Rev. Latino-Am. Enfermagem
2014 Mar.-Apr.;22(2):234-40
DOI: 10.1590/0104-1169.3262.2407
www.eerp.usp.br/rlae

Original Article

Reliability of indicators of nursing care quality: testing interexaminer agreement and reliability[1]

Dagmar Willamowius Vituri[2]
Yolanda Dora Martinez Évora[3]

Objective: this study sought to test the interexaminer agreement and reliability of 15 indicators of nursing care quality. Methods: this was a quantitative, methodological, experimental, and applied study conducted at a large, tertiary, public teaching hospital in the state of Paraná. For data analysis, the Kappa (k) statistic was applied to the categorical variables – indicators 1 to 11 and 15 – and the interclass correlation coefficient (ICC) to the continuous variables – indicators 12, 13, and 14, with the corresponding 95% confidence intervals. The categorical data were analyzed using the Lee software, elaborated by the Laboratory of Epidemiology and Statistics of Dante Pazzanese Institute of Cardiology – Brazil, and the continuous data were assessed using BioEstat 5.0. Results: the k-statistic results indicated excellent agreement, which was statistically significant, and the values of the ICC denoted excellent and statistically significant reproducibility/agreement relative to the investigated indicators. Conclusion: the investigated indicators exhibited excellent reliability and reproducibility, thus showing that it is possible to formulate valid and reliable assessment instruments for the management of nursing care.

Descriptors: Nursing; Nursing Audit; Quality Indicators, Health Care; Validation Studies as Topic; Reproducibility of Results.

[1] Paper extracted from doctoral dissertation “Evaluation as the principle of Quality Management: testing the interrater reliability of indicators for assessing the quality of nursing care” presented to Escola de Enfermagem de Ribeirão Preto, Universidade de São Paulo, WHO Collaborating Centre for Nursing Research Development, Ribeirão Preto, SP, Brazil.

[2] Doctoral student, Escola de Enfermagem de Ribeirão Preto, Universidade de São Paulo, WHO Collaborating Centre for Nursing Research Development, Ribeirão Preto, SP, Brazil. RN, Hospital Universitário, Universidade Estadual de Londrina, Londrina, PR, Brazil.

[3] PhD, Full Professor, Escola de Enfermagem de Ribeirão Preto, Universidade de São Paulo, WHO Collaborating Centre for Nursing Research Development, Ribeirão Preto, SP, Brazil.

Corresponding Author:
Dagmar Willamowius Vituri
Universidade Estadual de Londrina. Hospital Universitário
Gerência de Riscos
Av. Robert Koch, 60
Vila Operária
CEP: 86038-350, Londrina, PR, Brasil
E-mail: dagmar@uel.br

Introduction

The quality of products and services is a concern for various types of organizations, particularly for those in the healthcare sector, because in addition to the influence on the economy of the services they provide, their clients are increasingly demanding high quality in the satisfaction of their needs[1].

Nurses are important contributors to healthcare organizations' quest for quality, and as a consequence, managers have made explicit their expectations concerning the role of nurses in the management of care in the hospital setting. These high expectations are warranted because direct contact with clients is one of the specific characteristics of nursing care, which allows for the identification of the clients' needs and expectations[1], in addition to the crucial role the nursing profession plays in general care assistance.

Management demands accurate knowledge of an institution's performance relative to its mission and goals[2], which in turn requires the implementation of appropriate systems of assessment and indicators allowing for the (re)formulation of guidelines[3]. However, the quantitative assessment of the quality of health care provided may pose challenges related to selecting the appropriate statistical measures[4].

Indicators are units of measurement of activities, and they can be applied to measure qualitative and quantitative features of healthcare organizations, including their structure, processes, and results[5-6]. Although countless indicators have been formulated, it is difficult to select an appropriate measure that bears high validity relative to the domain of interest[7], i.e., to the healthcare feature one intends to measure.

