UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

SCUOLA POLITECNICA E DELLE SCIENZE DI BASE

Ph.D Degree in Industrial Engineering

Ph.D Thesis:

Assessing and inferring intra and inter-rater agreement

Thesis Advisor
Prof. Amalia Vanacore

PhD. Candidate:
Maria Sole Pellegrino

Academic Year 2018/2019

Contents

List of Tables

List of Symbols

Symbol	**Meaning**
1-β	statistical power of the conducted hypothesis test
a	acceleration parameter for BCa bootstrap confidence interval
b	bias-correction parameter for BC and BCa bootstrap confidence interval
B	number of bootstrap replications
D_e	disagreement that would be expected by chance in Krippendorff's α among the ratings assigned to the items
d_l	deviation of true score from the mean for item l when modelling the evaluation
D_o	observed disagreement in Krippendorff's α among the ratings assigned to the items
e_{lr}	measurement error component in the evaluation of item l from rater r when modelling the evaluation
G	cumulative distribution function of the bootstrap distribution of the kappa coefficient
H	number of evaluation sessions replicated over time
h	code for indexing the generic session
i	code for indexing the generic classification category of the first session
$\mathrm{I}\left[\cdot\right]$	indicator function
j	code for indexing the generic classification category of the second session
k	number of classification categories of the adopted categorical rating scale
l	code for indexing the generic item
$\mathrm{LB}_{\mathrm{Asymp}}$	lower bound of the asymptotic confidence interval based on the assumption of asymptotic normality
$\mathrm{LB}_{\mathrm{BC}}$	lower bound of the Bias Corrected bootstrap confidence interval
$\mathrm{LB}_{\mathrm{BCa}}$	lower bound of the Bias Corrected and Accelerated bootstrap confidence interval
LB_p	lower bound of the percentile bootstrap confidence interval
$\mathbf{M_w}$	weighted misclassification rate
${}_{metric}\delta^2_{ij}$	distance metric in Krippendorff's α
n	sample size dimension

n_i frequencies for category i in Krippendorff's α

n_{ij} number of items classified into category i during the first session and into category j during the second session

$n_{i\cdot}$ number of items classified into category i during the first session independently of the second session

$n_{\cdot j}$ number of items classified into category j during the second session independently of the first session

n_{ri} number of items that rater r classified into category i

o_{ij} frequencies of values of coincident matrices in Krippendorff's α

p_a proportion of observed agreement among the evaluations provided on a nominal scale by the same rater in two sessions

$p_{a|c}$ proportion of agreement expected by chance for nominal evaluations provided by the same rater in 2 sessions

$p_{a|c}^{AC_1}$ proportion of agreement expected by chance of Gwet's AC$_1$ for nominal evaluations provided by the same rater in 2 sessions

$p_{a|c}^{K}$ proportion of agreement expected by chance of Cohen's K for nominal evaluations provided by the same rater in 2 sessions

$p_{a|c}^{U}$ proportion of agreement expected by chance of Uniform kappa for nominal evaluations provided by the same rater in 2 sessions

$p_{a|c}^{\pi}$ proportion of agreement expected by chance of Scott's π for nominal evaluations provided by the same rater in 2 sessions

p_{a_W} weighted version of p_a

$p_{a|c_W}$ weighted version of $p_{a|c}$

$p_{a|c_W}^{AC_2}$ weighted proportion of agreement expected by chance of Gwet's AC$_2$ for ordinal evaluations provided by the same rater in 2 sessions

$p_{a|c_W}^{K}$ weighted proportion of agreement expected by chance of Cohen's K for ordinal evaluations provided by the same rater in 2 sessions

$p_{a|c_W}^{U}$ weighted proportion of agreement expected by chance of Uniform kappa for ordinal evaluations provided by the same rater in 2 sessions or simultaneously by R raters

$p_{a|c_W}^{\pi}$ weighted proportion of agreement expected by chance of Scott's π for ordinal evaluations provided by the same rater in 2 sessions

$p_{a(R)}$ proportion of observed agreement among the evaluations simultaneously provided on a nominal scale by multiple raters

$p_{a|c(R)}$ proportion of agreement expected by chance among the evaluations simultaneously provided on a nominal scale by multiple raters

$p_{a|c(R)}^{AC_1}$ proportion of agreement expected by chance of AC$_1$ among the evaluations simultaneously provided on a nominal scale by multiple raters

$p_{a|c(R)}^{C}$ proportion of agreement expected by chance of Conger's K among the evaluations simultaneously provided on a nominal scale by multiple raters

$p_{a|c(R)}^{F}$ proportion of agreement expected by chance of Fleiss's K among the evaluations simultaneously provided on a nominal scale by multiple raters

$p_{a|c(R)}^{U}$ proportion of agreement expected by chance of Uniform kappa among the evaluations simultaneously provided on a nominal scale by multiple raters

$p_{a(R)_W}$	weighted version of $p_{a(R)}$
$p_{a\|c(R)_W}$	weighted version of $p_{a\|c(R)}$
$p^{AC_2}_{a\|c(R)}$	weighted version of $p^{AC_1}_{a\|c(R)}$
$p^{C}_{a\|c(R)_W}$	weighted version of $p^{C}_{a\|c(R)}$
$p^{F}_{a\|c(R)_W}$	weighted version of $p^{F}_{a\|c(R)}$
p_{ri}	proportion of items classified into category i by rater r
$\bar{p}_{\cdot i}$	mean value of the proportions p_{ri}
R	number of raters involved/employed in the study
r	code for indexing the generic rater
$\mathbf{r}$	number of Monte Carlo replications
$\mathbf{r}'$	code for indexing the generic Monte Carlo replication
rc_{lr}	interaction effect between rater r and item l
r_i	probability of classifying an item into category i
r_{li}	number of raters who classified item l into category i
S^*	generic bootstrap data set
s_i^2	sample variance of the proportions p_{ri}
s_{ij}^2	weighted variant of s_i^2
$\mathrm{UB}_{\mathrm{Asymp}}$	upper bound of the asymptotic confidence interval based on the assumption of asymptotic normality
$\mathrm{UB}_{\mathrm{BC}}$	upper bound of the Bias Corrected bootstrap confidence interval
$\mathrm{UB}_{\mathrm{BCa}}$	upper bound of the Bias Corrected and Accelerated bootstrap confidence interval
UB_p	upper bound of the percentile bootstrap confidence interval
$X_{\mathbf{r}'}$	benchmark of the generic $\mathbf{r}'$ Monte Carlo data set
Y_{lh}	evaluation provided by a rater to item l during session h
Y_{lr}	evaluation provided to item l by rater r
Y'_{lr}	ranking provided to item l by rater r
Y_{lrh}	evaluation provided to item l by rater r during the h^{th} session
T_w	sum of weights w_{ij} across the cells of the contingency table
w_{ij}	symmetric agreement weight for mismatch of ordinal classification between categories i and j
$w_{\omega\omega'}$	misclassification weight for mismatch of benchmarking classification between agreement categories ω and ω'
z_α	α percentile of the standard normal distribution
α	statistical significance of the conducted hypothesis test
Φ	cumulative distribution function of the normal distribution
κ	symbol of the general kappa-type agreement coefficient for 2 sessions and nominal data
κ_l^j	jackknife (j) estimate of κ obtained deleting item l
$\bar{\kappa}^j$	average out of all n jackknife estimate of κ_l^j
κ_C	tested critical value of κ in the hypothesis test

κ_W	symbol of the general weighted kappa-type agreement coefficient for 2 sessions and ordinal data
$\kappa(S^*)$	κ coefficient of the bootstrap data set $S*$
μ	mean value of the evaluation
π_i	proportion of items classified into category i whatever the session
σ_c^2	variance of the component c_r
σ_d^2	variance of the component d_l
σ_e^2	variance of the component e_{lr}
σ_l^2	variance of the component rc_{lr}
$\hat{\sigma}_\kappa^2$	sample variance of κ
ω	code for indexing the generic agreement category

Glossary

ANOVA	Analysis of Variance
BC	Bias Corrected bootstrap confidence interval
BC$_\mathrm{a}$	Bias Corrected and Accelerated bootstrap confidence interval
ICC	Intraclass Correlation Coefficient
IRR	Inter Rater Reliability
p	percentile bootstrap confidence interval
R&R	Repeatability and Reproducibility

Summary

The research work wants to provide a scientific contribution in the field of subjective decision making since the assessment of the consensus, or equivalently the degree of agreement, among a group of raters as well as between more series of evaluations provided by the same rater, on categorical scales is a subject of both scientific and practical interest. Specifically, the research work focuses on the analysis of measures of agreement commonly adopted for assessing the performance (evaluative abilities) of one or more human raters (i.e. a group of raters) providing subjective evaluations about a given set of items/subjects. This topic is common to many contexts, ranging from medical (diagnosis) to engineering (usability test), industrial (visual inspections) or agribusiness (sensory analysis) contexts.

In the thesis work, the performance of the agreement indexes under study, belonging to the family of the kappa-type agreement coefficients, have been assessed mainly regarding their inferential aspects, focusing the attention on those scenarios with small sample sizes which do not satisfy the asymptotic conditions required for the applicability of the standard inferential methods. Those scenarios have been poorly investigated in the specialized literature, although there is an evident interest in many experimental contexts.

The critical analysis of the specialized literature highlighted two criticisms regarding the adoption of the agreement coefficients: 1) the degree of agreement is generally characterized by a straightforward benchmarking procedure that does not take into account the sampling uncertainty; 2) there is no evidence in the literature of a synthetic index able to assess the performance of a rater and/or of a group of raters in terms of more than one evaluative abilities (for example repeatability and reproducibility).

Regarding the former criticism, an inferential benchmarking procedure based on non parametric confidence intervals, build via bootstrap resampling techniques, has been suggested. The statistical properties of the suggested benchmarking procedure have been investigated via a Monte Carlo simulation study by exploring many scenarios defined by varying: level of agreement, sample size and rating scale dimension. The simulation study has been carried out for different agreement coefficients and building different confidence intervals, in order to provide a comparative analysis of their performances.

Regarding the latter criticism, instead, has been proposed a novel composite index able to assess the rater abilities of providing both repeatable (i.e. stable over time) and reproducible (i.e. consistent over different rating scales) evaluations. The inferential benchmarking procedure has been extended also

to the proposed composite index and their performances have been investigated under different scenarios via a Monte Carlo simulation.
The proposed tools have been successfully applied to two real case studies, about the assessment of university teaching quality and the sensory analysis of some food and beverage products, respectively.

Outline of the thesis

The remainder of the thesis is as follows: Chapter 2 introduces the family of κ-type agreement coefficients and describes the most common coefficients adopted for assessing the degree of intra- and inter-rater agreement when the evaluations are provided either with nominal or ordinal rating scales.
The inferential benchmarking procedures adopted for characterizing the extent of rater agreement are presented in Chapter 3, together with the Monte Carlo simulation study conducted for investigating their statistical properties.
Two real case studies are described in Chapter 4 and finally conclusions are summarized in Chapter 5.
All the published and forthcoming papers are collected in the Appendix.

Chapter 1

A brief overview of agreement models and measures for quantitative data

In many fields, ranging from business and industrial system to medical, social and behavioral contexts, the research is based on data generated by human beings (in the specialized literature also defined observers, raters, assessors or judges, and hereafter referred to as human raters or simply raters) who are asked to judge some objects (or items or subjects). In content analysis, for example, people are employed in the systematic interpretation of textual, visual or audible matter; in industrial contexts the operators classify the production faults into defect types or are involved in pass/fail inspection; whereas in medical sciences they provide clinical diagnosis. Despite the huge involvement of human judgments in research studies, when relying on human raters, researchers must worry about the quality of the data and, specifically, about their reliability because of the common premise that *only reliable raters can provide fair evaluations.*

Three kinds of reliability can be analyzed: precision (or intra-rater reliability or repeatability), reproducibility (or inter-rater reliability) and accuracy. The former refers to rater ability of providing the same evaluations try after try under the same conditions (i.e. in different occasions over time); reproducibility refers to raters' ability of providing the same evaluations, on average, as the others of an homogeneous group of raters; whereas accuracy refers to the closeness of the provided evaluation to the true or accepted value. It is worthy to note that being subjective, rater evaluations lack a gold standard against which to check their accuracy, therefore their reliability is related only to precision and reproducibility.

To deal with these issues, a number of theoretical and methodological approaches have been proposed over the years in different disciplines.

The key to rater reliability is the agreement observed within rater and among independent raters, respectively: the more raters agree on the evaluations (or ratings) they provide, the more comfortable we can be that he/she is precise and that their evaluations are reproducible and exchangeable with those provided by other raters [41] and thus trustworthy. Although precision and reproducibility are equally important, the reproducibility is, by far, the ability most frequently tested.

The currently available methods for the assessment of rater repeatability and reproducibility can be grouped in two main families: index-based approach and model-based approach. The former quantifies the level of repeatability and reproducibility in a single number and does not provide insight into the structure and nature of reliability differences; the latter overcomes this criticism and models the ratings provided by each rater to each item focusing on the association structure between repeated evaluations.

1.1 Model-based approach

Of interest in a model-based approach is the joint distribution of the data (i.e. evaluations provided by each rater) and in particular the association structure between the repeated evaluations since a lack of association implies a poor level of Repeatability and Reproducibility (R&R). Specifically, in a R&R study, the n items are evaluated H times by each of R raters into one of the k classification categories. The data are denoted Y_{lrh}, with l indexing items, r indexing raters, and h indexing evaluation sessions.
Among all the available methods for modelling the data, the main alternatives are latent variable model (e.g. [2, 18]) and log-linear models [1].
Both of them typically model typically model cell counts rather than the individual outcomes of measurements. Particularly, latent variable model assumes the existence of an underlying latent dimension, widely used in the popular latent variable modeling framework when dealing with categorical dependent variables (e.g. [58]); log-linear models, instead, are a class of generalized linear models that describe the means of cell counts in a multidimensional table and the cell counts are treated as independent observations of a Poisson random component.

Recently, instead, De Mast and Van Wieringen suggested to evaluate the R&R of nominal data on the basis of heterogeneous appraisers model [19] and that of ordinal data on the basis of models borrowed from Item Response Theory methodology [51], and particularly using Master's Partial Credit model [53] in the generalized form proposed by Muraki [57], estimating —from the experimental data— the parameters of the model by means of the maximum likelihood method [21]. The approach they proposed models the individual outcomes Y_{lrh} and provides insight into the workings of a rating process, which is vital information for fixing an unreliable rater, and analyses (via both graphic and diagnostic instruments) the nature of the differences among the raters.

1.2 Index-based approach

A widely adopted methodology for the assessment of inter-rater reliability (IRR) was developed by Fisher [26]; the method, based on the analysis of variance (ANOVA), leads to the Intraclass Correlation Coefficient (ICC) which expresses the reliability as a ratio of the variance of interest over the total variance. The ICC, developed to deal with several ratings, has emerged as a universal and widely accepted reliability index [52, 64, 67].

In a typical IRR study, each of a random sample of n items from a population of items is rated independently by R raters belonging to a population of raters. Therefore, among the possible effects are those for the r^{th} rater, for the l^{th} item, for the interaction between rater and item, for the constant level of ratings, and for a random error component. Among the many variants of ICC proposed in the literature over the years, two will be hereafter presented, each corresponding to a study design [4, 66] and thus to a standard ANOVA model:

1. each item is rated by a different set of R raters, randomly selected from a larger population of rater (one-way random effects ANOVA model);
2. each item is rated by the same random sample of R raters selected from a larger population (fully-crossed design; two-way random effects ANOVA model).

One-way random effects ANOVA model

Let Y_{lr} and $Y_{lr'}$ be the ratings provided to item l by rater r and r', respectively; the ICC estimates the correlation between different ratings (Y_{lr} and $Y_{lr'}$) of a single item (l) as:

$$\text{ICC} = \frac{\text{Cov}(Y_{lr}, Y_{lr'})}{\sqrt{\text{Var}(Y_{lr}) \cdot \text{Var}(Y_{lr'})}} \tag{1.1}$$

It is worthy to specify that the ICC is based on the assumption that the ratings (Y) provided from multiple raters for a set of items are composed of a true score component (T) and of a measurement error component (E):

$$\text{Rating=True Score + Measurement Error} \tag{1.2}$$

or in abbreviated symbols:

$$Y = T + E \tag{1.3}$$

This can be rewritten in the form:

$$Y_{lr} = \mu + d_l + e_{lr} \tag{1.4}$$

where μ is the mean of the ratings for variable Y, d_l is the deviation of the true score from the mean for item l, and e_{lr} is the measurement error. It is assumed that the component $d_l \sim N(0;\sigma_d^2), (l = 1, ..., n)$, $e_{lr} \sim N(0;\sigma_e^2), (l = 1, ..., n; r = 1, ..., R)$ and that the d_l are independent of e_{lr}.

The ICC is computed adopting the variances of the components in Eq. 1.4 as follows:

$$\text{ICC} = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_e^2} \tag{1.5}$$

Higher ICC values indicate greater IRR: when there is perfect reliability, the ICC estimate equals 1; vice-versa when the measurement system is no more consistent than chance so that the agreement — and thus the reliability — is random, the coefficient is null. In presence of reliability the coefficients return positive values, whereas negative ICC estimates indicate systematic disagreement, with values less than -1 only when there are three or more raters. In order to qualify the extent of positive reliability, Cicchetti [12] provided four ranges of ICC values corresponding to as many categories of reliability: poor reliability for coefficient values less than 0.4; fair, good and excellent for coefficient values ranging between 0.4 and 0.6, 0.6 and 0.75, 0.75 and 1.00, respectively, as reported in Table 1.1.

Table 1.1. Cicchetti benchmark scale for interpreting reliability values

Coefficient	**Reliability**
ICC < 0.40	Poor
$0.40 <$ ICC ≤ 0.60	Fair
$0.60 <$ ICC ≤ 0.75	Good
$0.75 <$ ICC ≤ 1.00	Excellent

From a computational point of view, the variance components in Eq. 1.5 can be estimated via one-way ANOVA. Denoting by WMS and BMS the within and between groups mean squares, respectively, a biased but consistent estimator of ICC is:

$$\widehat{\mathrm{ICC}} = \frac{\mathrm{BMS} - \mathrm{WMS}}{\mathrm{BMS} + (R-1)\mathrm{WMS}} \tag{1.6}$$

Note that this estimate is only acceptable if the items l are sampled randomly from the population. If this is not the case, σ_p^2 should be estimated from a historical sample.

Two-way ANOVA model

Suppose now that each item is rated by the same random sample of R raters selected from a larger population (fully-crossed designs). A two-way ANOVA model can be used to represent the data Y_{lr} because there is a systematic source of variation between items and raters; the component e_{lr} may also be modelled by revising Eq. 1.4 such that

$$Y_{lr} = \mu + d_l + c_r + rc_{lr} + e_{lr} \tag{1.7}$$

where c_r represents the deviation of the ratings of rater r from the overall mean and rc_{lr} represents the degree to which the rater r departs from his/her usual rating tendencies when confronted to item l (interaction effect). It is assumed that $c_r \sim N(0; \sigma_c^2)$, $rc_{lr} \sim N(0; \sigma_l^2)$ and that all components $rc_{lr}(l = 1, ..., n; r = 1, ..., R)$ are mutually independent. Under model 1.7, ICC is defined as:

$$\mathrm{ICC} = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_e^2 + \sigma_c^2 + \sigma_l^2} \tag{1.8}$$

Gauge Repeatability and Reproducibility study, R&R

The standard methods for the assessment of the reliability of a human rater, who in subjective evaluation systems acts as measurement instrument [6, 59, 60, 61], include also Gauge Repeatability and Reproducibility (Gauge R&R) studies [10, 54, 55, 78]. The model underlying the Gauge R&R equals model 1.4 of the ICC method; the main difference is that ICC expresses the ratio of rating variation and total variation in terms of variances whereas the Gauge R&R in terms of standard deviations or (Pearson) correlations. The rating variation is compared to the total variation (including rating variation), as is done by the Gauge R&R statistic:

$$\text{Gauge R\&R} = \frac{\sigma_e}{\sigma_{total}} \tag{1.9}$$

with $\sigma_{total} = \sqrt{\sigma_d^2 + \sigma_e^2}$.

The standard deviation is split into a component due to the measurement instrument (i.e. human rater; repeatability) and a component due to additional sources of variation (reproducibility). Proportions suggest that the numerator plus its complement add up to the denominator. This holds for variances ($\sigma_d^2 + \sigma_e^2 = \sigma_{total}^2$), but not for standard deviations (in general $\sigma_d + \sigma_e \neq \sigma_{total}$), which makes ICC the more natural choice [79]. Anyway, the ICC and the Gauge R&R are essentially the same:

$$\text{ICC} = 1 - (\text{Gauge R\&R})^2 \tag{1.10}$$

A drawback of the standard ANOVA methods is that despite their popularity, they are not suitable for subjective evaluations which — because of their "qualitative" nature — are mainly expressed on categorical (i.e. nominal or ordinal) rating scales; however ANOVA approach could take great advantage from a recently proposed unifying approach for assessing variation over every scale of measurement [32, 33].

Alternative non-parametric methods for the assessment of rater reliability with categorical data are Kendall's coefficient of concordance W, Goodman and Kuskal's gamma (γ) and Krippendorff's α.

Kendall's W

Kendall's W, a generalization of Spearman's ρ [45] correlation coefficient, is a measure of rank correlation between R rankings. Here, the idea is to transform ratings into rankings, and treat these rankings as though they were on an interval or ratio scale (with equidistant classes), and apply ANOVA-like techniques (such as sums of squares). The coefficient assesses the correlation as the sum of squares of the differences in rank number of R rankings for each item. In this form, it is defined as:

$$W = \frac{\sum_{l=1}^{n} r_l - \frac{1}{2}R(n+1))^2}{\frac{1}{12}R^2(n^3 - n)} \tag{1.11}$$

being $r_l = \sum_{r=1}^{R} Y'_{lr}$ and Y'_{lr} the ranking of item l by rater r.
The main criticism of this coefficient is its analogy to the analysis of variance because the rankings are treated as independent of each other although they are assigned in conjunction with each other.

Goodman and Kuskal's γ

Goodman and Kruskal's γ is based on the notion of concordance, which refers to the extent to which raters are consistent in ordering items relative to each other. Specifically, γ is defined for pairs of ratings (that is ratings provided during two evaluation sessions) and expresses the rater reliability as a difference between the probability of observing a concordant pair of ratings and the probability of observing a discordant pair of ratings. It is defined as:

$$\begin{aligned}\gamma &= \frac{P_c - P_d}{1 - P_{tie}} \\ &= P(\text{concordance}|\text{no ties}) - P(\text{discordance}|\text{no ties}).\end{aligned} \tag{1.12}$$

Following the formulas in Goodman and Kruskal [35], γ is estimated as follows:

$$\hat{\gamma} = \frac{\hat{C} - \hat{D}}{\hat{C} + \hat{D}} \tag{1.13}$$

where

$$\hat{C} = 2\frac{1}{n^2}\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}\sum_{i'=i+1}^{k}\sum_{j'=j+1}^{k} n_{ij}n_{i'j'} \tag{1.14}$$

$$\hat{D} = 2\frac{1}{n^2}\sum_{i=1}^{k-1}\sum_{j=2}^{k}\sum_{i'=i+1}^{k}\sum_{j'=1}^{j-1} n_{ij}n_{i'j'} \tag{1.15}$$

and n_{ij} counts the items assigned to category i during one evaluation session and to category j during a second evaluation session on a rating scale with k classification categories.
Using Eq. 1.13, it is possible to estimate $\hat{\gamma}$ (for a fixed couple of raters) for each pair of evaluation sessions and taking the average of the $\hat{\gamma}$-values as the IRR level. A value of $\hat{\gamma} = 1$ implies perfect consistency in order and thus concordance, whereas a value of $\hat{\gamma} = 0$ means that the ratings are done at random and hence they are completely uninformative.

Krippendorff's α

Krippendorff's α is a reliability coefficient developed to measure agreement between raters, or generally between measuring instruments [47]. α emerged in content analysis but is widely applicable whenever two or more methods of generating data (namely, evaluations provided by human raters or measurements made by measuring instruments) are applied to the same set of items and the question is how much the resulting data can be trusted to represent

something real.
The general form of α is as follows:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1.16}$$

where D_o is the observed disagreement among the ratings assigned to the items:

$$D_o = \frac{1}{2n} \sum_{i=1}^{k} \sum_{j=1}^{k} o_{ij} \; {}_{metric}\delta_{ij}^2 \tag{1.17}$$

and D_e denotes the disagreement that would be expected when the rating is attributable to chance rather than to the properties of the rated items:

$$D_e = \frac{1}{2n(2n-1)} \sum_{i=1}^{k} \sum_{j=1}^{k} n_i \cdot n_j \; {}_{metric}\delta_{ij}^2. \tag{1.18}$$

The arguments o_{ij}, n_i and n_j in the two disagreement measures in Eq. 1.17 and 1.18 refer to the frequencies of values in coincidence matrices whereas ${}_{metric}\delta_{ij}^2$ is a metric function. See [49] for all computational details.
If the observed disagreement D_o is null, $\alpha = 1$ and thus the reliability will be perfect. When raters agree by chance then $D_o = D_e$ and $\alpha = 0$, which indicate the absence of reliability.
The Krippendorff's α can be considered a good index of IRR because when assessing the agreement between more than two raters, independently evaluating each item, it is both independent of the number of employed raters and invariant to the permutation and selective participation of raters [42].

Another widely applied method to assess the precision of categorical evaluations is related to the concept of agreement. Specifically, according to this approach, the rater precision is assessed in terms of the similarity of the provided evaluations with respect to the true value of the rated items; otherwise, namely when the true value does not exist and it is not possible to check ratings' trueness, the rater precision is assessed as her/his ability of providing repeatable evaluations over different occasions (i.e. intra-rater agreement) and/or as the raters' ability of providing the same evaluations, on average, as the whole group of raters (i.e. inter-rater agreement); in such circumstances, the whole group of raters plays the role of a gold standard.
Among the several agreement indexes proposed in the literature over the years, the most common ones belong to the family of the kappa-type agreement coefficients. Typically the kappa coefficient, denoted κ, is defined as a population parameter, and the statistical models for the data are not provided. The focus of this thesis is on the index-based approach and specifically on the family of the most widely applied κ coefficients that will be widely discussed in the following Chapter 2.

It is evident that the index-based approach describes the rater reliability in a single number generally ranging between -1 and 1. Whereas the extreme values of disagreement/unreliability and perfect agreement/reliability have clear interpretations, the intermediate values are hard to give a tangible meaning different from *the rater is somewhere in between unreliable and perfectly reliable.* Such a single number may be useful for comparing raters (or generally measurement instruments) relative to each other, but it is hard to see its practical value when the aim is to characterize the evaluation performance of a single rater (or generally measurement instrument).

Although the model-based approach gives more information than the single estimate, output of the index-based approach, this latter is the easiest to implement and thus the most widely applied, especially by practitioners.

Chapter 2

Kappa-type agreement coefficients

The kappa-type agreement coefficients are sample statistics which rescale the difference between the proportion of observed agreement and the proportion of agreement expected by chance alone.

A typical agreement study involves n items classified H times (i.e. H evaluation sessions) by one or more raters on a categorical scale with $k \geq 2$ classification categories.
In the case of one rater classifying the n items in two evaluation sessions, the kappa-type coefficient estimates the level of intra-rater agreement (also defined precision or repeatability); whereas when two or more raters simultaneously (i.e. $H = 1$ evaluation session) classify the n items, the coefficient estimates the level of inter-rater agreement (also defined reproducibility). The two study designs will be separately analysed in Section 2.1 and 2.2 for intra- and inter-rater agreement, respectively.

2.1 Intra-rater agreement

Let Y_{lh} denote the evaluation provided by one rater during the h^{th} evaluation session to item l. Of interest for the assessment of intra-rater agreement is the joint distribution of the Y_{lh}.
In the simplest case of two evaluation sessions (i.e. $h = 1, 2$), the data can be arranged in a $k \times k$ contingency table $(n_{ij})_{k \times k}$ (Table 2.1), where the generic (i, j) cell contains the joint frequency n_{ij} that counts the number of items classified into i^{th} category in the first session and into j^{th} category in the second session. Specifically the cells along the main diagonal represent the perfect match between the evaluations provided in the two sessions, whereas the off-diagonal cells represent mismatch.

The traditional formula of kappa-type agreement coefficient is:

$$\kappa = \frac{p_a - p_{a|c}}{1 - p_{a|c}} \tag{2.1}$$

Table 2.1. $k \times k$ contingency table

		2^{nd} session					
	Category	1	...	j	...	k	**Total**
1^{st} session	1	n_{11}	...	n_{1j}	...	n_{1k}	$n_{1\cdot}$
	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
	i	n_{i1}	...	n_{ij}	...	n_{ik}	$n_{i\cdot}$
	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
	k	n_{k1}	...	n_{kj}	...	n_{kk}	$n_{k\cdot}$
	Total	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot k}$	n

where p_a is the proportion of agreement observed among the provided evaluations and $p_{a|c}$ is the proportion of agreement expected by chance alone, corresponding to the agreement observed when raters assign evaluations randomly, independently of the true value of the rated item.

2.1.1 Agreement coefficients for nominal data

When the n items are evaluated on a nominal scale with k classification categories, independent, mutually exclusive and exhaustive, the proportion of observed agreement is given by:

$$p_a = \sum_{i=1}^{k} \frac{n_{ii}}{n} \tag{2.2}$$

On the other hand, different notions of agreement expected by chance alone are advocated in the literature, corresponding to as many kappa-type coefficients, leading to quite some controversy.

Scott's π

The pioneer coefficient of the kappa-type family coefficients has been proposed by Scott in 1955 [63] in the context of content analysis.

According to Scott, the proportion of agreement expected by chance alone depends not only on the number of classification categories but also on the frequency with which each of them is used in the two sessions (i.e. marginal frequencies). Minimum chance agreement occurs when all categories are used (in the two replications) with equal frequency. Any deviation from a rectangular distribution of frequencies across categories will increase the agreement expected by chance. Moreover, he assumes that the distribution of the marginal frequencies (known for the population) is equal for the two replications. Thus, the total probability of agreement expected by chance alone is the sum over all categories of the squared (since the categories are mutually exclusive and the probabilities of using any one of the categories are assumed equal in the two replications) proportion π_i:

$$p^{\pi}_{a|c} = \sum_{i=1}^{k} \pi_i^2 \tag{2.3}$$

Being $n_{i\cdot}$ the number of items classified into i^{th} category in the first replication (independently of the evaluations of the second replication) and $n_{\cdot j}$ the number of items classified into j^{th} category in the second replication (independently of the evaluations of the first replication), π_i is the proportion of items classified into i^{th} category (independently of the replication) and is equal to:

$$\pi_i = \frac{n_{i\cdot} + n_{\cdot i}}{2n} \tag{2.4}$$

Cohen's K

However, the family of the kappa-type coefficients borrowed its name from the Cohen's Kappa coefficient, proposed by Cohen in 1960 [15]. Cohen formulates the proportion of agreement expected by chance alone assuming that:

1. the n rated items are independent to each other;
2. the k classification categories are independent, mutually exclusive and exhaustive;
3. the two series of evaluations are independent to each other.

