



When coders are reliable: The application of three measures to assess inter-rater reliability/agreement with doctor–patient communication data coded with the VR-CoDES

Ian Fletcher^{a,*}, Mariangela Mazzi^b, Matthias Nuebling^c

^a Division of Clinical Psychology, University of Liverpool, UK

^b Department of Public Health and Community Medicine and Public Health, Section of Clinical Psychology, University of Verona, Italy

^c GEB: Gesellschaft für Empirische Beratung mbH (Empirical Consulting), Denzlingen, Germany

ARTICLE INFO

Article history:

Received 1 September 2010

Received in revised form 21 December 2010

Accepted 7 January 2011

Keywords:

Inter-rater study

Kappa

Intraclass correlation coefficient

Sensitivity and specificity

VR-CoDES

ABSTRACT

Objective: To investigate whether different measures of inter-rater reliability will compute similar estimates with nominal data commonly encountered in communication studies. To make recommendations how reliability should be computed and described for communication coding instruments.

Methods: The raw data from an inter-rater study with three coders were analysed with; Cohen's κ , sensitivity and specificity measures, Fleiss's multirater κ_j , and an intraclass correlation coefficient (ICC).

Results: Minor differences were found between Cohen's κ and an ICC model across paired data (largest margin = 0.01). There were negligible differences between the multirater estimates e.g. κ_j (0.52) and ICC (0.53). Sensitivity analyses were in general agreement with the multirater estimates.

Conclusion: It is more practical to analyse nominal data with >2 raters with an appropriate model ICC for inter-rater studies, and little difference exists between Cohen's κ or an ICC.

Practice implication: Alternatives to Cohen's κ are readily available, but researchers need to be aware of the different ICC definitions. An ICC model should be fully described in reports. Investigators are encouraged to supply confidence limits with inter-rater data, and to revisit guidance regarding the relative strengths of agreement of reliability coefficients.

© 2011 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

The readers of this journal will be familiar with articles reporting observational studies that code interactions between health professionals and patients, which have been instrumental in the development of patient-centred communication. It is usual to train coders and to report whether they have been coding in the same manner i.e. reliably. The value of assessing reliability has been succinctly expressed, "If the measure is not reliable it cannot be expected to show lawful relationships with the other variables being measured" [1]. In other words, any subsequent analyses may produce results that do not accurately reflect reality, and the bedrock of these investigations rests on the reliability of the coders' judgements. A central issue of reliability focuses on the question of how to define and obtain 'true' scores. The problem of dealing with the impossibility of ever obtaining exact (true) scores has been addressed by Classical Test Theory which differentiated a

measurement into two components: its true score T , and the error e associated with the true score [2]. The formal definition of reliability calculates the ratio between the variance of the subjects ($\sigma_{\text{Subjects}}^2$) to the total variance ($\sigma_{\text{Subjects}}^2 + \sigma_{\text{Error}}^2$), with a score of one indicating no measurement error (ideal reliability) and zero no reliability [3] e.g.

$$\text{Reliability ratio} = \frac{\sigma_{\text{Subjects}}^2}{\sigma_{\text{Subjects}}^2 + \sigma_{\text{Error}}^2}$$

Investigators will also refer to 'agreement' and 'consistency' under the general heading of reliability [4,5], although they have different definitions. Agreement is understood to mean that raters have awarded identical codes or raw scores to the same subject. Whereas consistency infers that raters have ranked the subjects in the same order, but their codes/raw scores do not need to have been the same [6]. This article will focus on three measures that can be used to assess reliability from the perspective of agreement and will report the results of applying each statistic to a dataset generated from a specific coding scheme. These findings will be relevant to many coding schemes within doctor–patient commu-

* Corresponding author at: Division of Clinical Psychology, University of Liverpool, Whelan Building, Liverpool L69 3GB, UK. Tel.: +44 151 7945530; fax: +44 151 7945537.

E-mail address: ian.fletcher@liverpool.ac.uk (I. Fletcher).

nication. The research question was: RQ: Are similar levels of inter-rater reliability obtained from three measures of agreement applied to the same dataset?

