International Journal of Research in Medical Sciences
Nurjannah I et al. Int J Res Med Sci. 2017 Apr;5(4):1169-1175
www.msjonline.org
pISSN 2320-6071 | eISSN 2320-6012

DOI: http://dx.doi.org/10.18203/2320-6012.ijrms20171220

Review Article

Guidelines for analysis on measuring interrater reliability of nursing outcome classification

Intansari Nurjannah[1]*, Sri Marga Siwi[2]

[1]Basic and Emergency Nursing Department, School of Nursing, Faculty of Medicine, Universitas Gadjah Mada, Yogyakarta, Indonesia
[2]School of Nursing, Faculty of Medicine, Universitas Gadjah Mada, Yogyakarta, Indonesia

Received: 09 February 2017
Accepted: 09 March 2017

***Correspondence:**
Dr. Intansari Nurjannah,
E-mail: intansarin@ugm.ac.id

ABSTRACT

Indicators in nursing outcome classification (NOC) need to be tested for their validity and reliability. One method to measure reliability of NOC is by using interrater reliability. Kappa and percent agreement are common statistic analytical methods to be used together in measuring interrater reliability of an instrument. The reason for using these two methods at the same time is that those statistic analytical methods have easy reliability interpretation. Two possible conflicts may possibly emerge when there are asynchronies between kappa value and percent agreement. This article is aimed to provide guidance when a researcher faces these two possible conflicts. This guidance is referring to interrater reliability measurement using two raters.

Keywords: Interrater reliability, Kappa, Percent agreement

INTRODUCTION

There are several methods in measuring the result of nursing interventions. Nursing outcome classification (NOC) is one of the tools to measure the efficacy of nursing interventions. Even though NOC have been continuously developed since the first edition in 1997, publications related to this subject still need to be conducted.[1]

There are several methods in developing valid and reliable nursing outcome classification. One type of indicator in nursing outcome classification can be measured through observation. One of method to ensure the reliability of this observation instrument is by using interrater reliability with two statistical analysis methods namely kappa value and percent agreement. This article firstly, will review the usage of kappa and percent agreement for measuring interrater reliability and secondly, to provide a guidance to solve problem when this two-analysis statistic shows an opposite result.

Nursing outcome classification (NOC) is one of nursing outcome measurement developed by Mosby in 1997.[1] Outcome in NOC is stated in a concept of variables which represent patient or family caregiver's status in terms of behavior or perception. This behavior or perception is measured along a continuum in response to nursing interventions.[1] Each outcome in NOC has its definition and indicators which consist of likert scale from 1 to 5. There are 24 class and 109 outcomes of NOC. The 24 classes are divided into 6 domains namely functional health, physiologic health, psychosocial health, health knowledge and behaviour, perceived health, family health, community health.[2] Both of practice standard, quality and outcome are depend on validity and reliability of patient outcome to measure efficacy and effectivity of nursing intervention.[2]

Even though NOC considered as a complete and comprehensive outcome classification for nursing practice, this classification will continuously be developed.[1] In order to develop this classification, research related to validity and reliability need to be conducted.

There are several methods for measuring reliability which are internal consistency reliability, test-retest reliability, parallel forms reliability, intrarater reliability and interrater reliability.[3,4] As observation may be one of method to measure indicators of NOC, then consistency of rater becomes an important issue in reliability of NOC.[5] Recommended reliability measurement for consistency of raters is by using interrater reliability. Kappa coefficient together with percent agreement are suggested as a statistic test for measuring interrater reliability.[6-9] Morris et al also mentioned the benefit of percent agreement when it is used together with kappa.[6] The benefit is that the result of percent agreement will show whether there is any problem or not with kappa value.[6]

Although using two statistic method is recommended and is easy in calculation, several studies found there is sometimes a conflict in using kappa and percent agreement at the same time.[6,9-11] This conflict of value leads to confusion for researchers to determine whether kappa or percent agreement needs to be chosen for measuring reliability. There are two phenomena that could possibly occur regarding the result of kappa value and percent agreement. The first phenomenon is when the result of kappa can be accepted ($\kappa \geq 0{,}41$) but the percent agreement is unacceptable (<80%).[6,9] The second phenomenon is when the result of kappa is unacceptable ($\kappa < 0.41$) but the result of percent agreement is acceptable (>80%).[11,12]

METHODS

Literature search and results

Science direct data base was used to search for answer of the problem identified. Key words used were high agreement and low kappa and percent agreement and interrater reliability without limit time. The results of this search hit 260 articles. Researchers chose articles which were relevant to the problem.