The $p_{a|c}$ is thus formulated in terms of marginal frequencies as follows:

$$p^{K}_{a|c} = \frac{1}{n^2} \sum_{i=1}^{k} n_{i\cdot} n_{\cdot i} \tag{2.5}$$

Despite their popularity, two well-documented effects can substantially cause Scott's π and Cohen's Kappa to misrepresent the degree of agreement of a measurement system [24, 37].

The first effect appears when the marginal distributions of observed ratings fall under one classification category at a much higher rate over another, called the *prevalence problem*, which typically causes κ estimates to be unrepresentatively low; in other words, for a fixed value of observed agreement, tables with marginal asymmetry produce lower values of κ than tables with homogeneous marginal. Prevalence problems may exist within a set of ratings due to the nature of the rating instrument used in a study, the tendency for raters to identify one or more classification categories more often than others, or due to truly unequal frequencies of items/events occurring within the population under study.

The second effect appears when the marginal distributions of specific ratings are substantially different between raters, called the *bias problem*, which typically causes κ estimates to be unrepresentatively high.

These criticisms were firstly observed by Brennan and Prediger in 1981 [9] although they are widely known as "Kappa paradoxes" as referred to by Feinstein and Cicchetti [14, 25].

Two κ variants developed to accommodate these effects are the uniform kappa, proposed by Bennett, Alpert and Goldstein [7] and advocated by Brennan and Prediger [9] and others [36, 43, 44], and the Agreement Coefficient AC_1 proposed by Gwet [38].

Uniform kappa

The uniform kappa formulates the agreement expected by chance alone adopting the notion of uniform chance measurement [7] which assigns equal probability to any classification category and thus is the most non-informative measurement system given a certain rating scale [20, 23]. $p_{a|c}$ under the assumption of uniform chance measurement is formulated as follows:

$$p_{a|c}^{U} = \sum_{i=1}^{k} \frac{1}{k^2} = \frac{1}{k} \tag{2.6}$$

Gwet's AC_1

The AC_1 agreement coefficient, instead, formulates the agreement expected by chance alone as the probability of the concomitance that a rater performs a random rating R and that the two raters agree G:

$$p_{a|c}^{AC_1} = P(G \cap R) = P(G|R) \cdot P(R) \tag{2.7}$$

where $P(G|R)$ is given assuming the uniform distribution for chance measurements and is formulated as in Eq. 2.6; and $P(R)$ is approximated with a normalized measure of randomness defined by the ratio of the observed variance Var, to the variance expected under the assumption of totally random ratings Var_M:

$$P(R) = \frac{\text{Var}}{\text{Var}_M} = \frac{\sum_{i=1}^{k} \pi_i(1-\pi_i)}{(k-1)/k} \tag{2.8}$$

being π_i the proportion of items classified into i^{th} category formulated as in Eq. 2.4. Thus:

$$p_{a|c}^{AC_1} = \frac{1}{k-1} \sum_{i=1}^{k} \pi_i(1-\pi_i) \tag{2.9}$$

The formulation of observed agreement of Eq. 2.2 and those of agreement expected by chance alone of Eq. 2.3, 2.5, 2.6 and 2.9 are useful only when the rater provides her/his evaluations on a nominal rating scale because these formulations treat all disagreements as homogeneous and there is agreement only in the case of perfect match between evaluations, being the nominal classification categories mutually exclusive and exhaustive.

2.1.2 Weighted agreement coefficients for ordinal data

When the evaluations are provided on an ordinal rating scale it is undoubtful that some disagreements are more serious than others. In this case, the introduction of either a distance metric or a weighting scheme enables to account that disagreement on two distant categories should be considered more relevant than disagreement on neighbouring categories.
Different kinds of distance metrics and weighting schemes appropriate for various practical situations have been proposed and discussed in the literature. Typically, these metrics are expressed as non decreasing functions of $|i-j|$ when assessing the degree of disagreement among the provided evaluations (e.g. loss matrix [5]) or, vice-versa, as non-increasing function of $|i-j|$ when assessing the degree of agreement (e.g. linear agreeing weights and quadratic agreeing weights).

The weighted version of the kappa-type coefficients, including symmetric weights (i.e. $w_{ij} = w_{ji}$) a priori assigned to each pair (i, j) of ratings, is formulated as follows:

$$\kappa_W = \frac{p_{a_W} - p_{a|c_W}}{1 - p_{a|c_W}} \tag{2.10}$$

where p_{a_W} is the weighted proportion of observed agreement and is given by:

$$p_{a_W} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\frac{n_{ij}}{n} \tag{2.11}$$

and $p_{a|c_W}$ is the weighted proportion of agreement expected by chance, differently formulated for each κ_w coefficient.

w_{ij} is the symmetrical agreement weight ranging between 0 and 1, with the minimum value 0 assigned to maximally disagreeing pairs of ratings, (1, k) and (k, 1), and the maximum value 1 assigned to pairs of concident ratings (i, i). It is worthwhile to pinpoint that although the weights can be arbitrary defined, the linear (w^L_{ij}) [13] and quadratic (w^Q_{ij}) [29] weights are the most commonly used weighting schemes for kappa-type coefficients and are formulated as follows:

$$w^L_{ij} = 1 - \frac{|i-j|}{k-1}; \quad w^Q_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2} \tag{2.12}$$

Weighted Scott's π

The weighted variant of Scott's π formulates the weighted proportion of agreement expected by chance as:

$$p^{\pi}_{a|c_W} = \sum_{i=1}^{k}\sum_{j=1}^{k} \pi_i\pi_j w_{ij} \tag{2.13}$$

Weighted Cohen's K

The Cohen's weighted Kappa [16] formulates the weighted proportion of agreement expected by chance as:

$$p^{K}_{a|c_W} = \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}w_{ij} \tag{2.14}$$

Weighted Uniform kappa

The weighted proportion of agreement expected under the assumption of uniform chance measurement is given by:

$$p^{U}_{a|c_W} = \frac{T_w}{k^2} \tag{2.15}$$

where T_w is the sum of weights across the cells of the contingency table (Table 2.1):

$$T_w = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} \tag{2.16}$$

Gwet's AC$_2$

Gwet, instead, proposed the weighted variant of the AC$_1$ coefficient in [39] and named it AC$_2$. AC$_2$ formulates the weighted proportion of agreement expected by chance as:

$$p^{AC_2}_{a|c} = \frac{T_w}{k(k-1)}\sum_{i=1}^{k}\pi_i(1-\pi_i) \tag{2.17}$$

It is worthy to note that the coefficients proposed in this Section 2.1 can be adopted also for assessing the degree of inter-rater agreement between $R = 2$ raters. In this case, the rows of the contingency table 2.1 refer to the first rater, whereas its columns to the second rater.

2.2 Inter-rater agreement

Kappa-type coefficients were extended to the general case of three raters or more by a number of authors. Most generalized versions are formulated as in Eq. 2.1, where p_a is the observed agreement, and $p_{a|c}$ the agreement expected by chance.
In the case of $R \geq 3$ raters assessing the same set of n items in the same session, the data can be arranged into a $n \times k$ table $(r_{li})_{n\times k}$, where the generic (l, i) cell contains the number of raters r_{li} who classified item l into category i (Table 2.2).

While the majority of kappa-type agreement coefficients, generalized for at least 3 raters, share the same formulation of observed agreement, they still differ on their expression used to compute the agreement expected by chance.

Table 2.2. $n \times k$ table for classifying the ratings provided by R raters in the same session

		Category					
		1	...	i	...	k	**Total**
Product	1	r_{11}	...	r_{1i}	...	r_{1k}	R
	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	R
	l	r_{l1}	...	r_{li}	...	r_{lk}	R
	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	R
	n	r_{n1}	...	r_{ni}	...	r_{nk}	R

2.2.1 Inter-rater agreement coefficients for nominal data

The proportion of observed agreement among the R raters, assessing n items on a nominal scale with k classification categories, is given by:

$$p_{a(R)} = \frac{1}{n} \cdot \frac{1}{R(R-1)} \sum_{l=1}^{n} \sum_{i=1}^{k} r_{li} \left(r_{li} - 1 \right) \tag{2.18}$$

and represents the average of all $R(R-1)$ pairwise agreement percentages.

Fleiss's kappa

To define the multiple-rater version of the proportion of agreement expected by chance of Eq. 2.5, Fleiss in 1971 [27] assumed that the knowledge of the ratings from one rater does not affect those of the others (i.e. the classification of an item into a category is a random process). Under this assumption of independence, he defined the agreement expected by chance as the probability that any pair of raters classify an item into the same category. Specifically, assuming that a randomly selected rater —selected randomly with replacement from the population of R raters— classifies a randomly selected item —randomly selected from the population of n items—, the proportion of agreement expected by chance is defined as follows:

$$p^{F}_{a|c(R)} = \sum_{i=1}^{k} r_i^2 \tag{2.19}$$

where r_i is the estimate of the probability of classifying an item into i^{th} category and is given by:

$$r_i = \frac{1}{n} \cdot \sum_{l=1}^{n} \frac{r_{li}}{R} \tag{2.20}$$

Conger's kappa

The Fleiss's generalized kappa has been criticized by Conger because it does not reduce to Cohen's Kappa when the raters are two. To resolve this generalization problem, Conger in 1980 [17] suggested to estimate the proportion of

agreement expected by chance among R raters by averaging all $R(R-1)/2$ Cohen's Kappa pairwise chance agreement estimates (Eq. 2.5). However, Fleiss and Conger's coefficients get closer as the number of raters increases.
The merit of Conger's coefficient is to be a more natural extension of Cohen's Kappa to the case of al least three raters solving the generalization problem of Fleiss' kappa, but on the other hand averaging all pairwise chance agreement becomes time-consuming when $R \geq 3$.

Fortunately, an alternative method more efficient and with direct calculation exits. Let n_{ri} be the number of items that rater r classifies into category i, $p_{ri} = n_{ri}/n$ the relative proportion and s_i^2 the sample variance of the R proportions $p_{1i}, ..., p_{Ri}$. This variance is given by:

$$s_i^2 = \frac{1}{R-1} \sum_{r=1}^{R} (p_{ri} - \bar{p}_{\cdot i}) \tag{2.21}$$

where $\bar{p}_{\cdot i}$ is the mean value of the proportions $p_{1i}, ..., p_{Ri}$: $\bar{p}_{\cdot i} = 1/R \sum_{r=1}^{R} p_{ri}$. The multiple-rater proportion of agreement expected by chance of Conger's kappa is given by:

$$p^{C}_{a|c(R)} = \sum_{i=1}^{k} \bar{p}_{\cdot i}^2 - \sum_{i=1}^{k} \frac{s_i^2}{R} \tag{2.22}$$

Uniform kappa

The uniform kappa can be easily generalized to the case of multiple raters; the agreement expected by chance among R raters under the assumption of uniform chance measurement is still formulated as in Eq. 2.6, since it does not depend on the subjective evaluations provided by the raters but only on the rating scale dimension:

$$p^{U}_{a|c(R)} = \sum_{i=1}^{k} \frac{1}{k^2} = \frac{1}{k} \tag{2.23}$$

Gwet's AC$_1$

A natural way for generalizing the proportion of agreement expected by chance to the case of three raters or more consists of replacing in Eq. 2.9 the proportion of items classified into i^{th} category with the corresponding value for multiple raters. Using the approach already proposed by Fleiss [27], $p^{AC_1}_{a|c(R)}$ is estimated as follows:

$$p^{AC_1}_{a|c(R)} = \frac{1}{k-1} \sum_{i=1}^{k} r_i (1 - r_i) \tag{2.24}$$

where r_i is the estimate of the probability of classifying an item into i^{th} category (Eq. 2.20).

2.2.2 Weighted inter-rater agreement coefficients for ordinal data

As for intra-rater agreement, in the case of ordinal classifications the degree of inter-rater agreement can be assessed adopting the weighted versions of the coefficients. For a given set of evaluations, the final level of agreement is determined not only by looking at the number of raters who classify item l into category i, but also by looking at the other categories j that represent partial agreement with the category i: two evaluations of the same item are in partial agreement when the weight corresponding to the categories i and j is nonzero. The weighted proportion of observed agreement among the evaluations provided by at least 3 raters, common to all the analysed multiple-raters kappa-type coefficients, is given as:

$$p_{a(R)_W} = \frac{1}{n} \cdot \sum_{l=1}^{n} \sum_{i=1}^{k} \frac{r_{li} \left(\sum_{j=1}^{k} r_{lj} w_{ij} - 1 \right)}{R(R-1)} \tag{2.25}$$

Weighted Fleiss's kappa

The weighted variant of Fleiss' kappa is obtained by computing the weighted proportion of agreement expected by chance as follows:

$$p^F_{a|c(R)_W} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} r_i r_j \tag{2.26}$$

Weighted Conger's kappa

The weighted proportion of agreement expected by chance among R raters proposed by Conger is given by:

$$p^C_{a|c(R)_W} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} \left(\bar{p}_{\cdot i} \bar{p}_{\cdot j} - \frac{s^2_{ij}}{R} \right) \tag{2.27}$$

where

$$s^2_{ij} = \frac{1}{R-1} \left(\sum_{r=1}^{R} p_{ri} p_{rj} - R \bar{p}_{\cdot i} \bar{p}_{\cdot j} \right) \tag{2.28}$$

Weighted Uniform kappa

The weighted variant of the uniform kappa for multiple raters, instead, assesses the weighted observed proportion of agreement with Eq. 2.25 and the weighted proportion of agreement expected by chance under the assumption of uniform chance measurement with Eq. 2.16.

Gwet's AC$_2$

The generalized version of AC$_2$ for the case of multiple raters formulates the proportion of agreement expected by chance as:

$$p^{AC_2}_{a|c(R)} = \frac{T_w}{k(k-1)} \sum_{i=1}^{k} r_i (1 - r_i) \tag{2.29}$$

It is worthy to note that the Table 2.2 could be adopted for arranging the evaluations provided by the same rater in more than two (i.e. $H = R$) different evaluation sessions and the corresponding degree of intra-rater agreement can be assessed adopting one of the coefficients presented in this Section 2.2.

Chapter 3

Characterization of the extent of rater agreement

The different notions adopted for defining the agreement expected by chance lead to as many agreement coefficients. The point now is to interpret the meaning of this empirical number (estimated value of agreement) and to understand its real value. An intra- or inter-rater agreement coefficient is useful only if it is possible to interpret its magnitude: even though the coefficient quantifies the extent of rater/s agreement, this estimate does not tell how valuable that information is.
It is thus clear that a rule of thumb is needed to characterize the extent of agreement, that is to relate the magnitude of the estimated coefficient to the notion of extent of agreement.

3.1 Straightforward benchmarking procedure

Practitioners look for a threshold value for κ, beyond which the extent of agreement can be considered "good". This process of comparing the estimated coefficient to a predetermined threshold value for deciding whether the extent of agreement is good or bad is called *Benchmarking*.

Many benchmark scales have been proposed in the literature over the years. According to Hartmann [40], for example, acceptable values for κ should exceed 0.6 (Table 3.1(a)). The most common benchmark scale, instead, has been proposed by Landis and Koch in 1977 [50]; the Landis and Koch's scale consists in six ranges of κ values corresponding to as many categories of agreement: Poor agreement for coefficient values less than 0, Slight, Fair, Moderate, Substantial and Almost perfect agreement for coefficient values ranging between 0 and 0.2, 0.2 and 0.4, 0.4 and 0.6, 0.6 and 0.8 and 0.8 and 1.0, respectively (Table 3.1(b)). This scale was simplified by Fleiss [31] and Altman [3], with three (Table 3.1(c)) and five (Table 3.1(d)) ranges, respectively, and by Shrout [65] who collapsed the first three ranges of values into two agreement categories (Table 3.1(e)). Munoz and Bangdiwala [56], instead, proposed guidelines for interpreting the values of a kappa-type agreement coefficient with respect to the raw proportion of agreement (Table 3.1(f)).

The above described benchmarking procedure is straightforward since the

Table 3.1. Some benchmark scales for interpreting kappa-type agreement coefficients

(a) Hartmann

Coefficient	**Agreement**
$\kappa > 0.6$	Good

(b) Landis and Koch

Coefficient	**Agreement**
$\kappa \leq 0.0$	Poor
$0.00 < \kappa \leq 0.20$	Slight
$0.20 < \kappa \leq 0.40$	Fair
$0.40 < \kappa \leq 0.60$	Moderate
$0.60 < \kappa \leq 0.80$	Substantial
$0.80 < \kappa \leq 1.00$	Almost perfect

(c) Fleiss

Coefficient	**Agreement**
$\kappa \leq 0.40$	Poor
$0.40 < \kappa \leq 0.75$	Intermediate to Good
$0.75 < \kappa \leq 1.00$	Excellent

(d) Altman

Coefficient	**Agreement**
$\kappa \leq 0.20$	Poor
$0.20 < \kappa \leq 0.40$	Fair
$0.40 < \kappa \leq 0.60$	Moderate
$0.60 < \kappa \leq 0.80$	Good
$0.80 < \kappa \leq 1.00$	Very good

(e) Shrout

Coefficient	**Agreement**
$0.00 < \kappa \leq 0.10$	Virtually none
$0.10 < \kappa \leq 0.40$	Slight
$0.40 < \kappa \leq 0.60$	Fair
$0.60 < \kappa \leq 0.80$	Moderate
$0.80 < \kappa \leq 1.00$	Substantial

(f) Munoz and Bengdiwala

Coefficient	**Agreement**
$\kappa \leq 0.00$	Poor
$0.00 < \kappa \leq 0.20$	Fair
$0.20 < \kappa \leq 0.45$	Moderate
$0.45 < \kappa \leq 0.75$	Substantial
$0.75 < \kappa < 1.00$	Almost perfect
$\kappa = 1.00$	Perfect

agreement is qualified, whatever the benchmark scale that best fits the aims of the study, according to the range of values where the estimated coefficient falls.

Although commonly adopted by practitioners, this straightforward benchmarking procedure relies on the limited information provided by the estimated agreement coefficient, a single summary measure of agreement; it is thus evident that the straightforward benchmarking can be misleading for two main reasons:

- it fails to consider that an agreement coefficient, as any other sampling estimate, is imprecise (i.e. the sample statistic is affected by sampling uncertainty): almost certainly a different agreement estimate will be obtained if the study is repeated under identical conditions on different samples drawn from the same population of items [34];
- it does not allow to compare the extent of agreement across different studies, unless they are carried out under the same experimental conditions (i.e. number of rated items, number of classification categories or distribution of items across the categories).

3.2 Inferential benchmarking procedure

In order to overcome these criticisms and take into account also the uncertainty due to sampling process, researchers (e.g. [48, 62, 80]) recommend to supplement the agreement coefficient with information on statistical uncertainty and suggest the use of the lower confidence bound for agreement benchmarking purpose, that is to test for significance the magnitude of the κ coefficient against desirable levels of agreement.

3.2.1 Standard methods

The standard methods for building confidence interval for kappa-type coefficients (e.g. [8, 30]) are generally based on the assumption of asympotic normality and require large sample sizes of more than 50 items [28]. Assuming the asymptotic normal approximation, the lower and upper bounds of the two-sided $(1-2\alpha)\%$ two-sided confidence interval for κ are given by:

$$\mathrm{LB}_{\mathrm{Asym}} = \kappa - z_\alpha \hat{\sigma}_\kappa \tag{3.1a}$$

$$\mathrm{UB}_{\mathrm{Asym}} = \kappa + z_\alpha \hat{\sigma}_\kappa \tag{3.1b}$$

where z_α is the α percentile of the standard normal distribution and $\hat{\sigma}_\kappa$ the sample standard error of the κ coefficient.

These methods are commonly adopted in many research studies where large samples of items are easily obtainable; but when the asymptotic conditions cannot be reached — as in the most affordable agreement studies — they are not the methods of choice because of their poor performance with small (of no more than 30 items) or moderate (approximately 50 or less) samples [46].

3.2.2 Non-parametric methods based on bootstrap resampling

An alternative to the standard method for building confidence intervals, that picks up where the former leaves off, is the bootstrapping resampling technique, which leads to build non-parametric confidence intervals, independent of the assumption of asymptotic normality and suitable for both small and large samples. The only assumptions required for use of the non-parametric bootstrap resampling are that the data are independent and identically distributed and governed by an unknown cumulative distribution function [11].

Among the available methods to build bootstrap confidence intervals [11, 22], the percentile bootstrap is surely the simplest and the most popular one. It uses the α and $1-\alpha$ percentiles of the bootstrap distribution as the lower (LB_p) and upper (UB_p) bounds of the $(1-2\alpha)\%$ two-sided p bootstrap confidence interval:

$$\mathrm{LB}_P = G^{-1}(\alpha) \tag{3.2a}$$

$$\mathrm{UB}_P = G^{-1}(1-\alpha) \tag{3.2b}$$

where G is the cumulative distribution function of the bootstrap distribution of the κ coefficient.

Another method based on bootstrap resampling is the Bias Corrected or BC method that adjusts for any bias in the distribution through the bias-correction parameter b. The lower and upper bounds of the BC bootstrap confidence interval are:

$$\mathrm{LB}_{\mathrm{BC}} = G^{-1}(\Phi(-2b - z_\alpha)) \tag{3.3a}$$

$$\mathrm{UB}_{\mathrm{BC}} = G^{-1}(\Phi(-2b + z_\alpha)) \tag{3.3b}$$

Specifically, let $S = \left\{ Y_{lrh}, n \right\}$ be the sample of the n evaluations provided for each item; the detailed algorithm for building the $(1-2\alpha)\%$ two-sided BC bootstrap CI for a κ coefficient is:

1. sample n sets of evaluations randomly with replacement from S to obtain a bootstrap data set, denoted S^*;

2. for each bootstrap data set, compute $\kappa(S^*)$ according to the generic Eq. 2.1;

3. repeat B times steps 1 and 2 in order to obtain B estimates $\kappa(S^*)$; count the number of bootstrap estimates $\kappa(S^*)$ that are less than the coefficient value calculated from the original data set. Call this number p and set $b = \Phi^{-1}(p/B)$, being Φ^{-1} the inverse cumulative distribution function of the normal distribution;

4. estimate the lower ($\mathrm{LB}_{\mathrm{BC}}$) and upper ($\mathrm{UB}_{\mathrm{BC}}$) bounds of the two-sided $(1-2\alpha)\%$ two-sided BC bootstrap CI for κ using Eq. 3.3a and 3.3b, respectively.

For severely skewed distribution, instead, the Bias-Corrected and Accelerated bootstrap or BCa confidence interval is recommended, since it adjusts for any bias and lack of symmetry of the bootstrap distribution through the acceleration parameter a and the bias correction parameter b. The lower and upper bound of the $(1-2\alpha)\%$ two-sided BCa confidence-interval are defined as:

$$\mathrm{LB}_{\mathrm{BCa}} = G^{-1}\left(\Phi\left(b - \frac{z_\alpha - b}{1 + a(z_\alpha - b)}\right)\right) \tag{3.4a}$$

$$\mathrm{UB}_{\mathrm{BCa}} = G^{-1}\left(\Phi\left(b + \frac{z_\alpha + b}{1 + a\left(-z_\alpha - b\right)}\right)\right) \tag{3.4b}$$

$\mathrm{LB}_{\mathrm{BCa}}$ and $\mathrm{UB}_{\mathrm{BCa}}$ can be computed as follows:

1. calculate the bias correction parameter b as for BC method (steps 1 through 3);

2. calculate the acceleration parameter a using the jack-knife κ estimates, κ_l^j:

$$a = \frac{\sum_{l=1}^{n}\left(\overline{\kappa^j} - \kappa_l^j\right)^3}{6\Big[\sum_{l=1}^{n}\left(\overline{\kappa^j} - \kappa_l^j\right)^2\Big]^{3/2}} \tag{3.5}$$

   being $\overline{\kappa^j}$ the average out of all n jack-knife estimates κ_l^j.

3. estimate the lower ($\mathrm{LB}_{\mathrm{BCa}}$) and upper ($\mathrm{UB}_{\mathrm{BCa}}$) bounds of the two-sided $(1-2\alpha)\%$ two-sided BCa bootstrap CI for κ using Eq. 3.4a and 3.4b, respectively.

Despite the higher computational complexity, the BCa confidence intervals have generally smaller coverage errors than the others bootstrap intervals, decreasing for $\alpha < 0.025$ as α tends to 0 [11].

3.3 A Monte Carlo simulation study

In order to analyse the statistical properties of the inferential procedure for benchmarking purpose, a Monte Carlo simulation study was developed considering two replications (evaluations provided by the same rater over time in the case of intra-rater agreement or simultaneously by two raters for inter-rater agreement) of n items into one of the k classification categories. Specifically, this research study focused on two bootstrap confidence intervals: the easiest and most common percentile and the more accurate BCa and aims at:

- investigating whether the lower bound of the bootstrap confidence intervals can be effectively used to characterize the extent of agreement with small sample sizes;

- comparing the performance of the proposed benchmarking procedures in order to recommend the method that best fits each specific scenario.

The performances of the benchmarking procedures under comparison were evaluated in terms of weighted misclassification rate (hereafter, $\mathbf{M}_{\mathrm{w}}$) and statistical significance (α) and power ($1-\beta$), computed for the case of null and non-null inference on rater agreement (Table 3.2).

Table 3.2. Null and non null inference cases

Inference case	H_0	H_1
Null	*Chance agreement* $\kappa_W = 0.00$	*Positive agreement* $\kappa_W > 0.00$
Non Null	*No more than Fair* $\kappa_W \leq 0.40$	*Moderate* $\kappa_W > 0.40$
	No more than Moderate $\kappa_W \leq 0.60$	*Substantial* $\kappa_W > 0.60$

The null inference case tests the hypothesis that the rater agreement is positive against the null hypothesis of chance agreement; the non-null inference cases tests the hypothesis that the rater agreement is at least Moderate against the null hypothesis of no more than Fair agreement as well as the hypothesis that the rater agreement is at least Substantial against the null hypothesis of no more than Moderate agreement. Specifically, for the null inference case, 5 alternative hypotheses of positive rater agreement (starting from $\kappa_W = 0.50$ with step size 0.10) were tested against the hypothesis of chance agreement; for the first case of non-null inference, the above 5 alternative hypotheses were tested against the hypothesis of at least Moderate agreement; for the second case of non-null inference, 4 alternative hypotheses (starting from $\kappa_W = 0.70$ with step size 0.10) were tested against the hypothesis of at least Substantial agreement.

Let I $[\cdot]$ be an indicator taking value 1 if the argument is true and 0 otherwise, $\{X_{\mathbf{r}'};\mathbf{r}\}$ be a Monte Carlo data set containing $\mathbf{r}$ benchmarks $X_{\mathbf{r}'}$ obtained for a population value taken as reference for a specific agreement category ω and $w_{\omega\omega'}$ a linear misclassification weight adopted to account that on an ordinal benchmarking scale some misclassifications are more serious than others; the weighted misclassification rate $\mathbf{M}_{\mathrm{w}}$ is evaluated as the weighted proportion of misclassified $X_{\mathbf{r}'}$:

$$\mathbf{M}_{\mathrm{w}} = \frac{1}{\mathbf{r}} \sum_{\omega=1,\Omega} w_{\omega\omega'} \cdot I\left[X_{\mathbf{r}'|\omega} \in \omega'\right]; \quad \omega' \neq \omega \tag{3.6}$$

The Monte Carlo estimate of the statistical significance α, and statistical power $1-\beta$, are respectively given by:

$$\alpha = \frac{1}{\mathbf{r}} \sum_{\mathbf{r}'=1}^{\mathbf{r}} I\Big[LB > \kappa_C |\ \mathrm{H}_0\Big] \tag{3.7}$$

$$1-\beta = \frac{1}{\mathbf{r}} \sum_{\mathbf{r}'=1}^{\mathbf{r}} I\Big[LB > \kappa_C |\ \mathrm{H}_1\Big] \tag{3.8}$$

where LB is the lower bound of the $(1-2\alpha)\%$ two-sided confidence interval obtained from the $\mathbf{r}'$ specific Monte Carlo data set and κ_C is the tested critical value of rater agreement.

The statistical properties of the benchmarking procedure applied to the two paradox-resistant agreement coefficients κ_W^U and AC_2 were studied under several scenarios differing for both sample size and rating scale dimension; for each scenario $\mathbf{r} = 2000$ Monte Carlo data sets were sampled and for each data set the bootstrap confidence intervals were computed on $B = 1500$ bootstrap replications.
The data sets were sampled from a multinomial distribution with parameters n and $\boldsymbol{p} = (\pi_{11}, ..., \pi_{ij}, ..., \pi_{kk})$ where π_{ij} were chosen so as to obtain the desired true population values for the agreement coefficients.
The simulation algorithm was implemented using Mathematica (Version 11.0, Wolfram Research, Inc., Champaign, IL, USA).

The statistical significance and power obtained via Monte Carlo simulation study for the benchmarking procedures based on percentile and BCa bootstrap confidence intervals in the case of $n = 20, 30, 40, 50$ items and $k = 4$ ordinal classification categories applied to the linear weighted Uniform kappa are published in:

> Vanacore A.; Pellegrino M. S.: *Characterizing the extent of rater agreement via a non-parametric benchmarking procedure.* In: Proceedings of the Conference of the Italian Statistical Society SIS 2017. Statistics and Data Science: new challenges, new generations. Firenze University Pres, 2017. p. 999-1004 [70].

The power analysis of the benchmarking procedures based on percentile and BCa bootstrap confidence intervals was then extended to $n = 10, 30, 50, 100$ items, $k = 2, 3, 5, 7$ ordinal classification categories and to the two linear weighted paradox-resistant agreement coefficients, κ_W^U and AC_2 in:

> Vanacore A.; Pellegrino M. S.: *Inferring rater agreement with ordinal classification.* Forthcoming in: Post-conference volume Convegno della Società Italiana di Statistica "New Statistical Developments in Data Science" PROMS (Springer), 2017 [71].

