Available online at www.sciencedirect.com
ScienceDirect
Research in Social and Administrative Pharmacy 9 (2013) 330–338
RESEARCH IN SOCIAL & ADMINISTRATIVE PHARMACY

Commentary

Interrater agreement and interrater reliability: Key concepts, approaches, and applications

Natasa Gisev, B.Pharm. (Hons.)[a,*], J. Simon Bell, Ph.D.[a,b,c], Timothy F. Chen, Ph.D.[a]

[a]*Faculty of Pharmacy, Pharmacy and Bank Building (A15), The University of Sydney, New South Wales 2006, Australia*
[b]*Quality Use of Medicines and Pharmacy Research Centre, Sansom Institute, School of Pharmacy and Medical Sciences, University of South Australia, Adelaide, South Australia 5001, Australia*
[c]*School of Pharmacy, Faculty of Health Sciences, University of Eastern Finland, 7011 Kuopio, Finland*

Summary

Evaluations of interrater agreement and interrater reliability can be applied to a number of different contexts and are frequently encountered in social and administrative pharmacy research. The objectives of this study were to highlight key differences between interrater agreement and interrater reliability; describe the key concepts and approaches to evaluating interrater agreement and interrater reliability; and provide examples of their applications to research in the field of social and administrative pharmacy. This is a descriptive review of interrater agreement and interrater reliability indices. It outlines the practical applications and interpretation of these indices in social and administrative pharmacy research. Interrater agreement indices assess the extent to which the responses of 2 or more independent raters are concordant. Interrater reliability indices assess the extent to which raters consistently distinguish between different responses. A number of indices exist, and some common examples include Kappa, the Kendall coefficient of concordance, Bland-Altman plots, and the intraclass correlation coefficient. Guidance on the selection of an appropriate index is provided. In conclusion, selection of an appropriate index to evaluate interrater agreement or interrater reliability is dependent on a number of factors including the context in which the study is being undertaken, the type of variable under consideration, and the number of raters making assessments.

Keywords: Health services research; Research design; Reproducibility of results; Observer variation

Introduction

Evaluating interrater agreement (IRA) or interrater reliability (IRR), either as the primary focus or as a secondary component of a study, is a common objective of many social and administrative pharmacy research studies. The concept of IRA/IRR is fundamental to the design and evaluation of research instruments. However, many variations and statistical tests exist, and as a result, there is often confusion surrounding their appropriate use. This may lead to incomplete and inconsistent reporting of results. Consequently, a set of guidelines for reporting

* Corresponding author. Tel.: +61 2 9036 7081; fax: +61 2 9351 4391.
E-mail address: natasa.gisev@sydney.edu.au (N. Gisev).

1551-7411/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
doi:10.1016/j.sapharm.2012.04.004

reliability and agreement studies has recently been developed to improve the scientific rigor in which IRA/IRR studies are conducted and reported.[1]

The objective of this commentary is to outline key concepts in relation to IRA/IRR and to describe commonly used approaches to evaluating IRA and IRR. The emphasis will be on the practical aspects surrounding their use in social and administrative pharmacy research, rather than the mathematical derivation of the indices. Examples have been provided from the social and administrative pharmacy literature where possible.

Clarifying terminology

Although often used interchangeably, there is a technical distinction between the terms agreement and reliability and therefore IRA and IRR.[2-4] Fundamentally in the context of research studies, agreement is defined as the degree to which scores/ratings are identical, whereas reliability relates to the extent of variability and error inherent in a measurement (Fig. 1, Equation 1).[5]

IRA indices, therefore, relate to the extent to which different raters assign the same precise value for each item being rated. In contrast, IRR indices relate to the extent to which raters can consistently distinguish between different items on a measurement scale. The general trend in ratings is important, not the absolute value assigned by each of the raters, and the variation between ratings and measurement error is accounted for in IRR. Depending on the context of the study, terms such as interobserver and intercoder are also used.

If the ratings are performed by the same person on multiple occasions, *intra*rater agreement or reliability is instead assessed, although many of the indices still apply.