DISCUSSION

Review of interrater reliability.

There are several authors who define reliability, for example van der Vleuten states that "reliability refers to the precision of measurement or the reproducibility of the scores obtained with the examination".[13] Interrater reliability is an agreement on the same data as a result of measurement from raters, by using scale classification on the same instrument or procedures.[14] Interrater reliability will be able to predict the number of errors in each procedure by using a rating or scoring.[14] Higher interrater reliability refers to stronger agreement between raters' results.[9] This method can be used to measure the accuracy of a skills' measurement instrument. This is supported by Rushforth who states that interrater reliability is the accuracy between two raters toward student's performance in specific skills when objective structured clinical examination (OSCE) is conducted.[15]

REVIEW OF PERCENT AGREEMENT AND KAPPA

Percent agreement

Percent agreement is one of the statistical tests to measure interrater reliability.[9] A researcher simply "calculates the number of times raters agree on a rating, then divides by the total number of ratings".[16]

Percent agreement formula is as follows,[17]

$$\text{Percent Agreement} = \frac{\text{agreement}}{\text{agreement+disagreement}} \text{ x } 100\% \qquad (1)$$

Acceptable percent agreement occurs only if the value is>80%.[9]

Kappa

Kappa statistic can also be used to measure interrater reliability, beside percent agreement.[9] Kappa was firstly introduced by Jacob Cohen in 1960 as a revision of percent agreement.[9] Formula of kappa created by Jacob Cohen is as follows:

$$\kappa = \frac{\text{po}-\text{pe}}{1-\text{pe}} \qquad (2)$$

The symbol κ was kappa coefficient. P0 represents actual observed agreement between raters, Pe represents chance agreement between raters.[9,18]

Table 1: Interrater reliability total item between raters.

	Rater 1			
		Pass	Not pass	
Rater 2	Pass	A	B	g1
	Not Pass	C	D	g2
		f1	f2	N

P0 and Pe are obtained from the results of scoring by two raters which is entered into 2x2 contingency table (see Table 1) as follows:

$$Po = \frac{A+D}{N} \quad (3) \text{ and } \quad \text{Pe} = \frac{\text{f1xg1}}{\text{N}} + \frac{\text{f2xg2}}{\text{N}} \quad (4)$$

Po formula in kappa equivalent with percent agreement.

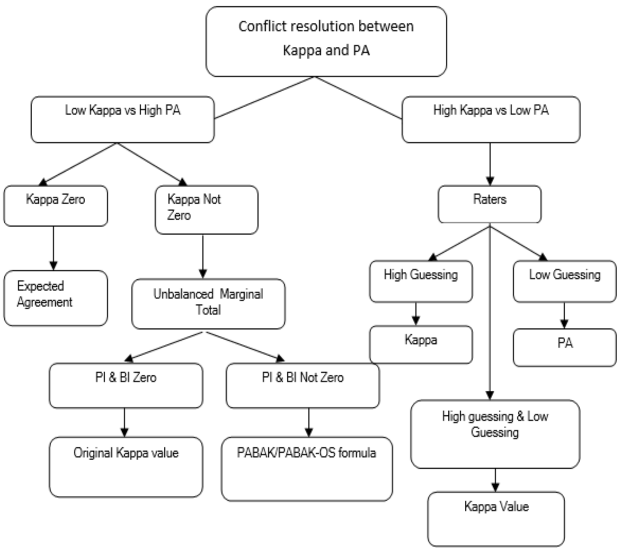
There are several different interpretations of kappa coefficient based on different authors, such as Landis and Koch, Fleiss, and Altman.[19-22] Those three interpretations can be seen below in Table 2.