1.1. Cohen’s κ

A commonly reported reliability statistic in doctor–patient communication is Cohen’s κ coefficient [7], which was devised for nominal data to correct for the inflation of κ due to chance agreement. The κ coefficient refers to the (chance-corrected) actual agreement as a proportion of the possible (chance-corrected) maximum agreement, with chance-corrected calculated from the proportion observed (p_o) minus the proportion expected (p_e) agreements and can be considered as “... another correlation, explaining some percentage of the variance” [8] e.g.

$$k = \frac{p_o - p_e}{1 - p_e}$$

Cohen’s κ has been extended to quantify levels of disagreements between codes with a weighted κ_w , such that partial agreement can be taken into consideration [9] i.e. coders who record either a patient cue or concern for an utterance have partial agreement, but, they would both fully disagree with a third rater who did not believe that the patient expressed any hint or explicit emotion in the same utterance. It has been shown that the results generated from a weighted κ_w , with quadratic weights, will be identical to those produced from an equivalent intraclass correlation coefficient [10] (ICC). Although rarely reported there are different forms of κ for multiple raters [10], such as κ_j which is available in some statistical computer packages. The issues of consistency and agreement do not arise with κ , because the simplest form, a 2×2 table, will only have two levels i.e. total agreement or disagreement. A problem with κ depends on the proportion of rated subjects in each category, and two datasets may show the same level of proportional agreement, but, the κ value can vary if there are different proportions in the categories [11]. Therefore, comparisons across inter-rater studies using κ are problematical unless information on the prevalence of the categories is also available.

1.2. Intra-class correlation coefficient

Investigators working with quantitative data have assessed inter-rater reliability with the intra-class correlation coefficient (ICC) based on the analysis of variance (ANOVA) [12]. Prior to calculating an ICC the assumptions of the different available definitions of ICC have to be considered. The objective for this study was to assess levels of agreement between coders at the level of the individual coder, which requires a two-way random effects model with the definition of absolute agreement and single

measures. This particular ICC has been classified variously as ICC (2, 1) [14], ICC (A, 1) [13], and ICC2 (A, 1) [6] and it enables the reliability estimates to be generalized to other populations of coders e.g.

$$r = \frac{\sigma_{Subjects}^2}{\sigma_{Subjects}^2 + \sigma_{Coders}^2 + \sigma_{Interaction}^2 + \sigma_{Error}^2}$$

1.3. Sensitivity and specificity

In the previous examples with κ and ICC all the coders were assumed to have equal importance. The third method of estimating inter-rater agreement compared codes from inexperienced coders to an expert coder who was familiar with the scheme. This is an analogous situation to examples from the medical literature in which a new measure is compared to a gold standard to assess its accuracy [15]. In this instance the expert coder may be regarded as the gold standard, because the novice coders were expected to agree with the expert’s coding decisions. In these situations researchers use the term ‘positive’ to refer to the presence of the condition of interest and ‘negative’ to the absence of the condition [16], which corresponds to coding the presence or absence of a behaviour of interest e.g. the presence/absence of a patient cue/concern. Sensitivity in this instance refers to the proportion of patient cues/concerns (positives) correctly identified by the novice coders in comparison to an expert coder, and specificity to the proportion of no cues/concerns (negatives) correctly detected by the inexperienced coders in comparison to the expert’s assessments [17] e.g.

$$\text{Sensitivity} = \frac{a}{a + b}, \quad \text{Specificity} = \frac{d}{d + c}$$

Data have to be coded to reflect the numbers of agreements (1) and disagreements (0) between the expert and novice coders [16] e.g. Fig. 1.

An advantage of the sensitivity and specificity compared to κ is that they are unaffected by the true prevalence of the condition [15].

2. Methods

2.1. Participants

A researcher experienced in coding doctor–patient communication with the coding scheme selected for the study and two post-graduate students studying for a doctorate in clinical psychology at the University of Liverpool.