The distinction between IRR and IRA is further illustrated in the hypothetical example in Table 1. Three raters have independently rated the communication skills of 10 pharmacists on a scale of 1 to 10. For each case, comparison of the ratings across the rows provides an indication of IRA, whereas comparison of each set of ratings down the columns provides an indication of IRR. In the first case, all 3 raters having applied the same ranked order for the individual pharmacists, indicating that they are applying the same ranking system to assess the pharmacists, and therefore IRR is high. The raters also consistently provide identical scores for the pharmacists, resulting in high IRA. High IRR is again observed in case 2 despite there being differences in the actual scores provided. The general ranking provided for the pharmacists is consistent across all 3 raters, and they are able to differentiate between the pharmacists' communication skills. The lowest score is matched with the lowest scores across all 3 raters, and the rank order remains the same so that high scores are also matched. The ordering of ranks is therefore important to establish IRR. When the rankings are tightly clustered and there is no ordering of scores as observed in the example in case 3, IRR is dramatically decreased. However, IRA remains high because there is little variation in the ranking provided by the 3 raters; they have similar numerical values. Whereas IRR is sensitive to the ordering of ratings, IRA is sensitive to the variation in ratings or differences in rating levels.[4] High IRR can exist with low IRA, and thus the level of reliability does not provide an indication of the level of agreement between raters. Although seldom reported together, both indices provide valuable information when evaluating ratings by different raters.

Equation 1 $$\text{Reliability} = \frac{\text{Subject variability}}{\text{Subject variability} + \text{Measurement error}}$$

Equation 2 $$\text{Percent agreement} = \frac{\text{Number of concordant responses} \times 100\%}{\text{Total number of responses}}$$

Equation 3 $$\text{Kappa } (\kappa) = \frac{\text{Proportion observed agreement} - \text{Proportion expected chance agreement}}{1 - \text{Proportion expected chance agreement}}$$

Equation 4 $$\text{Intraclass correlation coefficent (ICC)} = \frac{\text{Between subjects variance}}{\text{Between subjects variance} + \text{Within subjects variance}}$$

Fig. 1. Equations relating to IRA and IRR calculations.

Table 1
Hypothetical ratings of communication skills (on a scale of 1 to 10) of 10 pharmacists illustrating different levels of IRR and IRA for interval-scaled data

Pharmacist	Case 1 (high IRR and high IRA)			Case 2 (high IRR and low IRA)			Case 3 (low IRR and high IRA)			Case 4 (low IRR and low IRA)		
	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
A	1	1	1	1	3	6	5	6	5	1	5	10
B	2	2	2	1	3	6	4	4	4	1	6	9
C	3	3	3	2	4	7	6	6	6	2	5	10
D	4	4	4	2	4	7	4	5	6	2	6	1
E	5	5	5	3	5	8	5	4	4	3	6	5
F	6	6	6	3	5	8	6	6	5	7	4	9
G	7	7	7	4	6	9	4	4	5	3	5	8
H	8	8	8	4	6	9	5	5	4	3	5	2
I	9	9	9	5	7	10	4	5	3	1	6	7
J	10	10	10	5	7	10	6	6	6	6	3	9

Adapted from Tinsley and Weiss.

Reviewing levels of measurement

Most of the IRA/IRR indices are suitable for the evaluation of particular variables, and it is necessary to establish the nature of the data to apply the correct technique. To assist in the selection of an index, the 4 levels of measurement are reviewed below.[6] For the purpose of data analysis, interval and ratio variables are generally treated in an identical manner.

Nominal/categorical—A variable assigned meaning through a name or category. Examples include sex and diagnostic categories.

Ordinal—A specific type of nominal variable in which there is a natural ordering or hierarchy of categories. Examples include Likert-type items and ranking in a competition.

Interval—An ordered variable in which there is an equal distance between each possible value. An example is temperature.

Ratio—An interval variable that also has a true 0 point. Examples include weight and length.

Selecting an appropriate index

There is debate in the statistical literature about the applications and appropriateness of the different IRA/IRR indices and their derivatives. Although some are strictly only valid as IRA measures (eg, Bland-Altman plots), others have been used in the literature as measures of both IRA and IRR (eg, Cohen's kappa). In many cases though, the difference in end values is not great. A suggested framework to guide the selection of an appropriate IRA/IRR index is provided in Table 2 for each of the 4 levels of measurement, differentiated by the number of raters. The main questions to consider when selecting an IRA/IRR index are:

1. What is the purpose of the analysis?
2. Is the absolute value or trend in ratings important?
3. What type of variable is being analyzed?
4. How many raters are involved?