The performances of the benchmarking procedures applied to characterize the extent of rater agreement were evaluated in terms of $\mathbf{M}_{\mathrm{w}}$ and compared each other in the case of $n = 10, 30, 50, 100$ items with $k = 2, 3, 4$ ordinal classification categories in:

> Vanacore A.; Pellegrino M. S.: *Benchmarking rater agreement: probabilistic versus deterministic approach.* Advanced Mathematical and Computational Tools in Metrology and Testing XI, **89**, 365-374, December 2018 [72].

The analysis and comparison in terms of $\mathbf{M}_{\mathrm{w}}$ was then extended to $k = 5, 7$ ordinal classification categories with the same sample sizes in:

> Vanacore A.; Pellegrino M. S.: *A comparative study of benchmarking procedures for interrater and intrarater agreement studies.* In: Proceedings of the 49th Scientific Meeting of the Italian Statistical Society 2018 "Book of short Papers SIS 2018" Pearson [77].

The research study about the man acting as measurement instrument reveals the importance of the employment of reliable raters in order to not compromise the quality of the decision making process.
The uncertainty that does not pertain the inherent variability of the process under study and that arises from imperfect knowledge and/or incomplete information can be reduced by selecting the *right* raters, where *right* means *accurate and precise.*

Anyway, as introduced in Section 1, the subjective evaluations lack a gold standard against which to check their trueness so that only the precision can be assessed.
In "*RRep: A composite index to assess and test rater precision*" is suggested a new composite index to assess rater precision in terms of her/his ability of consistently score the same set of items both in different occasions and using different rating scales; these abilities, respectively defined repeatability over time and reproducibility over scales, are then properly combined in a synthetic index, denoted RRep.
All details as well as the main statistical properties — investigated via a Monte Carlo simulation study — of the proposed RRep index and of the recommended inferential benchmarking procedure, useful for characterizing the extent of rater precision, can be found in [76].

All the published and forthcoming papers are attached in the Appendix.

Chapter 4

Real case studies

The usefulness of the proposed inferential benchmarking procedure in practical situations has been demonstrated by applications to real case studies involving human beings acting as measurement instruments.

The first conducted case study involved classes of more than 20 university students who evaluated the teaching quality of the same university course over three successive academic years (from 2013 to 2016). The students played the role of teaching quality assessors and rated $n = 20$ statements about teaching quality during 3 evaluation sessions adopting 3 different rating scales.
The evaluations simultaneously provided by the whole class of students were used to estimate the level of inter-student agreement; whereas those provided by each student for estimating the level of intra-student agreement. Particularly, the single student's evaluations collected during two successive sessions on the same rating scale were used to assess the student ability of providing stable evaluations in time, whereas those collected during the third session on different rating scales were used to assess the student ability of providing consistent evaluations over scales (i.e. adopting different rating instruments).

Preliminary steps of the analysis regarding the estimated level of intra- and inter-student agreement are presented in:

> Vanacore A.; Pellegrino M. S. (2017, June): *An agreement-based approach for reliability assessment of Students' Evaluations of Teaching.* In: Proceedings of the 3rd International Conference on Higher Education Advances (pp. 1286-1293). Editorial Universitat Politècnica de València [69].

Further analysis supplemented the preliminary steps by inferring the extent of intra-student agreement via benchmarking procedure based on BCa bootstrap confidence intervals. The study results are reported in:

> Vanacore A.; Pellegrino M. S. (2018): *How reliable are Students' Evaluations of Teaching (SETs)? A study to test student's reproducibility and repeatability.* Forthcoming in: Social Indicator Re-

search [73].

In the second case study consumers were employed as sensory panellists and evaluated five sensory dimensions of different food and beverage products in two evaluation sessions. The quality of sensory data were assessed in terms of panelist repeatability and panel reproducibility via the proposed inferential benchmarking procedure.
The case study details and its results are published in:

> Vanacore A.; Pellegrino M. S. (2017, September): *Checking quality of sensory data by assessing intra/inter panelist agreement.* In Proceedings of 8th Scientific Conference on INNOVATION & SOCIETY, Statistical Methods for Evaluation and Quality - IES 2017. p. 1-4 [68].
>
> Vanacore A.; Pellegrino M. S. (2018): *Checking quality of sensory data via an agreement-based approach.* Quality & Quantity, 1–12 [74].

Chapter 5

Conclusions

The research work focused on the assessment of the degree of agreement between series of subjective evaluations provided on ordinal rating scales by the same rater in different occasions (intra-rater agreement) or by different raters in the same occasion (inter-rater agreement).

A short review about approaches proposed in the literature for assessing the degree of inter/intra-rater agreement is provided in the Introduction; however, the main corpus of the thesis work is devoted to the index-based approach for the measurement of rater agreement on categorical scales and particularly to two paradox-resistant coefficients belonging to the family of the kappa-type.

The crucial point is the inferential benchmarking procedure based on non-parametric bootstrap confidence interval, adopted for characterizing the extent of rater agreement. Its statistical properties were investigated via a Monte Carlo simulation study under different scenarios differing from each other in sample size, rating scale dimension and bootstrap confidence interval. The benchmarking procedures were compared in terms of weighted misclassification rate and statistical significance and power, referred to both null and non-null inference cases.

Simulation results reveal that the proposed benchmarking procedures are adequately powered in detecting differences in the extent of rater agreement that are of practical interest for agreement studies. They can be suitably applied for the characterization of the extent of agreement over a small or moderate number of subjective evaluations provided by human raters.

Further analysis regarding the unbiasness and robustness of some kappa-type agreement coefficients were conducted and presented at the 18th Annual Conference of the European Network for Business and Industrial Statistics, ENBIS 2018 [75] and will be published in the future.

Additional research efforts aimed at the development of new tools to estimate and characterise rater precision. A novel composite index, the RRep, was formulated in such a way that both rater abilities of providing evaluations stable over time and consistent over rating scales are accounted for.

The Monte Carlo simulation results, conducted for studying the performances of the index — in terms of percent bias and relative standard deviation —

and those of its inferential procedure — in terms of statistical significance and power —, show that their performance is satisfactory in distinguishing even between adjacent categories of precision.

The usefulness and the effectively applicability in many industrial contexts of both the new RRep index and the proposed inferential benchmarking procedure are worthy to note; as a matter of fact, they were proved to be valid tools for characterizing the extent of agreement and precision, for selecting inspectors able to provide precise diagnosis as well as raters providing precise subjective evaluations, but especially for testing the efficacy of rater training programs.

Appendix A

Published and forthcoming papers:

1. *Characterizing the extent of rater agreement via a non-parametric benchmarking procedure.* In: Proceedings of the Conference of the Italian Statistical Society SIS 2017. Statistics and Data Science: new challenges, new generations. Firenze University Pres, 2017. p. 999-1004 [70];

2. *Inferring rater agreement with ordinal classification.* Forthcoming in: Post-conference volume Convegno della Società Italiana di Statistica "New Statistical Developments in Data Science" PROMS (Springer) 2017 [71];

3. *Benchmarking rater agreement: probabilistic versus deterministic approach.* Advanced Mathematical and Computational Tools in Metrology and Testing XI, **89**, 365-374, December 2018 [72];

4. *A comparative study of benchmarking procedures for interrater and intrarater agreement studies.* In: Proceedings of the 49th Scientific Meeting of the Italian Statistical Society 2018 "Book of short Papers SIS 2018" Pearson [77];

5. *An agreement-based approach for reliability assessment of Students' Evaluations of Teaching.* In: Proceedings of the 3rd International Conference on Higher Education Advances, 1286-1293, june 2017. Editorial Universitat Politècnica de València [69];

6. *How reliable are Students' Evaluations of Teaching (SETs)? A study to test student's reproducibility and repeatability.* Forthcoming in: Social Indicator Research [73];

7. *Checking quality of sensory data by assessing intra/inter panelist agreement.* In Proceedings of 8th Scientific Conference on INNOVATION & SOCIETY, Statistical Methods for Evaluation and Quality - IES 2017. p. 1-4 [68];

8. *Checking quality of sensory data via an agreement-based approach.* Quality & Quantity, 1–12, 2018 [74];

9. *RRep: A composite index to assess and test rater precision.* Quality and Reliability Engineering International, **34**(7), 1352-1362, 2018 [76].

Characterizing the extent of rater agreement via a non-parametric benchmarking procedure

Caratterizzazione del grado di accordo intra/inter-valutatore mediante una procedura non parametrica di benchmark

Amalia Vanacore[1] and Maria Sole Pellegrino[2]

Abstract In several context ranging from medical to social sciences, rater reliability is assessed in terms of intra (-inter) rater agreement. The extent of rater agreement is commonly characterized by comparing the value of the adopted agreement coefficient against a benchmark scale. This *deterministic* approach has been widely criticized since it neglects the influence of experimental conditions on the estimated agreement coefficient. In order to overcome this criticism, in this paper a statistical procedure for benchmarking is presented. The proposed procedure is based on non parametric bootstrap confidence intervals. The statistical properties of the proposed procedure have been studied via a Monte Carlo simulation.

Abstract *In numerosi contesti applicativi, dal medico al sociale, l'affidabilità di un valutatore è valutata in funzione del grado di accordo intra (-inter) valutatore. La caratterizzazione del grado di accordo è tipicamente effettuata confrontando la stima del coefficiente di accordo adottato con una scala di riferimento (benchmark). Questo approccio "deterministico" è stato spesso criticato in letteratura in quanto non tiene in conto l'influenza delle condizioni sperimentali sul processo di stima. In questo lavoro è presentata una procedura di benchmark basata su intervalli di confidenza bootstrap. Le proprietà statistiche della procedura proposta sono state studiate mediante simulazione Monte Carlo.*

Key words: rater reliability, kappa-type agreement index, statistical power, Monte Carlo simulation

[1] Amalia Vanacore, Department of Industrial Engineering, University of Naples Federico II; email: amalia.vanacore@unina.it

[2] Maria Sole Pellegrino, Department of Industrial Engineering, University of Naples Federico II; email: mariasole.pellegrino@unina.it

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

1. Introduction

In many context of research (*e.g.,* cognitive and behavioural science, quality science, clinical epidemiology, diagnostic imaging, content analysis), there is frequently a need to assess the performance of human instruments (*i.e.,* raters) providing subjective measurements, expressed on a dichotomous, nominal or ordinal rating scale. Rater reliability is often evaluated in terms of the extent of agreement between two or more series of ratings provided by two or more raters (inter-rater agreement) or by the same rater in two or more occasions (intra-rater agreement). Specifically, inter-rater agreement is concerned about the reproducibility of measurements provided by different raters, whereas intra-rater agreement is concerned about self-reproducibility (also known as repeatability).

The easiest way of measuring agreement between ratings is to calculate the overall percentage of agreement; nevertheless, this measure does not take into account the agreement that would be expected by chance alone [11]. A reasonable alternative is to adopt the widespread kappa-type index that was introduced by Cohen in 1960 as a rescaled measure of the probability of observed agreement corrected with the probability of agreement expected by chance alone. A main issue for the correct definition of a kappa-type index regards the notion of expected proportion of agreement: chance measurements are conceived as blind (that is, uninformative about the rated items) and any distributional assumption for them is likely to be arbitrary. A solution is to adopt the notion of uniform chance measurement [2] that — given a certain rating scale — can be assumed as a reasonable model for the maximally non-informative measurement system. This uniform version of Kappa is often referred to as Brennan-Prediger coefficient [3].

The extent of a kappa-type index is generally qualified through a benchmark scale [*e.g.* 1, 8, 10]: threshold values against which compare the estimated agreement coefficient for deciding whether the extent of agreement is good or poor. Although commonly adopted, this deterministic benchmarking approach does not consider that the value of the information provided by an agreement coefficient is unknown since, being computed on a sample of items, its estimate is subject to sampling error. In order to identify a suitable neighbourhood of the truth (*i.e.,* the true population value), sampling error has always to be considered.

In this paper a benchmarking procedure based on bootstrap resampling is proposed in order to take into account the sampling uncertainty when characterizing the extent of rater agreement. The main statistical properties of the proposed procedure have been assessed via a Monte Carlo simulation study.

The remainder of this paper is organized as follows: in Section 2 the weighted Brennan-Prediger coefficient is introduced; Sections 3 is devoted to coefficient estimation and inference; in Section 4 the simulation design is described and the main results are discussed; finally, conclusions are summarized in Section 5.

2. Weighted Brennan-Prediger Coefficient

Let n be the number of items rated twice (*i.e.,* two replications) on an ordinal k-points rating scale (with $k > 2$), n_{ij} the number of items classified into i^{th} category in the first replication and into j^{th} category in the second replication and w_{ij} the corresponding symmetrically weight (*i.e.,* $w_{ij} = w_{ji}$) introduced in order to consider that on an ordinal rating scale, some disagreements are more serious than others (*i.e.,* disagreement on two distant categories are more relevant than disagreement on neighbouring categories).

The weighted Brennan-Prediger coefficient [9] is defined as:

$$\hat{K}_W^U = (\hat{p}_a - p_{a|c}^U)\big/(1 - p_{a|c}^U)$$

where $\hat{p}_a = \sum_{i=1}^{k}\sum_{j=1}^{K} w_{ij}\left(n_{ij}/n\right)$ and $p_{a|c}^U = \left(1/k^2\right)\sum_{I=1}^{k}\sum_{j=1}^{k} w_{ij}$.

The $\hat{K}_W^U$ coefficient ranges from -1 to +1 and it can be assumed asymptotically normal distributed [9] with mean K_W^U and variance $\hat{\sigma}^2_{\hat{K}_W^U}$ given by:

$$\hat{\sigma}^2_{\hat{K}_W^U} = \frac{1}{n\left(n-1\right)}\sum_{l=1}^{n} a_h^2 \Big/ \left(1 - p_{a|c}\right)^2 \qquad (2)$$

where h refers to the generic rated item and $a_h = \sum_{i,j=1}^{k} w_{ij}(\delta_{ij}^{(h)} - p_{ij})$ with $\delta_{ij}^{(h)} = 1$ if the rater rated item h into i^{th} and j^{th} category in the first and second replication, respectively, and $\delta_{ij}^{(h)} = 0$ otherwise.

3. Characterization of rater agreement

The approach currently adopted to characterize the extent of agreement is based upon a straight comparison between the estimated coefficient and an adopted benchmark scale. The most widespread benchmark scale for interpreting the magnitude of agreement coefficients was proposed by Landis and Koch [10]. According to this scale, there are 5 categories of agreement corresponding to as many ranges of coefficient values: slight, fair, moderate, substantial and almost perfect agreement for coefficient values ranging between 0 and 0.2, 0.21 and 0.4, 0.41 and 0.6, 0.61 and 0.8 and 0.81 and 1.0, respectively.

Although benchmark scales are widely adopted for relating the magnitude of the coefficient to the notion of extent of agreement, some researchers question their validity and give advice that their uncritical application may lead to practically questionable decisions [11]. Actually, as argued in [9] the choice of the benchmark scale is less important than the way it is used for characterizing the extent of agreement.

A deterministic approach to benchmarking does not account for the influence of experimental conditions on the estimated coefficient and, thus, it does not allow for a statistical characterization of the extent of rater agreement. This criticism may be overcame by benchmarking the lower bound of the confidence interval of the agreement coefficient rather than its point estimate.

Assuming the asymptotic normal approximation, the lower and upper bound of a two-sided $(1-\alpha)\%$ confidence interval for K_W^U are given by:

$$\hat{K}_W^U \pm z_{\alpha/2}\,\hat{\sigma}_{\hat{K}_W^U}$$

The accuracy of the above confidence interval depends on the asymptotic normality of $\hat{K}_W^U$ and on the asymptotic solution for $\hat{\sigma}^2_{\hat{K}_W^U}$ which are both questionable for small sample sizes.

Resampling, which is generally considered the approach of choice when the assumptions of classical statistical methods are not met, may yield more accurate confidence limits and thus it can be usefully adopted to characterize the extent of rater agreement.

Among the available methods to build bootstrap confidence intervals, the percentile bootstrap (hereafter, *p*) is the simplest and the most popular one. The lower and upper bounds of the $(1-\alpha)\%$ two-sided *p* confidence interval are, respectively, the $(\alpha/2)$ and $(1-\alpha/2)$ percentiles of the cumulative distribution function G of the bootstrap replications of $\hat{K}_W^U$. On the other hand the Bias-Corrected and Accelerated bootstrap (hereafter, BCa) confidence interval is recommended for severely non normal data [4, 6]. Despite the high computational complexity needed, BCa confidence intervals have generally smaller coverage errors than the others. The lower and upper bounds of the $(1-\alpha)\%$ two-sided BCa confidence interval are defined as:

$$\mathrm{G}^{-1}\left(\Phi\left(b \pm \left(z_{\alpha/2} \pm b\right)\big/\left[1 + a\left(\mp z_{\alpha/2} - b\right)\right]\right)\right) \tag{4}$$

being Φ the cumulative distribution function of the normal distribution, $z_{\alpha/2}$ the $\alpha/2$ percentile of the standard normal distribution, b the bias correction parameter and a the acceleration parameter.

4. Simulation study

In order to analyse the statistical properties of the proposed benchmarking procedure in terms of Type I error and statistical power, a Monte Carlo simulation study has been developed considering one rater who classifies n items into one of k possible ordinal rating categories. The simulation has been conducted by sampling $r = 2000$ Monte Carlo data sets from a multinomial distribution with parameters n and $\mathbf{p} = (\pi_{11}, ..., \pi_{ij}, .., \pi_{kk})$; the π_{ij} values have been chosen so as to obtain nine true population values of K_W^U (*viz.,* 0, 0.4, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 0.95, 1.0), assuming a linear weighting scheme [4]. The BCa confidence interval for each $\hat{K}_W^U$ has been built on 1500 bootstrap replications. The statistical properties of the benchmarking procedure have been studied for a $k = 4$ points rating scale and for $n = 20, 30, 40, 50$ which are the most affordable sample sizes in many experimental contexts and also the most critical ones for statistical inference.

Simulation results in terms of statistical significance and power are reported in Table 1, organized in four distinct sections each corresponding to a null hypothesis of rater agreement, which is tested against several specific alternative hypotheses.

Table 1: S*tatistical significance (in bold) and power for different true population values of* K_W^U

		n=20		***n=30***		***n=40***		***n=50***	
		p	*BC_a*	*p*	*BC_a*	*p*	*BC_a*	*p*	*BC_a*
$\boldsymbol{K_W^U} = 0.00$	$\boldsymbol{K_W^U = 0.00}$	**0.046**	**0.046**	**0.038**	**0.029**	**0.028**	**0.027**	**0.030**	**0.027**
	$K_W^U = 0.50$	0.645	0.622	0.813	0.768	0.887	0.878	0.950	0.940
	$K_W^U = 0.60$	0.870	0.852	0.972	0.956	0.991	0.991	0.998	0.997
	$K_W^U = 0.70$	0.958	0.946	0.994	0.992	1.000	0.999	1.000	1.000
	$K_W^U = 0.80$	0.997	0.996	0.999	0.999	1.000	1.000	1.000	1.000
	$K_W^U = 0.90$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$\boldsymbol{K_W^U} \leq 0.40$	$\boldsymbol{K_W^U = 0.40}$	**0.043**	**0.034**	**0.045**	**0.034**	**0.039**	**0.032**	**0.033**	**0.023**
	$K_W^U = 0.50$	0.091	0.087	0.124	0.117	0.131	0.123	0.156	0.132
	$K_W^U = 0.60$	0.242	0.233	0.358	0.344	0.379	0.357	0.468	0.407
	$K_W^U = 0.70$	0.460	0.427	0.636	0.612	0.721	0.705	0.828	0.781
	$K_W^U = 0.80$	0.774	0.749	0.915	0.906	0.966	0.958	0.987	0.987
	$K_W^U = 0.90$	0.958	0.933	0.993	0.993	0.999	0.999	1.000	1.000
$\boldsymbol{K_W^U} \leq 0.60$	$\boldsymbol{K_W^U = 0.60}$	**0.058**	**0.055**	**0.045**	**0.043**	**0.043**	**0.037**	**0.033**	**0.032**
	$K_W^U = 0.70$	0.172	0.166	0.184	0.180	0.221	0.189	0.243	0.218
	$K_W^U = 0.80$	0.407	0.393	0.484	0.460	0.573	0.533	0.648	0.648
	$K_W^U = 0.85$	0.694	0.681	0.823	0.806	0.914	0.890	0.953	0.953
	$K_W^U = 0.90$	0.747	0.735	0.870	0.854	0.941	0.916	0.974	0.969
	$K_W^U = 0.95$	0.932	0.936	0.979	0.980	0.995	0.988	1.000	1.000
$\boldsymbol{K_W^U} \leq 0.80$	$\boldsymbol{K_W^U = 0.80}$	**0.140**	**0.125**	**0.069**	**0.078**	**0.064**	**0.060**	**0.061**	**0.055**
	$K_W^U = 0.85$	0.329	0.346	0.246	0.254	0.312	0.241	0.291	0.253
	$K_W^U = 0.90$	0.425	0.452	0.380	0.363	0.451	0.394	0.452	0.405
	$K_W^U = 0.95$	0.716	0.747	0.714	0.682	0.801	0.799	0.834	0.761
	$K_W^U = 1.00$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

As foreseen, the statistical properties of the proposed benchmarking procedure improve as the sample size increases being satisfactory even for relatively small sample size. Specifically, the significance level is generally slightly better for BC$_a$ bootstrap confidence interval; it decreases with increasing sample size but it grows up for increasing true population value of K_W^U; it is close to the nominal level $(\alpha = 0.025)$ only in the case of null rater agreement for $n \geq 40$; however, it is always less than 0.10 except for $n = 20$ when testing an high rater agreement level. The statistical power, instead, is generally slightly better for *p* bootstrap confidence interval; for $n \geq 30$, it is less than 80% only in testing hypotheses referring to adjacent agreement categories (*e.g.,* poor vs slight, moderate vs substantial).

5. Conclusions

The proposed benchmarking procedure can be suitably applied for the characterization of the extent of agreement over a small or moderate number of subjective ratings provided by one or more raters. The procedure shows satisfactory statistical properties in testing both null and non-null cases of rater agreement, being adequately powered in detecting differences in the extent of rater agreement that are of practical interest for agreement studies (*i.e.,* differences of at least 0.2).

It is worthwhile to note that the proposed benchmarking procedure can be also adopted to characterize the extent of inter-rater agreement which, in the case of more than two raters, could be estimated using a suitable variant of kappa coefficient, such as the Fleiss' kappa.

References

1. Altman, D. G.: Practical Statistics for Medical Research. Chapman and Hall (1991)
2. Bennett, E. M., Alpert, R., Goldstein, A. C.: Communications through limited response questioning. Public Opinion Quarterly. 18.3, 303--308 (1954). DOI: 10.1086/266520
3. Brennan, R. L., Prediger, D. J.: Coefficient Kappa: Some Uses, Misuses, and Alternatives. EPM (1981), 41, 687–699
4. Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statist. Med. 19, 1141--1164. (2000),
5. Cicchetti, D. V., Allison, T.: A new procedure for assessing reliability of scoring EEG sleep recordings. Amer. J. EEG Technol. 11.3, 101--110. (1971)
6. Cohen, J.: A coefficient of agreement for nominal scale. EPM. 20.1, 37-46 (1960).
7. Efron, B., Tibshirani, R. J.: An introduction to the bootstrap. CRC press (1994)
8. Fleiss, J. L.: Statistical Methods for Rates and Proportions. John Wiley & Sons (1981)
9. Gwet, K. L.: Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC (2014)
10. Landis, J. R., Koch, G. G.: The measurement of observer agreement for categorical data. Biometrics. 33.1, 159--174 (1977)
11. Sim, J., Wright, C. C.: The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys. Ther. 85.3, 257--268 (2005).

Benchmarking rater agreement: probabilistic versus deterministic approach

Amalia Vanacore[†] and Maria Sole Pellegrino

Department of Industrial Engineering, University of Naples "Federico II", Naples, 80125, Italy
[†]*E-mail: amalia.vanacore@unina.it*

In several industries strategic and operational decisions rely on subjective evaluations provided by raters who are asked to score and/or classify group of items in terms of some technical properties (*e.g.* classification of faulty material by defect type) and/or perception aspects (*e.g.* comfort, quality, pain, pleasure, aesthetics). Because of the lack of a gold standard for classifying subjective evaluations as "true" or "false", rater reliability is generally measured by assessing her/his precision via inter/intra-rater agreement coefficients. Agreement coefficients are useful only if their magnitude can be easily interpreted. A common practice is to apply a straightforward procedure to translate the magnitude of the adopted agreement coefficient into an extent of agreement via a benchmark scale. Many criticisms have been attached to this practice and in order to solve some of them, the adoption of a probabilistic approach to characterize the extent of agreement is recommended. In this study some probabilistic benchmarking procedures are discussed and compared via a wide Monte Carlo simulation study.

Keywords: rater agreement, kappa-type coefficient, probabilistic benchmarking procedures, Monte Carlo simulation

1. Introduction

Agreement coefficients are widely adopted for assessing the precision of subjective evaluations provided by human raters to support strategic and operational decisions in several contexts (*e.g.* manufacturing and service industries, food, healthcare, safety, among many others).

Subjective evaluations are typically provided on categorical rating scale for which the common statistical tools, that work readily for continuous data, are not applicable. For this reason, rater precision is generally assessed in terms of the extent of agreement between two or more series of evaluations on the same sample of items (subjects or objects) provided by two or more raters (inter-rater agreement) or by the same rater in two or more occasions (intra-rater agreement). Specifically, inter-rater agreement is concerned about the reproducibility of measurements by different raters, whereas intra-rater agreement is concerned about self-reproducibility (also known as repeatability).

The degree of inter/intra-rater agreement for categorical rating scale is commonly assessed using kappa-type agreement coefficients that, originally introduced by Cohen [1], are rescaled measures of agreement corrected with the probability of agreement expected by chance alone. It is common practice to qualify the magnitude of a kappa-type agreement coefficient by comparing it against an arbitrary benchmark scale; by applying this straightforward benchmarking, practitioners relate the magnitude of the coefficient to an extent of agreement and then decide whether it is good or poor. Although widely adopted, the straightforward benchmarking has some drawbacks. As demonstrated for example by Thompson and Walter [2] and Gwet [3], the magnitude of an agreement coefficient may strongly depend on some experimental factors such as the number of rated items, rating scale dimension, trait prevalence and marginal probabilities. Thus, interpretation based on the straightforward benchmarking should be treated with caution especially for comparison across studies when the experimental conditions are not the same.

A proper characterization of the extent of rater agreement should rely upon a *probabilistic* benchmarking procedure that allows to identify a suitable neighborhood of the truth (*i.e.* the true value of rater agreement) by taking into account sampling uncertainty.

The most simple and intuitive way to accomplish this task is by building a confidence interval of the agreement coefficient and comparing its lower bound against an adopted benchmark scale.

A different approach to probabilistic benchmarking is the one recently proposed by Gwet [3] which, under the assumption of asymptotically normal distribution, evaluates the likelihood that the estimated agreement coefficient belongs to any given benchmark level.

The above benchmarking approaches will be in the following fully discussed and their performances will be evaluated and compared via a Monte Carlo simulation study with respect to the ability to correctly interpreting the magnitude of the agreement coefficient in terms of weighted misclassification rate.

The paper focuses on agreement on ordinal rating scale, thus in the following we will deal with weighted kappa-type coefficients that allow to consider that disagreement on two distant categories are more serious than disagreement on neighboring categories.

The remainder of the paper is organized as follows: in Section 2 two well-known paradox-resistant kappa-type agreement coefficients are discussed; the commonly adopted benchmark scales are presented in Section 3; four characterization procedures based on a probabilistic approach to benchmarking are discussed in Section 4; in Section 5 the simulation design is described and the main results are fully discussed; finally, conclusions are summarized in Section 6.

2. Weighted Kappa-type agreement coefficients

Let n be the number of items rated by two raters on an ordinal k-points rating scale (with $k > 2$), n_{ij} the number of items classified into i^{th} category by the first rater but into j^{th} category by the second rater and w_{ij} the corresponding symmetrical weight, $n_{i\cdot}$ be the total number of items classified into i^{th} category by the first rater and $n_{\cdot i}$ be the total number of items classified into i^{th} category by the second rater. The weighted Cohen's Kappa coefficient [4] can be computed as:

$$\hat{K}_W = \left(p_a - p_{a|c}\right) \big/ \left(1 - p_{a|c}\right) \tag{1}$$

where

$$p_{aw} = \sum\nolimits_{i=1}^{k} \sum\nolimits_{j=1}^{k} w_{ij}\, n_{ij} \big/ n\,; \quad p_{a|c} = \sum\nolimits_{i=1}^{k} \sum\nolimits_{j=1}^{k} w_{ij} \cdot n_{i\cdot} \big/ n \cdot n_{\cdot j} \big/ n \tag{2}$$

Despite its popularity, researchers have pointed out two main criticisms with Cohen's Kappa: it is affected by the degree to which raters disagree (bias problem); moreover, for a fixed value of observed agreement, tables with marginal asymmetry produce lower values of Kappa than tables with homogeneous marginal (prevalence problem). These criticisms were firstly observed by Brennan and Prediger [5] although they are widely known as "*Kappa paradoxes"* as referred to by Feinstein and Cicchetti [6].

A solution to face the above paradoxes is to adopt the uniform distribution for chance measurements, which — given a certain rating scale — can be defended as representing the maximally non-informative measurement system. The obtained weighted uniform kappa, referred to as Brennan-Prediger coefficient (although proposed also by several other authors [5, 7, 8, 9, 10]), is formulated as:

$$\widehat{\mathrm{BP}}_w = \left(p_{aw} - p_{a|c}^{U}\right) \big/ \left(1 - p_{a|c}^{U}\right) \tag{3}$$

where p_{aw} is defined as in equation (2) and $p_{a|c}^{U} = T_w \big/ k^2$, being T_w the sum over all weight values w_{ij}.

Another well-known paradox-resistant agreement coefficient alternative to Cohen's Kappa is the AC coefficient (proposed by Gwet [11]), whose weighted version (AC_2) is formulated as:

$$\widehat{\mathrm{AC}}_2 = \left(p_{aw} - p_{a|c}^{G}\right) \Big/ \left(1 - p_{a|c}^{G}\right) \tag{4}$$

where p_{aw} is defined as in equation (2) and the probability of chance agreement is defined as the probability of the simultaneous occurrence of random rating (R) by one of the raters and rater agreement (G): $P(G \cap R) = P(G \mid R) \cdot P(R)$. Specifically, $P(R)$ is approximated with a normalized measure of randomness defined by the ratio of the observed variance $\sum_{i=1}^{k} p_i(1-p_i)$ to the variance expected under the assumption of totally random rating $1/(k-1)$; whereas the conditional probabilities of agreement P(G|R) is given by $P(G \mid R) = T_w / k^2$:

$$p_{a|c}^{G} = T_w \Big/ \left(k(k-1)\right) \sum_{i=1}^{k} p_i \,(1-p_i) \tag{5}$$

being p_i=($n_{i\cdot}$+$n_{\cdot i}$)/2n the estimate of the propensity that a rater classifies an item into i^{th} category.

