Var\left(\frac{1}{n_R(n_R-1)}\sum_{j=1}^{n_R}\sum_{j'=1}^{n_R}|X_{ij}-X_{ij'}|\right) \\ &= \left(\frac{n_R-1}{n_R}\right)^2 \frac{1}{n_R(n_R-1)}(4\sigma^2 + 4(n_R-2)J - 2(2n_R-3)D^2) \\ &= \left(\frac{1}{n_R^2}-\frac{1}{n_R^3}\right)(4\sigma^2 + 4(n_R-2)J - 2(2n_R-3)D^2). \end{aligned} \tag{18}$$

□

Remark 3 For n_R sufficiently large, we have

$$Var(\widehat{D}_i) \approx \frac{4(J-D^2)}{n_R}. \tag{19}$$

As an estimator of d index (6) we consider

Journal : Small-ext 10260 | Dispatch : 19-11-2020 | Pages : 19
Article No. : 551 | □ LE | □ TYPESET
MS Code : SMAP-D-19-00262R2 | ☑ CP | ☑ DISK

$$\widehat{d} = \frac{\widehat{\widehat{D}}}{D_{max}} = \frac{1}{D_{max}}\left(\frac{1}{n_T}\sum_{i=1}^{n_T}\widehat{D}_i\right). \tag{20}$$

where $\widehat{\widehat{D}}$ is an estimator of D obtained averaging the n_T estimates $\widehat{D}_1, \ldots, \widehat{D}_{n_T}$.

In Proposition 3 both the sampling properties and the asymptotic distribution of $\widehat{d}$ 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* $\widehat{d}$ *has expectation*

$$E(\widehat{d}) = \left(\frac{n_R - 1}{n_R}\right)d \tag{21}$$

and variance

$$V_d = \left(\frac{1}{D_{max}}\right)^2 \frac{V}{n_T} \tag{22}$$

where V *is given in* (14). *Furthermore, since* $\widehat{D}_1, \ldots, \widehat{D}_{n_T}$ are *i.i.d., for the central limit theorem, as* n_T *goes to infinity the random variable* $\widehat{d}$ *tends to a standard normal distribution with mean and variance given by* (21) *and* (22),*respectively.*

Remark 4 If the homogeneity assumption is violated then the X_{ij} random variables are independent but not identically distributed. The main result in this area is the Liaponouv's Theorem, (see Billingsley 1995). The theorem strengthens the requirement of finite variance requiring that the X_{ij} have finite moments of order $(2 + \delta)$, for some $\delta > 0$. Clearly, the convergence to normal distribution could be slower.

In Proposition 4 an unbiased estimator of d is proposed and its asymptotic distribution is evaluated.

Proposition 4 *From* (21), *an unbiased estimator of* d *can be defined as follows*

$$\widehat{d^*} = \frac{n_R}{n_R - 1}\widehat{d}. \tag{23}$$

As a consequence of Proposition(3), *the distribution of* $\widehat{d}^*$ *is approximately normal with mean* d *and variance*

$$V_{d^*} = \left(\frac{n_R}{n_R - 1}\right)^2 \left(\frac{1}{D_{max}}\right)^2 \frac{V}{n_T}. \tag{24}$$

The proof of Proposition 4 follows from Proposition 3. The above results are useful to construct point and interval estimates of d. They are also useful for testing both the statistical significance of the index (that is the null hypotheses $H_0 : d = 0$) and null hypothesis such as $H_0 : d \leq d_0$, where d_0 be a real number in [0, 1]. Consider the hypothesis problem

Journal : Small-ext 10260 Dispatch : 19-11-2020 Pages : 19
Article No. : 551 □ LE □ TYPESET
MS Code : SMAP-D-19-00262R2 ☑ CP ☑ DISK

$$\begin{cases} H_0: & d \leq d_0 \\ H_1: & d > d_0 \end{cases} \tag{25}$$

As a consequence of Proposition 4, a test with an asymptotic significance level α consists in accepting H_0 whenever

$$\widehat{d}^* \leq d_0 + z_\alpha \sqrt{\widehat{V}_{d^*}} \tag{26}$$

where z_α is the α-th quantile of the standard normal distribution and $\widehat{V}_{d^*}$ is an estimate of variance (24).