Worked example—research scenario 1

An audit tool was developed to evaluate adherence to pain management protocols and correct charting of analgesia orders on a pediatric ward of a large teaching hospital following recent incidents that occurred on the ward. The audit tool requires judgments to be made on the presence or absence of certain criteria. After establishing the face and content validity of the audit tool, 2 clinical pharmacists were assigned to pilot the tool and collect a total of 10 cases (5 cases each). The same 2 pharmacists will be conducting the audit over 2 weeks.

Piloting a research tool to establish validity and reliability on a small number of cases is important before implementation in a study. In this example, the 2 pharmacists are independently using the audit tool, and the purpose of the analysis is to determine the intercoder reliability of the audit tool as the consistency of the pharmacists' judgments is of interest. The variable to be assessed is categorical as the presence or absence of certain criteria is recorded. Therefore, according to Table 2, either Cohen's kappa or the intraclass correlation coefficient (ICC) are appropriate IRR indices in this scenario.

Table 2
Examples of interrater indices suitable for use for various types of data[a]

	Level of measurement					
	Nominal/categorical		Ordinal		Interval and ratio	
	2 raters	>2 raters	2 raters	>2 raters	2 raters	>2 raters
Interrater indices	Cohen's kappa	Fleiss' kappa	Weighted kappa	Kendall coefficient of concordance	Bland-Altman plots	ICC
	ICC Weighted kappa	ICC	ICC	ICC	ICC	

[a] Table is not exhaustive and represents a summary of some of the indices and the contexts in which they can be used only.

Worked example—research scenario 2

An expert panel consisting of a pharmacist, pharmacologist, and physician, was convened to judge the potential clinical significance of pharmacists' interventions in a collection of reports of "near-miss" events detected in community pharmacies. Each panelist independently scored each event on a 5-point Likert-type scale with a score of 1 assessed as "not harmful" to 5 being "life threatening."

The purpose of this study was to determine the level of agreement among the 3 panelists and how similar the actual scores are to one another. Likert-type scales are treated as an ordinal variable, and with 3 individuals scoring each report, Table 2 indicates that either the Kendall coefficient of concordance or ICC would be suitable IRA indices.

Explanation of key indices

Percent agreement

Percent (or proportion) agreement is a basic IRA index that offers a measure of raw agreement (Fig. 1, Equation 2).[7] There is no limit to the number of raters that can be assessed; however, the effect of chance in achieving agreement between raters is not accounted for. Nevertheless, it is still used and is sometimes reported in conjunction with indices that are chance-corrected. Mathematically, percent agreement is the simplest IRA index to calculate. An example calculation is provided in Holdford's[7] review of content analysis methodology.

The kappa index (κ)

Kappa calculations are one of the original and most commonly used IRA/IRR indices. Arising from Cohen's[8] seminal work in 1960, kappa describes an index for nominal/categorical variables assessed by 2 raters. Since then, a number of variations have been proposed, and the term *kappa* now refers to a group of indices. These indices provide a chance-corrected index of IRA/IRR and are based on the ratio of the proportion of times agreement is observed, to the maximum proportion of times that the raters could agree (both corrected for chance agreement) (Fig. 1, Equation 3).[9] An example calculation is provided by Holdford.[7]

Kappa can take any value between −1 and +1, where +1 indicates perfect agreement. However, mathematically, a value of ±1 is difficult to achieve and is only observed in extreme circumstances.[8] Furthermore, the lower limit of kappa varies and is dependent on the number of categories.[8] Negative values indicate that the observed agreement is less than that expected from chance alone; a value of 0 indicates exactly chance agreement, and positive values indicate that the observed agreement is greater than that expected from chance.[8]

There is wide variation in the interpretation of kappa values, and several attempts have been made to assign practical meaning to calculated kappa values. The most comprehensive and widely cited interpretation was proposed by Landis and Koch[10] (Table 3). This classification is often simplified into 3 categories such that a kappa value of 0.75 or greater is considered to represent an excellent level of agreement, a value of 0.40 or less is indicative of poor agreement, and values between 0.40 and 0.75 represent fair to good agreement.[11] However, because of the inherent properties of the kappa formula, it has been suggested that this upper limit is unnecessarily high and realistically may not be achievable in the context of some research studies.[3] Hence, a low kappa value may not always be indicative of low agreement.