3. Aid to the characterization of the extent of agreement: benchmark scales

After computing an agreement coefficient, a common question is ‘how good is the agreement?’

In order to provide an aid to qualify the magnitude of kappa-type coefficients, a number of benchmark scales have been proposed mainly in social and medical sciences over the years. The best known benchmark scales are reviewed below and reported in Table 1.

According to Hartmann [12], acceptable values for kappa should exceed 0.6. The most widely adopted benchmark scale is the one with six ranges of values proposed by Landis and Koch [13], which was simplified by Fleiss [14] and Altman [15], with three and five ranges, respectively, and by Shrout [16] who collapsed the first three ranges of values into two agreement categories. Munoz and Bangdiwala [17], instead, proposed guidelines for interpreting the values of a kappa-type agreement coefficient with respect to the raw proportion of agreement.

Whatever the adopted scale, the benchmarking procedure is generally straightforward since the coefficient magnitude is qualified as the extent of agreement (*e.g.* good) associated to the range of values where the estimated agreement coefficient falls.

Table 1. Benchmark scales for kappa-type coefficients.

Hartmann (1977)		Landis and Koch (1977)		Fleiss (1981)	
Kappa coefficient	Strength of agreement	Kappa coefficient	Strength of agreement	Kappa coefficient	Strength of agreement
> 0.6	Good	< 0.0	Poor	< 0.4	Poor
		0.0 to 0.20	Slight	0.40 to 0.75	Intermediate to Good
		0.21 to 0.40	Fair	> 0.75	Excellent
		0.41 to 0.60	Moderate		
		0.61 to 0.80	Substantial		
		0.81 to 1.00	Almost perfect		

Altman (1991)		Shrout (1998)		Munoz and Bengdiwala (1997)	
Kappa coefficient	Strength of agreement	Kappa coefficient	Strength of agreement	Kappa coefficient	Strength of agreement
< 0.2	Poor	0.00 to 0.10	Virtually none	< 0.00	Poor
0.21 to 0.40	Fair	0.11 to 0.40	Slight	0.00 to 0.20	Fair
0.41 to 0.60	Moderate	0.41 to 0.60	Fair	0.21 to 0.45	Moderate
0.61 to 0.80	Good	0.61 to 0.80	Moderate	0.46 to 0.75	Substantial
0.81 to 1.00	Very good	0.81 to 1.00	Substantial	0.76 to 0.99	Almost perfect
				1.00	Perfect

4. Probabilistic benchmarking procedures

Despite its popularity, the straightforward benchmarking procedure can be misleading for two main reasons:

- it does not associate the interpretation of the coefficient magnitude with a degree of certainty failing to consider that an agreement coefficient, as any other sampling estimate, is exposed to statistical uncertainty;
- it does not allow to compare the extent of agreement across different studies, unless they are carried out under the same experimental conditions (*i.e.* the number of observed items, the number of categories or the distribution of items among the categories).

In order to have a fair characterization of the extent of rater agreement, the benchmarking procedure should be probabilistic so as to associate a degree of certainty to the interpretation of the kappa coefficient [18].

Under asymptotic conditions, the magnitude of the kappa type coefficient can be related to the notion of extent of agreement by benchmarking the lower bound of its asymptotic normal (1 − 2α)% CI:

$$\mathrm{LB}_N = \hat{K} - z_\alpha \cdot se(\hat{K}) \tag{6}$$

where $\hat{K}$ is the estimated kappa coefficient, $se(\hat{K})$ its standard error and z_α the α percentile of the standard normal distribution.

Recently, Gwet [3] proposed a probabilistic benchmarking procedure based on the Interval Membership Probability (IMP). Gwet's procedure characterizes the magnitude of agreement by benchmarking the lowest value K_{L} such that the probability that K exceeds K_{L} is equal to 1–2α.

The above two benchmarking procedures rely on the asymptotically normal distribution assumption and thus they can work well only for reasonable large sample sizes. Vice-versa, under non-asymptotic conditions, bootstrap resampling can be adopted for building approximated as well as exact non-parametric CIs [19, 20].

Among the available bootstrap methods, in this study we focus on percentile CI and Bias-Corrected and Accelerated (BCa) CI [21]. The former is by far the easiest and most widespread method, the latter is recommended for severely skewed distribution.

Being G the cumulative distribution function of the bootstrap replications of the kappa-type coefficient, the lower bound of the (1–2α)% percentile CI is:

$$\mathrm{LB}_p = \mathrm{G}^{-1}\left(\alpha\right) \tag{7}$$

whereas, being Φ the standard normal CDF, b the bias correction parameter and a the acceleration parameter, the lower bound of the (1–2α)% BCa bootstrap CI is:

$$\mathrm{LB}_{\mathrm{BCa}} = \mathrm{G}^{-1}\left(\Phi\left(b - \frac{z_\alpha - b}{1 + a\left(z_\alpha - b\right)}\right)\right) \tag{8}$$

5. Simulation study

The statistical properties of the above-discussed probabilistic benchmarking procedures have been investigated via a Monte Carlo simulation study across 72 different settings defined by varying three parameters: number of rated items (*i.e.* n = 10, 30, 50, 100), number of categories (*i.e.* k = 2, 3, 4) and strength of agreement (low, moderate and high), represented by six levels of agreement ranging from 0.4 to 0.9, computed assuming a linear weighting scheme [22]:

$$w_{ij} = 1 - \frac{|i-j|}{k-1} \tag{9}$$

Specifically, the simulation study has been developed considering two raters classifying n items into one of k possible ordinal rating categories. The data have been simulated by sampling r = 2000 Monte Carlo data sets from a multinomial distribution with parameters n and $\boldsymbol{p} = (\pi_{11}, ..., \pi_{\mathrm{ij}}, ... \pi_{\mathrm{ik}})$, being π_{ij} the probability that an item is classified into category i^{th} by the first rater and into j^{th} category by the second rater. For each rating scale dimension and assuming a linear weighting scheme, the values of the joint probabilities π_{ij} have been chosen so as to obtain the six true population values of agreement (*viz.* 0.4, 0.5, 0.6, 0.7, 0.8, 0.9) for a total of 18 different vectors $\boldsymbol{p}$ for each sample size; for example $\boldsymbol{p}$ = (0.22, 0.06, 0.05, 0.06, 0.22, 0.06, 0.05, 0.06, 0.22) for obtaining an agreement level equal to 0.5 with k = 3 rating categories. The four probabilistic benchmark procedures have been applied to characterize the simulated AC_2 and BP_w agreement coefficients across all different settings and their performances have been evaluated in terms of the weighted proportion of misclassified benchmarks (weighted misclassification rate, M_w).

The simulation results obtained for each coefficient and each combination of n and k values are represented in the bubble chart in Figure 1, where the size of each bubble expresses the M_w value. Since the parametric benchmarking procedures apply only under asymptotic conditions, the Figure 1 is divided into 2 sections by a dashed line: on the left side only the non-parametric procedure are compared each other (i.e. two overlapping bubbles for LB_BCa and LB_p, respectively), whereas the right side refers to all the benchmarking procedures under comparison (i.e. four overlapping bubbles for LB_BCa, LB_p, K_L and LB_N, respectively). The bubble chart displays all the 24 analyzed comparisons: for each of them, the foreground bubble represents the benchmarking procedure with the best performance (i.e. the one with the smallest M_w), whereas the background bubble represents the procedure with the worst performance (i.e. the one with the highest M_w, whose value is reported in the label).

For small sample sizes M_w slightly differs across non-parametric benchmarking procedures — with a difference no more than 5% — and agreement coefficients. The results seem to suggest that the best choice is benchmarking the lower bound of the percentile CI for AC_2 and benchmarking the lower bound of the BCa CI for BP_w. For large sample sizes, instead, M_w is comparable across benchmarking procedures and agreement coefficients.

Specifically, it is worthwhile to pinpoint that the differences in M_w across non-parametric benchmarking procedures and agreement coefficients get smaller as n increases because of the decreasing skewness in the distributions of the agreement coefficients: if the distribution is symmetric, the BCa and percentile CIs agree.

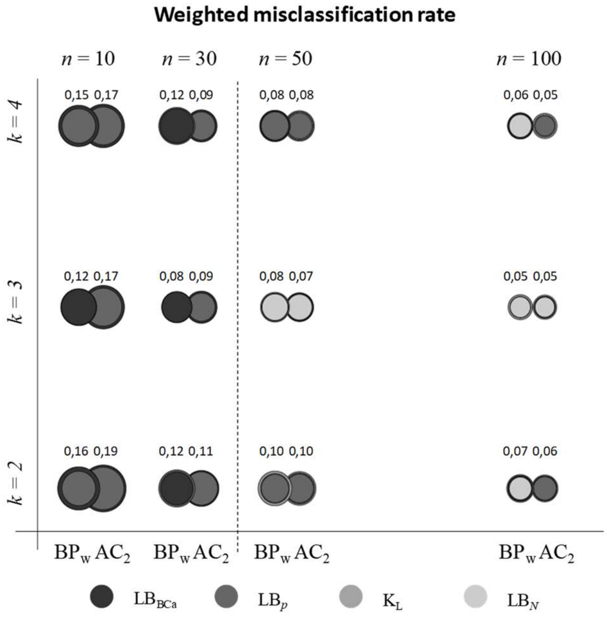


Fig. 1. M_w for BP$_w$ and AC$_2$, for different benchmarking procedures, n and k values

6. Conclusions

One of the main issues related to the widely adopted agreement coefficients regards the characterization of the extent of agreement.

Most research studies characterize the extent of agreement by comparing the obtained agreement coefficient against well-known threshold values, like 0.5 or 0.75, whereas only few research studies adopt probabilistic approaches, overcoming the straightforward comparison. Anyway, the probabilistic approaches commonly adopted in the literature are generally based on parametric asymptotic CIs that, by definition, are only applicable for large sample sizes so that small sample sizes become the most critical for statistical inference, although the most affordable in many experimental contexts.

The conducted Monte Carlo simulation study suggests that the non-parametric probabilistic benchmarking procedures based on bootstrap resampling have satisfactory and comparable (with a difference up to 5%) properties for moderate or small number of rated items. Specifically, with n = 30 the performances of the procedures based on bootstrap CIs differ from each other at most for 2%, therefore benchmarking the lower bound of the percentile bootstrap CI could be suggested because of the less computation burden. Otherwise, with large sample sizes, being the performances indistinguishable across all benchmarking procedures, parametric procedures should be preferred because of their lower computational complexity.

References

1. J. Cohen, A coefficient of agreement for nominal scale, EPM **20(1)**, 37-46. (1960).
2. W. D. Thompson and S. D. Walter, A reappraisal of the kappa coefficient, J Clin Epidemiol **41(10),** 949–58 (1988).
3. K. L. Gwet, Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters (Advanced Analytics, LLC 2014).
4. J. Cohen, Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, Psychological Bulletin **70(4)** 213--219. (1968)
5. R. L.Brennan and D. J. Prediger, Coefficient Kappa: Some Uses, Misuses, and Alternatives, EPM **41**, 687–699 (1981)
6. A. Feinstein and D. Cicchetti, High agreement but low kappa: I. The problems of two paradoxes, J Clin Epidemiol **43(6)**, 543--549. (1990)
7. L. Guttman, The test-retest reliability of qualitative data, Psychometrika **11**, 81-95. (1945)
8. E. M. Bennett, R. Alpert and A. C. Goldstein, Communications through limited response questioning. Public Opin Q. **18(3)**, 303–308. (1954)
9. J.W. Holley and J. P. Guilford, A note on the G index of agreement, EPM **24**, 749-753. (1964)
10. S. Janson and J. Vegelius, On generalizations of the G index and the Phi coefficient to nominal scales, Multivariate Behav Res **14(2)**, 255–269. (1979)
11. K. L. Gwet, Computing inter-rater reliability and its variance in the presence of high agreement, Br. J. Math. Stat. Psychol **61**, 29--48. (2008)
12. D. Hartmann, Considerations in the choice of interobserver reliability estimates, J Appl Behav Anal **10(1)**, 103–116 (1977).

13. J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data, Biometrics **33(1)**, 159-174 (1977).
14. J. L. Fleiss, Statistical Methods for Rates and Proportions (John Wiley & Sons. 1981).
15. D. G. Altman, Practical Statistics for Medical Research (Chapman and Hall, 1991).
16. P. E. Shrout, Measurement reliability and agreement in psychiatry, Stat Methods Med Res **7(3)**, 301-317 (1998).
17. S. R. Munoz and S. I. Bangdiwala, Interpretation of Kappa and B statistics measures of agreement, J Appl Stat, **24(1)**, 105-112 (1997).
18. J. Kottner, L. Audige, S. Brorson, A. Donner, B. J. Gajewski, A. Hrobjartsson, C. Roberts, M. Shoukri, D. L. Streiner, Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed, Int J Nurs Stud, **48,** 661-671 (2011)
19. N. Klar, S. Lipsitz, M. Parzen and T. Leong, An exact bootstrap confidence interval for κ in small samples. J R Stat Soc Ser D (The Statistician), **51(4),** 467–478 (2002).
20. J. Lee and K. P. Fung, Confidence interval of the kappa coefficient by bootstrap resampling [letter]. Psychiatry Research **49**, 97–98 (1993).
21. J. Carpenter and J. Bithell, Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians, Statist Med. **19**,1141-64 (2000).
22. D. V. Cicchetti and T. Allison, A new procedure for assessing reliability of scoring EEG sleep recordings. Amer. J. EEG Technol. **11(3)**, 101-10. (1971).

A comparative study of benchmarking procedures for interrater and intrarater agreement studies

Valutazione comparativa di procedure di benchmarking per l'analisi dell'accordo inter e intra valutatore

Amalia Vanacore[1] and Maria Sole Pellegrino[2]

Abstract Decision making processes typically rely on subjective evaluations provided by human raters. In the absence of a gold standard against which check evaluation trueness, the magnitude of inter/intra-rater agreement coefficients is commonly interpreted as a measure of the rater's evaluative performance. In this study some benchmarking procedures for characterizing the extent of agreement are discussed and compared via a Monte Carlo simulation.

Abstract *In numerosi contesti, le decisioni strategiche sono affidate a valutazioni soggettive, fornite da valutatori umani, per le quali non esiste un gold standard che permetta di valutarne la veridicita'. L'affidabilita' del valutatore viene quindi spesso misurata in termini di precisione attraverso coefficienti di accordo inter- e intra-valutatore, che risultanto utili se interpretabili. Nel lavoro proponiamo uno studio Monte Carlo per analizzare e confrontare le prestazioni di alcune procedure di benchmarking.*

Key words: rater agreement, kappa-type coefficient, benchmarking procedures, Monte Carlo simulation

1 Introduction

Agreement coefficients are widely adopted for assessing the precision of subjective evaluations provided by human raters to support strategic and operational decisions in several fields (e.g. manufacturing and service industries, food, healthcare and risk management). Specifically, the agreement between the evaluations provided on the same sample of items by two or more raters (i.e. inter-rater agreement) or by

[1] Dept. of Industrial Engineering, University of Naples "Federico II", p.le Tecchio 80, 80125 Naples; email: amalia.vanacore@unina.it

[2] Dept. of Industrial Engineering, University of Naples "Federico II", p.le Tecchio 80, 80125 Naples; e-mail: mariasole.pellegrino@unina.it

the same rater in two or more occasions (i.e. intra-rater agreement) is commonly measured using a kappa-type agreement coefficient.

In order to qualify the extent of agreement as good or poor the computed coefficient is compared against an arbitrary benchmark scale. However, the magnitude of an agreement coefficient may strongly depend on some experimental factors such as number of rated items, rating scale dimension, trait prevalence and marginal probabilities [13, 9]. Thus, interpretation based on the straightforward benchmarking should be treated with caution especially for comparison across studies when experimental conditions are not the same.

A proper characterization of the extent of rater agreement should rely upon a benchmarking procedure that allows to identify a suitable neighborhood of the true value of rater agreement by taking into account sampling uncertainty. The most simple and intuitive way to accomplish this task is by building a confidence interval of the agreement coefficient and comparing its lower bound against an adopted benchmark scale. A different approach is the one recently proposed by Gwet [9] which, under the assumption of asymptotically normal distribution, evaluates the likelihood that the estimated agreement coefficient belongs to each benchmark category.

The above benchmarking approaches will be in the following discussed and their performances will be evaluated and compared in terms of weighted misclassification rate via a Monte Carlo simulation study.

The remainder of the paper is organized as follows: in Section 2 two well-known paradox-resistant kappa-type agreement coefficients are discussed; the commonly adopted benchmark scales and some characterization procedures based on parametric and non-parametric approaches to benchmarking are presented and discussed in Section 3; in Section 4 the simulation design is described and the main simulation results are fully discussed; finally, conclusions are summarized in Section 5.

2 Paradox-resistant agreement coefficients

The kappa-type agreement coefficients are rescaled measures of the observed agreement corrected with the probability of agreement expected by chance. The most common kappa-type coefficient is that proposed by Cohen [5]. Despite its popularity, it is affected by two paradoxes [4]: the degree to which raters disagree (bias problem) and the marginal distribution of the evaluations independently provided by each rater (prevalence problem). A solution to face the above paradoxes is to adopt the uniform distribution for chance measurements, which given a certain rating scale can be defended as representing the maximally non-informative measurement system [6].

Specifically, let n be the number of items rated by two raters on an ordinal k-point rating scale (with $k > 2$), n_{ij} the number of items classified into i^{th} category by the first rater and into j^{th} category by the second rater and w_{ij} the corresponding symmetrical weight, introduced in order to consider that, on an ordinal rating scale, disagreement on two distant categories is more serious than disagreement on neighboring categories. The weighted version of the uniform kappa, often referred to as Brennan-Prediger coefficient [9], is formulated as:

$$\widehat{BP}_w = \frac{p_{a_w} - p_{a|c}^{BP_w}}{1 - p_{a|c}^{BP_w}} \tag{1}$$

where p_{a_w}, the weighted observed proportion of agreement, and $p_{a|c}^{BP_w}$, the weighted proportion of agreement expected under the assumption of uniform chance measurements, are respectively given by:

$$p_{a_w} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\frac{n_{ij}}{n}; \quad p_{a|c}^{BP_w} = \frac{T_w}{k^2} \tag{2}$$

being T_w the sum over all weight values w_{ij}.

Another well-known paradox-resistant agreement coefficient alternative to Cohen's Kappa is the AC_1 coefficient proposed by Gwet [8], whose weighted version AC_2 [9] is formulated as:

$$\widehat{AC}_2 = \frac{p_{a_w} - p_{a|c}^{AC_2}}{1 - p_{a|c}^{AC_2}} \tag{3}$$

where the probability of chance agreement $p_{a|c}^{AC_2}$ is given by:

$$p_{a|c}^{AC_2} = \frac{T_w}{k(k-1)} \cdot \sum_{i=1}^{k} \pi_i(1-\pi_i) \tag{4}$$

Specifically, $p_{a|c}^{AC_2}$ is defined as the probability of the simultaneous occurrence of two events, one rater provides random rating (R) and the two raters agree (G):

$$p_{a|c}^{AC_2} = P(G \cap R) = P(G|R) \cdot P(R) \tag{5}$$

where $P(G|R) = T_w/k^2$ and $P(R)$ is approximated with a normalized measure of randomness defined by the ratio of the observed variance to the variance expected under the assumption of totally random ratings:

$$P(R) = \frac{\sum_{i=1}^{k} p_i(1-p_i)}{(k-1)/k} \tag{6}$$

with p_i denoting the propensity that a rater assigns score i to an item which is estimated by $p_i = (n_{i\cdot} + n_{\cdot i})/2n$ being $n_{i\cdot}$ (resp. $n_{\cdot i}$) the total number of items classified into i^{th} category by the first (resp. second) rater.

3 Benchmarking procedures for characterizing the extent of agreement

After computing an agreement coefficient, a common question is ”how good is the extent of agreement?” As a general rule kappa values greater than 0.6 are generally considered acceptable [10]. In order to provide an aid to qualify the magnitude of kappa-type coefficients, a number of benchmark scales have been proposed mainly in social and medical sciences over the years. The scale proposed by Landis and Koch [11] is by far the most widely adopted benchmark scale; it consists of six ranges of values corresponding to as many categories of agreement: poor, slight, fair, moderate, substantial and almost perfect agreement for coefficient values ranging between -1 and 0, 0 and 0.2, 0.21 and 0.4, 0.41 and 0.6, 0.61 and 0.8 and 0.81 and 1.0, respectively. This scale was then simplified by Fleiss [7] and Altman [1], with three and five ranges, respectively, and by Shrout [12] who collapsed the first three ranges of values into two agreement categories.

Despite its popularity, the straightforward benchmarking can be misleading because it does not associate the interpretation of the extent of agreement with a degree of uncertainty and it does not allow to compare the extent of agreement across different studies, unless they are carried out under the same experimental conditions. In order to have a fair characterization of the extent of rater agreement, it is necessary to associate a degree of uncertainty to the interpretation of the coefficient.

Under asymptotic conditions, the magnitude of the kappa type coefficient can be related to the notion of extent of agreement by benchmarking the lower bound of its asymptotic $(1-2\alpha)\%$ confidence interval (CI). Recently, Gwet [9] proposed a parametric benchmarking procedure based on Interval Membership Probability (IMP) that is the probability that the coefficient falls into each benchmark category.

Under non-asymptotic conditions, two non-parametric CIs based on bootstrap resampling are the percentile (p) CI and, for severely skewed distribution, the Bias-Corrected and Accelerated (BCa) CI [2]. Being free from distributional assumptions, the benchmarking procedure based on bootstrap CIs fits also the cases of moderate and small sample sizes.

4 Simulation study

The above-discussed benchmarking procedures have been applied to characterize the extent of both BP_w and AC_2 across 72 different settings. Their statistical properties have been investigated via a Monte Carlo simulation study developed considering two raters classifying $n = 10, 30, 50, 100$ items into one of $k = 2, 5, 7$ possible ordinal rating categories. The data have been simulated by sampling $r = 2000$ Monte Carlo data sets from a multinomial distribution with parameters n and $\mathbf{p} = (\pi_{11}, \ldots, \pi_{ij}, \ldots, \pi_{ik})$; the π_{ij} values have been set so as to obtain six true population values of agreement (viz. 0.4, 0.5, 0.6, 0.7, 0.8, 0.9), assuming a linear weighting scheme [3].

The performances of the benchmarking procedures under comparison have been evaluated in terms of weighted misclassification rate (hereafter, $\mathbf{M}_\mathrm{w}$). Specifically, let $\{X_h; r\}$ be a Monte Carlo data set containing r benchmarks X_h obtained for a population value taken as reference for a specific agreement category ω. $\mathbf{M}_\mathrm{w}$ is evaluated as the weighted proportion of misclassified X_h:

$$\mathbf{M}_\mathrm{w} = \frac{1}{r} \sum_{\omega=1,\Omega} w_{\omega\omega'} \cdot I\left[X_{h|\omega} \in \omega'\right]; \quad \omega' \neq \omega \tag{7}$$

where $I[\cdot]$ is an indicator taking value 1 if the argument is true and 0 otherwise and $w_{\omega\omega'}$ is a linear misclassification weight adopted to account that on an ordinal benchmarking scale some misclassifications are more serious than others. The best and worst values of $\mathbf{M}_\mathrm{w}$ obtained across the analysed benchmarking procedures for BP_w and AC_2 are reported in Table 1 for each combination of n and k values. Specifically, while the benchmarking procedure based on bootstrap CIs are suitable for all the analysed sample sizes, the parametric procedures work only under asymptotic conditions being thus applied only to large samples of $n \geq 50$; therefore the parametric and non-parametric procedures are compared each other only for $n \geq 50$.

Table 1 Best and worst $\mathbf{M}_\mathrm{w}$ across the four benchmarking procedures (Standard: Parametric CI; Underlined: IMP; *Italics*: p CI; **Bold**: BCa CI) for BP_w and AC_2 for different n and k values

(a) Best $\mathbf{M}_\mathrm{w}$ for BP_w

	$n=10$	$n=30$	$n=50$	$n=100$
$k=2$	*0.102*	**0.096**	*0.068*	0.049
$k=5$	**0.123**	*0.081*	0.056	0.034
$k=7$	**0.087**	*0.066*	0.048	0.027

(b) Worst $\mathbf{M}_\mathrm{w}$ for BP_w

	$n=10$	$n=30$	$n=50$	$n=100$
$k=2$	**0.160**	*0.118*	0.080	*0.058*
$k=5$	*0.131*	**0.088**	0.072	0.051
$k=7$	*0.089*	0.072	0.060	0.044

(c) Best $\mathbf{M}_\mathrm{w}$ for AC_2

	$n=10$	$n=30$	$n=50$	$n=100$
$k=2$	*0.159*	*0.098*	*0.072*	0.046
$k=5$	*0.111*	*0.073*	0.051	0.030
$k=7$	*0.085*	**0.031**	0.044	*0.026*

(d) Worst $\mathbf{M}_\mathrm{w}$ for AC_2

	$n=10$	$n=30$	$n=50$	$n=100$
$k=2$	**0.193**	**0.099**	**0.084**	**0.055**
$k=5$	**0.130**	**0.092**	**0.066**	0.046
$k=7$	**0.092**	*0.058*	**0.056**	0.042

For small and moderate samples (i.e. $n \leq 30$), $\mathbf{M}_\mathrm{w}$ slightly differs across non-parametric benchmarking procedures and agreement coefficients: specifically, the highest difference in $\mathbf{M}_\mathrm{w}$ is 9%, observed for $n = 10$ and $k = 2$. Moreover, for increasing sample sizes, $\mathbf{M}_\mathrm{w}$ becomes quite indistinguishable across procedures and coefficients with a difference always no more than 2%. It is worthwhile to pinpoint that the differences in $\mathbf{M}_\mathrm{w}$ across non-parametric benchmarking procedures and agreement coefficients get smaller as n increases because of the decreasing skewness in the distributions of the coefficients: if the distribution is symmetric, the BCa and p CIs agree.

5 Conclusions

The results of the Monte Carlo simulation suggest that for small samples the non-parametric benchmarking procedures based on bootstrap resampling have satisfactory and comparable properties in terms of weighted misclassification rate. Moreover, with $n \geq 30$ the performances of the procedures based on bootstrap CIs differ from each other at most for 2%, therefore benchmarking the lower bound of the percentile bootstrap confidence interval could be suggested — because of its less computational burden — for characterizing the extent of rater agreement, both for BP_w and AC_2. For large samples, the performances are indistinguishable across all benchmarking procedures, thus benchmarking the lower bound of the parametric confidence interval would be preferred being the easiest method to implement.

References

1. Altman, D.G.: Practical statistics for medical research. CRC press (1990)
2. Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med **19**(9), 1141–1164 (2000)
3. Cicchetti, D.V., Allison, T.: A new procedure for assessing reliability of scoring EEG sleep recordings. Am J EEG Technol **11**(3), 101–110 (1971)
4. Cicchetti, D.V., Feinstein, A.R.: High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. **43**(6), 551–558 (1990)
5. Cohen, J.: A coefficient of agreement for nominal scales. Educ Psychol Meas **20**(1), 37–46 (1960)
6. De Mast, J., Van Wieringen, W.N.: Measurement system analysis for categorical measurements: agreement and kappa-type indices. J Qual Technol **39**(3), 191–202 (2007)
7. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol Bull **76**(5), 378–382 (1971)
8. Gwet, K.L.: Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol **61**(1), 29–48 (2008)
9. Gwet, K.L.: Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC (2014)
10. Hartmann, D.P.: Considerations in the choice of interobserver reliability estimates. J. Appl. Behav. Anal. **10**(1), 103–116 (1977)
11. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics pp. 159–174 (1977)
12. Shrout, P.E.: Measurement reliability and agreement in psychiatry. Stat. Methods Med. Res. **7**(3), 301–317 (1998)
13. Thompson, W.D., Walter, S.D.: A reappraisal of the kappa coefficient. J Clin Epidemiol **41**(10), 949–958 (1988)

3rd International Conference on Higher Education Advances, HEAd'17
Universitat Politècnica de València, València, 2017
DOI: http://dx.doi.org/10.4995/HEAd17.2017.5583

An agreement-based approach for reliability assessment of Students' Evaluations of Teaching

Vanacore, Amalia; Pellegrino, Maria Sole
Department of Industrial Engineering, University of Naples "Federico II", Italy

Abstract

Students' Evaluations of Teaching (SETs) are the most common way to measure teaching quality in Higher Education: they are assuming a strategic role in monitoring teaching quality, becoming helpful in taking the major formative and summative academic decisions. The majority of studies investigating SETs reliability focus on the instruments and the procedures adopted to collect students' evaluations rather than on the capability of the students as teaching quality assessors. In order to overcome this lack, a study has been carried out with the aim of measuring SETs reliability in terms of inter-student agreement and intra-student agreement. The results of our study show that the majority of students provided substantially repeatable evaluations whereas only a few students provided almost perfectly repeatable evaluations; the evaluations provided by different students generally slightly agreed, which means that the students did not share the same opinions and beliefs on teaching quality.

Keywords: *teaching quality assessment; reliability; inter-student agreement; intra-student agreement.*

1. Introduction

Measuring the student experience is assuming increasingly importance in Higher Education (hereafter, HE) representing a widespread method for evaluating teaching quality whose importance is relevant for taking the major formative and summative academic decisions (Berk, 2005; Gravestock & Gregor-Greenleaf, 2008; Onwuegbuzie *et al.*, 2009).

Student ratings, also known as Student Evaluations of Teaching (SETs), have dominated as the primary measure of teaching quality over the past 40 years (*e.g.*, Centra, 1979; Seldin, 1999; Emery *at al.*, 2003; Gaertner, 2014) forming the basis for the rankings of HE institutions. Although widely used, SETs are one of the most controversial and highly-debated measures of teaching quality: many researchers argue that there is no better option that provides the same sort of quantifiable and comparable data on teaching quality (McKeachie, 1997; Abrami, 2001) but, on the opposite, others point out significant biasing factors for SETs.