2.2. Materials

A set of eight transcribed interviews that were randomly selected from a pool of 278 anonymised interviews for an inter-

		Expert rater		Total
		Behaviour Present (ER+)	Behaviour absent (ER-)	
Novice rater	Behaviour present (NR+)	a	b	a+b
	Behaviour absent (NR-)	c	d	c+d
Total		a+c	b+d	N=a+b+c+d

Fig. 1. Sensitivity and specificity of expert (ER) and novice rater (NR).

rater study between the three participants. The interviews were conducted between first year medical students enrolled at the Medical School in Liverpool in 2003, as part of a longitudinal investigation investigating the impact of communication skills training on patient-centredness. The eight selected transcripts featured three females (mean age 23.3, range 18–27) and five males (mean age 25, range 18–34) medical students. The simulators were a female in her 30s and a male in their 60s, and they were instructed to portray a patient attending a GP surgery with symptoms of anxiety or depression. The students were invited to interview the simulator and elicit their reasons for attending the surgery prior to a meeting with a GP, and each interview was timed to last for 5 min. The study was ethically approved by the Research Ethics Committee at the University of Liverpool.

2.3. Procedure

The two post-graduate students were training to use the VR-CoDES [18,19] to code a dataset of 278 interviews collected for the longitudinal study, and can be regarded as novice coders (A and B). They were coached by an experienced researcher who had contributed to the development of the VR-CoDES and was familiar with its application to medical interviews, and who was viewed as an 'expert' (rater C). A series of familiarization and training sessions had taken place prior to the inter-rater study, and there were no interim coding meetings/discussions during the inter-rater coding phase of the investigation i.e. a blind inter-rater study. The novice raters were instructed to only code simulator utterances in the transcripts with the relevant patient focused codes of the VR-CoDES (patient cues and concerns), as it was their initial experience with this coding scheme. The definition of a patient cue is, "... a verbal or non-verbal hint which suggests an underlying unpleasant emotion but lacks clarity", with a patient concern as "a clear and unambiguous expression of an unpleasant current or recent emotion that is explicitly verbalized with or without a stated issue of importance" [18]. Examples of patient cues are "My work is very stressful" or "I cannot concentrate", and patient concerns "I am depressed" or "I feel disgusted" [19]. The definitions and coding of patients' cues and concerns have been supported in a validation study with patients [20]. Previous practice experience with training novice coders had demonstrated that it was optimal to focus on and become reliable with patient behaviour codes followed by training to code health provider responses.

2.4. Data analysis

Simulator utterances that did not contain patient cues or concerns were coded '1' and utterances with cues or concerns were coded '2' and '3', respectively. The collapsed category that contained cues and concerns were all coded as '2'. The percentage of agreement and 95% confidence intervals for Fleiss's κ_j were calculated with SPSS (Statistical Package for the Social Sciences) syntax files available from <http://www.ccitonline.org/jking/homepage/interrater.html> using SPSS, Release Version 17.0.3. Inter-rater reliability/agreement was calculated with three different approaches; Cohen's κ for pairs of raters [7], Fleiss generalized κ_j for multi-raters [10], an ICC with a two-way random effects ANOVA model using the absolute agreement/single measure definition e.g. ICC2 (A, 1) [6], the sensitivity and specificity of the novice coders compared to the 'expert' rater [16]. The 95% confidence intervals for Cohen's κ were computed in Excel using the approximate standard error of κ and the standard calculation e.g. $\kappa - 1.96 se(\kappa)$ to $\kappa + 1.96 se(\kappa)$ [15]. Fleiss's κ_j was computed with Stata Version 11.1 and an ICC model was calculated with IBM

	A (novice) n (%)	B (novice) n (%)	C (expert) n (%)
Not coded	177 (78.0)	170 (74.9)	140 (61.7)
Cues	28 (12.3)	51 (22.5)	72 (31.7)
Concerns	22 (9.7)	6 (2.6)	15 (6.6)
Total	227	227	227

Fig. 2. Numbers of simulator utterances coded by rater.

SPSS for Windows, Release Version 18.0.2 (SPSS®, Inc., 2009, Chicago, IL, www.spss.com).