The use of valid and reliable measures permits monitoring the quality of the care provided to patients, identifying avoidable risks, and grounding the planning of corrective actions, in addition to orienting strategies and readjusting goals by means of educational actions and professional valorization. Moreover, the use of valid and reliable instruments can contribute to the advancement of professional knowledge, as well as the theory that underpins its practice[8].

Reliability and agreement are relevant features to be taken into account in the formulation of measuring instruments, as they provide information on the amount of error in the measures and thus their quality[9]. Agreement is the degree to which scores or ratings are identical[9], while reliability is the degree to which a measure reflects the true result, i.e., the degree to which a measure is free from random error variance[10]. Reliability may also be defined as the proportion of variance in measurement scores that is due to differences in the true score, rather than to random error[11-12]. Therefore, reliability assesses the consistency and stability of measures and increases together with a reduction in error[11].

Based on historical concerns of the nursing profession associated with the quality of assistance provided, which began with the work of Florence Nightingale, the crisis of credibility in current Brazilian healthcare services, and the potential for nurses to change the current situation through the measurement of the quality of care effectively afforded, the need for reliable measures of the quality of assistance is patent.

The elaboration of assessment instruments is complex and can be addressed by several disciplines. Within the context of the nursing practice, criticism has been raised against the use of such instruments inasmuch as they are intended to measure abstract and subjective constructs and notions. However, it is believed that such criticism is due to lack of knowledge of the process of conception and validation of assessment measures, to the point of discouraging the formulation of such measures and thus hindering scientific advances in this field.

Based on the situation described above, and with the goal of continuing a study that began with the analysis of the content validity of indicators of nursing care quality[13], the following research question was formulated: does the instrument comprising indicators of the quality of basic nursing care meet the standards of reliability required to assess the quality of the assistance provided to adult patients admitted to a medical-surgical unit at a public teaching hospital in northern Paraná? In particular, the aim of the present study was to test the interexaminer agreement and reliability of 15 indicators of nursing care quality.

Methods

This was a quantitative, methodological, experimental, and applied study that was conducted in 3 stages (face validity test, pilot test, and reliability test) at a large, tertiary, public teaching hospital in northern Paraná, Brazil. In particular, the present article reports and discusses the results of the reliability test to which the investigated measuring instrument was subjected.

The research project began by assessing the face validity of an instrument comprising 15 indicators of quality, with a minimum concordance percentage[14] of

80%. This instrument was based on the one previously elaborated by Vituri(13), which achieved satisfactory results in a content validity test performed by experts and was later reformulated as a function of the needs detected during its routine application in everyday practice.

To test the instrument's face validity, an intentional sample of judges was selected among trainees at the Nursing Care Quality Assessment Service of the investigated hospital. This service operates as a non-mandatory internship for students in the third and fourth terms of the institutional undergraduate nursing course and performs retrospective operational audits of the nursing care quality by means of the systematic application of Vituri's instrument(13).

The pilot test of the measuring instrument was performed by an intentional sample comprising 3 judges, including 2 nurses from the institution (one providing patient care, and the other performing managerial tasks at the institution's board of directors) and the investigator. The number of judges was selected based on studies by Crocker, Llabre, and Miller(15), which indicated that greater numbers of judges are associated with increased heterogeneity among the panel, which results in a lower degree of reliability and agreement(16).

The indicators of comprehension, assessment methods, and instrument applicability were tested on a random sample comprising 15 adult patients admitted to a medical-surgical unit.

For the assessment of reliability, among the 3 categories of methods for the estimation of agreement (consensus estimates, consistency estimates, and assessment/measurement estimates), the consensus estimates were used. These estimates were grounded on the assumption that examiners should be able to come to exact agreement on how to use a rating scale to score observable behaviors, thus sharing a common interpretation of the construct(17).