The performance $\widehat{d}^*$ has been evaluated in Sect. 4 by a simulation study and it has been compared with the bootstrap method. With regard to the size of d, the judgment depends on the application context. Researchers should gain experience using the proposed index to understand which values might be expected to be obtained for d in various situations and how to interpret these values. For instance, one of the main questions in multilevel data analysis is whether it is appropriate to aggregate data and to use the aggregated measures to make inferences about higher level units. A necessary precondition for aggregation is that there is an agreement among the individuals who form the group with regard to the aggregated construct. In this context, the problem is to evaluate if the degree of agreement justifies data aggregation. From this perspective, the hypothesis test (25) assumes a fundamental importance.

Remark 5 If the index $1-d$ introduced in Remark 1 is considered, as a consequence of Proposition 4, the distribution of $1-\widehat{d}^*$ is approximately normal with mean $1-d$ and variance given by (24).

4 Simulation study

In this section a simulation study has been performed. The aim is: (i) to evaluate the performance of $\widehat{d}^*$; (ii) to compare the normal approximation for the distribution of $\widehat{d}^*$ with the bootstrap method. Such a method is generally used in constructing confidence intervals of interrater agreement measures but its use is recommended when n_R is sufficiently large (e.g., $n_R > 20$), see Cohen et al. (2001). Alternative methods based on bootstrap to construct confidence intervals are compared in the simulation.

We focus on confidence intervals for the index d because confidence intervals indicate the range within which the population parameter d (the interrater agreement in the population) is likely to fall, as well as precision of this estimate (i.e., the size of the range).

A finite population of size $N_T = 150$ targets and $N_R = 28$ raters was generated from a multinomial model with parameters $N_R = 28$ and probabilities $(p_1, p_2, p_3, p_4, p_5) = (0.1, 0.2, 0.35, 0.25, 0.1)$. Then, the finite population consists in a matrix P of size $N_T \times N_R$. The value of d index (6) is 0.61.

Journal : Small-ext 10260 | Dispatch : 19-11-2020 | Pages : 19
Article No. : 551 | □ LE | □ TYPESET
MS Code : SMAP-D-19-00262R2 | ☑ CP | ☑ DISK

From the population, $S = 1000$ samples were drawn according to a simple random sampling without replacement on the basis of the following two-step procedure. First of all, a simple random sample of size $n_R = 7$ from the $N_R = 28$ raters has been selected. This is equivalent to select a simple random sampling without replacement of columns in the finite population matrix P, the result is a matrix P_R of size $N_T \times n_R$. Secondly, a simple random sampling of size $n_T = 50$ from $N_T = 150$ targets has been drawn. This means to draw a simple random sampling of $n_T = 50$ rows from P_R.

In order to construct confidence intervals for the index d, both the asymptotic result in Proposition 4 and bootstrapping procedures are used. The bootstrap methods are described in points (2)–(4) below, where we assume that $B = 1000$ bootstrap samples are drawn from each initial sample s. Formally, confidence intervals for d of level $1 - \alpha = 0.95$ have been constructed using the following methods:

(1) *Normal approximation* For the initial sample s (for $s = 1, \ldots, S$), the confidence interval $[L^s_{Norm}, U^s_{Norm}]$ based on the asymptotic normal approximation is given by

$$L^s_{Norm} = \widehat{d}^* - z_{1-\alpha/2}\sqrt{\widehat{V}_{d^*}}; \quad U^s_{Norm} = \widehat{d}^* - z_{\alpha/2}\sqrt{\widehat{V}_{d^*}} \tag{27}$$

where $\widehat{d}^*$ and $\widehat{V}_{d^*}$ are the estimates of d and V_{d^*}, respectively.

(2) *Percentile method* For the initial sample s (for $s = 1, \ldots, S$), the confidence interval $[L^s_{Perc}, U^s_{Perc}]$ is obtained by taking $\alpha/2$ and $1 - \alpha/2$ quantiles of the B bootstrap samples. Formally

$$L^s_{Perc} = Q_{\alpha/2}; \quad U^s_{Perc} = Q_{1-\alpha/2} \tag{28}$$

(3) *Bootstrap-t interval* For the initial sample s (for $s = 1, \ldots, S$), the confidence interval $[L^s_T, U^s_T]$ is computed as follows

$$L^s_{T-int} = \widehat{d^*} - t_{1-\alpha/2}\sqrt{\widehat{V}_{d^*}}; \quad U^s_{T-int} = \widehat{d^*} - t_{\alpha/2}\sqrt{\widehat{V}_{d^*}} \tag{29}$$

where t_α is the αth percentile of the distribution of z^*_b (for $b = 1, \ldots, B$) with

$$z^*_b = \frac{\widehat{d^*_b} - \widehat{d^*}}{\widehat{se}^*_b}. \tag{30}$$

In (30) $\widehat{d^*_b}$ is the estimate of d^* based on the bth bootstrap sample and $\widehat{se}^*_b$ is the standard error based on the data in the bth bootstrap sample.