Table 2: Interpretation of kappa coefficient.

Landis and Koch[20]	Fleiss[21]	Altman[22]
≤0: no agreement	k <0.40: poor agreement	<0.20: poor
0.01-0.20: none to slight	0.40<k<0.75: good	0.21-0.40: fair
0.21-0.40: fair	k>0.75: excellent agreement	0.41-0.60: moderate
0.41-0.60: moderate		0.61-0.80: good
0.61-0.80: substantial		0.81-1.00: very good
0.81-1.00: almost perfect agreement		

Algorithm for conflict resolution between kappa and percent agreement

The algorithm below (Figure 1) will guide researchers to solve conflicts between kappa and percent agreement.



PA=Percent Agreement, PI=Prevalence Index, BI=Bias Index, PABAK=Prevalence-Adjusted-Bias-Adjusted-Kappa, PABAK OS=Prevalence-Adjusted-Bias-Adjusted-Kappa Ordinal scale.

Figure 1. Conflict resolution between kappa and PA.

Conflict resolution when kappa value is acceptable and percent agreement is unacceptable

This conflict may occur if there is an influence of prevalence and bias.[23] This phenomenon however rarely occurs. When researchers find this phenomenon, they may be able to analyze it by examining the factors that influence interrater reliability.

Factors which influence interrater reliability include: subject to be observed, raters, atmosphere in measurement time and the instrument.[24] Rater is one aspect that is explored in detail in methodological research. Besar et al states that the different background of raters may influence the reliability of instruments.[25] McHugh states that when raters have high guessing characteristic in scoring, then kappa will be the best choice to determine reliability.[9] However, if raters are well trained and likely to have little guessing in scoring then percent agreement will be the best choice to determine reliability of the instruments.

Researchers may find a situation in which two raters have a different background. For example, one rater may have high guessing level in scoring and other raters have a low level of guessing in scoring. In this situation, probability that an agreement occurred only by chance (chance agreement) will be greater in this situation. Chance agreement could be corrected by kappa.[9] Based on this classical theory, all observable scores in measurement have two components[4] which are true score and error.[4] Generally, the reliability and quality of an instrument can be increased by decreasing the measurement error.[4] When percent agreement is unacceptable but kappa value is acceptable, what possibly happened is that in that "error" there is still a true score which percent agreement cannot detect, but kappa is able to detect this true score.

As shown above, kappa formula (Equation 1) has Pe as the most sensitive of kappa attributes when the raters change their scoring.[26] The changes in their scoring will affect kappa value as Pe represents chance agreement.[18] Kappa can correct chance agreement, but percent agreement is unable to correct chance agreement. Based on those considerations, kappa value is suggested to be a method to determine reliability in raters who have different backgrounds (i.e., one has high guessing and another has low guessing).

Table 3: Items with kappa value and PA.

Item	Kappa	PA
Wash hand	0.4441	76.60%
Greeting and call patient's name	0.4629	78.72%
Take clothes off (pants or skirt)	0.4445	67.02%

The example to determine reliability can be seen in research conducted by Siwi and Nurjannah.[27] They conducted research to measure interrater reliability of a checklist of an enema procedure for nursing training. This study involved 94 samples with two raters, conducted in OSCE in 2015.

In this study, raters have different backgrounds. The first one was a lecturer and the second one was a student. Considering that the first one has little guessing and the second one has high guessing, then kappa value is accepted as a reliability measurement by ignoring percent agreement (Table 3).