For this purpose, the equivalence method was used, whereby the interexaminer reliability is assessed through the application of the same instrument by various examiners for the measurement of the same phenomena (interexaminer reliability or examiner's precision)(11). This type of test is mainly indicated for clinical instruments that are strongly dependent on the examiner's judgment. Under such circumstances, there is evident variance among examiners, and thus, the estimation of the precision by means of the correlation of each examiner's individual results has paramount importance(11).

Indicator agreement and reliability were tested by the same judges who had participated in the pilot test because they were already trained in the use of the instrument, which was applied concomitantly and independently to a sample comprising 33 patients, including the 15 patients who had previously participated in the pilot test.

In the assessment of indicators 1 to 11 and 15, the judges read the descriptors defining the standards of quality and indicated in the assessment instrument whether the corresponding item was adequate or not according to the standard. In regard to indicators 12, 13, and 14, which were also based on the description of the standards of quality, the judges indicated the number of checking tasks and records of vital signs that were performed in an adequate or inadequate manner.

For data analysis, the Kappa statistic (*k*) was used for the categorical variables – indicators 1 to 11 and 15 – and the interclass correlation coefficient (ICC) was used for the continuous variables – indicators 12, 13, and 14, with the corresponding 95% confidence interval (CI)(12,18). The *k* coefficient is a measure of association, i.e., it measures the degree of consensus between examiners beyond the amount of agreement that might be expected by chance alone, while attributing equal weight to disagreement(17). Fleiss' *k* statistic is an extension of the k coefficient and is used to evaluate concordance or agreement between multiple raters, while no weighting is applied(19). The ICC, also known as the reproducibility coefficient (R), estimates the fraction of the total variability in measures that is due to variations among individuals(20-21). Although it does not provide detailed information on the structure of agreement or disagreement(22), the *k* statistic is more effective compared to simple percentage agreement measures(17). According to some authors, the *k* statistic and ICC are the most adequate methods to estimate interexaminer reliability(9). In addition to the *k* statistic and ICC, the simple percentage agreement measures were also calculated to obtain a more detailed picture of the interexaminer reliability and agreement(9).

The values of *k* were interpreted as follows: less than zero (0), *no agreement*; 0 to 0.19, *poor agreement*; 0.20 to 0.39, *fair agreement*; 0.40 to 0.59, *moderate agreement*; 0.60 to 0.79, *substantial agreement*; and 0.80 to 1.00, *almost perfect agreement*(23). The ICC results were classified as follows: 0.4 to 0.59, *reasonable reproducibility*; 0.6 to 0.74, *good reproducibility*; and above 0.74, *excellent reproducibility*(24).

Analysis of the categorical data (*k* statistic) was performed using the *Lee* software, provided by the Laboratory of Epidemiology and Statistics of Dante Pazzanese Institute of Cardiology – Brazil, and the continuous data (ICC) were analyzed using *BioEstat 5.0.* The significance level was established as a p-value <0.05.

This study complied with all applicable ethical principles and was approved by the institutional board of directors as well as the university's ethics committee for research involving human beings, ruling no. 126/10, CAAEE no. 0113.0.268.000-10. All of the participants, including judges and patients, signed an informed consent form.

Results

The results of the face validity assessment showed that all the indicators were apparently valid. In regard to the pilot test, the judges ruled the instrument comprehension and applicability to be adequate. As no doubts were raised and no suggestions were made, the instrument was considered appropriate to be subjected to the interexaminer reliability test.

Table 1 describes the results of the k-statistic analysis of indicators 1 to 11 and 15, in addition to the simple percentage agreement measure relative to this set of indicators.

