On the usefulness of interrater reliability coefficients

ten Hove, D.; Jorgensen, T.D.; van der Ark, L.A.

DOI
10.1007/978-3-319-77249-3_6

Publication date
2018

Document Version
Final published version

Published in
Quantitative Psychology

On the Usefulness of Interrater Reliability Coefficients

Debby ten Hove, Terrence D. Jorgensen and L. Andries van der Ark

Abstract For four data sets of different measurement levels, we computed 20 coefficients that estimate interrater reliability. The results show that the coefficients provide very different numerical values when applied to the same data. We discuss possible explanations for the differences among coefficients and suggest further research that is needed to clarify which coefficient a researcher should use to estimate interrater reliability.

Keywords Agreement · Interrater reliability coefficients · Estimates of interrater reliability

1 Introduction

Interrater reliability (IRR) entails the degree of agreement, consistency, or shared variance among two or more raters assessing the same subjects, expressed as a number between 0 (no agreement) and 1 (perfect agreement). On September 27, 2017, the term "inter-rater reliability"—including quotation marks—returned 173,000 hits on Google Scholar, which illustrates its academic importance. IRR also has societal relevance. For example, in the Netherlands an officer of Child Protection Services (Raad voor de Kinderbescherming) assesses the recidivism risks, risk factors, and protective factors of each juvenile delinquent (Van der Put et al. 2011). For the juvenile delinquent, the stakes are high because the assessment by the officer of Child Protection Services determines the district attorney's

D. ten Hove · T. D. Jorgensen · L. A. van der Ark (✉)
Research Institute of Child Development and Education, University of Amsterdam,
P. O. Box 15776, 1001 NG Amsterdam, The Netherlands
e-mail: L.A.vanderArk@uva.nl

D. ten Hove
e-mail: D.tenHove@uva.nl

T. D. Jorgensen
e-mail: T.D.Jorgensen@uva.nl

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 233, https://doi.org/10.1007/978-3-319-77249-3_6

sentencing recommendation. If the IRR of the assessment procedure were low, the sentencing recommendation would largely depend on the officer who did the assessment, which is highly undesirable.

In our experience, most researchers associate IRR with Cohen's (1960) kappa, but there is an abundance of coefficients available. Just for nominal data, Popping (1988) identified over 38 coefficients. Zhao et al. (2013) discussed 22 of these coefficients and found several were mathematically equivalent, resulting in 11 unique coefficients. The R package `irr` (Gamer et al. 2012) contains 17 different coefficients for various types of data that estimate the IRR. Some coefficients have different versions, which increases the number of coefficients even further. For example, the intraclass correlation coefficient (ICC) can be calculated using a one-way or two-way model, to estimate the consistency or agreement of either a single rating or the average across raters. Due to the abundance of coefficients, we found that preferring a particular coefficient to estimate IRR is hard to justify. Despite review articles on IRR (e.g., Gwet 2014; Hallgren 2012), it is unknown to what degree the estimated IRR depends on the coefficient.

It would be desirable if coefficients that can be applied to data with the same measurement level (e.g., nominal data) produce similar results. Therefore, this paper investigates to what degree the choice of coefficient affects the estimated IRR. In the discussion, we attempt to explain some of the differences among coefficients, and suggest research that is needed to answer the question: "Which coefficient should a researcher use to estimate interrater reliability?".

2 Methods

2.1 Data

We selected four datasets that are freely available from the R package `irr` (see Table 1; Gamer et al. 2012). Each dataset contained the ratings of R raters observing S subjects. The dataset *Diagnoses* (Fleiss 1971) consists of ratings by six psychiatrists classifying 30 patients into one of five nominal diagnostic categories:

Table 1 Characteristics of the four datasets

Dataset	S	R	NR	Min	Max	Level
Diagnoses	30	6	180	1	5	Nominal
Vision	7477	2	14954	1	3	Ordinal
Video	20	4	80	2	5	Interval
Anxiety	20	3	60	1	6	Interval

