Measures of interrater agreement when each target is evaluated by a different group of raters

Giuseppe Bove

1. Introduction

Measures of interrater agreement like *kappa* of Cohen (and its weighted versions) and intraclass correlations are usually defined for ratings regarding a group of targets (subjects or objects), each rated by the same group of raters. This happens when the agreement among clinical diagnoses provided by more physicians on the same set of patients is analysed for identifying the best treatment for the patients, or when the agreement among ratings of educators who assess on a new ordinal rating scale the language proficiency of a corpus of argumentative (written or oral) texts is considered to test reliability of the new scale.

In other situations, the agreement between ratings is analysed in a group of targets where each target is evaluated by a different group of raters, like for instance when teachers in a school are evaluated by a questionnaire administered to all the pupils (students) in the classroom. In these situations, it is important to analyse the reliability of the judgments by a measure of agreement between ratings, but since the ordering of the ratings assigned to each target is irrelevant, the measure can only be defined starting from the single target level.

In this paper, an index is proposed to evaluate the agreement between raters for each single target rated on an ordinal scale, and to obtain also a global measure of the interrater agreement for the whole group of targets evaluated. The main features of the proposal will be illustrated in a study for the assessment of the behaviour of student teachers in the classroom. Data were collected in a research conducted in 2018 at Roma Tre University with students of the degree course in Formazione Primaria, during their experience of internship ("tirocinio").

2. Target-specific measures of interrater agreement

When ratings provided on a quantitative (interval or ratio) scale are analysed in a group of targets where each target is evaluated by a different group of raters, a first approach available to measure the level of agreement for the whole group of targets is based on the ANOVA one-way random model (e.g., Shrout & Fleiss, 1979, McGraw & Wong, 1996). The intraclass correlation (ICC) for this model is the between-target variance divided by the sum of the between-target variance and the error variance (this sum is the ratings total variance). A high value of ICC indicates a good agreement among raters, because it is obtained when the between-target variance exceeds the error variance (that includes the within-target variance) by a wide margin. However, a low ICC value is not necessarily an indication of poor agreement, because a severe restriction in the range of ratings assigned in good agreement by the raters can cause low values of the between-target variance and low values of the ICC (the restriction of variance problem, LeBreton et al., 2003).

To overcome this problem of the ICC, target-specific measures of interrater agreement were proposed to work separately with each target i in the corresponding row of ratings in the targets × raters data matrix. James et al. (1984) proposed the index

$$r_{WG,i} = 1 - \frac{s_i^2}{\sigma_E^2},$$

Giuseppe Bove, Roma Tre University, Italy, giuseppe.bove@uniroma3.it, 0000-0002-2736-5697

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giuseppe Bove, *Measures of interrater agreement when each target is evaluated by a different group of raters*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0106-3.28, in Enrico di Bella, Luigi Fabbris, Corrado Lagazio (edited by), *ASA 2022 Data-Driven Decision Making. Book of short papers*, pp. 157-162, 2023, published by Firenze University Press and Genova University Press, ISBN 979-12-215-0106-3, DOI 10.36253/979-12-215-0106-3

where s_i^2 is the observed variance of the ratings in profile *i,* σ_E^2 is the variance obtained from a theoretical null distribution representing a complete lack of agreement among raters (e.g., the uniform distribution). For raters in perfect agreement, we have $s_i^2 = 0$, with a corresponding value $r_{WG,i} = 1$. For a total lack of agreement, the observed variance approaches the variance obtained from the theoretical null distribution. This leads $r_{WG,i}$ to approach 0.

A global measure of agreement for the whole group of targets can be defined as the arithmetic average of the $r_{WG,i}$ values ($\bar{r}_{WG} = \frac{1}{N}\sum_{i=1}^{N} r_{WG,i}$). The accuracy of the index depends strongly on the specification of the null distribution, and negative values could be obtained. Other possible indices for quantitative scales are reviewed, for instance, in LeBreton & Senter (2008). Recently, Bove (2022) has considered the normalised standard deviation and the coefficient of variation as possible alternatives to ICC and $r_{WG,i}$.