Table 1 - Assessment of the interexaminer agreement relative to indicators 1 to 11 and 15 by means of the k statistic, Londrina, PR, Brazil, 2013

Indicator*	N†	% agreement	Fleiss k	95% CI	p-value	Agreement
01	33	100	1.000	0.803-1.000	<0.001	Excellent
02	25	100	1.000	0.835-1.000	<0.001	Excellent
03	22	100	1.000	0.852-1.000	<0.001	Excellent
04	25	97	0.970	0.830-1.000	<0.001	Excellent
05a	10	97	0.956	0.795-1.000	<0.001	Excellent
05b	23	94	0.969	0.829-1.000	<0.001	Excellent
05c	04	100	1.000	0.847-1.000	<0.001	Excellent
06a	10	100	1.000	0.845-1.000	<0.001	Excellent
06b	23	97	0.969	0.829-1.000	<0.001	Excellent
06c	02	100	1.000	0.846-1.000	<0.001	Excellent
07	09	100	1.000	0.837-1.000	<0.001	Excellent
08	03	100	1.000	0.803-1.000	<0.001	Excellent
09	05	100	1.000	0.846-1.000	<0.001	Excellent
10	05	100	1.000	0.803-1.000	<0.001	Excellent
11	05	100	1.000	0.846-1.000	<0.001	Excellent
15	27	100	1.000	0.855-1.000	<0.001	Excellent

*Indicators 1 - 11; 15

1. Identification of the patient's bed.
2. Identification of the bed fall risk.
3. Identification of peripheral venous lines.
4. Verification of extravasation skin injuries.
5. a – Identification of intravenous infusion equipment (maintenance fluids).
b – Identification of intravenous infusion equipment (fluids to dilute medications).
c – Identification of intravenous infusion equipment (fluids – drug 1).
6. a – Identification of intravenous infusion flasks – label (maintenance fluids).
b – Identification of intravenous infusion flasks – label (fluids to dilute medications).
c – Identification of intravenous infusion flasks – label (fluids – drug 1).
7. Identification of infusion speed control – graduated scale.
8. Identification of gastric – oro- and nasogastric – tubes for drainage.
9. Indwelling urinary catheter fixation.
10. Position of urine drainage bag, indwelling urinary catheter.
11. Position of the urine drainage bag spigot.
15. Elaboration of complete daily prescription by nurse.

†Number of patients for whom indicator assessment applies.

According to the results described in Table 1, the simple percentage agreement was over 80%, which was therefore deemed adequate for all of the assessed indicators[14].

The results of the *k* statistic varied from 0.956 to 1.000, thus showing that the degree of agreement exhibited by indicators 1 to 11 and 15 was excellent and statistically significant (p-value <0.001)[23].

The following indicators did not achieve full agreement according to Fleiss' *k* statistic: indicator 4, verification of extravasation skin injuries (*k*= 0.970, 95% CI: 0.830-1.000); 5a, identification of intravenous infusion equipment (maintenance fluids) (*k*=0.956, 95% CI: 0.795-1.000); 5b, identification of intravenous infusion equipment (fluids to dilute medications) (*k*=0.969, 95% CI: 0.829 -1.000); and 6b, identification of intravenous infusion flasks (*k*=0.969, 95% CI 0.829-1.000).

The results for the ICC relative to indicators 12, 13, and 14 as well as those concerning the simple percentage agreement of this set of indicators are described in Table 2.

Table 2 - Assessment of the interexaminer agreement relative to indicators 12, 13, and 14 by means of the ICC, Londrina, PR, Brazil, 2013

Indicator*	N	% agreement	ICC	95% CI	p-value	Reproducibility†
12 - a	559	99.7	0.992	0.983-0.996	<0.001	Excellent
12 - b	81	98.1	0.980	0.959-0.990	<0.001	Excellent
13 - a	56	98.2	0.957	0.914-0.979	<0.001	Excellent
13 - b	44	97.8	0.951	0.903-0.976	<0.001	Excellent
14 - a	354	99.6	0.969	0.938-0.985	<0.001	Excellent
14 - b	64	97.5	0.859	0.732-0.929	<0.001	Excellent

*Indicators 12 – 14
12. a – Checking of nursing prescription procedures (adequate)
b – Checking of nursing prescription procedures (inadequate)
13. a – Record of verification of prescribed vital signs (adequate)
b – Record of verification of prescribed vital signs (inadequate).
14. a – Checking of nursing procedures in medical prescriptions (adequate)
b – Checking of nursing procedures in medical prescriptions (inadequate).
†Synonym of reliability

In regard to indicators 12, 13, and 14 (Table 2), the simple percentage agreement varied from 89.4% to 92.5%, thus denoting adequate agreement(14). As the ICC values varied from 0.859 to 0.992, the reproducibility/agreement of those indicators was excellent and statistically significant (p-value <0.001)(24).