Conflict resolution when kappa value is unacceptable and percent agreement is acceptable

Phenomenon in which kappa value is unacceptable and percent agreement is acceptable is called paradox kappa.[11] This paradox kappa occurs in several cases.[11,28] There are three opinions in which kappa value interpretation is unacceptable. Landis and Koch state that kappa value <0,00 is considered unacceptable, while Altman mentions <0,20.[2,20,22,26,27] Meanwhile Feinstein and Chiccheti and Morris state that kappa value is unacceptable if kappa≤0.41.[6,23]

Although researchers may use those three categories, however they may find negative value on kappa, and this sometimes can lead to a confusion because there is not much information regarding negative value of kappa.[20] However, the explanation of this negative kappa value can be seen from an article written by McHugh who states that negative kappa value indicates strong disagreement between raters and considered as a sign of poor reliability.[9] The example of negative kappa is found in the above-mentioned study by Siwi and Nurjannah.[27] One item of checklist (bringing tools to client) has negative kappa (-0.2381).

Beside a negative kappa value, researchers may find zero kappa. This type of kappa value also can be found in Siwi and Nurjannah's research study.[27] It concluded that kappa value which is unacceptable is kappa with value ≤0,41 including zero and negative kappa.

Three items in the study of interrater reliability of enema procedure have an acceptable kappa but unacceptable percent agreement which are hand washing, greeting to the patient, and take off the patient's cloth.[24] When researchers meet this conflict, kappa value is more recommended than percent agreement for measuring interrater reliability. This is because the raters have different characteristic so the level of guessing are also different.[9]

What researchers need to do when they find paradox kappa with kappa value not zero?

Researchers may find paradox kappa in which kappa value is not zero or negative. In this situation, researchers are suggested to use kappa value following by attributes of kappa and ignoring percent agreement score.[11] Kappa value becomes a more important value because kappa is considered to have more information than percent agreement.[29] Those attributes of kappa are proportion of observed agreement, proportion of expected agreement, proportion of positive agreement, proportion of negative agreement, prevalence index and bias index.[11,23,28] Formula for kappa attributes are as follow:

$$\text{proportion of expected agreement (Pe)} = \frac{(a+c)(a+b)+(b+d)(c+d)}{N} \quad (5)$$

$$\text{proportion of positive agreement (Ppos)} = \frac{2a}{N+a-d} \quad (6)$$

$$\text{proportion of negative agreement (Pneg)} = \frac{2d}{N-a+d} \quad (7)$$

$$\text{Prevalence Index (PI)} = \frac{a-d}{N} \quad (8)$$

$$\text{Bias Index (BI)} = \frac{b-c}{N} \quad (9)$$

As mentioned before, paradox kappa is influenced by prevalence and bias.[11] However, only values of prevalence and bias which is not zero will influence paradox kappa to occur.[30,31] Besides prevalence and bias, paradox kappa also can be influenced by unbalanced marginal totals.[23]

Researchers may find paradox kappa in which PI and Bl is not zero, and there is an unbalanced marginal total. In this situation, low kappa value needs to be corrected using PABAK (prevalence-adjusted-bias-adjusted-kappa).[30-32] This PABAK's formula however can be used to correct kappa value with PI and BI zero for nominal data.[12] PABAK has the following formula:[11]

$$\text{PABAK} = 2Po - 1 \quad (10)$$

One example to solve this phenomenon can be found in the research study of Siwi and Nurjannah.[27] Siwi and Nurjannah found paradox kappa with kappa value of the checklist total is 0.3071 and PA 80.85%.[27] Researchers calculated prevalence and bias and it showed that PI is 0.69 and B 0.11. PI and BI calculation was conducted through manual calculation from contingency table 2x2 (Table 3) as follows:

$$\text{PI} = \frac{71-6}{94} = 0.69 \quad (11)$$

$$\text{BI} = \frac{14-3}{94} = 0.11 \quad (12)$$

This result of PI and BI showed that paradox kappa occurred because of prevalence and bias.[30,31] This calculation also showed that there was an unbalance on marginal totals as can be seen below in the Table 4.

Table 4. Total item with unacceptable kappa (not zero value) and acceptable PA.