All the approaches described regard quantitative scales and are not appropriate for ordinal and nominal scales. Most of the indices of interrater agreement proposed for ratings on an ordinal scale (frequently averages of the weighted *kappa* of Cohen calculated for each of the possible pairs of raters) are not suitable for ratings regarding a group of targets, each rated by a different group of raters.

In order to propose a new index of interrater agreement for ordinal scales, the representation of the profile of the ratings for target *i* on a *K*-level ordinal scale in Table 1 is considered,

Table 1 – Profile of the ratings for target *i* on a *K*-level ordinal scale

Target	*Level 1*	*Level 2*	*..Level k ..*	*Level K*	*Total*
i	r_{i1}	r_{i2}	$\ldots.r_{ik}\ldots..$	r_{iK}	R_i

where, r_{ik} is the number of raters assigning level *k* to target *i* and R_i is the number of raters that rate target *i.* We propose a general approach that defines target-specific interrater agreement indices as normalised indices of variability for the distribution in profile *i,* according to the measurement level of the scale. A global measure of agreement can be defined as the arithmetic average of the target-specific values of the indices.

So, for ordinal scales, the following index of interrater agreement can be considered (analogous with the measure of dispersion for ordinal variables, e.g., Leti, 1983),

$$\delta_i = 1 - \frac{D_i}{D_{max}} = 1 - \frac{2\sum_{k=1}^{K-1} F_{ik}(1 - F_{ik})}{D_{max}}$$

where F_{ik} is the cumulative proportion associated with level *k* of the scale in the response profile *i*, for *k=1,2,....,K*, D_{max} is the maximum of $D_i = 2\sum_{k=1}^{K-1} F_{ik}(1 - F_{ik})$, and it is $D_{max} = (\frac{K-1}{2})$ as R_i is even, and D_{max}=$(\frac{K-1}{2})(1 - \frac{1}{R_i^2})$ as R_i is odd.

The index δ_i is always nonnegative, it is $\delta_i = 1$ in the case of maximum agreement and $\delta_i = 0$ in the case of maximum disagreement. Some simulations and experiences with real applications suggest the following thresholds for the interpretation of the values assumed by the δ_i index: values lower than 0.6 indicate low to moderate agreement, values between 0.6 and 0.8 good agreement, above 0.8 excellent agreement. The index allows for the identification of particular targets for which agreement is low: this is not possible with measures like *kappa* or intraclass correlations. Besides, a global measure of agreement can be defined as the arithmetic average of the δ_i values obtained for the *N* targets ($\bar{\delta} = \frac{1}{N}\sum_{i=1}^{N} \delta_i$). The index is not affected by the possible concentration of ratings in a few levels of the scale, like it happens for the measures based on the ANOVA approach or for the *kappa*-type indices, and it does not depend on the definition of a null distributions like $r_{WG,i}$.

In the next section, an application will be shown in which teachers in a school are evaluated by a questionnaire administered to all the pupils in the classrooms, so each teacher is evaluated by a different group of pupils. In this situation, it is interesting to analyse the level of dispersion of the ratings in the classrooms with respect to each question of the questionnaire, in order to investigate aspects of rating's reliability. Then, a matrix $\Delta = (\delta_{ij})$ is defined where each row corresponds to a teacher and each column to a question, and the entry δ_{ij} is the value of δ_i computed in the classroom of teacher *i* for question *j* (an example is provided in Table 2). Entries of matrix Δ can be considered as similarities between teachers and questions. The values δ_{ij} can be depicted in a diagram by the *unfolding* model (originally proposed by Coombs (1964) for rectangular matrices of preference scores). The model is

$$f(\delta_{ij}) = p_{ij} = \sqrt{\sum_{s=1}^{t}(a_{is} - b_{js})^2} + \varepsilon_{ij}\,, \qquad (1)$$

where f is a monotone transformation, mapping the similarities δ_{ij} into a set of dissimilarities p_{ij} (e.g., $p_{ij} = 1 - \delta_{ij}$), a_{is} and b_{js} are the coordinates respectively of row (teacher) *i* and column (question) *j* on dimension *s* in an *t-dimensional* space and ε_{ij} is a residual term. It is worth to notice that the Euclidean distance model usually used in multidimensional scaling for square dissimilarity matrices (e.g., Borg & Groenen 2005) is a constrained version of model (1), because for each *j* it is required $b_{js} = a_{js}$.