Table 3
Interpretation of kappa values suggested by Landis and Koch[10]

Kappa	Interpretation
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Furthermore, kappa is sensitive to bias between raters and the overall prevalence of responses.[12] In some instances, a relatively high proportion of observed agreement can result in a low kappa value and an unbalanced or biased distribution of responses can result in a higher kappa value than a more balanced distribution of responses.[13] To assist in the interpretation of kappa values and identify potential bias, the reporting of average proportions of agreement for positive and negative responses, in addition to the overall kappa value, is recommended.[14] These proportions, referred to as p_{pos} and p_{neg}, respectively, are calculated by dividing the number of positive (or negative) ratings observed by the mean number of positive (or negative) ratings. For example, p_{pos} and p_{neg} were reported in a study evaluating the IRR of physicians' responses to a tool identifying potentially inappropriate prescribing in older people.[15] In this study, raters agreed that in most cases, the particular criterion in question was not fulfilled, producing a heavily skewed distribution of responses. Reporting of p_{pos} and p_{neg} was therefore necessary to accurately interpret the results of the analyses. In another study, p_{pos} and p_{neg} were reported in the comparison of self-reported measures of agreement with those based on pharmacy records.[16] Another index, the *prevalence-adjusted bias-adjusted kappa* (denoted as PABAK) has also been proposed to correct for any potential bias in the kappa value for dichotomous variables assessed by 2 raters.[12] As an example, the PABAK was used by researchers comparing ratings on the presence or absence of specific drug-related problems by 2 pharmacists.[17]

Cohen's kappa (κ)

Cohen's unweighted kappa, often referred to as Cohen's kappa, remains the most widely used IRA/IRR index. It provides a chance-corrected index of IRA/IRR in studies employing the same 2 raters to rate items on a nominal scale.[8] Although theoretically possible, it is not recommended to compute a Cohen's kappa to assess agreement on ordinal scales because alternative measures are more appropriate.[18] Three assumptions must be met when using Cohen's kappa:[8]

1. The items to be rated are independent.
2. The categories are independent, mutually exclusive, and exhaustive.
3. The raters are independent.

Similar to other statistical tests, kappa values should be reported with the corresponding standard error and hypothesis testing undertaken to determine statistical significance.[9] The null hypothesis that kappa equals 0 against the alternative hypothesis that kappa is greater than 0 is tested. Rejection of the null hypothesis therefore indicates that any agreement observed is statistically significant. Information on hypothesis testing[8,19] and calculation of the standard error[8,19] and confidence interval[19] for Cohen's kappa are detailed elsewhere and commonly included in the output provided by statistical software.

Weighted kappa (κ_w)

Weighted kappa is an extension of Cohen's kappa and can be used in situations in which either nominal/categorical or ordinal variables are coded by 2 raters. Whereas Cohen's kappa considers only total ("all-or-none") agreement or disagreement, weighted kappa allows for the assignment of weights to different categories such that similar categories can be in partial agreement.[20] For example, on a scale of 1 to 5, scores of 1 and 2 have higher agreement than scores of 1 and 3. Similarly, in the case of nominal variables, if psychiatrists were to categorize patient diagnoses into 1 of 3 categories, personality disorder, neurosis, and psychosis, there would be higher agreement between neurosis and personality disorder than between psychosis and neurosis.[20]

Weighted kappa is suitable for rating items that have between 3 and 10 ordinal categories, and the minimum sample size required to approximate a normal distribution is $2x^2$ where x is the number of categories.[21] Weights are usually assigned on the basis of disagreement rather than agreement, and often a quadratic weighting system is used such that disagreement is weighted by the square of the number of levels separating the raters.[22]

Again, a standard error and confidence interval can be calculated and hypothesis testing performed.[11,20] The interpretation of weighted

kappa values is identical to Cohen's kappa,[20] and as expected, weighted kappa values tend to be higher than unweighted kappa values.[23]

Fleiss' kappa (κ_v)