Note S = number of subjects; R = number of raters; NR = number of ratings ($S \times R$); *Min* = minimum score; *Max* = maximum score depression, personality disorder, schizophrenia, neurosis, or other. The dataset *Vision* (Stuart 1953) consists of the distance-vision performance of 7477 subjects using their left eye and their right eye. The two eyes are considered the two instruments (i.e., two raters). The ratings were measured on a scale from 1 (*low performance*) to 4 (*high performance*), which we treat as ordinal. The dataset *Video* is an artificial dataset consisting of four raters rating the credibility of 20 videotaped testimonies. Ratings could vary from 1 (*not credible*) to 6 (*highly credible*), though observed scores only ranged from 2 to 5. Technically, rating scales cannot yield interval-level data unless it can be known that the distance between adjacent integers is equivalent for any pair of adjacent integers across the range of the scale; however, unbiased results may be obtained by treating Likert-type rating scales containing at least five points as interval-level rather than ordinal-level data (Rhemtulla et al. 2012). Therefore, we treated the ratings as interval-level data. The dataset *Anxiety* is also an artificial dataset, in which three raters rated the anxiety of 20 subjects on a scale from 1 (*not anxious at all*) to 6 (*extremely anxious*). The measurement level of these ratings was also treated as interval.

2.2 IRR Coefficients

We considered 20 IRR coefficients from the R package `irr` (version 0.84; Gamer et al. 2012). We considered nine coefficients for nominal ratings (Table 2, top panel). Cohen's kappa (κ; Cohen 1960) can be used only for nominal ratings with two raters. Weighted versions of κ have been derived that can also be used only for nominal ratings with two raters (Cohen 1968). The weights reflect the amount of disagreement between the raters. We calculated two weighted κ versions: κ with equal weights (κ_W) and with squared weights (κ_{W^2}). Three generalizations of κ were available to assess nominal data with more than two raters: Fleiss' kappa (κ_{Fleiss}; Fleiss 1971), Conger's exact kappa (κ_{Exact}; Conger 1980), and Light's kappa (κ_{Light}; Light 1971). The percent agreement, Krippendorff's (1980) alpha, and coefficient iota (Janson and Olson 2001) each have a version for several measurement levels, including nominal-level ratings. Their coefficients for nominal ratings are denoted PA_N, α_N, and ι_N, respectively.

We considered four coefficients for ordinal ratings (Table 2, central panel). Kendall's (1948) W and the mean of Spearman's rank-order correlation ($\bar{\rho}$; Spearman 1904) have been designed specifically for ordinal data, whereas the percent agreement and Krippendorff's (1980) alpha have a version for ordinal ratings. The latter two coefficients are denoted PA_O and α_O, respectively.

We considered seven coefficients for interval-level ratings (Table 2, bottom panel). Each coefficient can also be applied to ratio-level ratings. The percent agreement, Krippendorff's (1980) alpha, and coefficient iota (Janson and Olson 2001) have a version for interval ratings. These coefficients are denoted PA_I, α_I, and ι_I respectively. For the Finn (1970) coefficient and the ICC (Shrout and Fleiss

Table 2 Characteristics of the 20 IRR coefficients used in this study

Symbol	Name	*SE*	NHST	Miss	$R > 2$
Nominal level					
κ	Cohen's kappa	●	●		
κ_W	Weighted kappa (equal weights)	●	●		
κ_{W^2}	Weighted kappa (squared weights)	●	●		
κ_{Fleiss}	Fleiss' kappa	●	●		●
κ_{Exact}	Conger's exact kappa	●	●		●
κ_{Light}	Light's kappa	●	●		●
PA_N	Percent agreement	●			●
α_N	Krippendorff's alpha	●		●	●
ι_N	Coefficient iota				●
Ordinal level					
W	Kendall's *W*		●		●
$\bar{\rho}$	Mean Spearman's rank correlation				●
PA_O	Percent agreement	●			●
α_O	Krippendorff's alpha	●		●	●
Interval level					
PA_I	Percent agreement	●			●
α_I	Krippendorff's alpha	●		●	●
ι_I	Coefficient iota				●
Finn_2	Finn's coefficient (two-way)		●		●
ICC_2	Intraclass correlation coefficient (two-way)	●	●		●
$\bar{r}$	Mean Pearson's correlation				●
A	Robinson's *A*				●

Note SE = standard errors are available; *NHST* = null-hypothesis significance test is available; *Miss* = missing data can be handled by other methods than listwise deletion; $R > 2$ = the method can handle more than two raters

1979), we specified two-way models to treat both raters and subjects as each being randomly drawn from a population, which is often the case in social and behavioral research. In addition, for the ICC we computed the level of consistency rather than the level of absolute agreement. Furthermore, we computed the mean of Pearson's product-moment correlation coefficients ($\bar{r}$; Pearson 1895) and Robinson's measure of agreement (A; Robinson 1957).