So, a diagram for the pattern of relationships is obtained where each row (teacher) is represented as a point with coordinates a_{is} and each column (question) as a point with coordinates b_{js}. In the planar representation (*t*=2), the distance between row (teacher) i and column (question) j approximates the corresponding dissimilarity p_{ij} (so, for instance, we can detect in the diagram both the teachers and the questions with low/high levels of agreement of ratings in the classrooms). Distances within each of the two sets of the row-points and the column-points are only implicitly defined and do not have corresponding observed entries in the data matrix. Parameters in the model (1) are estimated by iterative algorithms that, starting from initial estimates of a_{is}^0, b_{js}^0 (*initial configuration*), iteratively decreases a least squares loss function moving vectors $\boldsymbol{a}_i^0 = (a_{i1}^0, a_{i2}^0, \ldots\ldots, a_{ir}^0)$ and $\boldsymbol{b}_j^0 = (b_{j1}^0, b_{j2}^0, \ldots\ldots, b_{jr}^0)$, until convergence to a minimum. An important point is picking a good initial configuration to avoid the problem of *local minima*.

3. Application

A reduced version for pupils of the Teachers' Educational Practices Questionnaire (TEP-Q, Catalano et al., 2014) was administered to evaluate a group of 24 female student teachers of Roma Tre University, during their training (internship) in several primary schools of the Italian region Lazio, in school year 2018. The questionnaire consists of the following 12 questions regarding teachers behaviour in the classroom: "In the class she was relaxed" (Q1),"Before each activity, she clearly explained what we had to do" (Q2), "When someone approached her, she turn to look at him" (Q3), "She help us to repeat one thing better if we were not so clear" (Q4), "When someone of us was saying something, she interrupted him" (Q5), "When she talked to us, she also used gestures (for example, she moved her hands)" (Q 6), "She yelled at the class when she get angry" (Q7), "If someone of us needed to be consoled, she has noticed it, even if he did not tell her" (Q8), "During the activities she told us we could help each other" (Q 9), "When she was tired, she complained in class" (Q 10), "She made us do group work" (Q 11), "She praised us when we deserved it" (Q 12). Answers were provided on a 4-levels Likert scale (1=almost never, 4=almost always).

For each student teacher, ratings were obtained from the pupils in the classroom (24 school classrooms, 418 pupils, 204 females, 214 males, aged between 7 and 12 years). For each student teacher i and each question j, the δ_{ij} value of the index was computed in order to analyse the reliability of the ratings provided by the pupils in the school classroom. Table 2 contains the matrix of the δ_{ij} values and in addition, in the last row, the average $\bar{\delta}_{.j}$ for each question.

Table 2 – Values δ_{ij} obtained for student teachers and questions in the twenty-four school classrooms.