3. Results

The eight transcripts contained a total of 464 coded utterances, of which 227 (48.9%) were simulator dialogue. There were differences in the numbers of patient cues/concerns identified by the raters and due to the relatively low numbers of concerns, <10% were coded by each rater, and the patient cues and concerns were collapsed into a single category (Fig. 2). The decision to collapse subcategories of similar behaviours is common in inter-rater studies faced with low numbers of events, and has face validity when the behaviours can be justifiably classified into a superordinate category.

It was apparent that the novice raters A and B were coding a lower percentage of simulator utterances as patient cues or concerns compared to the expert rater (C) e.g. 22%, 25.1%, and 38.3%, respectively.

The raw data contingency tables have been presented to enable interested readers to replicate the analyses if so desired (Fig. 3), and the percentage of actual and expected agreement data were also computed for the pairs of raters e.g. AB, BC, and AC (Fig. 4). The two novice raters had the highest levels of agreement (82.8%), although the most notable feature were the levels of expected agreement across the three pairs of raters which all exceeded 55%, probably due to the high correspondence across utterances not containing cues or concerns apparent in Fig. 3 i.e. cell 1, 1.

The data from the series of analyses with the measures (Cohen's κ , sensitivity, Fleiss's κ_j , and ICC) present similar point estimates of reliability/agreement with only marginal differences found between κ and ICC for pairs of raters, and a very minor difference between κ_j and IIC (Fig. 5). Not surprisingly, the sensitivity analysis demonstrated that the novices were only in approximately 50% agreement with the expert's coding of utterances with patient cues/concerns, which was slightly higher than the estimates from the other analyses. The specificity analyses revealed high levels of concordance (>90%) between the novice and expert with respect

Table 1		Rater A	
		no cue/concern (1)	cues/concerns (2)
Rater B	no cue/concern (1)	154	16
	cues/concerns (2)	23	34
Table 2		Rater A	
		no cue/concern (1)	cues/concerns (2)
Rater C	no cue/concern (1)	138	2
	cues/concerns (2)	39	48
Table 3		Rater B	
		no cue/concern (1)	cues/concerns (2)
Rater C	no cue/concern (1)	129	11
	cues/concerns (2)	41	46

Fig. 3. The raw data contingency tables.

Rater	Percentage agreement	Expected agreement
A, B	82.8	63.9
A, C	81.9	56.5
B, C	77.1	55.8
A, B, C	0.81	*

*not available

Fig. 4. The percentage of agreement and expected agreement.

Rater	Cohen's κ (95% CI)	Sensitivity	Specificity	Fleiss's κ_j (95% CI)	ICC2(A,1) (95% CI)
A, B	0.52 (0.39, 0.66)	*	*	*	0.53 (0.42, 0.61)
A, C	0.58 (0.48, 0.69)	0.55	0.99	*	0.59 (0.44, 0.69)
B, C	0.48 (0.37, 0.60)	0.53	0.92	*	0.48 (0.37, 0.58)
A, B, C	*	*	*	0.52 (0.45, 0.6)	0.53 (0.45, 0.61)

*not available

Fig. 5. Reliability estimates.

to identifying the absence of cues/concerns, which was expected from scrutinising the raw data tables in Fig. 4.

4. Discussion and conclusion

4.1. Discussion

The RQ addressed in this study concerned the degree of equivalence across three measures of inter-rater agreement and it was clear that very similar point estimates were generated from the analyses, in particular the point estimates from the Cohen's κ , Fleiss's κ_j , and ICC results. The minor differences between the tests are likely to be a result of the approximated standard error having to be calculated for the κ statistics. It should be acknowledged that the sensitivity analyses were restricted to pairs of raters and have to be considered as a special case, although they were also in general accord with the two measures that were able to utilize the data from the three raters. It is unlikely that investigators will present the results from sensitivity and specificity analyses for publication. However, this study demonstrates that these particular assessments can assist researchers by their capacity to easily illustrate potential areas of discrepancy between novices and an 'expert' coder. The sensitivity analyses in particular indicate that different training methods, other than coding transcripts, may be more salient and economical to 'tune in' novice coders to the characteristics of patient cues/concerns e.g. using flash cards. Whereas, the specificity analyses highlighted that the novice coders did not require further training to identify those patient utterances which do not have any cues or concerns. Readers may ask the question; should be a person who trains other researchers in a coding scheme be viewed as an expert? This is a matter of debate, but, it is likely that the rules of any coding scheme will require some degree of interpretation and novice coders will primarily seek clarification from the individual providing instruction, who by default implicitly adopts the role of an 'expert' in the eyes of the inexperienced coders. The results from this investigation have been echoed by other researchers who have reported virtually identical results after comparing the results of analyses with Cohen's κ and equivalent ICCs using nominal data with two categories, and κ_w for three levels of categorical data [6]. Readers should note the caveat that the smallest numerical value in the dataset should be 'one' and not 'zero' i.e. utterances with no