The fear that students cannot provide reliable teaching quality evaluations is, by far, one of the primary concerns about SETs. As a matter of fact, even highly motivated students can base their current evaluations on their past teaching experience, which can substantially vary depending on the college or university attended and/or on the student individual belief toward the degree (Ackerman *et al.*, 2009). Students who are generally satisfied/dissatisfied with the course and/or the instruction can bias the results upward/downward (Sliusarenko *et al.*, 2013). In addition, it is known that demographic (*e.g.*, gender and age; Thorpe, 2002; Fidelman, 2007; Kherfi, 2011) as well as logistic (*e.g.*, class size; Kuo, 2007) factors can influence SETs. The above considerations call into question the opportunity to consider the students as able to provide reliable evaluations on teaching quality. For this reason, differently from the majority of available studies, which rather focus on the instruments and the procedures adopted to collect SETs, our study aims at investigating the peculiar abilities of the students as teaching quality assessors by measuring SETs reliability in terms of inter-student and intra-student agreement. Particularly, the former allows evaluating the students' ability to provide the same score, on average, as the other students whereas the latter, also known as *repeatability*, allows evaluating the students' ability to score consistently a given quality item in different occasions.

2. Measuring inter-student and intra-student agreement: kappa-type indexes

The easiest approach for assessing the degree of agreement among repeated evaluations would be to simply calculate the observed agreement. This approach, however, provides a biased measure of agreement, especially when a rating scale with a few categories is adopted. In order to avoid this problem, inter-student and intra-student agreement will be assessed using the well-known kappa-type indexes, where the observed agreement is corrected for the agreement expected by chance. Specifically, the degree of inter-student agreement is assessed by calculating the s statistic proposed by Marasini *et al.* (2014), that is a rescaled measure of the probability of observed agreement p_a^s corrected with the probability of agreement expected by chance alone $p_{a|c}^s$:

$$s = (p_a^s - p_{a|c}^s) \big/ (1 - p_{a|c}^s) \tag{1}$$

Being r the number of students who rated twice (*i.e.* replications) the same n quality items on a $k \geq 3$ points ordinal scale, r_{hi} and r_{hj} the number of students who assigned the h^{th} quality item into i^{th} and j^{th} category during first and second replication, respectively; w_{ij} the corresponding weight, introduced in order to account that some disagreements (*i.e.* on categories that are at least two steps apart) are more serious than others (*i.e.* on neighboring categories), the observed proportion of agreement and the proportion of agreement expected by chance alone can be obtained as:

$$\hat{p}_a^s = \frac{1}{n}\sum_{h=1}^{n}\hat{p}_h; \quad p_{a|c}^s = \frac{1}{k} + \frac{1}{k^2}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} w_{ij} \tag{2}$$

where $\hat{p}_h$ is the proportion of agreement on h^{th} quality item given by:

$$\hat{p}_h = \left(\sum\nolimits_{i=1}^{k} r_{hi}\left(r_{hi} - 1\right) + 2\sum\nolimits_{i=1}^{k-1}\sum\nolimits_{j=i+1}^{k} w_{ij} r_{hi} r_{hj}\right) \Big/ \left(r\left(r-1\right)\right) \tag{3}$$

The degree of intra-student agreement, instead, is assessed using the weighted version of Brennan-Prediger coefficient (1981) proposed by Gwet (2014), that is a rescaled measure of the probability of observed agreement p_a corrected with the probability of agreement expected by chance alone $p_{a|c}$:

$$K_W^U = (p_a - p_{a|c}) \big/ (1 - p_{a|c}) \tag{4}$$

The chance measurement system adopted in Brennan-Prediger coefficient is the uniform one. Being n the number of quality items rated twice on a $k \geq 3$ points ordinal scale by the same student, n_{ij} the number of quality items classified into i^{th} category in the first replication and into j^{th} category in the second replication, the observed proportion of agreement $\hat{p}_a$ and the proportion of agreement expected by chance alone $p_{a|c}$ are:

$$\hat{p}_a = \sum\nolimits_{i=1}^{k}\sum\nolimits_{j=1}^{1} n_{ij} w_{ij}; \quad p_{a|c} = \left(\sum\nolimits_{i=1}^{k}\sum\nolimits_{j=1}^{1} w_{ij}\right) \Big/ k^2 \tag{5}$$

The values of kappa-type indexes range between -1 and 1, with negative values meaning disagreement. The index magnitude can be interpreted by adopting the Landis and Koch (1977) benchmark scale. According to this scale, there are 5 categories of agreement corresponding to as many ranges of coefficient values: slight, fair, moderate, substantial and almost perfect agreement for coefficient values ranging between 0 and 0.2, 0.21 and 0.4, and 0.41 and 0.6, 0.61 and 0.8 and 0.81 and 1.0, respectively.

3. Case Study

The case study was conducted at the Department of Industrial Engineering of University of Naples "Federico II" and consisted of 3 supervised experiments (hereafter, E.1, E.2, E.3) carried out on classes of students attending the course of Statistical Quality Control (SQC) in 3 successive academic years. All three involved classes included more than 20 students; all of them obtained the first level degree in Management Engineering from the University of Naples "Federico II" and thus they can be reasonably assumed homogeneous in curriculum and instruction.

Students were asked to fill two evaluation sheets (each with a specific rating scale) in order to collect their quality evaluation for a set of $n = 20$ items (regarding, for example, organization, workload and readings) of the SQC course they were attending. The first evaluation sheet used a Numeric Rating Scale (NRS) with scores ranging from 0 to 10 whereas the other used a Verbal Rating Scale (VRS) with agreement grades: "strongly disagreeing with the statement", "slightly agreeing with the statement", "quite agreeing with the statement" and "strongly agreeing with the statement". For comparability purposes, students' evaluations on the NRS were rescaled to the 4-points VRS using the following cut-off ranges: 0 to 2, 3 to 5, 6 to 8 and 9 to 10.

Each experiment consisted of two sessions: the first evaluation session (*i.e.*, S.I) took place at mid-term course and the second evaluation session (*i.e.*, S.II) took place the following lesson. Between S.I and S.II there was no new lesson and no interaction with the teacher, therefore no change in quality evaluation was expected. In order to guarantee evaluation traceability while preserving anonymity, each student signed her/his evaluation sheets with a nickname, which enabled to match student's ratings provided in the two evaluation sessions in order to estimate intra-student agreement. Only those students who rated all quality items in both experimental sessions were retained as participants in the study (*viz.* 17 students in E.1, 18 students in E.2 and 17 students in E.3).

The collected data were used to estimate the inter-student and intra-student agreement on NRS (hereafter, $\hat{s}_{\text{NRS}}$ and $\hat{K}^{U}_{W|\text{NRS}}$, respectively) and the inter-student and intra-student agreement on VRS (hereafter, $\hat{s}_{\text{VRS}}$ and $\hat{K}^{U}_{W|\text{VRS}}$); the intra-student agreement coefficients were both computed adopting the linear weighing scheme (Cicchetti & Allison, 1971).

3.1. Study results

The value of $\hat{s}_{\text{NRS}}$ and $\hat{s}_{\text{VRS}}$ for E.1, E.2 and E.3 are reported in Table 1.

Table 1. Inter-student agreement on NRS and VRS

Experiment	E.1	E.2	E.3
$\hat{s}_{\text{NRS}}$	0.395	0.300	0.600
$\hat{s}_{\text{VRS}}$	0.380	0.528	0.277

The results for intra-student agreement for each student participating in E.1, E.2 and E.3, are reported in Table 2 and plotted in Figures 1 against the 5 regions of intra-student agreement on NRS and intra-student agreement on VRS identified according to the Landis and Koch's benchmark scale.

Results in Table 1 highlight that the inter-student agreement is at most moderate, so that it is not possible to assume that the involved students shared the same opinions about teaching quality; the difference between the two rating scales is irrelevant only for students of E.1, however results do not allow preferring a rating scale over the other.

The intra-student agreement was generally higher than the inter-student agreement: 73% of students were at least substantially repeatable on both NRS and VRS whereas 19% of them were even almost perfectly repeatable on both NRS and VRS. In addition, the majority of students show over the years values of $\hat{K}^{U}_{W|\text{VRS}}$ higher than those of $\hat{K}^{U}_{W|\text{NRS}}$ although for about half of them the repeatability on the two rating scales belong to the same agreement categories and only for few (*i.e.*, 10) students $\hat{K}^{U}_{W|\text{NRS}}$ and $\hat{K}^{U}_{W|\text{VRS}}$ belong to no-adjacent categories of agreement.

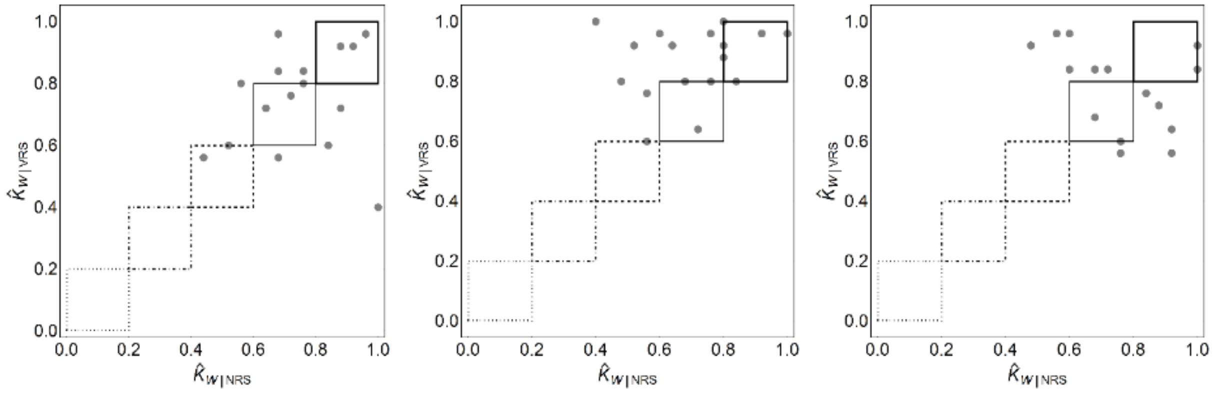


Figure 1. Intra-student agreement on NRS (as abscissa) and VRS (as ordinate) for each student participating in E.1. (on the left), E.2. (in the middle) and E.3. (on the right)

Table 2. Intra-student agreement on NRS ($\hat{K}^{U}_{W|\mathrm{NRS}}$) and VRS ($\hat{K}^{U}_{W|\mathrm{VRS}}$)

	E.1		E.2		E.3	
Student	$\hat{K}^{U}_{W\|\mathrm{NRS}}$	$\hat{K}^{U}_{W\|\mathrm{VRS}}$	$\hat{K}^{U}_{W\|\mathrm{NRS}}$	$\hat{K}^{U}_{W\|\mathrm{VRS}}$	$\hat{K}^{U}_{W\|\mathrm{NRS}}$	$\hat{K}^{U}_{W\|\mathrm{VRS}}$
1	0.76	0.80	0.56	0.76	0.92	0.56
2	0.88	0.72	0.80	0.92	0.56	0.96
3	0.68	0.96	0.48	0.80	0.72	0.84
4	1.00	0.40	0.84	0.80	0.60	0.96
5	0.68	0.84	0.52	0.92	0.84	0.76
6	0.76	0.84	1.00	0.96	0.88	0.72
7	0.92	0.92	0.64	0.92	0.68	0.68
8	0.96	0.96	0.76	0.80	0.68	0.84
9	0.64	0.72	0.60	0.96	0.60	0.84
10	0.44	0.56	1.00	0.96	0.76	0.60
11	0.72	0.76	0.56	0.60	0.48	0.92
12	0.84	0.60	0.92	0.96	0.92	0.64
13	0.76	0.84	0.80	1.00	0.72	0.84
14	0.56	0.80	0.72	0.64	0.76	0.56
15	0.68	0.56	0.80	0.88	1.00	0.84
16	0.52	0.60	0.68	0.80	1.00	0.92
17	0.88	0.92	0.40	1.00	1.00	0.92
18			0.76	0.96		

4. Conclusions

This research aimed at investigating the reliability of Students' Evaluations of Teaching by evaluating intra- and inter-student agreement.

With respect to intra-rater agreement, the results of our study highlight that, on average, the 65% of involved students could be considered repeatable assessors of teaching quality, since they provided quality evaluations that were consistent over time. Specifically, for NRS, the percentage of at least substantially repeatable students ranges, across the three experiments, between 66% and 82%, whereas, for VRS, the percentage of at least substantially repeatable students ranges between 71% and 94%. These results seem to suggest that even if the NRS is the most common rating scale, the students were able to express their opinion more consistently using a verbal rather than a numeric rating scale.

On the other hand, focusing on inter-student agreement, results seem to suggest that the whole class of students could not be considered homogeneous in terms of beliefs and/or opinions and/or knowledge about teaching quality, being the inter-student agreement at most moderate, independently of the specific class of students and the adopted rating scale.

The obtained results cannot of course be generalized since, although the experiments were repeated over three academic years, they involved only students attending the same course. In order to overcome this weakness, an interesting development could be to conduct the same experiment on different university courses.

References

Abrami, P. C. (2001). Improving Judgments About Teaching Effectiveness Using Teacher Rating Forms. *New Directions for Institutional Research,* 2001(109), 59-87.

Ackerman D., Gross B.L. & Vigneron F. (2009). Peer Observation Reports and Student Evaluations of Teaching: Who Are the Experts?. *The Alberta Journal of Educational Research*, 55(1), 18-39.

Berk R. A. (2005). Survey of 12 Strategies to Measure Teaching Effectiveness. *International Journal of Teaching and Learning in Higher Education*, 17(1), 48-62.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement*, 41, 687–699.

Centra, J. A. (1979). *Determining Faculty Effectiveness. Assessing Teaching, Research, and Service for Personnel Decisions and Improvement*. Jossey-Bass Publications, ERIC Number: ED183127.

Cicchetti, D. V., & Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology,* 11(3), 101-110.

Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: a critique of student evaluations of teaching effectiveness. *Quality Assurance in Education*, 11(1), 37–46.

Fidelman, C.G. (2007). *Course Evaluation Surveys: In-class Paper Surveys Versus Voluntary Online Surveys.* ProQuest.

Gaertner, H. (2014). Effects of student feedback as a method of self-evaluating the quality of teaching. *Studies in Educational Evaluation*, 42, 91-99.

Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends.* Toronto: Higher Education Quality Council of Ontario.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* Advanced Analytics, LLC.

Kherfi, S. (2011). Whose Opinion Is It Anyway? Determinants of Participation in Student Evaluation of Teaching. *The Journal of Economic Education*, 42(1), 19-30.

Kuo, W. (2007). Editorial: How reliable is teaching evaluation? The relationship of class size to teaching evaluation scores. *IEEE Transactions on Reliability*, 56(2), 178-181.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Marasini, D., Quatto, P., & Ripamonti, E. (2014). A Measure of Ordinal Concordance for the Evaluation of University Courses. *Procedia Economics and Finance*, 17, 39-46.

McKeachie, W. J. (1997). Student ratings: Their validity of use. *American Psychologist*, 52(11), 1218–1225.

Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. MT (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity*, 43(2), 197-209.

Seldin, P. (1999). *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions (Vol. 10).* Jossey-Bass.

Sliusarenko, T., Ersbøll, B. K. & Clemmensen, L. K. H. (2013). *Quantitative assessment of course evaluations.* Doctoral dissertation, Technical University of Denmark Danmarks Tekniske Universitet, Department of Informatics and Mathematical Modeling Institut for Informatik og Matematisk Modellering.

Thorpe, S. W. (2002). *Online student evaluation of instruction: An investigation of non-response bias.* 42nd Annual Forum of the Association for Institutional Research.

Checking quality of sensory data by assessing intra/inter panelist agreement

Amalia Vanacore[1] and Maria Sole Pellegrino[2]

Abstract This study aims at checking the quality of sensory data by evaluating and testing both panelist precision and panel reproducibility via an agreement index-based approach which has been already adopted for the assessment of rater reliability but is almost unexplored in the field of sensory analysis. The approach has been applied to a case study concerning the assessment of sensory characteristics induced by different food and beverages.

Key words: panelist precision, panel reproducibility, agreement coefficient

1 Introduction

In many contexts (*e.g.*, food and beverage, cosmetics, service) product development and quality improvement processes rely on sensory data provided by a panel of (expert or novice) assessors. Testing panelist/panel reliability is a common methodological requirement in order to guarantee the quality of sensory data.

Different methodologies, based on univariate or multivariate analysis, have been proposed in the specialized literature to assess panel/panelist reliability in terms of consonance [7], sensitivity and discrimination ability [1, 4] or inconsistency [10, 12]. A quite different approach evaluates panel/panelist reliability in terms of repeatability and reproducibility measures defined through variability indexes [11] or correlational (or ANOVA) reliability indexes — such as the ICC [2] — or agreement coefficients [6, 9].

Following the index-based approach, this study aims at checking the quality of sensory data by evaluating and testing both panelist precision and panel reproducibility. Specifically, the former is measured as the degree of agreement within evaluations provided by the panelist in different occasions, whereas the latter is

[1] Amalia Vanacore, Department of Industrial Engineering, University of Naples Federico II; email: amalia.vanacore@unina.it

[2] Maria Sole Pellegrino, Department of Industrial Engineering, University of Naples Federico II; email: mariasole.pellegrino@unina.it

measured as the degree of agreement across the evaluations provided by the whole panel. The usefulness of the above measures of precision and reproducibility is illustrated through a case study concerning the assessment of sensory dimensions induced by similar food and beverage products.

2 Method

The weighted Brennan-Prediger coefficient [3] is a rescaled measure of the proportion of observed agreement corrected for the agreement expected under the assumption of uniform chance measurements.

Using the appropriate formulation for observed agreement, the weighted Brennan-Prediger coefficient can be adopted as a measure of panelist precision as well as panel reproducibility. Specifically, for panelist precision:

$$\hat{K}_{W|p} = \frac{\sum_{i,j=1}^{k} w_{ij}\, n_{ij} \big/ n - 1\big/k^2 \sum_{i,j=1}^{k} w_{ij}}{1 - 1\big/k^2 \sum_{i,j=1}^{k} w_{ij}} \tag{1}$$

whereas for panel reproducibility:

$$\hat{K}_{W|r} = \frac{1\big/n \sum_{l=1}^{n} \sum_{i=1}^{k} \left(r_{li} \left(\sum_{j=1}^{k} r_{lj} w_{ij} - 1 \right) \right) \big/ \left(r \left(r-1 \right) \right) - 1\big/k^2 \sum_{i,j=1}^{k} w_{ij}}{1 - 1\big/k^2 \sum_{i,j=1}^{k} w_{ij}} \tag{2}$$

being n the number of items rated twice (*i.e.,* two replications) on an ordinal k-points rating scale (with $k > 2$) by each panelist, n_{ij} the number of items classified into i^{th} category in the first replication and into j^{th} category in the second replication, w_{ij} the corresponding symmetrically weight (*i.e.,* $w_{ij=}$ w_{ji}), r the number of panelists, r_{li} the number of panelists who rated items l into category i.

In order to account for sampling uncertainty, a reliable characterization of the extent of panelist precision and panel reproducibility can be obtained by building a Bias-Corrected and Accelerated bootstrap confidence interval (hereafter, BC$_a$ CI) for $\hat{K}_{W|p}$ and $\hat{K}_{W|r}$ [5, 13]. The lower and upper bounds of the (1-α)% two-sided BC$_a$ confidence interval are defined as:

$$\boldsymbol{G}^{-1}\left(\Phi\left(\boldsymbol{b} \pm \left(\boldsymbol{z}_{\alpha/2} \pm \boldsymbol{b} \right) \big/ \left[1 + \boldsymbol{a} \left(\mp \boldsymbol{z}_{\alpha/2} - \boldsymbol{b} \right) \right] \right) \right) \tag{3}$$

being Φ the cumulative distribution function of the normal distribution, $z_{\alpha/2}$ the $\alpha/2$ percentile of the standard normal distribution, b the bias correction parameter and a the acceleration parameter.

3 Case Study

The data refer to a discrimination test conducted on different food and beverage products, involving a panel of untrained consumers who were asked to evaluate in two different occasions 5 sensory dimensions (*viz.*, appearance, taste, smell, texture and general impression) on an ordinal rating scale.

Panelist precision and panel reproducibility were characterized by calculating $\hat{K}_{W|p}$ and $\hat{K}_{W|r}$ for every dimension together with their BCa CIs, all represented in Figure 1.

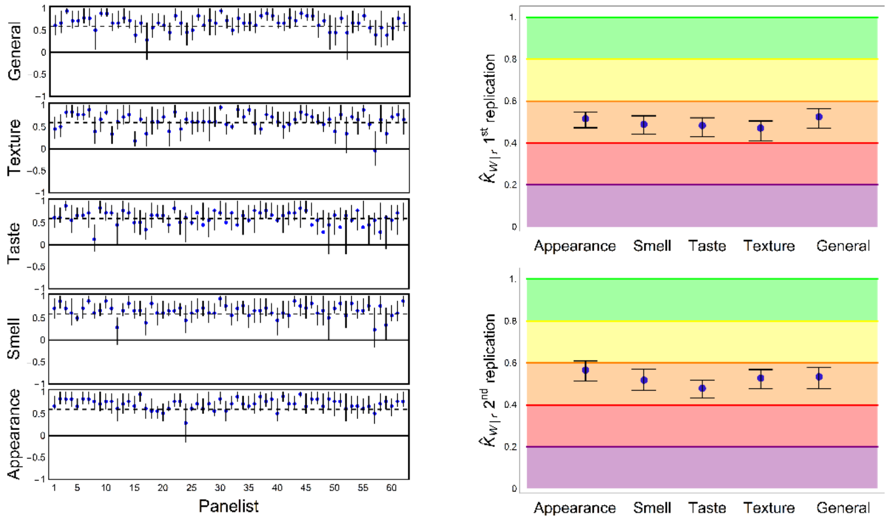


Figure 1: *Point estimates and BCa CIs of* $\hat{K}_{W|p}$ *(left side)* and $\hat{K}_{W|r}$ *(right side) for each sensory dimension*

In order to test for panelist precision and panel reproducibility, the lower bound of BCa CIs of $\hat{K}_{W|p}$ and $\hat{K}_{W|r}$, were benchmarked against the well-known Landis and Koch scale [8]. Figure 1 shows the benchmark results: the 53% of panelists can be reasonably assumed as substantially precise assessors for *Appearance* (*i.e.*, the lower bound of BCa CI of $\hat{K}_{W|p}$ is greater than the threshold value 0.60 represented in the diagram by the dashed line) whereas only the 24% of them can be reasonably assumed as substantially precise in assessing *Taste*. On the other hand, the panel is always moderately reproducible because the BCa CIs of $\hat{K}_{W|r}$ — built for all dimensions and both replications — always belong to moderate agreement category.

The Cochran's Q test uncovered significant differences among the proportions of substantially precise panelists for the five sensory dimensions (**Q** = 17.176, *d.f.* = 4, *p* = 0.001786). Pairwise comparisons using continuity-corrected McNemar's tests with Bonferroni correction revealed that significantly more panelists resulted substantially precise for *Appearance* than for *Taste* (*p*-adjusted = 0.005202). Specifically, the effect-size, measured by an approximate 95% confidence interval for the difference between marginal proportions, highlighted that the proportion of substantially precise panelists for *Appearance* may be up to 33% higher than for *Taste*.

4 Conclusions

The agreement index-based approach supported by a non-parametric inferential procedure can be usefully adopted as an effective strategy for checking the quality of sensory data. Indeed it provides useful information in order to decide whether a panelist can be assumed as a substantially precise assessor and, moreover, it allows to pinpoint the sensory dimension(s), if any, for which panelists need to be trained.

References

1. Bi, J.: Agreement and reliability assessments for performance of sensory descriptive panel. J Sens Stud, 18, 61–76 (2003)
2. Bi, J., Kuesten, C.: Intraclass correlation coefficient (ICC): A framework for monitoring and assessing performance of trained sensory panels and panelists. J Sens Stud (2012), 27.5, 352-364.
3. Brennan, R. L., Prediger, D. J.: Coefficient Kappa: Some Uses, Misuses, and Alternatives. EPM, 41, 687–699 (1981)
4. Brockhoff, P. B.: Statistical testing of individual differences in sensory profiling. Food Qual Prefer, 14.5, 425-434 (2003)
5. Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statist. Med, 19, 1141—1164 (2000)
6. Falahee, M., MacRae, A. W.: Perceptual variation among drinking waters: The reliability of sorting and ranking data for multidimensional scaling. Food Qual Prefer, 8(5–6), 389–394 (1997)
7. Kermit, M., Lengard, V.: Assessing the performance of a sensory panel – Panelist monitoring and tracking. Nedre Vollgate, 8, 0185 (2006)
8. Landis, J. R., Koch, G. G.: The measurement of observer agreement for categorical data. Biometrics. 33.1, 159--174 (1977)
9. Latreille, J., *et al.*: Measurement of the reliability of sensory panel performances. Food Qual Prefer, 17, 369–375 (2006)
10. Lundahl, D. S., McDaniel, M. R.: Use of contrasts for the evaluation of panel inconsistency. J Sens Stud (1990), 5(4), 265–277
11. Rossi, F.: Assessing sensory panelist performance using repeatability and reproducibility measures. Food Qual Prefer, 12, 467–497 (2001)
12. Sivertsen, H., Risvik, E.: A study of sample and assessor variation: a multivariate study of wine profiles. J Sens Stud, 9, 293–312 (1994)
13. Vanacore, A., Pellegrino, M. S.: Characterizing the extent of rater agreement via a non-parametric benchmarking procedure. In: Proceedings of SIS2017 "Statistics and Data Science: new challenges, new generations", University of Florence, June 28-30, 2017.

Quality & Quantity
https://doi.org/10.1007/s11135-018-0807-5

Checking quality of sensory data via an agreement-based approach

Amalia Vanacore[1] · Maria Sole Pellegrino[1]

Abstract
Sensory evaluations are adopted in many fields for measuring and comparing sensory properties of products and improving their quality. The selection of panelists able to provide precise evaluations is a crucial issue to perform reliable sensory analysis. An agreement-based approach is here suggested in order to assess the quality of sensory data in terms of both panelist repeatability and panel reproducibility. The approach has been applied to two case studies involving untrained sensory panelists and trained teaching quality assessors, respectively. The results of the case studies show that although reproducibility can be assumed *moderate* for both groups of raters, repeatability is generally higher for the group of trained raters.

Keywords Panelist repeatability · Panel reproducibility · Kappa-type coefficient · Quality assessment of sensory data

1 Introduction

Sensory data are obtained by collecting with a scientific method humans perceptions expressed with respect to some product characteristics. The expressed perception is the result of a human decision, reasonable assumed as the outcome of complex interactions conditioned by personal history, environmental variables, subjective factors (or covariates), product characteristics and also survey conditions (e.g. survey design and format adopted for data collection) (Manisera et al. 2011).

The evaluative abilities of the sensory panel are of paramount importance in order to guarantee the quality and thus the validity of the provided sensory evaluations (Bi 2003; Kermit and Lengard 2005; Latreille et al. 2006; Iannario et al. 2012): only a good sensory panel provides accurate, precise and discriminating data that reflect some intrinsic and true values associated with the products (Piggott 1995; King et al. 2001; Kermit and Lengard 2005; Pinto et al. 2014). In sensory analysis accuracy refers to panel ability to score products the same, on average, as the other panel members (i.e. reproducibility, Rossi 2001);

✉ Amalia Vanacore
amalia.vanacore@unina.it

[1] Department of Industrial Engineering, University of Naples "Federico II", p.le Tecchio 80, 80125 Naples, Italy

Published online: 14 August 2018

precision refers to panelist ability to provide consistent ratings for the same product during different replications over time (i.e. repeatability, Rossi 2001; Pinto et al. 2014); whereas discrimination ability refers to panelist ability to recognize very small (i.e. just noticeable) differences between products (Pinto et al. 2014).

Different approaches have been proposed over the years to check the quality of sensory data by monitoring individual panelist performance as well as the panel as a whole. The most widespread approach uses standard analysis of variance (ANOVA) to investigate panelist accuracy, precision and discrimination ability (Næs and Solheim 1991; Schlich 1994; Lea et al. 1995; Brockhoff 2003) or to evaluate panel inconsistency (Lundahl and McDaniel 1990, 1991).

Despite its popularity, standard ANOVA is not suitable for sensory evaluations which—because of their "qualitative" nature—are mainly expressed on nominal or ordinal rating scales; however ANOVA approach could take great advantage from a recently proposed unifying approach for assessing variation over every scale of measurement (Gadrich and Bashkansky 2012; Gadrich et al. 2015).

A quite different approach for the assessment of the quality of sensory data is adopted in the Repeatability and Reproducibility (R&R) study proposed by Rossi (2001) which, exploiting the analogy between analytical laboratory measures and sensory panelist evaluations, assesses the performance of sensory panelist in terms of accuracy and precision through descriptive statistics.

Starting from the definition of repeatability and reproducibility proposed by Rossi (2001), this paper follows an agreement-based approach (Cohen 1960; Ludbrook 2002; Vanbelle 2009) to assess and characterize panelist performance in providing sensory evaluations on ordinal rating scales.

The choice of assessing panelist repeatability and panel reproducibility in terms of agreement is coherent with the definitions of precision and accuracy provided by International Organization for Standardization and with those adopted in sensory science. Specifically, ISO 5725 (1994) defines precision as *the closeness between independent test results obtained under stipulated conditions* and accuracy as *the closeness between the new measurement and the truth or true value*. Therefore the precision of sensory data can be assessed as the degree of agreement between replicated (i.e. over different times) evaluations provided by the same panelist under the same conditions, where "same conditions" means that nothing changed other than the times of the evaluations. Vice-versa, the common concept of accuracy cannot be straightforward operationalized for sensory data because panelists' evaluations, being subjective, lack a gold-standard against which to check their trueness. In such circumstances, the gold-standard is replaced by the whole panel perception and the accuracy can be thus assessed as the degree of agreement among the evaluations simultaneously provided by the whole panel under the same conditions.

Though widespread in many fields of research (e.g. medicine, psychological and educational measurement), the agreement-based approach is still almost unexplored in sensory analysis. This approach estimates repeatability and reproducibility via kappa-type agreement coefficients, which are rescaled measures of the observed proportion of agreement corrected with the agreement expected by chance alone.

Several kappa-type agreement coefficients have been proposed in the literature differing from each other only in the definition, and thus formulation, of the agreement expected by chance alone. This paper adopts the Brennan–Prediger agreement coefficient (Brennan and Prediger 1981) which formulates the agreement expected by chance alone assuming a uniform distribution for chance measurement, that is the most non-informative measurement system given a certain rating scale (Mast 2007; Erdmann et al. 2015).

The magnitude of agreement is commonly qualified by comparing the estimated coefficient against an arbitrary benchmark scale. However, this straightforward benchmarking procedure does not consider estimate uncertainty and, moreover, it should be treated with caution for comparison across studies when the experimental conditions are not the same (Gwet 2014).