We excluded three coefficients of the R package `irr` from our analyses, because they clearly measured something different than the IRR: the Stuart-Maxwell coefficient (Maxwell 1970) and the Bhapkar (1966) coefficient assess homogeneity in marginal distributions, and the coefficient of Eliasziw et al. (1994) estimates intrarater reliability (i.e., consistency of repeated ratings from the same rater).

2.3 Analyses

For the nominal dataset (*Diagnoses*), we applied only nominal IRR coefficients. For the ordinal dataset (*Vision*), we applied all ordinal, nominal, and interval-level IRR coefficients, with the exception of α_N and α_I. The results of interval-level coefficients are interesting because researchers frequently treat Likert-type scales as though they are continuous. The results of nominal IRR coefficients are interesting when the ordering is not of primary interest in the application at hand. Therefore, for the interval-level datasets (*Video* and *Anxiety*), we also computed all nominal, ordinal, and interval-level IRR coefficients, with the exception of $PA_N, PA_O, \alpha_N, \alpha_O,$ and ι_N.

We investigated the range of values obtained by these coefficients. We also investigated whether the choice of coefficient affects the conclusion about the IRR using the heuristic labels suggested by Landis and Koch (1977) for the use of κ: negative values indicate a poor IRR, values between 0 and 0.20 indicate a slight IRR; values between 0.21 and 0.40 indicate a fair IRR; values between 0.41 and 0.60 indicate a moderate IRR; values between 0.61 and 0.80 indicate a substantial IRR, and values between 0.81 and 1.00 indicate an almost perfect IRR.

Furthermore, we investigated the following aspects of the IRR coefficients in Table 2, by checking the literature and the functions of the package `irr`: Are standard errors available? Is it possible to conduct null-hypothesis significance testing? Are missing data allowed? And if so, how can missing data be handled? How many raters are allowed?

3 Results

Table 3 shows the variability of the evaluated IRR coefficients as estimated for the four datasets. For the nominal-level dataset *Diagnoses*, the six available IRR coefficients ranged from 0.17 (*PA*) to 0.46 (κ_{Light}; $M = 0.40$, $SD = 0.11$). For the ordinal-level dataset *Vision*, the IRR coefficients ranged from 0.60 (several coefficients) to 0.85 (*W*; $M = 0.69$, $SD = 0.09$), but from 0.71 (several coefficients) to 0.85 if only ordinal IRR coefficients are considered. For the interval-level dataset *Video*, the IRR coefficients ranged from 0.04 (κ_{Fleiss}) to 0.92 (Finn; $M = 0.26, SD = 0.24$), but from 0.10 (α_I) to 0.92 if only interval-level IRR coefficients are considered. For the interval-level dataset *Anxiety*, the IRR coefficients ranged from −0.04 (κ_{Fleiss}) to 0.54 (*W*; $M = 0.22$, $SD = 0.21$), but from 0.00 (PA_I) to 0.50 (Finn$_2$) if only interval-level IRR coefficients are considered.

Table 3 (cf. the asterisks next to the values) also shows that the interpretation of the IRR of a dataset by means of the benchmarks of Landis and Koch (1977) depends on the choice of coefficient. For the dataset *Diagnoses*, the IRR could be labelled either slight, fair, or moderate; for the dataset *Vision*, the IRR could be labelled either moderate, substantial, or almost perfect; for the dataset *Video*, the