behaviour(s) of interest are numerically recorded as '1'. In terms of considering the most appropriate measure to assess inter-rater reliability; this has been an issue that has received limited attention with the majority of the debate focused on the nature of the data i.e. nominal, ordinal, etc. Clearly, there are unlikely to be objections to disregarding the percentage of agreement as a suitable measure to assess and report inter-rater agreement, and as highlighted sensitivity analysis will be limited. Some authors have recommended that investigators should focus on ICCs for anything other than "... the most simple 2×2 tables ..." [6], or that ICCs alone should assess inter-rater data [21]. As researchers we are left with the choice between Cohen's κ and an ICC to analyse inter-rater reliability with datasets that only contain two categories. An advantage of an ICC analysis compared to the κ is that 95% confidence limits are routinely calculated with an ICC in statistical software, which is not the case with Cohen's κ . Calculating the appropriate confidence bounds for Cohen's κ is straightforward because programmes routinely present the approximate standard error of κ in the analysis output, but they are rarely reported. The argument for ICC strengthens with codes that have more than two categories. For instance, Fleiss's κ_j was only readily available in Stata which does not compute confidence limits for this test, and running the equivalent SPSS syntax file was problematical and only possible with SPSS Version 17.0.3. Investigators may be deterred by the problems in locating statistical software that successfully implements Fleiss's κ_j and calculates the associated confidence limits. It is appropriate to highlight the value of reporting confidence intervals, as it is usual practice for research articles to only present the point estimates of inter-rater studies. The point estimate is a very useful statistic. But, it does not inform readers of the range of values that include the point estimate, which can assist investigators to more accurately determine the value of an inter-rater study. It was clear from this study that despite the point estimates ranging from 0.48 to 0.59 (Cohen's κ , Fleiss's κ_j , ICC) the lower confidence limits were 0.37–0.45, which were low and not deemed acceptable. Considerable attention has been paid to trying to judge whether the estimates generated from inter-rater studies are sufficiently high to demonstrate raters are reliable/in agreement with each other. Many authors and investigators refer to guidelines outlined in the 1970s specific to Cohen's κ , which present bands of point estimates with judgements about the relative strength of agreement e.g. 0.41–0.60 'moderate'. The

authors did describe the benchmarks as, “. . . clearly arbitrary, they do provide useful “benchmarks” for the discussion of the specific example in Table 1” [22]. The guidelines have become widely reported and established in the literature despite criticism for being too generous [23,24], and a recent summary of the guidelines has stated that estimates below 0.60 should be simply disregarded and researchers should aim for 0.75 as a lower limit [6]. Clearly, the inter-rater results from this study fell outside of these limits, and consensus meetings were held with two further inter-rater studies before the two novice coders featured in this article were sufficiently reliable to code independently. The major limitation in this investigation was the reliance on a single dataset, although we believe that it reflects the reality of many investigators experiences when running an inter-rater study.

4.2. Conclusion

The results from the estimates generated from Cohen's κ and the ICC model in this study were in-line with other reports in the literature which gives assurance to their accuracy. Researchers coding doctor–patient interactions with schemes similar to the VR-CoDES may use the illustrated ICC model in preference to Cohen's κ secure in the knowledge that it will give the same point estimate, with the advantage that it also accommodates data generated from more than two coders.