Another extension of kappa was developed by Fleiss[24] for use when nominal categories are assessed by multiple raters. Fleiss' kappa is also chance-corrected; however, unlike Cohen's kappa, Fleiss' kappa does not assume that the same raters have assessed all items. Formulas for calculating the standard error and for testing the hypothesis that the observed agreement equals chance agreement have also been described.[24] A version of Fleiss' kappa specifically for dichotomous variables has also been proposed.[25] Fleiss' kappa was applied to a study assessing the IRA of 5 experts evaluating adverse drug effect causality for a number of adverse event-drug pairs.[26]

Kendall coefficient of concordance (W)

The Kendall coefficient of concordance is suitable for ordinal variables assessed by multiple raters.[9] A W score provides an indication of the strength of agreement and is interpreted with its corresponding P-value. Unlike kappa that can take negative values, W scores range between 0 and 1, where 0 signifies no agreement and 1 signifies complete agreement. Negative W values are impossible because complete disagreement cannot be achieved with more than 2 raters.[9] It becomes increasingly harder to achieve high W scores as the number of raters increases, and consequently, low W scores can also be significant.[27] When testing the significance of the W score, the null hypothesis that the ratings of the different judges are independent of one another is tested. Rejection of the null hypothesis (using a 1-tailed test) therefore enables one to conclude that any agreement observed between the judges is statistically significant.

Various interpretations and reporting of the results of tests using the Kendall coefficient of concordance have been reported in the literature. The Landis and Koch[10] interpretation of kappa categories has been extended to the interpretation of W scores.[28] Furthermore, an interpretation linking the W score and confidence in rankings has been proposed.[27] Additionally, the W score has also been interpreted analogously to the correlation coefficient.[29]

In practice, the Kendall coefficient of concordance has been used in social and administrative pharmacy studies employing a panel of experts to make judgments on rankings of an item. In 1 study, 5 experts were asked to rank the medication regimen complexity of 6 medication regimens, which was compared with the newly developed Medication Regimen Complexity Index.[30] In another study, a 4-member multidisciplinary expert panel was employed to assess the expected outcomes of comprehensive medication reviews for clients of community mental health teams.[29] Using a 5-point Likert-type scale, each panelist independently assessed review findings, review recommendations, likelihood of recommendation implementation, and the overall expected clinical impact. Agreement among panelists was established with W scores for each of the scales.

Similarly, another useful application of the Kendall coefficient of concordance is in the conduct of multiple-round Delphi surveys.[27] The W score can be used to determine whether consensus has been reached, whether consensus is increasing between rounds, and also the relative strength of consensus.

Bland-Altman plots

Calculating a Pearson's product-moment correlation coefficient (r) seems a logical choice to assess the level of agreement of 2 raters for interval or ratio data. However, use of the Pearson's correlation coefficient is inappropriate as an IRA index.[31] To clarify this point, it is useful to consider the representation of data visually. When data are plotted, perfect correlation is evident when points lie on any straight line; however, perfect agreement is only evident when points lie on the line of equality (ie, $y = x$). Thus, a Pearson's correlation coefficient is indicative of the strength of the relationship for 2 variables and not agreement. A high correlation coefficient may be observed even though agreement is poor.

To overcome these limitations and accurately evaluate IRA, Bland and Altman proposed an alternative approach that relies on graphically plotting scores.[31] Originally developed to measure agreement in method comparison studies, this approach can also be applied to IRA studies.[31,32] Each point on the line is derived by plotting the difference in scores of the 2 raters (x) against the average of the 2 scores (y). The magnitude of disagreement and any outliers and trends in scores can then be determined from the graph. Additionally, the *95% limits of agreement* can be estimated by calculating the mean difference ± 1.96 multiplied by the standard deviation of the differences, providing

an interval in which 95% of the differences in ratings are expected to lie, provided that the differences are normally distributed.[31,32] A nonparametric alternative has also been described.[32]

In the context of social and administrative pharmacy research, Bland-Altman plots have been used to determine agreement between blood pressure measurements taken by community pharmacists and values obtained through ambulatory and home blood pressure recordings.[33] In another example, the accuracy of using pharmacy sales data to monitor broad-spectrum antibiotic use was compared with a pharmacist conducting weekly ward stock counts.[34]

Intraclass correlation coefficient (ICC)

The ICC is widely reported in the literature and relates to the proportion of variance attributable to subjects being measured (Fig. 1, Equation 4).[35,36] Based on analysis of variance (ANOVA) models, the ICC was originally applied to the evaluation of differences between interval or ratio variables.[35,36]

The ICC is able to handle situations of multiple raters, the evaluation of different sets of scores, and situations when there are missing ratings. Similar to the other tests described, ICC values should be reported with corresponding *P*-values or confidence intervals.