The lowest ICC value was 0.859 (95% CI: 0.732 -0.929), corresponding to indicator 14b (checking of nursing procedures in medical prescriptions) (adequate), and the highest value was 0.992 (95% CI: 0.983-0.996), which corresponded to indicator 12a (checking of nursing prescription procedures) (adequate). These values indicated that the reproducibility/agreement of indicators 12, 13, and 14 was excellent.

Discussion

The values of interexaminer reliability assessed with the k statistic revealed excellent agreement among the judges relative to the construct, descriptors, and assessment of the investigated indicators.

It should be noted that it was not the validity of the results that was measured but rather the degree of error in the measures, which is due to differences in the true score(8,10-11). On these grounds, because indicators 4, 5a, 5b, 6b, 12a, 12b, 13a, 13b, 14a, and 14b did not achieve Fleiss' k or an ICC of 1.000, one may infer that they did not exhibit a precision of 100%. However, it is worth stressing that no matter how accurate a measuring instrument might be, the resulting scores will never be fully free from error(11).

The types of error to which measures are liable are classified in 2 groups: random errors, which are due to factors that affect the measurement of a variable in the full sample in an accidental manner, and although they increase the variability of the data, do not affect the average performance of the sample; and systematic errors, which are due to any factor that systematically affects the measurement of a variable in the full sample and thus tend to exert a consistently positive or negative effect, for which reason they are sometimes considered as measurement bias(11).

It was proposed that certain factors may have interfered with the measures' precision and may have contributed to cause measurement error, including transitory and personal(25) factors such as haste and fatigue; for example, one of the judges performed the assessment immediately after a full regular workday. In addition, 2 judges mentioned the poor legibility of the reports for checking medical and nursing prescriptions, which raised doubts as to the full compliance with the established standard of quality.

The *k* and ICC values of the indicators represent the degree to which the results obtained by the application of the measuring instrument reflect the true result(10). Although the indicators evaluated did not achieve a Fleiss' *k* value of 1.000, the degree of agreement exhibited by indicators 4, 5a, and 5b was excellent, i.e., *almost perfect* (*k* of 0.80 to 1.000)[23]. Additionally, the agreement exhibited by indicators 12a, 12b, 13a, 13b, 14a, and 14b was excellent, as the ICC values were above 0.74[24].

The results reported here indicate adequate interexaminer reliability relative to the investigated indicators and measuring instrument, which indicates their precision and potential for use in the assessment of nursing care quality. It is also worth noting that reliability and agreement are not inherently fixed characteristics of the measuring instruments but rather are a product of the interaction among instruments/tools, subjects/ objects, and the context of the assessment[9].

For a given characteristic to be liable to contextual influence, controlling the variables that interfere with the process of measurement has paramount importance. To this end, examiners require training related to the construct, descriptors, ideal conformity index, assessment criteria, and standard procedure for assessment.

Conclusion

Based on the study results, the 15 investigated indicators of nursing care quality previously validated by means of the content validity strategy exhibited excellent reliability and reproducibility. This agreement and reliability of the indicators, as tested by the k statistic and the ICC, highlight the relevance of this instrument for the assessment of nursing care quality in clinical practice.

This confirmation of indicator reliability increases the available evidence showing that it is possible to elaborate valid and reliable instruments to assess nursing care quality. Such instruments are indispensable for effective and efficacious management of nursing care, as they allow for the identification of avoidable risks, ground the planning of corrective actions, and orient strategies for goal readjustment.