4.3. Practice implications

The study has demonstrated the worth of considering an ICC analysis for inter-rater studies for nominal as well as interval/scale data, as long as an appropriate ICC model is selected and reported in subsequent publications. The discussion regarding confidence limits and judging whether an inter-rater study has adequately demonstrated reliability highlights the value of reporting confidence limits, and reconsidering the value of qualitative guidelines about relative strengths of agreement.

4.4. Relevance and application of comparing three inter-rater reliability coefficients for health communication research

We believe that the data from this study will assist investigators to select and report the most appropriate inter-rater reliability coefficient when coding doctor–patient interviews. The results demonstrate that researchers no longer have to base their decision solely on the level of the measurement i.e. nominal, interval, etc. that commonly restricts their selection to Cohen's κ , when suitable alternatives are readily available.

Conflict of interest

The authors are aware of no conflict of interest arising from this work.

Role of funding source

There was no dedicated funding given or sought to support this study.

Acknowledgements

We wish the acknowledge that this study would not have been possible without the time given by Rachel Hick and Tony Roach to code the transcripts.

References

- [1] Mitchell SK. Interobserver agreement, reliability, and generalisability of data collected in observational studies. *Psychol Bull* 1979;86:376–90.
- [2] Dunn G. Statistical evaluation of measurement errors, 2nd ed., London: Arnold; 2004.
- [3] Traub RE. Reliability for the social sciences: theory and applications. London: Sage; 1994.
- [4] Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. *Stat Med* 1994;13:2465–76.
- [5] Bakeman R, Gottman JM. Observing interaction: an introduction to sequential analysis, 2nd ed., Cambridge: Cambridge University Press; 1997.
- [6] Streiner DL, Norman GR. Health measurement scales, 4th ed., Oxford: Oxford University Press; 2008.
- [7] Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
- [8] Streiner DL, Norman GR. Biostatistics: the bare essentials, 3rd ed., London: BC Decker Inc; 2008.
- [9] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213–20.
- [10] Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;76:378–82.
- [11] Brennan RL, Prediger DJ. Coefficient kappa: some uses, misuses, and alternatives. *Educ Psychol Meas* 1981;41:687–99.
- [12] Bako JJ. The intraclass correlation coefficient as a measure of reliability. *Psychol Rep* 1966;19:3–11.
- [13] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996;1:30–46.
- [14] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420–8.
- [15] Altman DG. Practical statistics for medical research. London: Chapman Hall; 1991
- [16] Campbell MJ, Machin D. Medical statistics: a commonsense approach, 3rd ed., Chichester: John Wiley; 1999.
- [17] Rothman KJ, Greenland S, Lash TL. Modern epidemiology, 3rd ed., Philadelphia: Lippincott Williams & Wilkins; 2008.
- [18] Zimmermann C, Del Piccolo L, Bensing J, Bergvik S, De Haes H, Eide H, et al. Coding patient emotional cues and concerns in medical consultations: the verona coding definitions of emotional sequences (VR-CoDES). *Patient Educ Couns* 2010. doi: [10.1016/j.pec.2010.03.017](https://doi.org/10.1016/j.pec.2010.03.017).
- [19] Zimmermann C, Del Piccolo L, Finset A, Verona Network Verona Coding Definitions of Emotional Sequences (VR-CoDES). Cue and Concern Manual European Association for Communication in Health Care; 2009. Available from: <http://www.each.nl/>.
- [20] Eide H, Eide T, Rustoen T, Finset A. Patient validation of cues and concerns identified according to Verona coding definitions of emotional sequences (VR-CoDES): a video- and interview-based approach. *Patient Educ Couns* 2010. doi: [10.1016/j.pec.2010.04.036](https://doi.org/10.1016/j.pec.2010.04.036).
- [21] Berk RA. Generalizability of behavioural observation: a clarification of interobserver agreement and interobserver reliability. *Am J Ment Def* 1979;83:460–72.
- [22] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
- [23] Dunn G. Design and analysis of reliability studies: the statistical evaluation of measurement errors. London: Edward Arnold; 1989.
- [24] Shrout PE. Measurement reliability and agreement in psychiatry. *Stat Methods Med Res* 1998;7:301–17.