(4) *Pivotal method* For the initial sample s (for $s = 1, \ldots, S$), the confidence interval $[L^s_{Pivot}, U^s_{Pivot}]$ is computed as follows

$$L^s_{Pivot} = 2\widehat{d^*} - Q_{1-\alpha/2}; \quad U^s_{Pivot} = 2\widehat{d^*} - Q_{\alpha/2} \tag{31}$$

where $Q_{\alpha/2}$ and $Q_{1-\alpha/2}$ are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the B bootstrap estimates $\widehat{d^*_b}$, for $b = 1, \ldots, B$.

Journal : Small-ext 10260 | Dispatch : 19-11-2020 | Pages : 19
Article No. : 551 | □ LE | □ TYPESET
MS Code : SMAP-D-19-00262R2 | ☑ CP | ☑ DISK

As far as the methods described in steps (2)–(4) are concerned, from each of the $S = 1000$ initial samples, the $B = 1000$ bootstrap samples were selected according to the following methods:

1 *Nonparametric bootstrap* From each initial sample s, the bth bootstrap sample is selected as follows: (i) a simple random sample with replacement of $r = 7$ raters has been selected from the original sample of raters; (ii) a simple random sampling with replacement of $n = 50$ writers has been drawn from the original sample of writers. Then, bootstrap is applied to the raters sample as well as the targets sample in order to take into account the variability in $\widehat{d}^*$ due to the two-way random sampling design (where the sampling design involves a sample of raters and a sample of targets). Clearly, when the sampling design involves only the raters the proposed methodology resembles that used in literature, see Cohen et al. (2001) and reference therein.

2 *Parametric bootstrap* From each initial sample s, the bth bootstrap sample is generated according the multinomial model specified in Proposition 1.

3 *Pseudo-Nonparametric bootstrap* The nonparametric bootstrap described in point (1), is based on the assumption that the data are *i.i.d.*, see Efron (1979). Since survey data are not necessarily *i.i.d.*, many bootstrap resampling methods have been proposed in the context of survey sampling. These methods are obtained after making some modifications to the classical *i.i.d.* bootstrap in order to adapt it for survey data. For a review of bootstrap methods in the context of survey data, see Mashreghi et al. (2016). The class of pseudo-population bootstrap methods consists in creating a pseudo-population by repeating the units of the initial sample and drawing from such a pseudo-population bootstrap samples with the same design as the initial one. In order to illustrate how a pseudo-population is constructed, let us assume that a simple random sample without replacement has been selected from a finite population of size N. A pseudo-population of size N can be created by repeating the selected sample, N/n times. This method, was first introduced by Gross (1980). In practice N/n is rarely an integer, in this case a method to build a pseudo-population of size N was proposed by Booth et al. (1994). In this method, a pseudo-population is first constructed by replicating $k = \lfloor N/n \rfloor$ times each unit of the original sample s. Then, the pseudo-population is completed by taking a simple random sample of size $N - nk$ without replacement from s. Taking into account the two-way sampling design of both targets and raters, the pseudo-population has been generated according the following two step procedure:

Step 1 the ratings of $N_R = 28$ raters have been reconstruted replicating the columns of the original sample s, $k_R = N_R/n_R = 28/7 = 4$ times. As a consequence, this first step generates a sample s_R of size $n_T = 50$ and $n_R = N_R = 28$;

Step 2 the points of $N_T = 150$ targets have been reconstruted replicating the rows of the sample s_R obtained in Step 1, $k_T = N_T/n_T = 150/50 = 3$ time.

Journal : Small-ext 10260 | Dispatch : 19-11-2020 | Pages : 19
Article No. : 551 | ☐ LE | ☐ TYPESET
MS Code : SMAP-D-19-00262R2 | ☑ CP | ☑ DISK

The accuracy of confidence intervals has been evaluated by the following indicators.