Based on the reported reliability of these 15 investigated indicators, their use in other healthcare institutions may considerably improve the management of nursing care and, consequently, also the quality of the assistance provided and the safety of patients.

The methods used for the elaboration and validation of assessment systems are widely discussed and employed in social and behavioral sciences, and their application in the present study shows that their underlying principles could potentially be adapted to the elaboration of assessment instruments for nursing practices.

The potential limitations of the present study are related to the intentional sampling technique and the sample size of judges. The intentional sampling technique was selected because it was not feasible to randomly select nurses during their normal work shifts. The established minimum of 3 judges was based on the difficulty in assessing, in a concomitant and independent manner, all of the patients admitted to hospital wards comprising 3 to 6 beds. Such an alternative may have been disruptive and could have caused discomfort and embarrassment, in addition to favoring the occurrence of measurement error.

In regard to the investigated indicators, although they most likely do not cover all features pertinent to nursing care, they encompass care measures relevant to the prevention of risks and are highly sensitive to improvement through the application of simple actions, such as educational strategies. Unfortunately, the limited selection of these particular 15 indicators implied the exclusion of other relevant features of assistance, which might thus represent one further limitation of the present study.

References

1. Rocha ESB, Trevizan MA. Quality management at a hospital's nursing service. Rev. Latino-Am. Enfermagem. 2009;17(2):240-5.

2. Kuwabara CCT, Évora YDM, Oliveira MMB. Risk Management in Technovigilance: construction and Validation of a Medical-Hospital Product Evaluation Instrument. Rev. Latino-Am. Enfermagem. 2010;18(5):943-51.

3. D'Innocenzo MANP, Cunha ICKO. O movimento pela qualidade nos serviços de saúde e enfermagem. Rev Bras Enferm. [Internet]. 2006. [acesso 23 abr 2013];59(1):84-8. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-71672006000100016&lng=en&nrm=iso&tlng=pt

4. Collier R. The challenges of quantifying quality. CMAJ. [Internet]. 2010. [acesso 23 abr 2013];182(5):E250. Disponível em: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2842849/pdf/182e250.pdf

5. Takahashi AA, Barros ALBL, Michel JLMS, Mariana F. Difficulties and facilities pointed out by nurses of a university hospital when applying the nursing process. Acta Paul Enferm. [Internet]. 2008. [acesso 23 abr 2013];21(1):32-8. Disponível em: http://www.scielo.br/pdf/ape/v21n1/04.pdf

6. Vieira APM, Kurcgant, P. Quality indicators of the management of human resources in nursing: point of

view of registered nurses. Acta Paul Enferm. [Internet]. 2010. [acesso 23 abr 2013];23(1):11-5. Disponível em: http://www.scielo.br/pdf/ape/v23n1/02.pdf

7. Nakrem S, Vinsnes AG, Harkless GE, Paulsen B, Seim A. Nursing sensitive quality indicators for nursing home care: international review of literature, policy and practice International. J Nurs Stud. 2009; 46:848-57.

8. Gillespie BM, Polit DF, Hamlin L, Chaboyer W. Developing a model of competence in the operating theatre: Psychometric validation of the Perceived Perioperative Competence Scale-Revised. Int J Nurs Stud. [Internet]. 2012. [acesso 23 abr 2013];49(1):90-101. Disponível em: http://www.journalofnursingstudies.com/article/S0020-7489(11)00301-4/fulltext

9. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. [Internet]. 2011. [acesso 23 abr 2013];64(1):96-106. Disponível em: http://www.rygforskning.dk/sites/default/files/files/articles/Kottner%20el%20al%20%202011.pdf

10. Hora HRM, Monteiro GTR, Arica J. Confiabilidade em questionários para qualidade: um estudo com o Coeficiente Alfa de Cronbach. Prod Prod. [Internet]. 2010. [acesso 23 abr 2013];11(2):85-103. Disponível em: http://seer.ufrgs.br/ProdutoProducao/article/view/9321/8252