The proposed agreement-based approach is fully exploited through a case study aimed at evaluating the quality of a sensory panel, in terms of panelist repeatability and panel reproducibility, assessing five sensory dimensions of eight food products. Moreover, in order to demonstrate the applicability of the proposed approach to different evaluation contexts involving human raters, a second case study, concerning the assessment of service quality by trained assessors, is also discussed.

The remainder of this paper is organized as follows: two linear weighted Brennan–Prediger agreement coefficients—one for estimating panelist repeatability and the other for panel reproducibility—and a statistical benchmarking procedure for both the adopted coefficients are introduced in Sect. 2; Section 3 is devoted to the discussion of the two case studies aimed at illustrating the usefulness of the proposed procedure; finally, conclusions are summarized in Sect. 4.

2 Methods

2.1 Assessment of panelist repeatability and panel reproducibility via linear weighted Brennan–Prediger coefficient

Let n be the number of products rated two or more times (i.e. replications) by a panelist on an ordinal k-point rating scales, with $k > 2$. Panelist's ratings are denoted Y_{hr}, with $h = 1, \dots, n$ indexing products and r indexing replications. Of interest for the assessment of panelist repeatability is the joint distribution of the Y_{hr}, which in the special case of two replications can be cross-classified into a $k \times k$ contingency table $(n_{ij})_{k \times k}$, where the generic (i, j) cell contains the joint frequency n_{ij} that counts the number of products classified into ith category over the first replication and into jth category over the second replication (Table 1).

It is clear that the cells along the main diagonal represent the perfect match between the evaluations provided in the two replications, whereas the off-diagonal cells represent mismatch. The introduction of a distance metric or of a weighting scheme enables to

Table 1 $k \times k$ table for classifying the ratings of a panelist in two replications

		2^{nd} replication					
	Category	1	...	j	...	k	**Total**
1^{st} replication	1	n_{11}	...	n_{1j}	...	n_{1k}	$n_{1\cdot}$
	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
	i	n_{i1}	...	n_{ij}	...	n_{ik}	$n_{i\cdot}$
	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
	k	n_{k1}	...	n_{kj}	...	n_{kk}	$n_{k\cdot}$
	Total	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot k}$	n

account that in the case of ordinal rating scale some disagreements are more serious than others, that is disagreement on two distant categories should be considered more important than disagreement on neighbouring categories. Different kinds of distance metrics and weighting schemes, appropriate for various practical situations, have been proposed and discussed in the literature. Typically, these metrics are expressed as non decreasing functions of $|i-j|$ when assessing the degree of disagreement among the provided evaluations [e.g. loss matrix (Bashkansky et al. 2008)] or, vice-versa, as nonincreasing function of $|i-j|$ when assessing the degree of agreement [e.g. linear agreeing weights (Cicchetti and Allison 1971), quadratic agreeing weights (Fleiss et al. 2013)]. Adopting the linear weighting scheme, panelist repeatability will be assessed using the weighted Brennan–Prediger agreement coefficient (Brennan and Prediger 1981) which formulates the weighted observed proportion of agreement p_{a_w} as:

$$p_{a_w} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\frac{n_{ij}}{n} \tag{1}$$

where w_{ij} is the symmetrical ($w_{ij} = w_{ji}$) agreement weight a priori assigned to each pair (i, j) of ratings. Specifically, w_{ij} ranges between 0 and 1: the minimum value 0 is assigned to maximally disagreeing pairs of ratings (i.e. classified in cells (1, k) and (k, 1) in Table 1); the maximum value 1 is assigned to pairs of coincident ratings (i.e. classified in cells (i, i) along the main diagonal in Table 1). It is worthwhile to pinpoint that although the weights can be arbitrary defined, the linear and quadratic weights are the most commonly used weighting schemes for kappa-type coefficients and are formulated as follows:

$$w_{ij}^{L} = 1 - \frac{|i-j|}{k-1}; \quad w_{ij}^{Q} = 1 - \frac{(i-j)^2}{(k-1)^2}. \tag{2}$$

It is worthy to note that when the classification provided by a human rater is compared against a gold-standard, the observed proportion of agreement (Eq. 1) corresponds to the weighted percentage of correct classification, which is a common measure of prediction success (Veall and Zimmermann 1992).

The weighted agreement expected by chance alone $p_{a|c_w}$, formulated assuming a uniform distribution for chance measurement, is given by:

$$p_{a|c_w} = \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}. \tag{3}$$

Panelist repeatability can be thus assessed adopting the following formulation:

$$\widehat{BP}_w = \frac{\sum_{i,j=1}^{k} w_{ij}n_{ij}/n - 1/k^2 \cdot \sum_{i,j=1}^{k} w_{ij}}{1 - 1/k^2 \cdot \sum_{i,j=1}^{k} w_{ij}}. \tag{4}$$

Adopting the agreement-based approach, panel reproducibility is assessed in terms of the agreement across the evaluations provided by the whole panel via a proper multiple-raters' version of the linear weighted Brennan–Prediger coefficient (Gwet 2014).

Let r be the total number of panelists who rated the n products over the same k-point ordinal rating scale. Of interest for the assessment of panel reproducibility are the proportions of pairwise agreement, which can be obtained by classifying the Y_{hr} into a $n \times k$ table

$(r_{li})_{n\times k}$, where the generic (l, i) cell contains the number of panelists r_{li} who classified product l into category i (Table 2).

The weighted observed proportion of agreement among multiple panelists is thus formulated as follows:

$$p^{r}_{a_w} = \frac{1/n \cdot \sum_{l=1}^{n} \sum_{i=1}^{k} r_{li}\left(r^{*}_{li} - 1\right)}{r(r-1)} \tag{5}$$

where

$$r^{*}_{li} = \sum_{j=1}^{k} w_{ij} r_{lj} \tag{6}$$

whereas the weighted agreement expected by chance alone is still formulated as in Eq. 3, since it depends only on the adopted weighting scheme and on the rating scale dimension.

Panel reproducibility can be thus assessed adopting the following formulation:

$$\widehat{BP}^{r}_{w} = \frac{\left(1/n \cdot \sum_{l=1}^{n} \sum_{i=1}^{k} r_{li}\left(r^{*}_{li} - 1\right)\right)/(r(r-1)) - 1/k^{2} \cdot \sum_{i,j=1}^{k} w_{ij}}{1 - 1/k^{2} \cdot \sum_{i,j=1}^{k} w_{ij}} \tag{7}$$

2.2 A statistical benchmarking procedure for characterizing the extent of panelist repeatability and panel reproducibility

The magnitude of agreement coefficient is most commonly related to the notion of extent of agreement by a straightforward comparison with a benchmark scale. A number of benchmarking scales have been proposed mainly in social and medical sciences over the years (e.g. Landis and Koch 1977; Altman 1990; Fleiss et al. 2013); among them the most widely adopted is the six range scale proposed by Landis and Koch (Table 3).

The straightforward procedure is commonly adopted for benchmarking purpose, nevertheless it can be misleading for two main reasons:

- it fails to consider that an agreement coefficient, as any other sampling estimate, is imprecise (i.e. the sample statistic is affected by sampling uncertainty): almost certainly a different agreement estimate will be obtained if the survey is repeated under identical

Table 2 $n \times k$ table for classifying the ratings provided by the whole panel in one replication

		Category					
		1	...	i	...	k	**Total**
Product	**1**	r_{11}	...	r_{1i}	...	r_{1k}	r
	⋮	⋮	...	⋮	...	⋮	r
	l	r_{i1}	...	r_{li}	...	r_{lk}	r
	⋮	⋮	...	⋮	...	⋮	r
	n	r_{n1}	...	r_{ni}	...	r_{nk}	r

Table 3 Landis and Koch benchmark scale for kappa-type coefficients

Coefficient Magnitude	Strength of agreement
≤ 0.00	*Poor*
0.01–0.20	*Slight*
0.21–0.40	*Fair*
0.41–0.60	*Moderate*
0.61–0.80	*Substantial*
0.81–1.00	*Almost perfect*

conditions on different samples drawn from the same population of items (Gardner and Altman 1986);

- it does not allow to compare the extent of agreement across different studies, unless they are carried out under the same experimental conditions (i.e. number of rated products, number of categories or distribution of products across the categories).

In order to overcome these criticisms, a statistical benchmarking procedure is recommended. The simplest and intuitive way to accomplish this task is by building a confidence interval of the agreement coefficient and comparing its lower bound against an adopted benchmark scale (Vanacore and Pellegrino 2017). A confidence interval suitable for both small (the most affordable size in many sensory experiments) and large samples can be obtained via bootstrap resampling. Among the available methods to build bootstrap confidence intervals (Carpenter and Bithell 2000), the percentile bootstrap is the simplest and the most popular one. On the other hand, the Bias-Corrected and Accelerated bootstrap (hereafter, BCa) confidence interval is recommended for severely skewed distribution. Despite the higher computational complexity, BCa confidence intervals have generally smaller coverage errors. The lower bound of the $(1-2\alpha)\%$ two-sided BCa confidence interval is defined as:

$$LB_{\mathrm{BCa}} = G^{-1}\left(\Phi\left(b - \frac{z_\alpha - b}{1 + a(z_\alpha - b)}\right)\right) \tag{8}$$

being G the cumulative distribution function of the bootstrap replications of the kappa-type coefficient, Φ the standard normal CDF, b the bias correction parameter and a the acceleration parameter.

3 Two illustrative case studies

Two case studies are hereafter presented involving untrained as well as trained raters. Specifically, in the first case study untrained consumers provided sensory evaluations about some food and beverage products, whereas in the second case study a class of university students, trained in evaluating teaching quality, rated a university teaching course.

3.1 Case study 1: consumers as sensory panelists

The first case study is aimed at checking the quality of sensory data by assessing panelist repeatability and panel reproducibility via the proposed procedure. The analyzed data have been published by Geier et al. (2016) and were obtained by performing some experiments of consumer sensory evaluation according to German standard DIN 10974.

Specifically, the data refer to a hedonic test involving a panel of $r = 62$ untrained consumers who were asked to evaluate in two different evaluation sessions 5 sensory dimensions (viz. appearance, taste, smell, texture and general impression) on a hedonic $k = 7$ -points rating scale (1: very bad–7: excellent). The panelists rated $n = 8$ different food and beverage products (i.e. four pairs of water, milk, bread, sugar) in identical product conditions (viz. means, temperature, dishes, portion sizes, and test-booth conditions).

For each sensory dimension panelist repeatability is assessed via $\widehat{BP}_w$ (Eq. 4) whereas panel reproducibility is assessed for each evaluation session and sensory dimension via $\widehat{BP}^r_w$ (Eq. 7); the extent of both panelist repeatability and panel reproducibility is then characterized by benchmarking the lower bound of each 95% BCa confidence interval (i.e. $BP_{w|l}$ and $BP^r_{w|l}$) against the Landis and Koch scale (Table 3).

The $\widehat{BP}_w$ estimates together with their BCa confidence interval are plotted in Fig. 1 for each panelist and sensory dimension, where the dashed lines represent the threshold value for *substantial* agreement.

The results show that benchmarking the lower bound of the BCa confidence interval for BP_w, the percentage of at least *substantially* repeatable panelists (i.e. $BP_{w|l} > 0.6$) ranges, across the 5 sensory dimensions, between 24 and 49%; the sensory dimensions with the best and worst panelist performance are Appearance and Taste, respectively. Specifically, the null hypothesis of *substantial* repeatability for Appearance and Taste can be rejected (with $\alpha = 0.025$) for 32 and 48 panelists, respectively.

Regarding panel reproducibility, $\widehat{BP}^r_w$ estimates and their 95% BCa confidence intervals for each sensory dimension and each evaluation session are plotted in Fig. 2 against the 5 categories of positive agreement of Landis and Koch scale (Table 3).

The results suggest that the panel is *moderately* reproducible, indeed, for every sensory dimension, the 95% BCa confidence intervals of panel reproducibility belong to the category of *moderate* agreement for both evaluation sessions.

It is worthy to note that in the first evaluation session the panel reproducibility is comparable across sensory dimensions; vice-versa, in the second session Appearance shows a significantly higher panel reproducibility than Taste.

3.2 Case study 2: students as teaching quality assessors

The second case study describes the results of an evaluation experiment involving a class of $r = 18$ university students homogeneous in curriculum and instruction, who assessed the quality of a teaching course scheduled at the last year of their carrier path. The students rated $n = 20$ teaching quality items using a verbal rating scale (VRS) with $k = 4$ grades “strong disagreement”, “disagreement”, “agreement” and “strong agreement”. Each student rated the same items twice, during two lessons, one week apart.

According to the purpose of the case study, the evaluations provided by each student during the two lessons are used to assess student’s repeatability via $\widehat{BP}_w$ (Eq. 4), whereas those collected during each lesson by the whole class of students are used to assess students reproducibility via $\widehat{BP}^r_w$ (Eq. 7); the extents of repeatability and reproducibility are then characterized by benchmarking the lower bound of the 95% BCa confidence interval (i.e. $BP_{w|l}$ and $BP^r_{w|l}$; Eq. 8) against the Landis and Koch scale (Table 3).

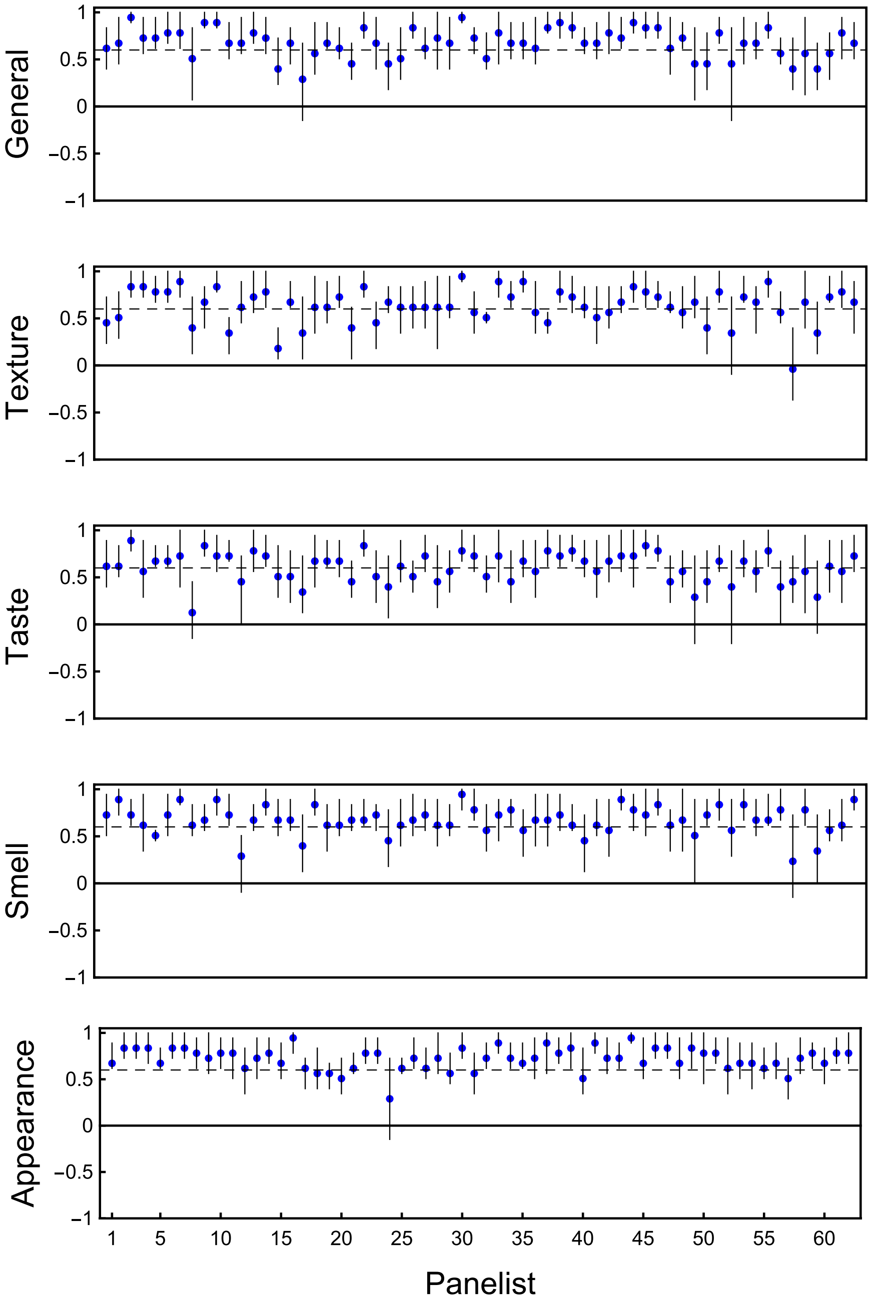
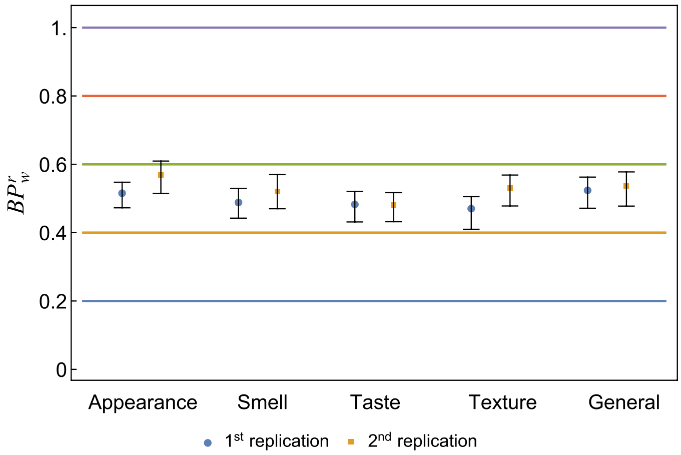


Fig. 1 Point estimates and 95% BCa confidence intervals of panelist repeatability $(\widehat{BP}_w)$ for each sensory dimension

Fig. 2 Point estimates and 95% BCa confidence intervals of panel reproducibility $(\widehat{BP}^r_w)$ for each sensory dimension and each evaluation session



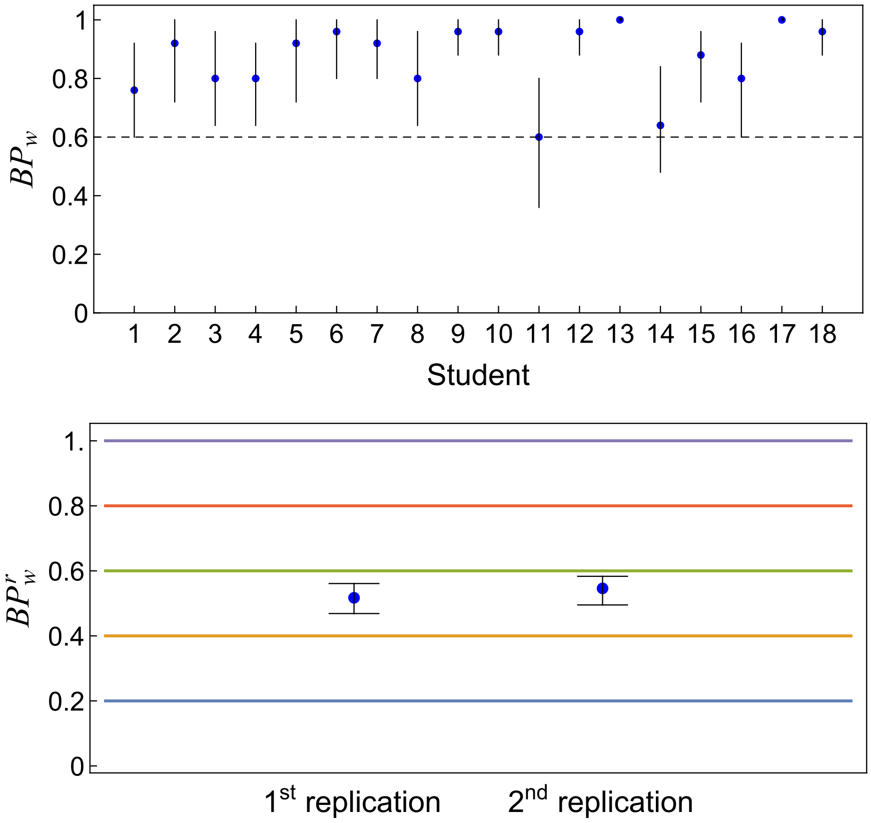


Fig. 3 Point estimates ($\widehat{BP}_w$ and $\widehat{BP}_w^r$) and 95% BCa confidence intervals for repeatability of each student (on the top) and for reproducibility of each replication (on the bottom)

In the top of Fig. 3 the $\widehat{BP}_w$ estimates and their 95% BCa confidence intervals are plotted for each student participating in the experiment; in the bottom of Fig. 3, instead, the 95% BCa confidence intervals for students reproducibility and the $\widehat{BP}_w^r$ estimates for both replications over time are graphically represented against the 5 categories of positive agreement of Landis and Koch scale (Table 3).

Assuming $\alpha = 0.025$, the hypothesis that the student repeatability is at least *moderate* can be rejected only for the student #11 whose $BP_{w|l}$ is lower than 0.4, whereas the hypothesis of at least *substantial* repeatability cannot be rejected for 14 out of 18 involved students (i.e. the 78% of students participating in the case study) since they have a $BP_{w|l}$ lying in the range 0.4–0.6.

Regarding the reproducibility level, the 95% BCa confidence intervals of BP_w^r belong to the category of *moderate* agreement, thus the analysed class of students can be assumed *moderately* reproducible.

3.3 Power analysis results

The results of both case studies achieve an adequate power as confirmed by the power analysis obtained via a Monte Carlo simulation study. Specifically, the simulation study has been developed considering one rater classifying $n = 8$ items into one of $k = 7$ possible ordinal rating categories and $n = 20$ items into $k = 4$ ordinal categories during two replications. The data have been simulated by sampling $r = 2000$ Monte Carlo data sets from a multinomial distribution with parameters n and $\mathbf{p} = (\pi_{11}, \ldots, \pi_{ij}, \ldots, \pi_{kk})$, with the π_{ij} values set so as to obtain six true population values of agreement (viz. 0.4, 0.5, 0.6, 0.7, 0.8, 0.9), assuming a linear weighting scheme (i.e. w_{ij}^L in Eq. 2).

The statistical power has been computed for three different hypothesis statements referring to null inference of at least *slight* agreement (i.e. $\mathrm{H}_1 : BP_w > 0$), non-null inference of at least *moderate* agreement (i.e. $\mathrm{H}_1 : BP_w > 0.40$) and non-null inference of at least *substantial* agreement (i.e. $\mathrm{H}_1 : BP_w > 0.60$). All the hypothesis tests have been conducted assuming a significance level $\alpha = 0.025$.

The Monte Carlo estimate of the statistical power is given by:

$$1 - \hat{\beta} = \frac{1}{r} \sum_{h=1}^{r} I\left[BP^h_{w|l} > BP^C_w | \mathrm{H}_1\right] \tag{9}$$

being I [·] an indicator taking value 1 if the argument is true and 0 otherwise, $BP^h_{w|l}$ the lower bound of the 95% BCa confidence interval obtained from the *h*th Monte Carlo data set and BP^C_w the tested critical value of panelist repeatability under the stated hypothesis.

The power curves obtained in the null inference and non-null inference cases of at least *moderate* and *substantial* agreement for both the analysed scenarios (i.e. $n = 8$, $k = 7$; $n = 20$, $k = 4$) are reported in Fig. 4. The power curves show that the statistical power of 80% is obtained when testing an agreement level at least *substantial* (i.e. $\mathrm{H}_1 : BP_w > 0.60$) against the null hypothesis of *poor* agreement (i.e. $\mathrm{H}_0 : BP_w = 0$), when testing an *almost perfect* agreement level (i.e. $\mathrm{H}_1 : BP_w > 0.80$) against the null hypothesis of no more than *fair* agreement (i.e. $\mathrm{H}_0 : BP_w \leq 0.40$) and, finally, when testing a very high agreement level of $BP_w > 0.90$ against the null hypothesis of no more than *moderate* agreement (i.e. $\mathrm{H}_0 : BP_w \leq 0.60$).

4 Conclusions

This study suggests the adoption of an agreement-based approach for the assessment of panelist repeatability and panel reproducibility. In order to demonstrate the applicability of the proposed approach to different evaluation contexts involving human raters, two illustrative case studies have been presented: the first case study analyses sensory evaluations provided by untrained consumers, instead the second case study deals with quality evaluations provided by trained assessors.

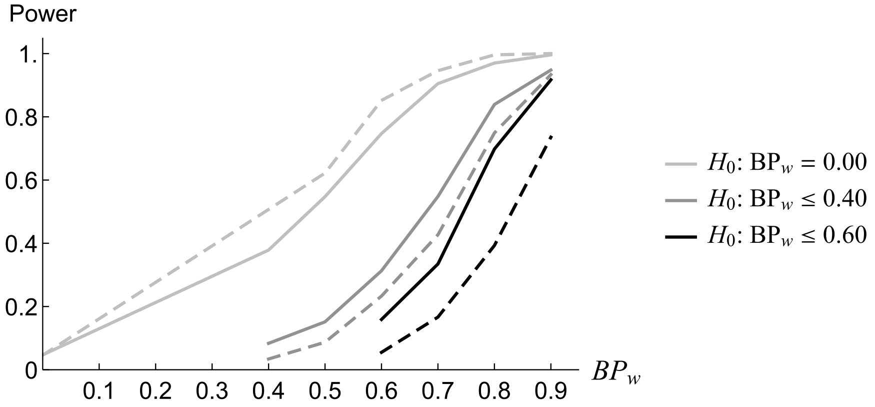


Fig. 4 Statistical power curves obtained benchmarking the lower bound of the BCa confidence interval in null and non-null inference conditions (solid curves for $n = 8$ and $k = 7$, dashed curves for $n = 20$ and $k = 4$)

The results of both case studies seem to suggest to be not completely confident in the accuracy of human raters' evaluations, being the reproducibility level always no more than *moderate*.

Differences between the two case studies concern rater precision, evaluated in terms of her/his repeatability. Specifically, the results of the first case study show that consumer's repeatability changes across the sensory dimensions being better for Appearance and worse for Taste; less than 50% of the consumers were able to provide *substantially* repeatable evaluations for at least one sensory dimension and only 9 of them provided *substantially* repeatable evaluations for all sensory dimensions. The results of the second case study would suggest that the higher expertise of students as quality assessors—achieved through their frequent involvement in evaluation processes on teaching quality—makes the majority of them at least *substantially* repeatable raters.

The different results obtained in terms of repeatability for untrained and trained raters highlight the effectiveness of the proposed approach in discriminating the performances of trained and untrained raters and its usefulness in identifying the most critical sensory dimension(s), if any, to which the training effort should be addressed.

Acknowledgements The authors deeply thank the two anonymous referees for their careful reading of the manuscript and helpful suggestions.

References

Altman, D.G.: Practical Statistics for Medical Research. CRC Press, Boca Raton (1990)

Bashkansky, E., Dror, S., Ravid, R., Grabov, P.: Effectiveness of a product quality classifier. Qual. Control Appl. Stat. **53**(3), 291–292 (2008)

Bi, J.: Agreement and reliability assessments for performance of sensory descriptive panel. J. Sens. Stud. **18**(1), 61–76 (2003)

Brennan, R.L., Prediger, D.J.: Coefficient kappa: some uses, misuses, and alternatives. Educ. Psychol. Meas. **41**(3), 687–699 (1981)

Brockhoff, P.B.: Statistical testing of individual differences in sensory profiling. Food Qual. Prefer. **14**(5–6), 425–434 (2003)

Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat. Med. **19**(9), 1141–1164 (2000)

Cicchetti, D.V., Allison, T.: A new procedure for assessing reliability of scoring eeg sleep recordings. Am. J. EEG Technol. **11**(3), 101–110 (1971)

Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**(1), 37–46 (1960)

De Mast, J.: Agreement and kappa-type indices. Am. Stat. **61**(2), 148–153 (2007)

Erdmann, T.P., De Mast, J., Warrens, M.J.: Some common errors of experimental design, interpretation and inference in agreement studies. Stat. Methods Med. Res. **24**(6), 920–935 (2015)

Fleiss, J.L., Levin, B., Paik, M.C.: Statistical methods for rates and proportions. Wiley (2013)

Gadrich, T., Bashkansky, E.: Ordanova: analysis of ordinal variation. J. Stat. Plan. Inference **142**(12), 3174–3188 (2012)

Gadrich, T., Bashkansky, E., Zitikis, R.: Assessing variation: a unifying approach for all scales of measurement. Qual. Quant. **49**(3), 1145–1167 (2015)

Gardner, M.J., Altman, D.G.: Confidence intervals rather than P values: estimation rather than hypothesis testing. Br. Med. J. (Clin Res Ed) **292**(6522), 746–750 (1986)

Geier, U., Büssing, A., Kruse, P., Greiner, R., Buchecker, K.: Development and application of a test for food-induced emotions. PLoS ONE **11**(11), 1–17 (2016)

Gwet, K.L.: Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Advanced Analytics, LLC, Gaithersburg (2014)

Iannario, M., Manisera, M., Piccolo, D., Zuccolotto, P.: Sensory analysis in the food industry as a tool for marketing decisions. Adv. Data Anal. Classif. **6**(4), 303–321 (2012)

International Organization for Standardization (ISO). Accuracy (Trueness and Precision) of Measurement Methods and Results Part 1: General Principles and Definitions (5725-1). Geneva: ISO (1994)

Kermit, M., Lengard, V.: Assessing the performance of a sensory panel–panellist monitoring and tracking. J. Chemom. **19**(3), 154–161 (2005)

King, M.C., Hall, J., Cliff, M.A.: A comparison of methods for evaluating the performance of a trained sensory panel. J. Sens. Stud. **16**(6), 567–581 (2001)

Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)

Latreille, J., Mauger, E., Ambroisine, L., Tenenhaus, M., Vincent, M., Navarro, S., Guinot, C.: Measurement of the reliability of sensory panel performances. Food Qual. Prefer. **17**(5), 369–375 (2006)

Lea, P., Rødbotten, M., Næs, T.: Measuring validity in sensory analysis. Food Qual. Prefer. **6**(4), 321–326 (1995)

Ludbrook, J.: Statistical techniques for comparing measurers and methods of measurement: a critical review. Clin. Exp. Pharmacol. Physiol. **29**(7), 527–536 (2002)

Lundahl, D.S., McDaniel, M.R.: Use of contrasts for the evaluation of panel inconsistency. J. Sens. Stud. **5**(4), 265–277 (1990)

Lundahl, D.S., McDaniel, M.R.: Influence of panel inconsistency on the outcome of sensory evaluations from descriptive panels. J. Sens. Stud. **6**(3), 145–157 (1991)

Manisera, M., Piccolo, D., Zuccolotto, P.: Analyzing and modelling rating data for sensory analysis in food industry. Quad. Stat. **13**, 68–81 (2011)

Næs, T., Solheim, R.: Detection and interpretation of variation within and between assessors in sensory profiling. J. Sens. Stud. **6**(3), 159–177 (1991)

Piggott, J.R.: Design questions in sensory and consumer science. Food Qual. Prefer. **6**(4), 217–220 (1995)

Pinto, F.S.T., Fogliatto, F.S., Qannari, E.M.: A method for panelists consistency assessment in sensory evaluations based on the cronbachs alpha coefficient. Food Qual. Prefer. **32**, 41–47 (2014)

Rossi, F.: Assessing sensory panelist performance using repeatability and reproducibility measures. Food Qual. Prefer. **12**(5), 467–479 (2001)

Schlich, P.: Grapes: a method and a sas® program for graphical representations of assessor performances. J. Sens. Stud. **9**(2), 157–169 (1994)

Vanacore, A., Pellegrino, M.S.: Characterizing the extent of rater agreement via a non-parametric benchmarking procedure. In: Proceedings of the Conference of the Italian Statistical Society, pp. 999–1004. Italian Statistical Society (2017)

Vanbelle, S.: Agreement between raters and groups of raters. Ph.D. thesis, Université de Liège, Belgique (2009)

Veall, M.R., Zimmermann, K.F.: Performance measures from prediction–realization tables. Econ. Lett. **39**(2), 129–134 (1992)

Received: 22 December 2017 | Revised: 27 April 2018 | Accepted: 1 May 2018

DOI: 10.1002/qre.2334

SPECIAL ISSUE ARTICLE

WILEY

RRep: A composite index to assess and test rater precision

Amalia Vanacore | Maria Sole Pellegrino

Department of Industrial Engineering, University of Naples "Federico II", Naples, Italy

Correspondence
Amalia Vanacore, Department of Industrial Engineering, University of Naples "Federico II", Naples, Italy.
Email: amalia.vanacore@unina.it

Abstract

In subjective evaluation systems, raters act as measurement instruments providing useful evaluations for taking strategic and/or operational decisions. The assessment of rater evaluative ability in terms of accuracy and precision is of critical importance since rater unreliability may compromise the quality of the decision-making process. The focus of this paper is on rater precision: we propose a novel composite index to assess the rater ability to provide evaluations repeatable over time and reproducible over scales. The extent of rater precision is qualified via a nonparametric benchmarking procedure. The properties of both proposed index and benchmarking procedure have been analysed via a Monte Carlo simulation study.