	Rater 2					
		Pass	Not Pass	Total	Kappa	PA
Rater 1	Pass	71	14	85 (g^1)		
	Not Pass	3	6	9 (g^2)	0.3071	80.85%
	Total	74 (f^1)	20 (f^2)	94		

Note: Scoring between rater 1 and rater 2 in the total item of the checklist of enema procedure before using PABAK resulted unacceptable kappa value and acceptable PA. $f^1$and f^2 showed the total column of "pass" and "not pass" category by rater 2, meanwhile g^1 and g^2 showed total row of pass" and "not pass" category by rater 1.

Unbalanced marginal total is shown by f1 which has a big interval with g1. This situation also can be seen from the ratio of f2 and g2.[23] Regarding the result of this calculation, the researchers then decided to use PABAK to correct the kappa value as mentioned above.

Sim and Wright states by substituting the mean of cell A and cell D for the actual cell value shows the prevalence effect towards kappa value.[12] The bias effect is referred by substituting the average of cell B and cell C for the actual cell. PABAK calculation for the total item (nominal data) resulted in kappa coefficient is higher than previous value. Kappa value becomes 0.9904 (Table 5).

Table 5: Scoring between rater 1 and rater 2 in total item.

	Rater 2					
		Pass	Not pass	Total	Kappa	PA
Rater 1	Pass	38	8	46		
	Not pass	9	39	48	0.9904	81.91%
	Total	47	47	94		

Note: Scoring between rater 1 and rater 2 in the total item of the checklist of enema procedure after using PABAK resulted in acceptable kappa value and acceptable PA.

This formula is the same for data that is ordinal, however, in ordinal data, researchers are suggested to correct kappa by using PABAK-OS. Online calculators be found in specific web address.[33] The result of PABAK or PABAK-OS calculation will be the mean to determine reliability by still ignoring whatever was the percent agreement score. Even though researchers can use this formula, researchers however still need to show previous kappa value before it is corrected by using PABAK or PABAK-OS. This formula also can be used for kappa with negative value. On the other hand, if researchers find paradox kappa with kappa value that is not zero and after that researchers find that one of PI or/and BI has a zero value, then researchers cannot use PABAK or PABAK-OS to correct kappa value. This is because zero value in PI or/and BI means that prevalence and bias do not influence paradox kappa. In this case, reliability value should be determined by original kappa value.

What researchers need to do when they find paradox kappa with zero kappa?

Kappa coefficient zero only occurs when Po=Pe.[34] This result also shows that observed agreement is less than better expected agreement.[12] Observed agreement is an agreement that occurred only by chance.[11]

Phenomenon of zero kappa also is explained by Krippendorf who tried to solve zero kappa in measuring interrater (intercoder) reliability.[26] Krippendorf is using his own Alpha Krippendorf formula. Based on Krippendorf, when the two raters agree consistently with categories measured and suddenly another rater disagrees in one category measurement, then reliability cannot be measured.[26] This rule is also supported by Xie who found when two raters 100% agree only in one category, Cohen's Kappa calculation cannot be identified.[35]

The Alpha Krippendorf formula is actually similar with kappa formula used by Kvålseth and Xie:[34,35]

$$\kappa = 1 - \frac{1-Pa}{1-Pe} = 1 - \frac{Pdo}{Pde} \quad (13)$$

This shows how Alpha Krippendorf's formula[1] is equivalent with Cohen's Kappa's formula.[2] Based on those explanations, Alpha Krippendorf's formula can be used to explain zero kappa coefficient.

The variability of agreement of raters in measurement will influence the value of kappa coefficient. When measurement of raters toward one item does not vary (Table 3), then kappa coefficient is low.[26]

One example is from the results of interrater reliability of the checklist of the enema procedure in the item "order verification" that has zero kappa value and this value is changed to become 1 (perfect agreement) (Table 2).

Table 2 shows that if researchers try to change zero kappa value to be 1, then it will be given in bracket.[1] This trial shows that expected agreement (Pe in percent) becomes important and it can be a method to measure reliability.[26] A slight change in score can have result on kappa coefficient from zero to become 1 (perfect).