11. Trochim WMK. Research methods: knowledge bases. [Internet]. 2nd. ed. Cincinatti (OH): Atomic Dog Publishing; 2006. [acesso 23 abr 2013]. Disponível em: http://socialresearchmethods.net/kb/

12. Wuensch KL. The Intraclass Correlation Coefficient. Karl Wuensch's Statistics Lessons [Internet]. Greenville (USA): East Carolina University – Department of Psychology; [2010]; [atualizada em 20 abr 2013; acesso 23 abr 2013]. Disponível em: http://core.ecu.edu/psyc/wuenschk/StatsLessons.htm

13. Vituri DW, Matsuda LM. Content validation of quality indicators for nursing care evaluation. Rev Esc Enferm USP. [Internet]. 2009. [acesso 23 abr 2013];43(2):429-37. Disponível em: http://www.scielo.br/pdf/reeusp/v43n2/en_a24v43n2.pdf

14. Westmoreland D, Wesorick B, Hanson D, Wyngarden K. Consensual validation of clinical practice model practice guidelines. J Nurs Care Qual. 2000;14(4):16-27.

15. Crocker L, Llabre M, Miller MD. The generalizability of content validity ratings. J Educ Measure. 1988;25(4):287-99.

16. Lilford RJ, Mohammed MA, Braunhoultz D, Hofer TP. The measurement of active errors: methodological issues, Qual Saf Health Care. [Internet]. 2003. [acesso 30 abr 2012];12 Suppl 2:8-12. Disponível em: http://qualitysafety.bmj.com/content/12/suppl_2/ii8.full.pdf+html

17. Stemler SE. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Prac Assess Res Eval. [Internet]. 2004. [acesso 15 abr 2012];9(4). Disponível em: http://pareonline.net/getvn.asp?v=9&n=4

18. Haley SM, Osberg JS. Kappa Coefficient Calculation Using Multiple Ratings Per Subject: A Special Communication. Phys Ther. [Internet]. 1989. [acesso 15 abr 2012];69:970-4. Disponível em: http://ptjournal.apta.org/content/69/11/970

19. Chang A. Cohen's and Fleiss's Kappa program Concordance in ordinal data. In: Chang A. StatTools Home Page [Internet]. Queensland: Austrália; 2011. [acesso 20 abr 2012]; Disponível em: http://www.stattools.net/CohenKappa_Pgm.php#Fleiss%27s%20Kappa%20from%20rating%20scores

20. Chang A. Intraclass correlation for parametric data Introduction and explanation. In: Chang A. StatTools Home Page [Internet]. Queensland: Austrália; 2011 [acesso 15 abr 2012]. Disponível em: http://www.stattools.net/ICC_Exp.php

21. Wilson-Genderson M, Broder HL, Phillips C. Concordance between caregiver and child reports of children´s oral health-related quality of life. Community Dent Oral Epidemiol. 2007;35 Suppl 1:32-40.

22. Perroca MG, Gaidzinki RR. Assessing the interrater reliability of an instrument for classifying patients - kappa quotient. Rev Esc Enferm USP. [Internet]. 2003. [acesso 23 abr 2013];37(1):72-80. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0080-62342003000100009

23. Zegers M, Bruijne MC, Wagner C, Groenewegen PP, Wal GVD, Vet HCW. The inter-rater agreement of retrospective assessments of adverse events does not improve with two reviewers per patient record. J Clin Epidemiol. 2010;63(1):94-112.

24. Fleiss JL. Reliability of measurement. In: Fleiss JL. The Design and Analysis of Clinical Experiments. New York (USA): John Wiley; 1999. p. 1-31.

25. Polit DF, Beck CT. Assessing data quality. In: Polit DF, Beck CT. Nursing research: principles and methods. 7th ed. Philadelphia (USA): Lippincott Williams & Wilkins; 2004. p. 413-48.

Received: Apr. 26th 2013
Accepted: Sept. 3rd 2013