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

$$ECP = \frac{100}{S}\sum_{s=1}^{S} I(L_t^s \leq d \leq U_t^s). \tag{32}$$

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

$$LE = \frac{100}{S}\sum_{s=1}^{S} I(L_t^s > d), \tag{33}$$

$$RE = \frac{100}{S}\sum_{s=1}^{S} I(U_t^s < d). \tag{34}$$

(3) Estimated average length (AL) of all 1000 simulated intervals given by

$$AL = \sum_{s=1}^{S} \frac{U_t^s - L_t^s}{S} \tag{35}$$

where $I(a) = 1$ if a is true and $I(a) = 0$ elsewhere, and $t = Norm, T - int, Perc, Pivot$.

4.1 Simulation results

Tables 2 presents the outcomes achieved in the simulation study. More specifically, 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 index d, when ($n_R = 7, n_T = 50$), are reported. The d value is equal to 0.61.

As reported in Table 2, the confidence intervals obtained with the normal approximation perform very well. Coverage probabilities are larger than 95% nominal value (99.4%) with an average length of 0.16. Furthermore, the normal confidence intervals construction is simple, as it does not require resampling from the initial sample. Figure 1 shows the kernel density of the d index estimated from the 1000 original samples. The bandwidth selection rule is as proposed by Sheather and Jones (1991).

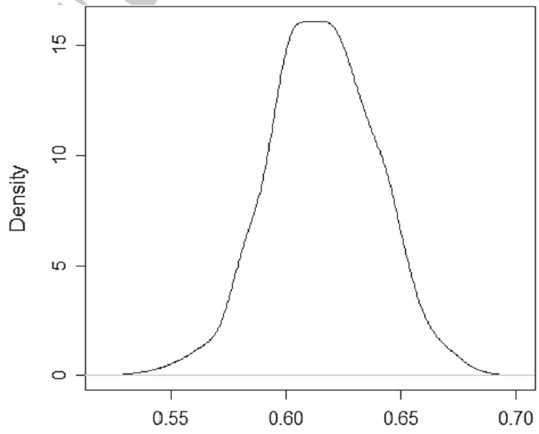
The percentile method has a good performance with coverage probability larger than 91%. The worst methods are the *Pivot* and *T*-int methods. The lower and upper error rates, giving us an idea of how skewed the distribution of the d estimator is, are not well balanced. With regard to the methods used to generate the bootstrat samples, the *parametric* approach performance is strictly related to the estimation of the multinomial probabilities. As previously stressed, each row in the inital sample s provides an estimate of $(p_1, p_2, p_3, p_4, p_5)$ and the mean of such estimates defines the

Journal : Small-ext 10260 | Dispatch : 19-11-2020 | Pages : 19
Article No. : 551 | □ LE | □ TYPESET
MS Code : SMAP-D-19-00262R2 | ☑ CP | ☑ DISK

Table 2 Performance of different confidence intervals for d when $n_R = 7$, $d = 0.61$

		$n_R = 7$		
Method	Indicators	Nonparametric	Parametric	Pseudo-Nonparametric
Normal	CP	99.4	99.4	99.4
	LE	0.6	0.6	0.6
	RE	0	0	0
	AL	0.16	0.16	0.16
T-int	CP	26.2	72.4	28.8
	LE	73.8	26.2	71.2
	RE	0	1.4	0
	AL	0.18	0.08	0.15
Perc	CP	92.8	91.2	92.8
	LE	0	8.8	0
	RE	7.2	0	7.2
	AL	0.23	0.10	0.18
Pivot	CP	27	79.2	30
	LE	73	19.6	70
	RE	0	1.2	0
	AL	0.23	0.10	0.18

Fig. 1 Kernel density estimate of d index from the 1000 original samples



estimated probabilities $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3, \widehat{p}_4, \widehat{p}_5)$ of the multinomial distribution used to generate the bootstrap samples as specified in Proposition 1. In Table 3, the minimum, the maximum, the mean and the standard deviation of the distribution of $\widehat{p}_k$ (for $k = 1, 2, 3, 4, 5$) estimated from the original 1000 samples are reported.