Table 3 IRR estimates for 20 coefficients on 4 datasets

Coefficient	Diagnoses	Vision	Video	Anxiety
Nominal level				
κ	[a]	0.60^{*}	[a]	[a]
κ_W	[a]	0.65^{**}	[a]	[a]
κ_{W^2}	[a]	0.60^{*}	[a]	[a]
κ_{Fleiss}	**0.43***	0.60^{*}	0.04	– 0.04
κ_{Exact}	**0.44***	0.60^{*}	0.10	– 0.02
κ_{Light}	**0.46***	0.60^{*}	0.07	– 0.02
PA_N	**0.17**	[b]	[b]	[b]
α_N	**0.43***	[b]	[b]	[b]
ι_N	**0.44***	[b]	[b]	[b]
Ordinal level				
W	[c]	**0.85*****	0.39	0.54^{*}
$\bar{\rho}$	[c]	**0.71****	0.24	0.34
PA_O	[c]	**0.71****	[b]	[b]
α_O	[c]	**0.71****	[b]	[b]
Interval level				
PA_I	[c]	[b]	**0.35**	**0**
α_I	[c]	[b]	**0.10**	**0.16**
ι_I	[c]	0.60^{*}	**0.15**	**0.19**
Finn$_2$	[c]	0.78^{**}	**0.92*****	**0.50***
ICC_2	[c]	0.70^{**}	**0.16**	**0.20**
$\bar{r}$	[c]	0.70^{**}	**0.24**	**0.28**
A	[c]	0.85^{***}	**0.40**	**0.48***
Ranges of values				
Range[d]	**0.17 – 0.46**	**0.71 – 0.85**	**0.10 – 0.92**	**0.00 – 0.50**
Range[e]	0.17 – 0.46	0.60 – 0.85	0.04 – 0.92	– 0.04 – 0.54

Note *coefficient greater than 0.40 (moderate IRR)

**coefficient greater than 0.60 (substantial IRR)

***coefficient greater than 0.80 (almost perfect IRR)

[a] coefficient cannot be computed because the number of raters is greater than 2

[b] coefficient was not computed because a version of the coefficient that applies to another measurement level was computed

[c] coefficient was not computed because the measurement level of data is nominal

[d] range of all IRR coefficients that match the measurement level of the ratings

[e] range of all IRR coefficients

Estimates that correspond to the correct measurement level are printed in boldface

IRR could be labeled anywhere from slight to almost perfect; and for dataset *Anxiety*, the IRR could be labelled either poor, slight, fair, or moderate.

For 13 of the 20 coefficients, standard errors were available (Table 2). To the best of our knowledge, for the other coefficients, standard errors are not available.

For nine coefficients, a test statistic is available that tests whether the coefficient equals zero.

Although no dataset contained missing values, it is worth noting that the package `irr` handles missing data differently for different coefficients. Coefficients α_N, α_O, and α_I use all available data by counting disagreements among any observed pair of ratings on the same subject (i.e., pairwise deletion). Coefficients ι_N and ι_I do not allow missing ratings (i.e., the software will return a missing value for the coefficient when any ratings are missing), whereas all other coefficients handle missing data by listwise deletion.

4 Discussion

The results showed that the coefficients provide very different numerical values when applied to the same dataset. Depending on the choice of the coefficient, the IRR label for a single dataset can range from poor to almost perfect. This seriously questions the usefulness of IRR coefficients. We limited ourselves to coefficients available in the R packages `irr` (Gamer et al. 2012), so the ranges may be even wider if more coefficients were included. This problem should be investigated further.

The usefulness of the coefficients in this paper can be investigated only if IRR has a sound definition; however, a clear definition seems to be absent. Some coefficients (e.g., the ICC) are based on variance decomposition, which is compatible with the framework of generalizability theory (e.g., Vangeneugden et al. 2005), whereas other coefficients (e.g., *PA*) are derived from the concept of literal agreement. Coefficients that stem from different conceptualizations of IRR cannot all measure the same thing. In a recent discussion with Feng (2015), Krippendorff (2016) wrote: “I contend Feng discusses reliability measures with seriously mistaken conceptions of what reliability is to assure us of” (p. 139). We need to distinguish the different theories behind the IRR coefficients and come up with a more accurate terminology to identify competing conceptualizations of IRR. Only if the theories and models behind IRR are sorted out, we can start investigating why some IRR coefficients produce higher values than others, and we can separate the wheat from the chaff. In that respect, we believe the work of Zhao et al. (2013) is a valuable contribution. They explain, for example, the flaws of chance-corrected coefficients such as κ. Once we have selected estimates for different conceptualizations of IRR, we can deal with other issues identified in this study.

Another major problem is that few coefficients can handle missing data. This is problematic because ratings in the social and behavioral sciences can be expensive. For example, an assessment of a juvenile delinquent by an officer of Child Protection Services in The Netherlands (see our Introduction) takes approximately 6–8 h. A study investigating the IRR must allow for planned missingness because it is financially and practically impossible to have all officers assess all juvenile delinquents. Hence, a useful coefficient must be estimable with missing data.