Several forms of the ICC exist, and choice is dependent on whether:

1. To treat the data as a 1-way or 2-way ANOVA model; and
2. The absolute value or consistency of ratings is important; and
3. The unit of analysis is an individual rating or the mean of several ratings.

Based on the above criteria, a comprehensive flowchart has been developed by McGraw and Wong[35] to assist in the selection of an appropriate ICC, each with their own specific formula for calculation. Each type of ICC can be explained by 1 of 3 underlying models that stem from the typical IRA/IRR scenario of a number of raters independently assessing a random sample of items:

1. *One-way random effects model*—Each item is assessed by a different set of randomly selected raters.
2. *Two-way random effects model*—Each item is assessed by all raters who have been randomly selected from a larger population of raters.
3. *Two-way mixed model*—Each item is assessed by all raters in the population of interest.

Under certain conditions, the ICC has shown to be equivalent to Cohen's kappa, weighted kappa, and the Kendall coefficient of concordance[19,37] It has also been argued that the ICC should replace Cohen's kappa and weighted kappa because it offers greater flexibility in data analysis.[22] However, in doing so, the fundamental rules regarding levels of measurement need to be disregarded, although the outcomes and interpretation of the results may not differ significantly. Furthermore, the application of the ICC to nonparametric contexts is a developing field of research and new indices continue to be developed.

An example of the use of the ICC to social and administrative pharmacy research includes establishing the IRR of a newly developed Medication-Based Disease Burden Index for the quantification of disease burden using chronic drug therapy data.[38] Additionally, the ICC was used to test the intrarater reliability and IRR of observer ratings using the visual analog scale in infants undergoing immunization injections.[39]

General considerations

IRA/IRR values can be interpreted in a number of ways and ranges, indicating that degrees of agreement and reliability are arbitrary. Therefore, it may not be possible to predefine an acceptable level of agreement or reliability. Rather, a judgment should be made regarding the interpretation of IRA/IRR values, considering the nature of the study and possible implications of the results.

Another point to acknowledge is that high IRA indicates that raters are concordant on a particular response and is independent of whether the choices made by the raters are correct. Thus all the raters may be applying the same (incorrect) reasoning when scoring items.

Furthermore, the results of an IRA/IRR analysis are unique to the individual study. They are a function of the population of interest and are dependent on the raters, the responses, and the rating scale used. IRA/IRR values are therefore not generalizable to other studies.[5,40]

Conclusion

IRA and IRR relate to 2 distinct concepts. The absolute value is important in the evaluation of IRA, whereas the consistency of ratings is important in the evaluation of IRR. Opinions on the appropriateness and suitable applications of IRA/IRR indices vary,

prompted by the fact that several indices produce similar results. Selection of an index therefore needs to be justified, bearing in mind the context and purpose of the study, as well as ease of calculation and interpretation of the results.