Journal : Small-ext 10260 | Dispatch : 19-11-2020 | Pages : 19
Article No. : 551 | □ LE | □ TYPESET
MS Code : SMAP-D-19-00262R2 | ☑ CP | ☑ DISK

Table 3 Descriptive statistics of $\widehat{p}_k$ distribution, for $k = 1, 2, 3, 4, 5$ and $d = 0.61$

Parameter	True value	Min	Max	Mean	Sd
p_1	0.10	0.05	0.15	0.10	0.01
p_2	0.20	0.14	0.25	0.19	0.02
p_3	0.35	0.29	0.41	0.35	0.02
p_4	0.25	0.20	0.33	0.26	0.02
p_5	0.10	0.06	0.16	0.10	0.02

As Table 2 shows, the *pseudo-nonparametric* approach taking into account the sample selection effects has a slightly better performance than the *nonparametric* approach both in terms of coverage probabilities and average lengths for all methods ($T - int$, $Perc$, $Pivot$).

Finally, note that in the *nonparametric* approach the resampling with replacement from $n_R = 7$ raters generates a replication of columns of the bootstrap sample introducing a false agreement between raters and as a consequence an underestimation of d. This fact is showed in Table 4 where the mean of the d estimates over both the 1000 original samples s and over the bootstrap replications b are reported.

Such means have been computed both for the original population with $d = 0.61$ and for a population with $d = 0.41$, showing as the magnitude of bias depends also on the original agreement degree between raters. That is, the higher the raters agreement (low values of d), the smaller the bias in the d estimator introduced by the resampling with replacement. Clearly, such a bias is also present in the *pseudo-nonparametric* approach but with a smaller magnitude, thank to the construction of the pseudo-population that mitigates such a phenomenon. As Table 4 shows, the *parametric approach* produces null bias estimates.

The simulation in Table 2 has been repeated for a population with $d = 0.41$. The results are reported in Table 5.

In conclusion, the most competitive method in terms of performance and computational time seem to be the normal. Finally, among the alternative methods based on bootstrap the percentile method in the *parametric* approach seems to perform better.

Table 4 The mean of $\widehat{d}$ over the initial samples s and over the bootstrap replications b

Approach	Mean of $\widehat{d}^*$ *(d=0.61)*	Mean of $\widehat{d}^*$ *(d=0.41)*
Nonparametric	0.53	0.36
Parametric	0.61	0.41
Pseudo-nonparametric	0.55	0.37

Journal : **Small-ext 10260** Dispatch : **19-11-2020** Pages : **19**
Article No. : **551** □ LE □ TYPESET
MS Code : **SMAP-D-19-00262R2** ☑ CP ☑ DISK

Table 5 Performance of different confidence intervals for d when $n_R = 7$, $d = 0.41$

		$n_R = 7$		
Method	Indicators	Nonparametric	Parametric	Pseudo-nonparametric
Normal	CP	98.2	98.2	98.2
	LE	1.8	1.8	1.8
	RE	0	0	0
	AL	0.13	0.13	0.13
T-int	CP	60.2	83.2	61.2
	LE	39.8	14.8	38.8
	RE	0	2	0
	AL	0.18	0.10	0.14
Perc	CP	93.2	93.8	93.2
	LE	0	5.8	0
	RE	6.8	0.4	6.3
	AL	0.19	0.10	0.15
Pivot	CP	64.8	84.6	65.4
	LE	35.2	12.6	34.6
	RE	0	2.8	0
	AL	0.19	0.10	0.15

5 An application on real data: the assessment of language proficiency

The aim of this section is to apply the methodology illustrated in the previous sections on an empirical data set, we have analysed ratings obtained in a research conducted at Roma Tre University [see (Nuzzo and Bove 2020), for a detailed description]. The main aim of the study was to investigate the applicability of a six-point Likert scale for functional adequacy (an aspect of language proficiency) developed by Kuiken and Vedder (2017) to texts produced by native and non-native writers, and to different task types (narrative, instruction, and decision-making tasks). The scale comprises four subscales, corresponding to the four dimensions of functional adequacy identified by the authors of the scale: content, task requirements, comprehensibility, coherence and cohesion [the reader is referred to Kuiken and Vedder (2017) for a detailed presentation of scales and descriptors]. 20 native speakers of Italian (L1) and 20 non-native speakers of Italian (L2) participated in the study as writers. All the texts produced by L1 and L2 writers (120 texts in total for the three tasks) were assessed by 7 native speakers of Italian on the Kuiken and Vedder six-point Likert scale. The raters did not have any specific experience in judging written texts, and can therefore be categorized as being non-expert. For our purposes, we have selected ratings concerning only the narrative task and the subscale comprehensibility. Just to give a general idea of the subscale, definitions of levels 1 and 6 are reported in the following:

Journal : Small-ext 10260 | Dispatch : 19-11-2020 | Pages : 19
Article No. : 551 | □ LE | □ TYPESET
MS Code : SMAP-D-19-00262R2 | ☑ CP | ☑ DISK

Level 1: The text is not at all comprehensible. Ideas and purposes are unclearly stated and the efforts of the reader to understand the text are ineffective.