Table 6: Change in scoring by rater 2.

	Rater 1	Rater 2				Kappa	PA
		0	1	2	Total		
Order verification (Item 1)	0	0	0	0	0		
	1	0	0 (1)	1 (0)	1	0 (1)	0.9894 (0.9780)
	2	0	0	93	93		
	Total	0	0 (1)	94(93)	94		

Note: The bold and italic number shows changes in scoring by rater 2 which causes changes in expected agreement value.

Below is calculation of Po, Pe and kappa made manually toward item 1 after there is a change of score by second rater.

$$\text{Po} = \frac{1+93}{94} = 1 \qquad (14)$$

$$\text{Pe} = \left(\frac{0}{94}\right)\left(\frac{0}{94}\right) + \left(\frac{1}{94}\right)\left(\frac{1}{94}\right) + \left(\frac{93}{94}\right)\left(\frac{93}{94}\right) = 0 + \frac{1}{8836} + \frac{8649}{8836} = 0.9780 \qquad (15)$$

$$\kappa = \frac{1-0{,}9780}{1-0{,}9780} = 1 \qquad (16)$$

In conclusion, for researchers who find paradox kappa with zero kappa, then it is suggested to use expected agreement as the best method to measure reliability by ignoring percent agreement and kappa value. This approach is more accurate because the score of expected agreement becomes the best standard when paradox kappa occurs accompanied by zero kappa.[26]

CONCLUSION

The backgrounds of raters need to be considered when kappa value is acceptable but percent agreement is unacceptable in interrater reliability measurement. In the situation when there is paradox kappa, then the kappa value needs to be considered for interrater reliability measurement. Additionally, other formulas have been recently developed to provide conflict resolution in contradictory results in interrater reliability measurements.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. Johnson M. Overview of the nursing outcomes classification (NOC). 2013. Available from: https://www.ojni.org/2_2/johnart.html (cited 2016 December 19th).
2. Moorhead S, Johnson M, Maas M, Swanson E. Nursing outcome classification (NOC). 5th edition. St Louis, Missouri: Elsevier Saunders; 2013.
3. Phelan C, Wren J. Exploring reliability in academic assessment. 2006. Available from: https://www.uni.edu/chfasoa/reliabilityandvalidity.htm. (cited 2015 May 5th).
4. Scholtes V, Terwee C, Poolman R. What makes a measurement instrument valid and reliable? Injury. 2011;42(3):236-40.
5. Kimberlin C, Winterstein AG. Validity and reliability of measurement instruments used in research. Am J Health-Syst Pharm. 2008;65(23):2276-84.
6. Morris R, MacNeela P, Scott A, Treacy P, Hyde A, O'Brien J, et al. Ambiguities and conflicting results: The limitations of the kappa statistic in establishing the interrater reliability of the Irish nursing minimum data set for mental health: A discussion paper. Int J Nurs Stud. 2008;45(4):645-7.
7. Craddock J. Interrater reliability of psychomotor skill assessment in athletic training: ProQuest; 2009.
8. Cargo M, Stankov I, Thomas J, Saini M, Rogers P, Mayo-Wilson E, et al. Development, interrater reliability and feasibility of a checklist to assess implementation (Ch-IMP) in systematic reviews: the case of provider-based prevention and treatment programs targeting children and youth. BMC Med Res Methodol. 2015;15(1):1.
9. McHugh M. Interrater reliability: the kappa statistic. Biochemia Medica. 2012;22(3):276-82.
10. O'Leary S, Lund M, Ytre-Hauge TJ, Holm SR, Naess K, Dailand LN, et al. Pitfalls in the use kappa when interpreting agreement between multiple raters in reliability studies. Physiotherapy. 2014;100(1):27-35.
11. Cunningham M, editor. More than just the kappa coefficient: a program to fully characterize interrater reliability between two raters. SAS global forum; 2009.
12. Sim J, Wright C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physic therap. 2005;85(3):257-68.
13. van der Vleuten C. Validity of final examinations in undergraduate medical training. Br Med J. 2000;321(7270):1217.
14. Kottner J, Audigé L, Brorson S, Donner A, Gajewski B, Hróbjartsson A, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. Int J Nurs Stud. 2011;48(6):661-71.
15. Rushforth H. Objective structured clinical examination (OSCE): review of literature and implications for nursing education. Nurse Education Today. 2007;27(5):481-90.
16. Graham M, Milanowski A, Miller J. Measuring and promoting interrater agreement of teacher and principal performance ratings. Online Submission. Center for Educator Compensation Reform. 2012.
17. House A, House B, Campbell M. Measures of interobserver agreement: Calculation formulas and distribution effects. J Behav Assess. 1981;3(1):37-57.
18. Viera A, Garrett J. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005;37(5):360-3.
19. McCray G, ed. Assessing interrater agreement for nominal judgement variables. Language Testing Forum; 2013.
20. Landis J, Koch G. The measurement of observer agreement for categorical data. biometrics. 1977:159-74.
21. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971;76(5):378-82.
22. Altman DG. Practical statistics for medical research. 1 ed. London; New York: Chapman and Hall. 1991.
23. Feinstein A, Cicchetti D. High agreement but low kappa: I. The problems of two paradoxes. J clin epidem. 1990;43(6):543-9.
24. Weiner J. Measurement: reliability and validity measures. Bloomberg School of Public Health, Johns Hopkins University, mimeo (Power Point Presentation) at http://ocw jhsph edu/ courses/ hsre/ PDFs/ HSRE_lect7_weiner pdf http://jae oxfordjournals org. 2007.