KEYWORDS

rater precision, rater repeatability and reproducibility index, subjective evaluations

1 | INTRODUCTION

In several business and industrial systems, as well as in many medical, social and behavioural contexts, diagnostic assessment relies on subjective evaluations provided by small groups of human raters, who may be —depending on the specific context— field experts (eg, physicians,[1,2] sensory panelists,[3,4] risk assessors[5]) or trained operators (eg, mystery shoppers, visual inspectors[6-8]). In subjective evaluation systems, human raters act as measurement instruments[9-12] and they can be a main source of epistemic uncertainty.[13] Indeed, differently from aleatory uncertainty, epistemic uncertainty does not pertain the inherent variability of the phenomenon under study since it arises from imperfect knowledge and/or incomplete information. Epistemic uncertainty can be reduced by selecting the right raters, able to provide accurate and precise evaluations.

Being subjective, rater evaluations lack a gold standard against which to check their accuracy. As a matter of fact, the reliability of subjective evaluations is related only to precision, assessed as the degree of agreement between repeated evaluations provided under the same conditions, where "same conditions" means that nothing changed other than the times of the evaluations.

In order to assess the precision of the evaluations provided by the same rater in different occasions, a number of theoretical and methodological approaches have been proposed over the years. Among them, the most widely adopted is the Intraclass Correlation Coefficient (ICC).[14] Although ICC is accepted as a universal reliability index,[15] it can properly evaluate reliability between quantitative measurements provided on continuum scale rather than on categorical (ie, nominal or ordinal) scales, the ones typically adopted in subjective evaluations.

A widely applied method to assess the precision of categorical measurements consists in quantifying (intra/inter) rater agreement[16,17] via a kappa-type coefficient, that is a rescaled measure of the observed proportion of agreement corrected with the proportion of agreement expected by chance alone. The kappa-type coefficients proposed in the literature (eg, Fleiss' kappa,[18] Conger's kappa,[19] Scott's pi,[20] Cohen's kappa,[21] Gwet's AC[22] and Brennan-Prediger coefficient—[23-26] also known as uniform kappa) differ from each other in the definition, and thus formulation, of agreement expected by chance alone. Among the available kappa-type coefficients, De Mast and VanWieringen[27] suggest to prefer the uniform kappa as a precision index since it separates precision from issues related to the accuracy of measurement system, while these issues are confounded in other kappa-type coefficients (eg, Fleiss' kappa[18] and Conger's kappa[19]).

This paper suggests (1) to estimate rater precision in terms of rater repeatability and reproducibility and (2) to characterise the extent of precision via a nonparametric inferential procedure, taking into account sampling uncertainty. Specifically, rater precision is defined as the rater ability of consistently score the same set of items not only in different occasions —as commonly done in the majority of applications— but also under different experimental conditions that, in the case of subjective evaluations, can be obtained by collecting rater evaluations using different rating scales. These rater abilities, referred to as repeatability over time and reproducibility over scales, are then properly combined in a synthetic index. In agreement studies, the magnitude of agreement is commonly qualified by a straightforward comparison of the calculated kappa-type coefficient against an arbitrary benchmark scale. The interpretation based on the straightforward benchmarking should be treated with caution, especially for comparison across studies when the experimental conditions are not the same. A proper characterization of the extent of precision should rely upon a benchmarking procedure that allows to identify a suitable neighbourhood of the truth by taking into account sampling uncertainty.

The main statistical properties of the proposed RRep index and those of the recommended benchmarking procedure have been assessed via a Monte Carlo simulation study.

The remainder of this paper is organised as follows: the linear weighted uniform kappa, the RRep index, and the benchmarking procedure are introduced in Section 2; design and results of the Monte Carlo simulation study are reported in Section 3; in Section 4, a real application aimed at illustrating the applicability and usefulness of the proposed procedure is fully described; finally, conclusions are summarised in Section 5.

2 | METHODS

2.1 | Measuring rater precision via agreement coefficients

Let n be the number of items rated 2 or more times (ie, replications) by a rater on an ordinal k-point rating scales, with $k>2$. Rater evaluations are denoted Y_{hr}, with $h=1,\dots,n$ indexing items and r indexing replications. Of interest for the evaluation of rater repeatability and reproducibility is the joint distribution of the Y_{hr}.

TABLE 1 $k \times k$ contingency table

		Second Replication					
	Category	**1**	…	***j***	…	***k***	**Total**
First replication	**1**	n_{11}	…	n_{1j}	…	n_{1k}	$n_{1\cdot}$
	⋮	⋮	…	⋮	…	⋮	⋮
	i	n_{i1}	…	n_{ij}	…	n_{ik}	$n_{i\cdot}$
	⋮	⋮	…	⋮	…	⋮	⋮
	k	n_{k1}	…	n_{kj}	…	n_{kk}	$n_{k\cdot}$
	Total	$n_{\cdot 1}$	…	$n_{\cdot j}$	…	$n_{\cdot k}$	n

In the simplest case of 2 replications (ie, $r=1,2$), the data can be arranged in a $k \times k$ contingency table $(n_{ij})_{k\times k}$ (Table 1), where the generic (i,j) cell contains the joint frequency n_{Tij} (n_{Sij}) that counts the number of items classified into i^{th} category in the first replication over time (over rating scales) and into j^{th} category in the second replication over time (over rating scales). Specifically, the cells along the main diagonal represent the perfect match between the evaluations provided in different replications, whereas the off-diagonal cells represent mismatch.

The degree of agreement between the series of ratings Y_{h1} and Y_{h2} is here estimated adopting the uniform kappa coefficient.

The uniform kappa formulates the agreement expected by chance alone adopting the notion of uniform chance measurement,[25] which assigns equal probability to any rating category and thus is the most noninformative measurement system given a certain rating scale.[27,28] The proportion of agreement expected by chance alone, $p_{a|c}$, under the assumption of uniform chance measurement is formulated as follows:

$$p^{U}_{a|c} = \sum_{i=1}^{k} \frac{1}{k^2} = \frac{1}{k}, \tag{1}$$

whereas the observed proportion of agreement, common to all kappa-type coefficients, is given by

$$p_a = \sum_{i=1}^{k} \frac{n_{ii}}{n}. \tag{2}$$

Although the generic kappa treats all disagreements as homogeneous, it is undoubtful that for ordinal rating scale some disagreements are more serious than others. In this case, the introduction of either a distance metric or a weighting scheme enables to account that disagreement on distant categories should be considered more relevant than disagreement on neighbouring categories. Different kinds of distance metrics and weighting schemes appropriate for various practical situations have been proposed and discussed in the literature. Typically, these metrics are expressed as nondecreasing functions of $|i - j|$ when assessing the degree of disagreement among the provided evaluations (eg, loss matrix[29]) or, vice versa, as nonincreasing function of $|i - j|$ when assessing the degree of agreement (eg, linear agreeing weights[30] and quadratic agreeing weights[31]). A weighted version of the uniform kappa, K_W^U, including symmetric weights (ie, $w_{ij} = w_{ji}$) a priori assigned to each pair (i,j) of ratings, has been proposed by Gwet,[32] and its statistical properties have been studied by Warrens.[33] The weighted uniform kappa is formulated as follows:

$$K_W^U = \frac{p_{a_W} - p^U_{a|c_W}}{1 - p^U_{a|c_W}}, \tag{3}$$

where p_{a_W} is the weighted observed proportion of agreement and is given by

$$p_{a_W} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\frac{n_{ij}}{n}. \tag{4}$$

$p^U_{a|c_W}$ is the weighted proportion of agreement expected under the assumption of uniform chance measurement and is given by

$$p^U_{a|c_W} = \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}, \tag{5}$$

and w_{ij} is the symmetrical agreement weight ranging between 0 and 1, with the minimum value 0 assigned to maximally disagreeing pairs of ratings, $(1, k)$ and $(k, 1)$, and the maximum value 1 assigned to pairs of coincident ratings (i, i). It is worthwhile to pinpoint that although the weights can be arbitrary defined, the linear[30] (w^L_{ij}) and quadratic[31] (w^Q_{ij}) weights are the most commonly used weighting schemes for kappa-type coefficients and are formulated as follows:

$$w^L_{ij} = 1 - \frac{|i-j|}{k-1}; \quad w^Q_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}. \tag{6}$$

The weighted uniform kappa can be assumed asymptotically normally distributed with mean $\mu_{K^U_W}$ and variance $\sigma^2_{K^U_W}$[32] estimated as follows:

$$\hat{\sigma}^2_{K^U_W} = \frac{1-f}{n(1-p_{a|c})^2}\left(\sum_{i=1}^{k}\sum_{j=1}^{k} w^2_{ij}\frac{n_{ij}}{n} - p^2_{a_W}\right), \tag{7}$$

where $f = n/N$ is the fraction of the sampled target population which in many studies is set equal to 0 being the size N of the item population unknown.

All kappa-type coefficients range from −1 to +1: when the observed proportion of agreement equals chance agreement, the coefficient is null; when the observed agreement is greater than chance agreement, the coefficients is positive; when the observed agreement is lower than chance agreement, the coefficient is negative and can be interpreted as disagreement. Being K^U_W used as a measure of agreement, the region of interest is $K^U_W > 0$,[16] thus in the following K^U_W will be coherently truncated at the lower limit $K^U_W = 0$ so as to obtain nonnegative measures of rater repeatability and reproducibility, defined as follows:

$$K^+_{WT} = \max\left(0, K^U_{WT}\right); \; K^+_{WS} = \max\left(0, K^U_{WS}\right), \tag{8}$$

being

$$K^U_{WT} = \frac{p_{a_{WT}} - p^U_{a|c_W}}{1 - p^U_{a|c_W}}; \; K^U_{WS} = \frac{p_{a_{WS}} - p^U_{a|c_W}}{1 - p^U_{a|c_W}}, \tag{9}$$

where $p_{a_{WT}}$ and $p_{a_{WS}}$ are given by

$$p_{a_{WT}} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\frac{n_{Tij}}{n}; \; p_{a_{WS}} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\frac{n_{Sij}}{n}. \tag{10}$$

In order to assess rater precision, we introduce a composite index of rater repeatability and reproducibility (hereafter, RRep). Specifically, because these rater abilities may be reasonably assimilated to the rings of a chain, where the weakest ring dictates the strength of the chain, a measure of rater precision can be defined as follows:

$$\text{RRep} = K^+_{WT} \cdot K^+_{WS}. \tag{11}$$

The formulation in Equation 11 links the components K^+_{WT} and K^+_{WS} of the RRep index by a series logical structure implying that RRep is null when the rater is either not repeatable or not reproducible or both. Since K^+_{WT} and K^+_{WS} range between 0 and 1, RRep too ranges between 0 and 1 and its value is always no more than the smallest of its components:

$$\text{RRep} \le \min\left(K^+_{WT}, K^+_{WS}\right), \tag{12}$$

meaning that the overall rater precision cannot be higher than the worst performance achieved on repeatability or reproducibility.

2.2 | A nonparametric confidence interval for RRep index

As for any point estimate, the meaning of RRep index alone is limited; it is thus recommended to build a $(1-2\alpha)\%$ confidence interval (CI) in order to provide a clearer understanding of sampling uncertainty and to test for significance the magnitude of the RRep index against a desirable level of precision.

Being defined as the product of 2 left truncated normal distributions, the RRep index is not normally distributed and, being the RRep bounded by 1, its sampling distribution (Figure 1) can be highly skewed. In such cases, the bias-corrected and accelerated (BCa) bootstrap CI is generally considered the method of choice to make inference.[34]

The lower (RRep_l) and upper (RRep_u) bounds of the 2-sided (1–2α)% BCa bootstrap CI are defined in terms of the cumulative distribution function G of B bootstrap replications and 2 numerical parameters: the bias correction b and the acceleration a. By definition, RRep_l and RRep_u are equal to

$$\begin{aligned} \mathrm{RRep}_l &= G^{-1}\left(\Phi\left(b - \frac{z_\alpha - b}{1 + a\left(z_\alpha - b\right)}\right)\right); \\ \mathrm{RRep}_u &= G^{-1}\left(\Phi\left(b + \frac{z_\alpha + b}{1 + a\left(-z_\alpha - b\right)}\right)\right), \end{aligned} \tag{13}$$

being Φ the cumulative distribution function of the normal distribution and z_α the α percentile of the standard normal distribution.

Specifically, let $T=\left\{\left(y_{h1}^{T}, y_{h2}^{T}\right), n\right\}$ be the sample of pairs of ratings provided by the rater during 2 replications over time with the same rating scale and $S=\left\{\left(y_{h1}^{S}, y_{h2}^{S}\right), n\right\}$ the sample of pairs of ratings provided by the rater over 2 rating scales, the detailed algorithm for building the (1–2α)% BCa bootstrap CI for RRep works as follows:

1. sample n pairs of rating randomly with replacement from T to obtain a bootstrap data set, denoted T^*; in the same way sample n pairs of rating randomly with replacement from S to obtain a bootstrap data set, denoted S^*;
2. for each bootstrap data set, compute $K_{WS}^{+}(S^*)$ and $K_{WT}^{+}(T^*)$ according to Equation 8 and then, according to Equation 11, calculate $\mathrm{RRep}\left(T^*, S^*\right)$, hereafter denoted as RRep^*;
3. repeat B times steps 1 and 2 in order to obtain B estimates RRep^*; count the number of bootstrap estimates RRep^* that are less than RRep calculated from the original data set. Call this number p and set $b = \Phi^{-1}\left(p/B\right)$, being Φ^{-1} the inverse cumulative distribution function of the normal distribution;
4. calculate the parameter a using the jackknife RRep estimates, RRep_i^j

$$a = \frac{\sum_{i=1}^{n}\left(\overline{\mathrm{RRep}^j} - \mathrm{RRep}_i^j\right)^3}{6\left[\sum_{i=1}^{n}\left(\overline{\mathrm{RRep}^j} - \mathrm{RRep}_i^j\right)^2\right]^{3/2}}, \tag{14}$$

being $\overline{\mathrm{RRep}^j}$ the average out of all n jackknife estimates RRep_i^j;

5. estimate the lower (RRep_l) and upper (RRep_u) bounds of the 2-sided (1–2α)% BCa bootstrap CI for RRep using Equation 13.

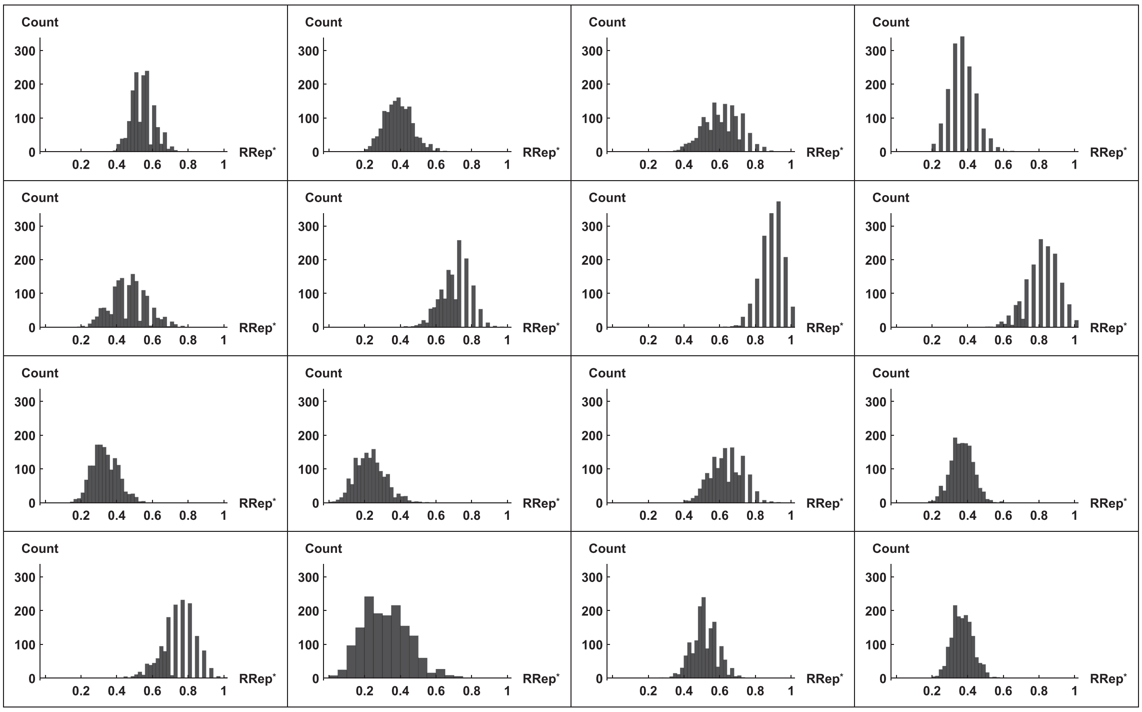


FIGURE 1 Some example of RRep bootstrap distribution. Abbreviation: RRep, rater repeatability and reproducibility

TABLE 2 Most common benchmark scales for kappa-type coefficients

Landis and Koch (1977)		Fleiss (1981)		Altman (1981)	
Kappa	**Agreement**	**Kappa**	**Agreement**	**Kappa**	**Agreement**
< 0.0	Poor	< 0.4	Poor	< 0.20	Poor
0.00 to 0.20	Slight	0.40 to 0.75	Intermediate to Good	0.21 to 0.40	Fair
0.21 to 0.40	Fair	> 0.75	Excellent	0.41 to 0.60	Moderate
0.41 to 0.60	Moderate			0.61 to 0.80	Good
0.61 to 0.80	Substantial			0.81 to 1.00	Very good
0.81 to 1.00	Almost perfect				

2.3 | A benchmarking procedure for interpreting the extent of rater precision

Any coefficient is useful only if its magnitude can be interpreted. Various benchmark scales have been proposed in the literature over the years for interpreting the magnitude of kappa-type coefficients. The most common benchmark scale is the one proposed by Landis and Koch,[35] which was simplified by Fleiss[36] and Altman[37] collapsing the 6 ranges into 3 and 5 ranges, respectively (Table 2).

Although some researchers question the validity of benchmark scales and give advice that their uncritical applications may lead to practically questionable decisions,[38] these scales are widely adopted for the interpretation of results from agreement studies (see, eg, Everitt,[39] Guillemin et al,[40] Blackman and Koval,[16] Altaye et al,[41] Klar et al,[42] Kraemer et al,[43] Hallgren,[44] Watson and Petrie[45]). As argued by Gwet,[32] the choice of a benchmark scale is less important than the way it is used for characterizing the extent of agreement. The approach currently adopted to characterise the extent of agreement is based upon a straight comparison between the estimated coefficient and the adopted benchmark scale. However, this straightforward benchmarking procedure does not account for uncertainty due to statistical error. To overcome this criticism, we characterise the extent of rater precision by comparing the lower bound of the RRep CI (RRep_l) against a benchmark scale adapted from that provided by Landis and Koch (Figure 2).

Figure 2 displays the 3 isoprecision curves, which are contour lines drawn through the set of points corresponding to the same precision level obtained by changing the levels of K^+_{WT} and K^+_{WS}. The isoprecision curves divide the domain space of the RRep index into 4 regions each corresponding to a specific rater precision level, as labelled in Figure 2.

According to the proposed benchmarking procedure, if the aim is to check for Perfect rater precision at a significance level $\alpha = 0.025$, the lower bound of the 2-sided 0.95% CI (RRep_l) has to be above 0.75; if a Slight precision is to be proven, RRep_l has to be below 0.25; all the other intermediate values assumed by RRep_l represent Moderate ($0.25 < \text{RRep}_l < 0.50$) and Substantial ($0.5 < \text{RRep}_l < 0.75$) precision, respectively.

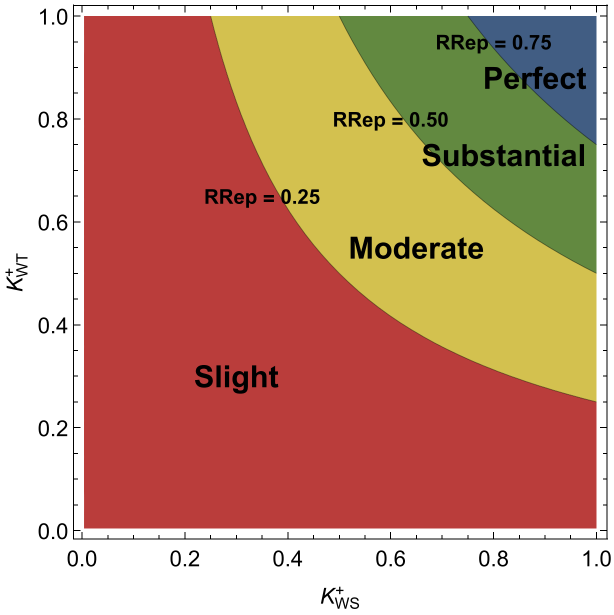


FIGURE 2 Benchmark scale for rater precision and isoprecision curves. Abbreviation: RRep, rater repeatability and reproducibility [Colour figure can be viewed at wileyonlinelibrary.com]

3 | MONTE CARLO SIMULATION STUDY

3.1 | Simulation design

An extensive Monte Carlo simulation study has been conducted in order to investigate the statistical properties of RRep and those of the adopted benchmarking procedure. In the simulation design we have considered one rater who classifies n items into one of the k possible rating categories and we have assumed 7 different repeatability and reproducibility levels (ie, $K^+_{WT,WS} = 0.6, 0.7, 0.8, 0.85, 0.9, 0.95, 1.00$) resulting in a total of 28 possible scenarios (ie, distinct couples of K^+_{WT} and K^+_{WS}) corresponding to as many different levels of rater precision. We simulated and analysed the statistical behaviour of RRep for 14 scenarios (Table 3), chosen in order to balance the number of those belonging to the categories of Substantial and Almost perfect/Perfect precision.

The precision in the estimate of RRep has been evaluated in terms of percent bias (Δ) and relative standard deviation (RSD), whereas the performance of the benchmarking procedure has been evaluated in terms of statistical significance ($\hat{\alpha}$) and statistical power ($1-\hat{\beta}$). Statistical significance and power have been computed for 2 different hypothesis statements. The first statement consists in testing the hypothesis that the rater precision at the population level is at least Substantial (ie, $H_1{:}RRep>0.50$) against the null hypothesis of no more than Moderate rater precision (ie, $H_0{:}RRep\leq 0.50$); the second hypothesis statement consists in testing that the rater precision is Almost perfect/Perfect (ie, $H_1{:}RRep>0.75$) against the null hypothesis of no more than Substantial rater precision (ie, $H_0{:}RRep \leq 0.75$). The sampling distributions of RRep under the null hypotheses of Moderate and Substantial rater precision correspond to the scenarios reported in Table 3 as #1 and #7, respectively. All the hypothesis tests have been conducted assuming a significance level $\alpha=0.025$.

Specifically, let *RRep* be the true population value of rater precision, r be the number of Monte Carlo data sets and RRep_h be the h^{th} Monte Carlo estimate for the index, the Monte Carlo estimate of the percent bias is given by

$$\Delta = \frac{1}{r}\sum_{h=1}^{r}\frac{\mathrm{RRep}_h - RRep}{RRep}\cdot 100. \tag{15}$$

The Monte Carlo estimate of the RSD is given by

$$\mathrm{RSD} = \frac{1}{\overline{\mathrm{RRep}}}\cdot\sqrt{\frac{\sum_{h=1}^{r}\mathrm{RRep}_h^2 - r\overline{\mathrm{RRep}}^2}{r}}\cdot 100 \tag{16}$$

where $\overline{\mathrm{RRep}}$ is the average out of all r Monte Carlo estimates RRep_h.

Let I [·] be an indicator taking value 1 if the argument is true and 0 otherwise, $\mathrm{RRep}_{l|h}$ be the lower bound of the $(1-2\alpha)$% BCa bootstrap CI obtained from the h^{th} specific Monte Carlo data set and RRep_C be the tested critical value of rater precision. The Monte Carlo estimate of the statistical significance is

$$\hat{\alpha} = \frac{1}{r}\sum_{h=1}^{r} I\left[\mathrm{RRep}_{l|h} > \mathrm{RRep}_C|\ \mathrm{H}_0\right]. \tag{17}$$

The Monte Carlo estimate of the statistical power is given by

$$1-\hat{\beta} = \frac{1}{r}\sum_{h=1}^{r} I\left[\mathrm{RRep}_{l|h} > \mathrm{RRep}_C|\ \mathrm{H}_1\right]. \tag{18}$$

For each scenario, $r=2000$ Monte Carlo data sets have been generated and for each data set the 95% BCa bootstrap CI has been computed on $B=1500$ bootstrap replications.

Accordingly to the definition of RRep (Equation 11), the Monte Carlo data sets used to obtain K_{WT}^{+} and those used to obtain K_{WS}^{+} have been independently drawn from multinomial distributions with parameters n and $\pi=\left(\pi_{11},\cdots,\pi_{1k},\cdots,\pi_{ij},\cdots,\pi_{k1},\cdots,\pi_{kk}\right)$, with the π_{ij} values set according to the desired levels of repeatability and/or reproducibility. For illustrative purpose, the π_{ij} values used

TABLE 3 True population values of K_{WT}^{+}, K_{WS}^{+} and RRep defining the simulated scenarios.

Scenario #	$K_{WT(WS)}^{+}$	$K_{WS(WT)}^{+}$	RRep	Rater Precision
1	≃0.70	≃0.70	≃0.49	Moderate
2	≃ 0.60	≃ 0.90	≃ 0.54	Substantial
3	≃ 0.70	≃ 0.80	≃ 0.56	Substantial
4	≃ 0.60	≃ 1.00	≃ 0.60	Substantial
5	≃ 0.70	≃ 0.90	≃ 0.64	Substantial
6	≃ 0.70	≃ 1.00	≃ 0.70	Substantial
7	≃0.85	≃0.85	≃0.72	Substantial
8	≃ 0.85	≃ 0.90	≃ 0.76	Almost perfect/Perfect
9	≃ 0.80	≃ 1.00	≃ 0.80	Almost perfect/Perfect
10	≃ 0.85	≃ 0.95	≃ 0.81	Almost perfect/Perfect
11	≃ 0.90	≃ 0.95	≃ 0.85	Almost perfect/Perfect
12	≃ 0.85	≃ 1.00	≃ 0.85	Almost perfect/Perfect
13	≃ 0.90	≃ 1.00	≃ 0.90	Almost perfect/Perfect
14	≃ 0.95	≃ 1.00	≃ 0.95	Almost perfect/Perfect

Abbreviation: RRep, rater repeatability and reproducibility.

for the simulation with $k=4$ rating categories are reported in Table 4. The statistical properties of the RRep index and those of the benchmarking procedure have been studied for $k=2, 4, 5, 7$ rating categories and for $n=10, 30, 50$ items, which are the most affordable sample sizes in many experimental contexts and also the most critical ones for statistical inference.

3.2 | Simulation results

For each sample size and each rating scale dimension simulation results in terms of percent bias and relative standard deviation are reported in Tables 5 and 6, respectively. The statistical significance and power for the first and second hypothesis statement are represented in Figure 3.