STUDENT TEACHER	Q 1	Q 2	Q 3	Q 4	Q 5	Q 6	Q 7	Q 8	Q 9	Q 10	Q 11	Q 12
1	1.00	0.93	0.63	0.57	0.67	0.34	0.81	0.34	0.60	0.93	0.57	0.53
2	0.63	0.89	0.51	0.57	0.65	0.41	0.61	0.51	0.57	0.73	0.77	0.51
3	0.46	0.70	0.32	0.42	0.26	0.25	0.79	0.50	0.42	0.88	0.61	0.54
4	0.60	0.66	0.43	0.38	0.60	0.45	0.25	0.43	0.35	1.00	0.27	0.56
5	0.82	0.79	0.52	0.69	0.65	0.40	0.59	0.72	0.49	0.32	0.70	1.00
6	0.69	1.00	0.71	0.67	0.87	0.19	0.23	0.27	0.37	0.62	0.52	0.13
7	0.62	0.84	0.73	0.62	0.83	0.42	0.62	0.77	0.65	0.49	0.71	0.50
8	0.87	0.70	0.53	0.53	0.47	0.50	0.81	0.23	0.25	0.67	0.42	0.73
9	0.95	0.80	0.53	0.58	0.83	0.44	0.77	0.35	0.50	1.00	0.61	0.58
10	1.00	1.00	0.51	1.00	0.61	0.49	0.67	0.90	0.44	1.00	0.91	0.77
11	0.81	0.90	0.49	0.56	0.54	0.30	0.38	0.31	0.71	0.33	0.67	0.56
12	0.40	1.00	0.59	0.61	0.71	0.21	0.24	0.46	0.28	1.00	0.61	0.61
13	0.40	1.00	0.51	0.44	0.36	0.57	0.33	0.29	0.36	0.78	0.59	0.62
14	0.53	1.00	0.87	0.73	0.31	0.57	0.72	0.86	0.86	1.00	0.93	0.93
15	0.56	0.86	0.40	0.71	0.59	0.31	0.32	0.24	0.40	0.71	0.65	0.53
16	0.32	0.81	0.83	0.76	0.43	0.33	0.48	0.48	0.35	1.00	0.83	0.44
17	0.78	0.92	0.67	0.61	0.75	0.40	0.32	0.36	0.51	0.85	0.47	0.59
18	0.62	0.67	0.62	0.61	0.55	0.61	0.69	0.29	0.48	0.83	0.51	0.46
19	0.89	1.00	0.71	0.51	0.84	0.45	1.00	0.48	0.30	1.00	0.52	0.67
20	0.32	0.81	0.38	0.27	0.31	0.27	0.13	0.29	0.34	0.67	0.41	0.10
21	0.88	1.00	0.75	0.75	0.62	0.54	1.00	0.94	0.61	0.81	0.62	0.94
22	0.94	0.84	0.73	0.62	0.94	0.32	0.88	0.55	0.22	0.93	0.40	0.73
23	0.80	0.67	0.44	0.38	0.78	0.36	0.33	0.51	0.09	0.67	0.73	0.69
24	0.77	0.83	0.63	0.43	0.46	0.24	0.88	0.48	0.30	0.70	0.67	0.45
Averages $\bar{\delta}_{.j}$	**0.69**	**0.86**	**0.59**	**0.58**	**0.61**	**0.39**	**0.58**	**0.48**	**0.43**	**0.79**	**0.61**	**0.59**

Different levels of reliability characterize the twelve questions. Questions 2 and 10 have high values of the average index (0.86 and 0.79, respectively), that means the pupils usually agree in the responses (in several classrooms it is $\delta_{ij} = 1$). On the contrary, questions 6 and 9 have low values of the average index (0.39 and 0.43, respectively), that means the pupils frequently have different opinions about the aspects of teacher's behaviour considered in the two questions. The remaining questions show low to moderate levels of agreement in the pupil's responses (average values between 0.48 and 0.69).

It is also interesting to analyse the values of the index δ_{ij} respect to each student teacher (rows of the matrix in Table 2). For instance, student teachers 10, 14, 19 and 21 have usually high levels of agreement between the pupil's responses in the twelve questions, on the contrary student teacher 20 has low values of agreement except for questions 2 and 10.

Model (1) was applied to analyse in a diagram the relationships between student teachers and questions. It is assumed $p_{ij} = 1 - \delta_{ij}$ in model (1), this means that distances are inversely proportional to the values δ_{ij}.