References

1. Kottner J, Audigé L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol* 2011;64: 96–106.
2. de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol* 2006;59:1033–1039.
3. LeBreton JM, Senter JL. Answers to 20 questions about interrater reliability and interrater agreement. *Organ Res Methods* 2008;11:815–852.
4. Tinsley HEA, Weiss DJ. Interrater reliability and agreement. In: Tinsley HEA, Brown SD, eds. *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. San Diego, CA: Academic Press; 2000. p. 95–124.
5. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 4th ed. New York, NY: Oxford University Press; 2008.
6. Stevens SS. On the theory of scales of measurement. *Science* 1946;103:677–680.
7. Holdford D. Content analysis methods for conducting research in social and administrative pharmacy. *Res Social Adm Pharm* 2008;4:173–181.
8. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
9. Siegel S, Castellen NJ Jr. *Nonparametic Statistics for the Behavioral Sciences*. 2nd ed. New York, NY: McGraw-Hill Book Co; 1988.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33: 159–174.
11. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*. 3rd ed. New York, NY: John Wiley & Son, Inc; 2003.
12. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol* 1993;46:423–429.
13. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990;43:543–549.
14. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 1990;43:551–558.
15. Gallagher P, Baeyens J-P, Topinkova E, et al. Interrater reliability of STOPP (Screening Tool of Older Persons' Prescriptions) and START (Screening Tool to Alert doctors to Right Treatment) criteria amongst physicians in six European countries. *Age Ageing* 2009;38:603–606.
16. Guénette L, Moisan J, Préville M, Boyer R. Measures of adherence based on self-report exhibited poor agreement with those based on pharmacy records. *J Clin Epidemiol* 2005;58:924–933.
17. LaFleur J, Larson BS, Gunning KM, et al. Agreement between pharmacists for problem identification: an initial quality measurement of cognitive services. *Ann Pharmacother* 2009;43:1173–1180.
18. Kraemer HC, Periyakoil VS, Noda A. Kappa coefficients in medical research. *Stat Med* 2002;21: 2109–2129.
19. Sheskin DJ. *Handbook of Parametric and Nonparametric Statistical Procedures*. 4th ed. Boca Raton, FL: Chapman & Hall/CRC; 2007.
20. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213–220.
21. Cicchetti DV. Testing the normal approximation and minimal sample size requirements of weighted kappa when the number of categories is large. *Appl Psychol Meas* 1981;5:101–104.
22. Streiner DL. Learning how to differ: agreement and reliability statistics in psychiatry. *Can J Psychiatry* 1995;40:60–66.
23. Soeken KL, Prescott PA. Issues in the use of kappa to estimate reliability. *Med Care* 1986;24:733–741.
24. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971;76:378–382.
25. Fleiss JL, Cuzick J. The reliability of dichotomous judgments: unequal numbers of judges per subject. *Appl Psychol Meas* 1979;3:537–542.
26. Arimone Y, Miremont-Salamé G, Haramburu F, et al. Inter-expert agreement of seven criteria in causality assessment of adverse drug reactions. *Br J Clin Pharmacol* 2007;64:482–488.
27. Schmidt RC. Managing delphi surveys using nonparametric statistical techniques. *Decis Sci J* 1997; 28:763–774.
28. Levitan B, Yee CL, Russo L, Bayney R, Thomas AP, Klincewicz SL. A model for decision support in signal triage. *Drug Saf* 2008;31:727–735.
29. Gisev N, Bell JS, O'Reilly CL, Rosen A, Chen TF. An expert panel assessment of comprehensive medication reviews for clients of community mental health teams. *Soc Psychiatry Psychiatr Epidemiol* 2010;45:1071–1079.
30. George J, Phun Y-T, Bailey MJ, Kong DCM, Stewart K. Development and validation of the Medication Regimen Complexity Index. *Ann Pharmacother* 2004;38:1369–1376.
31. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–310.
32. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135–160.
33. Sabater-Hernández D, De La Sierra A, Sánchez-Villegas P, et al. Agreement between community pharmacy and ambulatory and home blood pressure measurement methods to assess the effectiveness of

antihypertensive treatment: the MEPAFAR study. *J Clin Hypertens (Greenwich)* 2012;14:236–244.
34. Haug JB, Myhr R, Reikvam Å. Pharmacy sales data versus ward stock accounting for the surveillance of broad-spectrum antibiotic use in hospitals. *BMC Med Res Methodol* 2011;11:166.
35. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996;1:30–46.
36. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86: 420–428.
37. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas* 1973;33: 613–619.
38. George J, Vuong T, Bailey MJ, Kong DC, Marriott JL, Stewart K. Development and validation of the Medication-Based Disease Burden Index. *Ann Pharmacother* 2006;40:645–650.
39. Taddio A, O'Brien L, Ipp M, Stephens D, Goldbach M, Koren G. Reliability and validity of observer ratings of pain using the visual analog scale (VAS) in infants undergoing immunization injections. *Pain* 2009;147:141–146.
40. Tinsley HEA, Weiss DJ. Interrater reliability and agreement of subjective judgments. *J Couns Psychol* 1975;22:358–376.