Level 6: The text is very easily comprehensible and highly readable. The ideas and the purpose are clearly stated.

The results of the interrater agreement analysis for the subscale are summarized in Table 6, where the intraclass correlation $ICC(A, 1)$, as defined in McGraw and Wong (1996), and the average values of r_{WG}, as defined in LeBreton and Senter (2008), the coefficient of variation CV, $\widehat{d}$ and $\widehat{d}^*$ are shown for L1, L2 and total groups. The intraclass correlation $ICC(A, 1)$ provides a low-moderate level of agreement for the total group (0.67). The results for the average values of CV (12.16%) seems in accord with $ICC(A, 1)$, while the average value of $r_{WG} = 0.87$, $\widehat{d} = 0.22$ $(1 - \widehat{d} = 0.78)$ and $\widehat{d}^* = 0.25$ $(1 - \widehat{d}^* = 0.75)$ highlight a higher level of agreement. As it was observed in Bove et al. (2018), when the analysis focuses separately on the two subgroups of L1 and L2 students, results regarding the L1 group deserve particular attention. Interrater agreement measured by intraclass correlation is very low in the L1 group ($ICC(A, 1) = 0.14$). Analysing the dispersion of the ratings given to this subgroup, it comes out that most of the raters used almost exclusively levels 5 and 6 of the scale. Such a range restriction caused the very low value of the intraclass correlation, despite the substantial agreement among the raters that scored all the L1 texts in the same high levels. This problem does not regard the results for the other indices of Table 6: $r_{WG} = 0.90$; $CV = 8.12\%$; $\widehat{d} = 0.17$ $(1 - \widehat{d} = 0.83)$; $\widehat{d}^* = 0.19$ $(1 - \widehat{d}^* = 0.81)$. that show a very good level of absolute agreement. Finally, the standard deviation of $\widehat{d}^*$ computed on the basis of formula (24) is equal to 0.05. As a consequence, the $(1 - \alpha) = 0.95$ confidence interval using the normal approximation for the total group is [0.15, 0.35] and the error is at most 0.10.

6 Conclusions

In this paper a measure of interrater absolute agreement for ordinal scales is proposed. Such a measure is not affected by restriction of variance problems and does not depend on the choice of a particular null distribution. An unbiased estimator of the proposed measure is introduced and its sampling properties are investigated. In the simulation study confidence intervals for the proposed interrater agreement index are constructed using the normal approximation, the parametric

Table 6 $ICC(A, 1)$ and average of r_{WG}, CV, $\widehat{d}$ and $\widehat{d}^*$ for the comprehensibility subscale in the L1, L2 and the total groups

Group	N	ICC(A, 1)	r_{WG}	CV%	$\widehat{d}$	$\widehat{d}^*$
L1	20	0.14	0.90	8.12	0.17	0.19
L2	20	0.63	0.84	16.20	0.28	0.32
Total	40	0.67	0.87	12.16	0.22	0.25

Journal : Small-ext 10260 Dispatch : 19-11-2020 Pages : 19
Article No. : 551 □ LE □ TYPESET
MS Code : SMAP-D-19-00262R2 ☑ CP ☑ DISK

and nonparametric bootstrap. Furthermore, a pseudo-nonparametric bootstrap taking into account the sampling design is also implemented. As previously stressed, the resampling involves both raters and targets sample. Confidence intervals obtained with the normal approximation seem to perform very well both in terms of coverage probability and computational cost.

Acknowledgements The authors are grateful to the referees for very careful reading of the manuscript and thoughtful comments. We dedicate this paper to the memory of Professor Giovanni Battista Tranquilli for being a source of motivation with his scientific and human support.