We also found that for some coefficients, standard errors and confidence intervals cannot be computed and null-hypothesis testing is impossible. These standard errors, confidence intervals, and hypothesis tests should first be derived. Then the bias of all standard errors, the coverage of all confidence intervals, and the Type I error rate of all hypothesis tests should be investigated.

Finally, we used the benchmarks of Landis and Koch (1977). These benchmarks are considered to be the single most often used benchmarks (e.g., Gwet 2014, p. 164). The 42,000+ citations of the Landis and Koch paper on Google Scholar indicate at least their widespread use. A relevant question may be whether these benchmarks, which were designed for κ, can be used for coefficients stemming from different conceptualizations of IRR. In future research, it should be investigated whether different sets of heuristic rules should be provided for different types of coefficients.

References

Bhapkar, V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data. *Journal of the American Statistical Association, 61,* 228–235. https://doi.org/10.2307/2283057.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46. https://doi.org/10.1177/001316446002000104.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213–220. https://doi.org/10.1037/h0026256.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin, 88,* 322–328. https://doi.org/10.1037/0033-2909.88.2.322.

Eliasziw, M., Young, S. L., Woodbury, M. G., & Fryday-Field, K. (1994). Statistical methodology for the concurrent assessment of interrater and intrarater reliability: Using goniometric measurements as an example. *Physical Therapy, 74,* 777–788. https://doi.org/10.1093/ptj/74.8.777.

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology, 11,* 13–22. https://doi.org/10.1027/1614-2241/a000086.

Finn, R. H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement, 30,* 71–76. https://doi.org/10.1177/001316447003000106.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378–382. https://doi.org/10.1037/h0031619.

Gamer, M., Lemon, J., & Fellows, I., & Singh, P. (2012). irr: Various coefficients of interrater reliability and agreement [computer software]. https://CRAN.R-project.org/package=irr.

Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Gaithersburg, MD: Advanced Analytics, LLC.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*, 23–34. http://www.tqmp.org/RegularArticles/vol08-1/p023/p023.pdf.

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement, 61,* 277–289. https://doi.org/10.1177/00131640121971239.

Kendall, M. G. (1948). *Rank correlation methods*. London, UK: Griffin.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage.

Krippendorff, K. (2016). Misunderstanding reliability. *Methodology, 12,* 139–144. https://doi.org/10.1027/1614-2241/a000119.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174. https://doi.org/10.2307/2529310.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin, 76,* 365–377. https://doi.org/10.1037/h0031643.

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry, 116,* 651–655. https://doi.org/10.1192/bjp.116.535.651.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London, 58*, 240–242. http://www.jstor.org/stable/115794.

Popping, R. (1988). On agreement indices for nominal data. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric research* (pp. 90–105). London, UK: Palgrave Macmillan. https://doi.org/10.1007/978-1-349-19051-5_6.

Rhemtulla, M., Brosseau-Laird, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17,* 354–373. https://doi.org/10.1037/a0029315.

Robinson, W. S. (1957). The statistical measurement of agreement. *American Sociological Review, 22*, 17–25. http://www.jstor.org/stable/2088760.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101. https://doi.org/10.2307/1412159.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428. https://doi.org/10.1037/0033-2909.86.2.420.

Stuart, A. (1953). The estimation and comparison of strengths of association in contingency tables. *Biometrika, 40,* 105–110. https://doi.org/10.2307/2333101.

Van der Put, C. E., Spanjaard, H. J. M., van Domburgh, L., Doreleijers, T. A. H., Lodewijks, H. P. B., Ferwerda, H. B., et al. (2011). Ontwikkeling van het Landelijke Instrumentarium Jeugdstrafrechtketen (LIJ) [development of the national assessment procedure for youth criminal justice]. *Kind & Adolescent Praktijk, 10*, 76–83. http://www.tqmp.org/RegularArticles/vol08-1/p023/p023.pdf.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics, 61,* 295–304. https://doi.org/10.1111/j.0006-341X.2005.031040.x.

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association, 36,* 419–480. https://doi.org/10.1080/23808985.2013.11679142.