TABLE 4 Patterns of joint probabilities assumed to simulate different levels of repeatability and/or reproducibility with $k = 4$ rating categories

(a)$K_W^U = 0$					**(b)$K_W^U = 0.6$**				
Category	1	2	3	4	Category	1	2	3	4
1	0.180	0.020	0.020	0.040	1	0.200	0.013	0.014	0.030
2	0.030	0.180	0.020	0.020	2	0.015	0.200	0.013	0.014
3	0.020	0.020	0.180	0.020	3	0.013	0.014	0.205	0.015
4	0.030	0.020	0.020	0.180	4	0.027	0.014	0.013	0.200
(c)$K_W^U = 0.7$					**(d)$K_W^U = 0.8$**				
Category	1	2	3	4	Category	1	2	3	4
1	0.210	0.014	0.015	0.009	1	0.220	0.014	0.010	0.001
2	0.014	0.210	0.014	0.014	2	0.011	0.220	0.010	0.010
3	0.014	0.015	0.210	0.014	3	0.010	0.012	0.220	0.016
4	0.009	0.014	0.014	0.210	4	0.002	0.010	0.014	0.220
(e)$K_W^U = 0.85$					**(f)$K_W^U = 0.9$**				
Category	1	2	3	4	Category	1	2	3	4
1	0.230	0.007	0.007	0.005	1	0.240	0.005	0.000	0.000
2	0.007	0.230	0.007	0.007	2	0.000	0.240	0.010	0.000
3	0.007	0.007	0.230	0.007	3	0.010	0.005	0.235	0.000
4	0.005	0.007	0.007	0.230	4	0.005	0.000	0.010	0.240
(g)$K_W^U = 0.95$					**(h)$K_W^U = 1.00$**				
Category	1	2	3	4	Category	1	2	3	4
1	0.240	0.010	0.000	0.000	1	0.245	0.002	0.000	0.000
2	0.000	0.240	0.010	0.000	2	0.000	0.250	0.000	0.000
3	0.010	0.000	0.235	0.000	3	0.000	0.005	0.250	0.000
4	0.005	0.000	0.010	0.240	4	0.005	0.000	0.003	0.245

TABLE 5 Percent bias (Δ) for different rater precision levels with $k = 2$-, 4-, 5-, and 7-point scales and $n = 10, 30, 50$ items

	$n = 10$				$n = 30$				$n = 50$			
Scenario #	$k = 2$	$k = 4$	$k = 5$	$k = 7$	$k = 2$	$k = 4$	$k = 5$	$k = 7$	$k = 2$	$k = 4$	$k = 5$	$k = 7$
1	5.67	5.58	7.47	1.35	0.26	5.58	−0.56	−1.70	1.56	1.60	1.47	0.38
2	1.06	−0.83	−1.65	0.83	0.86	−0.83	v0.11	0.18	0.21	−0.34	−0.12	−0.07
3	3.64	−1.17	−2.42	−0.86	0.68	−0.78	−0.27	−1.14	0.88	0.00	0.37	−0.03
4	−1.35	−0.63	−2.48	−0.35	−0.42	−0.63	−0.60	−0.60	−0.33	−0.33	−0.08	−0.19
5	−1.19	−0.84	−1.83	−0.80	−0.93	−0.84	−1.00	−1.29	0.06	−0.06	0.12	−0.18
6	−2.65	−0.59	−2.87	−1.95	−1.88	−0.59	−1.71	−2.09	−0.33	−0.08	0.13	−0.29
7	2.14	2.36	1.97	2.24	2.13	2.36	2.15	1.77	1.59	0.44	6.13	6.24
8	2.00	1.69	2.11	2.20	2.44	2.20	2.74	v2.36	2.15	−0.02	4.08	4.03
9	−6.40	−3.14	−4.94	−8.60	−8.27	−3.14	−7.93	−9.51	−8.98	0.04	−8.26	−9.33
10	−2.46	−1.31	−2.57	−1.95	−1.86	−1.31	−1.42	−2.30	−0.19	0.10	0.23	−0.44
11	−0.12	−0.87	0.53	0.55	0.37	−0.87	0.51	0.11	−0.10	−0.01	0.02	−0.17
12	−2.56	−0.73	−2.96	−2.44	−2.18	−0.73	−2.11	−2.71	−0.13	0.04	0.23	−0.25
13	−0.27	−0.41	0.07	0.01	0.04	−0.41	−0.21	−0.30	−0.04	−0.07	0.02	0.02
14	−0.20	−0.52	−0.51	−0.64	−0.02	−0.52	−0.17	−0.70	−0.08	0.07	0.05	−0.29

The Monte Carlo simulation results exhibit good statistical properties for RRep, being the percent bias always no more than 9% and the relative standard deviation generally less than 30%. Specifically, they decrease as sample size, rating scale dimension and precision level increase; with small samples of $n = 10$ items, the relative standard deviation is less than 30% only for precision level higher than 0.7.

Regarding the performance of the benchmarking procedure, simulation results suggest that the statistical significance is closer to its nominal value $\alpha = 0.025$ when testing a level of rater precision at least Substantial, while it moves away from it when testing the highest level of rater precision especially with small samples (for $n = 10$ items, $\hat{\alpha} = 0.20$).

The statistical power (see Figure 3) increases with increasing rater precision level and rating scale dimension. It is satisfactory (ie, at least 80%) when referring to nonadjacent levels of rater precision (eg, Moderate against

TABLE 6 Relative standard deviation for different rater precision levels with $k = 2$-, 4-, 5-, and 7-point scales and $n = 10, 30, 50$ items

	$n = 10$				$n = 30$				$n = 50$			
Scenario #	$k = 2$	$k = 4$	$k = 5$	$k = 7$	$k = 2$	$k = 4$	$k = 5$	$k = 7$	$k = 2$	$k = 4$	$k = 5$	$k = 7$
1	0.55	0.59	0.55	0.38	0.58	0.59	0.27	0.23	0.28	0.26	0.22	0.18
2	0.48	0.59	0.46	0.30	0.29	0.59	0.18	0.18	0.22	0.18	0.14	0.14
3	0.48	0.47	0.41	0.25	0.30	0.43	0.19	0.14	0.24	0.17	0.15	0.11
4	0.44	0.54	0.42	0.29	0.25	0.54	0.17	0.18	0.19	0.17	0.13	0.14
5	0.39	0.47	0.41	0.22	0.23	0.47	0.16	0.13	0.18	0.15	0.13	0.10
6	0.33	0.41	0.36	0.21	0.19	0.41	0.14	0.12	0.15	0.13	0.11	0.09
7	0.28	0.27	0.29	0.12	0.16	0.24	0.12	0.07	0.13	0.12	0.10	0.07
8	0.24	0.28	0.30	0.11	0.14	0.28	0.12	0.07	0.11	0.09	0.09	0.05
9	0.35	0.21	0.21	0.15	0.22	0.12	0.12	0.08	0.18	0.09	0.10	0.07
10	0.23	0.25	0.27	0.10	0.13	0.25	0.11	0.06	0.10	0.07	0.08	0.05
11	0.19	0.24	0.21	0.10	0.11	0.24	0.08	0.06	0.08	0.07	0.06	0.05
12	0.20	0.21	0.25	0.09	0.12	0.21	0.10	0.05	0.09	0.06	0.07	0.04
13	0.16	0.20	0.17	0.09	0.09	0.20	0.07	0.05	0.07	0.06	0.06	0.04
14	0.11	0.14	0.11	0.06	0.06	0.14	0.04	0.04	0.05	0.04	0.03	0.03

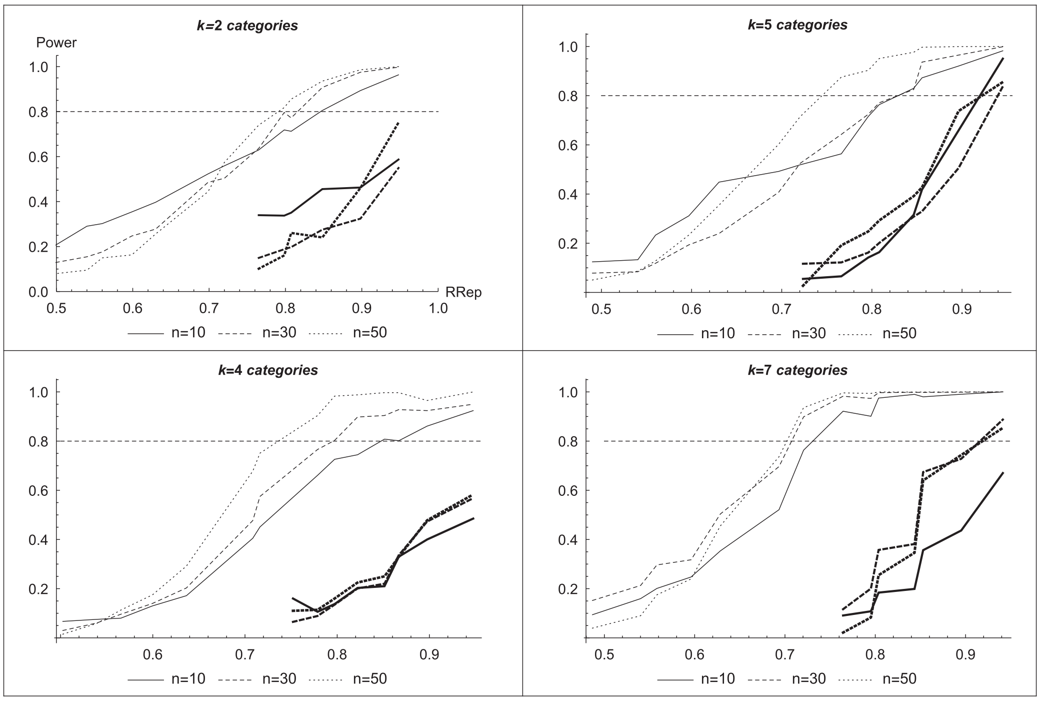


FIGURE 3 Statistical significance and power for different sample sizes and rating scale dimensions. (Thin lines for $H_0 : RRep < 0.50$ and in bold for $H_0 : RRep < 0.75$). Abbreviation: RRep, rater repeatability and reproducibility

Almost perfect) for samples of $n=50$ items and 2-, 4-, and 5-point scales and also for $n \geq 10$ items with 7-point scale. Moreover, in the first hypothesis statement, it is less than 80% when testing a precision level between 0.50 and 0.85 with 2-, 4-, and 5-point scales and $n=10$ items or when testing a precision level between 0.50 and 0.75 with 4- and 5-point scales and $n=50$ items; for a 7-point scale, $n=10$ items are enough for reaching 80% power when testing a precision level greater than 0.70. In the second hypothesis statement, instead, the statistical power reaches 80% only when testing a precision level greater than 0.90 with 5- and 7-point scales.

4 | EMPIRICAL ILLUSTRATION: A REAL CASE STUDY

In the following, a detailed walk-through example is presented showing step-by-step real application of the proposed procedure to test rater precision.

The data comes from an intrarater agreement study involving a class of university students who evaluated the teaching quality of the same university course. The whole experiment consisted of 3 evaluation sessions. During the first session, the students rated 20 quality statements about the teaching course using a 4-point verbal rating scale (VRS) with grades: strong disagreement, disagreement, agreement, strong agreement; during the second session (1 lesson after, 1 week apart), they rated the same university course using the same VRS; finally, during the third session (4 lessons after, 2 weeks apart), they rated again the course using the 4-point VRS and a visual analogue scale (VAS), with left anchor point labelled NO, right anchor point labelled YES, and 3 unlabelled thicks in between.

The evaluations collected during the first and second sessions on VRS have been used to assess student's repeatability, whereas those collected during the third session on VRS and VAS have been used to assess student's reproducibility. For practical purpose, the implementation procedure is fully exploited only for 2 students, labelled as #1 and #2.

The ratings provided on VRS during the first and second sessions and those provided on VRS and VAS during the third session have been classified in 2 different 4 × 4 contingency tables, as shown in Table 7.

Students' repeatability and reproducibility have been estimated via Equation 9 adopting the linear weighting scheme (Equation 6), RRep and RRep_l have been calculated according to Equation 11 and Equation 13, respectively (Table 8).

TABLE 7 4 × 4 contingency tables of students' ratings over different occasions (*a* and *b*) and with different rating scales (*c* and *d*)

(a)$K_{WT\|1}$						**(b)$K_{WT\|2}$**					
Category	1	2	3	4	Total	Category	1	2	3	4	
1	0	0	0	0	0	1	0	0	0	0	0
2	0	0	0	0	0	2	0	2	0	0	2
3	0	5	7	6	18	3	2	1	1	5	9
4	0	0	1	1	2	4	0	0	1	8	9
Total	0	5	8	7	20	Total	2	3	2	13	20
(c)$K_{WS\|1}$						**(d)$K_{WS\|2}$**					
Category	1	2	3	4	Total	Category	1	2	3	4	
1	0	0	0	0	0	1	0	0	0	0	0
2	0	0	0	0	0	2	2	0	0	0	2
3	0	0	7	6	13	3	2	0	1	5	8
4	0	0	1	6	7	4	0	0	0	10	10
Total	0	0	8	12	20	Total	4	0	1	15	20

TABLE 8 K_{WT}, K_{WS}, rater repeatability and reproducibility (RRep) and its 95% BCa CI for students #1 and #2

Student	**K_{WT}**	**K_{WS}**	**RRep**	**95% BCa CI**
#1	0.52	0.72	0.374	[0.282, 0.512]
#2	0.56	0.56	0.314	[0.102, 0.608]

The RRep_l for student #1 belongs to the region ranging from 0.25 to 0.50 whereas the RRep_l for student #2 belongs to the region ranging from 0.00 to 0.25. Thus, according to the proposed benchmarking procedure (Figure 3), with a significance level $\alpha=0.025$, the levels of precision can be assumed Moderate and Slight for student #1 and #2, respectively. It is worthwhile to pinpoint that the interpretation of the rater precision could be overestimated if the RRep values were straightly compared against the benchmark scale. Indeed, adopting this latter approach the extent of precision for student #2 would be Moderate rather than Slight because RRep=0.314 belongs to the range of Moderate rater precision.

5 | CONCLUSIONS

In this paper the problem of assessing the precision of subjective evaluations has been explored and new tools to estimate and characterise rater precision have been proposed.

Rater precision is commonly assessed as the rater ability of providing repeatable evaluations over different occasions under the same conditions. In this paper, instead, the concept of rater precision has been extended to the ability of providing reproducible evaluations in the same occasion under different settings (ie, with different rating scales).

The rater precision is thus coherently estimated via a novel composite index, RRep, formulated in such a way that both rater repeatability and reproducibility are accounted for; the characterization of the extent of rater precision is obtained via a nonparametric benchmarking procedure testing for significance RRep magnitude against desirable levels of precision.

The Monte Carlo simulation results show that the performance —in terms of percent bias, relative standard deviation, statistical significance and power— of the proposed tools are satisfactory.

Specifically, even with small sample sizes and rating scales with few categories, the benchmarking procedure shows satisfactory power in distinguishing between categories of precision that are at least 1-step apart, the ones of greater practical interest with differences of at least 0.25 (eg, when testing Almost perfect/Perfect precision level against the null hypothesis of no more than Moderate precision).

The proposed procedure can be effectively applied to characterise the extent of precision and selecting inspectors/raters able to provide precise subjective evaluations and/or diagnosis as well as for testing the efficacy of rater training programs.

REFERENCES

1. Nelson KP, Edwards D. Measures of agreement between many raters for ordinal classifications. *Stat Med*. 2015;34(23):3116-3132.
2. Tsai M-Y. Concordance correlation coefficients estimated by generalized estimating equations and variance components for longitudinal repeated measurements. *Stat Med*. 2017;36(8):1319-1333.
3. Rossi F. Assessing sensory panelist performance using repeatability and reproducibility measures. *Food Qual Preference*. 2001;12(5):467-479.
4. Geier U, Büssing A, Kruse P, Greiner R, Buchecker K. Development and application of a test for food-induced emotions. *PloS One*. 2016;11(11):1-17.
5. Hanea AM, McBride MF, Burgman MA, Wintle BC. Classical meets modern in the idea protocol for structured expert judgement. *J Risk Res*. 2018;21(4):1-17.
6. Maire J-L, Pillet M, Baudet N. Gage R2&E2: An effective tool to improve the visual control of products. *Int J Qual Reliab Manage*. 2013;30(2):161-176.
7. de Mast J, van Wieringen WN. Modeling and evaluating repeatability and reproducibility of ordinal classifications. *Technometrics*. 2010;52(1):94-106.
8. See JE. Visual inspection reliability for precision manufactured parts. *Hum Factors*. 2015;57(8):1427-1442.
9. Bashkansky E, Gadrich T, Knani D. Some metrological aspects of the comparison between two ordinal measuring systems. *Accredit Qual Assur*. 2011;16(2):63-72.
10. Pendrill LR, Fisher WPJr. Quantifying human response: linking metrological and psychometric characterisations of man as a measurement instrument. *J Phys: Conf Ser. IOP Publishing*. 2013;459(1):1-7.
11. Pendrill L. Man as a measurement instrument. *NCSLI Measure*. 2014;9(4):24-35.
12. Pendrill LR. Using measurement uncertainty in decision-making and conformity assessment. *Metrologia*. 2014;51(4): S206-S218.
13. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. *J Biopharm Stat*. 2007;17(4):529-569.
14. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep*. 1966;19(1):3-11.
15. Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. *Stat Methods Med Res*. 2004;13(4):251-271.
16. Blackman NJ-M, Koval JJ. Interval estimation for Cohen's kappa as a measure of agreement. *Stat Med*. 2000;19(5):723-741.
17. Ludbrook J. Statistical techniques for comparing measurers and methods of measurement: a critical review. *Clin Exp Pharm Physiol*. 2002;29(7):527-536.
18. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76(5):378-382.
19. Conger AJ. Integration and generalization of kappas for multiple raters. *Psychol Bull*. 1980;88(2):322-328.
20. Scott WA. Reliability of content analysis: the case of nominal scale coding. *Public Opin Q*. 1955;19(3):321-325.
21. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37-46.
22. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol*. 2008;61(1):29-48.
23. Guttman L. The test-retest reliability of qualitative data. *Psychometrika*. 1946;11(2):81-95.
24. Holley JW, Guilford JP. A note on the g index of agreement. *Educ Psychol Meas*. 1964;24(4):749-753.
25. Bennett EM, Alpert R, Goldstein AC. Communications through limited-response questioning. *Public Opin Q*. 1954;18(3):303-308.
26. Janson S, Vegelius J. On generalizations of the g index and the phi coefficient to nominal scales. *Multivariate Behav Res*. 1979;14(2):255-269.
27. De Mast J, Van Wieringen WN. Measurement system analysis for categorical measurements: agreement and kappa-type indices. *J Qual Technol*. 2007;39(3):191.
28. Erdmann TP, De Mast J, Warrens MJ. Some common errors of experimental design, interpretation and inference in agreement studies. *Stat Methods Med Res*. 2015;24(6):920-935.
29. Bashkansky E, Dror S, Ravid R, Grabov P. Effectiveness of a product quality classifier. *Qual Control Appl Stat*. 2008;53(3):291-292.
30. Cicchetti DV, Allison T. A new procedure for assessing reliability of scoring eeg sleep recordings. *Am J EEG Technol*. 1971;11(3):101-110.
31. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas*. 1973;33(3):613-619.
32. Gwet KL. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Gaithersburg, USA: Advanced Analytics, LLC; 2014.
33. Warrens MJ. Power weighted versions of Bennett, Alpert, and Goldstein's S. *J Math*. 2014;2014:9.
34. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med*. 2000;19(9):1141-1164.
35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
36. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. New York, NY: John Wiley & Sons; 2013.

37. Altman DG. *Practical Statistics for Medical Research*. London: CRC press; 1990.
38. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther*. 2005;85(3):257-268.
39. Everitt BS. *The Analysis of Contingency Tables*. Boca Raton, FL: CRC Press; 1992.
40. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol*. 1993;46(12):1417-1432.
41. Altaye M, Donner A, Eliasziw M. A general goodness-of-fit approach for inference procedures concerning the kappa statistic. *Stat Med*. 2001;20(16):2479-2488.
42. Klar N, Lipsitz SR, Parzen M, Leong T. An exact bootstrap confidence interval for κ in small samples. *J R Stat Soc: Ser D (The Statistician)*. 2002;51(4):467-478.
43. Chmura Kraemer H, Periyakoil VS, Noda A. Kappa coefficients in medical research. *Stat Med*. 2002;21(14):2109-2129.
44. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol*. 2012;8(1):23-34.
45. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. *Theriogenology*. 2010;73(9):1167-1179.

Amalia Vanacore is assistant professor in Statistics for Experimental and Technological Research at the Department of Industrial Engineering of the University of Naples "Federico II." She is the author of several scientific papers published in proceedings of international conferences and in international journals. She is member of the Italian Statistical Society and the ENBIS-European Network for Business and Industrial Statistics.

Maria Sole Pellegrino is a PhD student at the Department of Industrial Engineering of University of Naples "Federico II." She graduated cum laude in Management Engineering in 2014. Her research interest includes design and analysis of experiments, quality engineering and agreement studies.

How to cite this article: Vanacore A, Pellegrino MS. RRep: A composite index to assess and test rater precision. *Qual Reliab Engng Int*. 2018;1–11. https://doi.org/10.1002/qre.2334

Bibliography

[1] A. Agresti. A model for agreement between ratings on an ordinal scale. *Biometrics*, pages 539–548, 1988.

[2] A. Agresti and J. B. Lang. Quasi-symmetric latent class models, with application to rater agreement. *Biometrics*, pages 131–139, 1993.

[3] D. G. Altman. *Practical statistics for medical research.* CRC press, 1990.

[4] J. J. Bartko. The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, 19(1):3–11, 1966.

[5] E. Bashkansky, S. Dror, R. Ravid, and P. Grabov. Effectiveness of a product quality classifier. *Quality control and applied statistics*, 53(3): 291–292, 2008.

[6] E. Bashkansky, T. Gadrich, and D. Knani. Some metrological aspects of the comparison between two ordinal measuring systems. *Accreditation and Quality Assurance*, 16(2):63–72, 2011.

[7] E. M. Bennett, R. Alpert, and A. Goldstein. Communications through limited-response questioning. *Public Opinion Quarterly*, 18(3):303–308, 1954.

[8] N. J.-M. Blackman and J. J. Koval. Interval estimation for Cohen's kappa as a measure of agreement. *Statistics in medicine*, 19(5):723–741, 2000.

[9] R. L. Brennan and D. J. Prediger. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3):687–699, 1981.

[10] R. K. Burdick, C. M. Borror, and D. C. Montgomery. A review of methods for measurement systems capability analysis. *Journal of Quality Technology*, 35(4):342–354, 2003.

[11] J. Carpenter and J. Bithell. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in medicine*, 19(9):1141–1164, 2000.

[12] D. V. Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4):284, 1994.

[13] D. V. Cicchetti and T. Allison. A new procedure for assessing reliability of scoring eeg sleep recordings. *American Journal of EEG Technology*, 11 (3):101–110, 1971.

[14] D. V. Cicchetti and A. R. Feinstein. High agreement but low kappa: Ii. resolving the paradoxes. *Journal of clinical epidemiology*, 43(6):551–558, 1990.

[15] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[16] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[17] A. J. Conger. Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88(2):322, 1980.

[18] J. De Mast. Agreement and kappa-type indices. *The American Statistician*, 61(2):148–153, 2007.

[19] J. De Mast and W. N. Van Wieringen. Measurement system analysis for categorical measurements: Agreement and kappa-type indices. *Journal of Quality Technology*, 39(3):191–202, 2007.

[20] J. De Mast and W. N. Van Wieringen. Measurement system analysis for categorical measurements: agreement and kappa-type indices. *Journal of Quality Technology*, 39(3):191, 2007.

[21] J. de Mast and W. N. van Wieringen. Modeling and evaluating repeatability and reproducibility of ordinal classifications. *Technometrics*, 52(1): 94–106, 2010.

[22] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap.* CRC press, 1994.

[23] T. P. Erdmann, J. De Mast, and M. J. Warrens. Some common errors of experimental design, interpretation and inference in agreement studies. *Statistical methods in medical research*, 24(6):920–935, 2015.

[24] B. D. Eugenio and M. Glass. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101, 2004.

[25] A. R. Feinstein and D. V. Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6): 543–549, 1990.

[26] S. R. Fisher. Statistical methods for research workers.(13-th edition.) oliver and boyd. *Edinburgh, London*, 1958.

[27] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[28] J. L. Fleiss and D. V. Cicchetti. Inference about weighted kappa in the non-null case. *Applied Psychological Measurement*, 2(1):113–117, 1978.

[29] J. L. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973.

[30] J. L. Fleiss, J. Cohen, and B. Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological bulletin*, 72(5):323, 1969.

[31] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[32] T. Gadrich and E. Bashkansky. Ordanova: analysis of ordinal variation. *Journal of Statistical Planning and Inference*, 142(12):3174–3188, 2012.

[33] T. Gadrich, E. Bashkansky, and R. Zitikis. Assessing variation: a unifying approach for all scales of measurement. *Quality & Quantity*, 49(3):1145–1167, 2015.

[34] M. J. Gardner and D. G. Altman. Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal (Clin Res Ed)*, 292(6522):746–750, 1986.

[35] L. A. Goodman and W. H. Kruskal. Measures of association for cross cliassification. *Journal of the American Statistical Association*, 49:1732–1769, 1954.

[36] L. Guttman. The test-retest reliability of qualitative data. *Psychometrika*, 11(2):81–95, 1946.

[37] K. Gwet. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical methods for inter-rater reliability assessment*, 1(6):1–6, 2002.

[38] K. L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

[39] K. L. Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[40] D. P. Hartmann. Considerations in the choice of interobserver reliability estimates. *Journal of applied behavior analysis*, 10(1):103–116, 1977.

[41] A. F. Hayes. *Statistical methods for communication science*. Routledge, 2009.

[42] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[43] J. W. Holley and J. P. Guilford. A note on the g index of agreement. *Educational and psychological measurement*, 24(4):749–753, 1964.

[44] S. Janson and J. Vegelius. On generalizations of the g index and the phi coefficient to nominal scales. *Multivariate Behavioral Research*, 14(2): 255–269, 1979.

[45] M. Kendall and J. Gibbons. Rank correlation methods, trans. *JD Gibbons (5th edn ed.). Edward Arnold: London*, 1990.

[46] N. Klar, S. R. Lipsitz, M. Parzen, and T. Leong. An exact bootstrap confidence interval for κ in small samples. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 51(4):467–478, 2002.

[47] K. Klaus. Content analysis: An introduction to its methodology, 1980.

[48] J. Kottner, L. Audigé, S. Brorson, A. Donner, B. J. Gajewski, A. Hróbjartsson, C. Roberts, M. Shoukri, and D. L. Streiner. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International journal of nursing studies*, 48(6):661–671, 2011.

[49] K. Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[50] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[51] F. M. Lord. *Applications of item response theory to practical testing problems*. Routledge, 2012.

[52] F. M. Lord and M. R. Novick. *Statistical theories of mental test scores*. IAP, 2008.

[53] G. N. Masters. A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, 1982.

[54] D. C. Montgomery and G. C. Runger. Gauge capability and designed experiments. Part I: basic methods. *Quality Engineering*, 6(1):115–135, 1993.

[55] D. C. Montgomery and G. C. Runger. Gauge capability analysis and designed experiments. Part II: experimental design models and variance component estimation. *Quality Engineering*, 6(2):289–305, 1993.

[56] S. R. Munoz and S. I. Bangdiwala. Interpretation of kappa and b statistics measures of agreement. *Journal of Applied Statistics*, 24(1):105–112, 1997.

[57] E. Muraki. A generalized partial credit model: Application of an em algorithm. *ETS Research Report Series*, 1992(1):1–30, 1992.

[58] B. Muthen. Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, 22(1-2):43–65, 1983.

[59] L. Pendrill. Man as a measurement instrument. *NCSLI Measure*, 9(4): 24–35, 2014.

[60] L. Pendrill. Using measurement uncertainty in decision-making and conformity assessment. *Metrologia*, 51(4):S206–S218, 2014.

[61] L. Pendrill and W. P. Fisher Jr. Quantifying human response: Linking metrological and psychometric characterisations of man as a measurement instrument. *Journal of Physics: Conference Series. IOP Publishing*, 459 (1):1–7, 2013.

[62] C. Roberts and R. McNamee. A matrix of kappa-type coefficients to assess the reliability of nominal scales. *Statistics in medicine*, 17(4):471–488, 1998.

[63] W. A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325, 1955.

[64] M. M. Shoukri. *Measures of interobserver agreement and reliability.* CRC press, 2010.

[65] P. E. Shrout. Measurement reliability and agreement in psychiatry. *Statistical methods in medical research*, 7(3):301–317, 1998.

[66] P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[67] S. Siegel. Castellan. nonparametric statistics for the social sciences, 1988.

[68] A. Vanacore and M. S. Pellegrino. Checking quality of sensory data by assessing intra/inter panelist agreement. In *Proceedings of 8th Scientific Conference on INNOVATION & SOCIETY, Statistical Methods for Evaluation and Quality - IES 2017*, pages 1–4. University of Naples Federico II, 2017.

[69] A. Vanacore and M. S. Pellegrino. An agreement-based approach for reliability assessment of students' evaluations of teaching. In *Proceedings of the 3rd International Conference on Higher Education Advances*, pages 1286–1293. Editorial Universitat Politècnica de València, 2017.

[70] A. Vanacore and M. S. Pellegrino. Characterizing the extent of rater agreement via a non-parametric benchmarking procedure. In *Proceedings of the Conference of the Italian Statistical Society SIS 2017. Statistics and Data Science: new challenges, new generations*, pages 999–1004. Firenze University Pres, 2017.

[71] A. Vanacore and M. S. Pellegrino. Inferring rater agreement with ordinal classification. In *Convegno della Società Italiana di Statistica: New Statistical Developments in Data Science.* Springer, 2017.

[72] A. Vanacore and M. S. Pellegrino. Benchmarking rater agreement: probabilistic versus deterministic approach. In *Advanced Mathematical and Computational Tools in Metrology and Testing XI*, volume 89, pages 365–374. World Scientific, 2018.

[73] A. Vanacore and M. S. Pellegrino. How reliable are students' evaluations of teaching (sets)? a study to test student's reproducibility and repeatability. *Social Indicator Research*, 2018.

[74] A. Vanacore and M. S. Pellegrino. Checking quality of sensory data via an agreement-based approach. *Quality & Quantity*, pages 1–12, 2018.

[75] A. Vanacore and M. S. Pellegrino. Robustness of agreement in ordinal classifications. In *Proceedings of 18th Annual Conference of the European Network for Business and Industrial Statistics - ENBIS 2018*, page 86. ENBIS Communications and Multimedia Centre at the Faculty of Economics, University of Ljubljana, Slovenia, 2018.

[76] A. Vanacore and M. S. Pellegrino. Rrep: A composite index to assess and test rater precision. *Quality and Reliability Engineering International*, 2018.

[77] A. Vanacore and M. S. Pellegrino. A comparative study of benchmarking procedures for interrater and intrarater agreement studies. In *Book of short Papers SIS 2018*. Pearson, 2018.

[78] S. B. Vardeman and E. S. VanValkenburg. Two-way random-effects analyses and gauge r&r studies. *Technometrics*, 41(3):202–211, 1999.

[79] D. J. Wheeler. Problems with gauge r&r studies. In *ASQC Quality Congress Transactions*, volume 46, pages 179–185, 1992.

[80] A. Zapf, S. Castell, L. Morawietz, and A. Karch. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC medical research methodology*, 16(1):93, 2016.