25. Besar M, Siraj H, Manap R, Mahdy Z, Yaman M, Kamarudin M, et al. Should a single clinician examiner be used in objective structure clinical examination? Procedia-Social and Behavioral Sciences. 2012;60:443-9.
26. Krippendorff K. Agreement and information in the reliability of coding. Communication Methods and Measures. 2011;5(2):93-112.
27. Siwi S, Nurjannah I. Interrater Reliability pada Checklist Penilaian Pemberian Huknah di Program Studi Ilmu Keperawatan Fakultas Kedokteran Universitas Gadjah Mada. [Unpublished Thesis]. In press 2016.
28. Cicchetti D, Feinstein A. High agreement but low kappa: II. Resolving the paradoxes. J clinical epidem. 1990;43(6):551-8.
29. Joyce M, editor. Picking the best intercoder reliability statistic for your digital activism content analysis. Digital Activism Research Project: Investigating the Global Impact of Comment Forum Speech as a Mirror of Mainstream Discourse. 2013.
30. Xier L. Kappa—A Critical Review. 2010; Available from: http://www.diva-portal.org/ smash/ get/ diva2:326034/FULLTEXT01.pdf.
31. Flight L, Julious S. The disagreeable behaviour of the kappa statistic. Pharmaceutical statistics. 2015;14(1):74-8.
32. Byrt T, Bishop J, Carlin J. Bias, prevalence and kappa. Journal of clinical epidemiology. 1993;46(5):423-9.
33. Streiner DL, Geoffrey RN, Cairney J. Health Measurement Scales: A Practical Guide to Their Development and use. United Kongdom: Oxford University Press. 2015.
34. Kvålseth T. Measurement of Interobserver Disagreement: Correction of Cohen’s Kappa for Negative Values. J Probab Statist. 2015;2015.
35. Xie Q. Agree or Disagree? A Demonstration of An Alternative Statistics Cohen’s Kappa for Measuring the Extent and Reliability of Agreement between Observers. 2013 [cited 2016 March 12]; Available from: https://fcsm.sites.usa.gov/ files/ 2014/05/ J4_Xie_2013FCSM.pdf.

Cite this article as: Nurjannah I, Siwi SM. Guidelines for analysis on measuring interrater reliability of nursing outcome classification. Int J Res Med Sci 2017;5:1169-75.