References

Allsbrook WC, Mangold KA, Johnson MH, Lane RB, Lane CG, Amin MB (2001) Interobserver reproducibility of Gleason grading of prostatic carcinoma: urologic pathologists. Hum Pathol 32(1):74–80

Billingsley P (1995) Probability and measure, 3rd edn. Wiley, New York

Booth JG, Butler RW, Hall P (1994) Bootstrap methods for finite populations. J Am Stat Assoc 89(428):1282–1289

Bove G, Nuzzo E, Serafini A (2018) Measurement of interrater agreement for the assessment of language proficiency. In: Capecchi S, Di Iorio F, Simone R. ASMOD 2018: proceedings of the advanced statistical modelling for ordinal data conference. Università Federico II di Napoli. FedOAPress, Napoli pp 61–68

Burke MJ, Finkelstein LM, Dusig MS (1999) On average deviation indices for estimating interrater agreement. Organ Res Methods 2:49–68

Cohen A, Doveh E, Eick U (2001) Statistical properties of the r_{wg} index of agreement. Psychol Methods 6(3):297–310

Cumming A, Kantor R, Powers DE (2002) Decision making while rating ESL/EFL writing tasks: a descriptive framework. Mod Lang J 86:67–96

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

Grilli L, Rampichini C (2002) Scomposizione della dispersione per variabili statistiche ordinali [Dispersion decomposition for ordinal variables]. Statistica 62:111–116

Gross S (1980). Median estimation in sample surveys. In: Proceedings of the section on survey research methods. American Statistical Association, pp. 181–184

James LJ, Demaree RG, Wolf G (1984) Estimating within-group interrater reliability with and without response bias. J Appl Psychol 69:85–98

James LJ, Demaree RG, Wolf G (1993) rwg: an assessment of within-group interrater agreement. J Appl Psychol 78:306–309

James LR, Demaree RG, Wolf G (1984) Estimating within-group interrater reliability with and without response bias. J Appl Psychol 69:85–98

Kuiken F, Vedder I (2014) Rating written performance: What do raters do and why? Lang Test 31(3):329–348

Kuiken F, Vedder I (2017) Functional adequacy in L2 writing: towards a new rating scale. Lang Test 34:321–336

LeBreton JM, Burgess JRD, Kaiser RB, Atchley EK, James LR (2003) The restriction of variance hypothesis and interrater reliability and agreement: are ratings from multiple sources really dissimilar? Organ Res Methods 6:80–128

LeBreton JM, James LR, Lindell MK (2005) Recent issues regarding rwg, r^*wg, $rwg(j)$, and $r^*wg(j)$. Organ Res Methods 8(1):128–138

LeBreton JM, Senter JL (2008) Answers to 20 questions about interrater reliability and interrater agreement. Organ Res Methods 11(4):815–852

Leti G (1983) Statistica descrittiva. Il Mulino, Bologna

Lindell MK, Brandt CJ (1997) Measuring interrater agreement for ratings of a single target. Appl Psychol Meas 21:271–278

Lomnicki ZA (1952) The standard error of Gini's mean difference. Ann Math Stat 23(14):635–637

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

McGraw KO, Wong SP (1996) Forming inferences about some intraclass correlation coefficients. Psychol Methods 1:30–46

Nuzzo E, Bove G (2020) Assessing functional adequacy across tasks: a comparison of learners and native speakers' written texts. EuroAm J Appl Linguist Lang, 2. In print

Piccarreta R (2001) A new measure of nominal-ordinal association. J Appl Stat 28(1):107–120

Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for Kernel density estimation. J R Stat Soc Ser B 53:683–690

Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing reliability. Psychol Bull 86:420–428

Thompson I (1991) Foreign accents revisited: factors relating to transfer of accent from the first language to a second language. Lang Speech 24(3):265–272

von Eye A, Mun EY (2005) Analyzing rater agreement. Manifest variable methods. Lawrence Erlbaum Associates, Lawrence Erlbaum Associates, Mahwah, New Jersey

Journal : Small-ext 10260 | Dispatch : 19-11-2020 | Pages : 19
Article No. : 551 | □ LE | □ TYPESET
MS Code : SMAP-D-19-00262R2 | ☑ CP | ☑ DISK

Journal : **10260**
Article : **551**

Author Query Form

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query	Details Required	Author's Response
AQ1	Please check and confirm the author given name and family name is correct. Author [2] Given name [Pier Luigi] Family name [Conti]. Also, kindly confirm the details in the metadata are correct.	