In Figure 1, the solution for *t*=2 dimensions is provided (*Stress-I*=0.29). Distances between student teachers and questions represent the level of agreement of the responses for the questions in the classroom (the lower the distance the higher the agreement). Question 2, question 10 and, to a lesser extent, question 1 are located in the centre of the diagram, close to many points representing teachers, because they have usually high levels of agreement in the responses of the pupils in the school classrooms. Questions 6, 9 and 8 have high heterogeneity in many cases, so they are positioned far apart from many student teachers. Considering the student teachers, we observe that student teacher 20 is far from most questions because she has usually low values of agreement for the ratings obtained in her classroom. On the contrary, student teachers 10, 14 and 21 are near the centre of the diagram and close to many questions, a consequence of the homogeneity of ratings obtained on many questions.

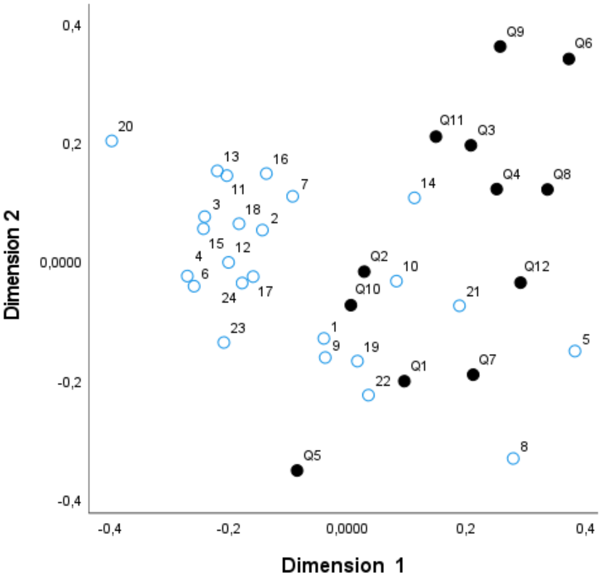

Figure 1: Unfolding of the δ_{ij} values for student teachers (empty circles) and questions (full black) in Table 2 (the higher δ_{ij} the smaller the distance)

4. Conclusion

A descriptive approach has been presented for the analysis of the agreement in ratings given to a group of targets, where each target is evaluated by a different group of raters. An index of interrater agreement defined at the single target level is proposed for ratings given on an ordinal scale, in a manner similar to the definition of the $r_{WG,i}$ index for ratings on a quantitative scale. Besides, a measure of agreement for the whole group of targets is obtained as the average of the target-specific values. The index presents some advantages respect to the methods based on ANOVA mean squares like intraclass correlation, and respect to many *kappa*-type indices. Besides, when the index is computed for a group of targets and more questions, it is shown that an unfolding model allows to analyse in a diagram the matrix of the values of the index obtained for each target-question pair.

The index proposed is mainly considered as a measure of size of the interrater agreement, therefore developments of this research may concern: 1) an accurate definition of reliable thresholds useful for the interpretation of the level of agreement in the applications; 2) the study of the sampling properties of the index.

References

Borg, I., Groenen, P.J.F. (2005). *Modern Multidimensional Scaling. Theory and Applications* (Second Edition). Springer, New York.

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

Catalano, M.G. Perucchini, P., Vecchio, G.M. (2014). The quality of teachers' educational practices: internal validity and applications of a new self-evaluation questionnaire. *Procedia-Social and Behavioral Sciences*, **141**, pp. 459-464.

Coombs, C.H. (1964). *A Theory of Data*. Wiley, New York.

James, L.R., Demaree, R.G., Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology,* **69***,* pp. 85–98.

LeBreton, J.M., Burgess, J.R.D., Kaiser, R.B., Atchley, E.K.P., & James, L.R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: are ratings from multiple sources really dissimilar?. *Organizational Research Methods*, **6** (1), pp. 80-128.

LeBreton, J.M., Senter, J.L. (2008). Answers to 20 questions about interrater reliability and interrater agreement, *Organizational Research Methods*, **11** (4), pp. 815-852.

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

McGraw, K.O., Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods,* **1**, pp. 30-46.

Shrout, P.E., Fleiss, J.L. (1979) Intraclass correlations: uses in assessing reliability. *Psychological Bullettin,* **86**, pp. 420–428