Journal of Aging and Health
Volume 20 Number 2
March 2008 183-197
© 2008 Sage Publications
10.1177/0898264307310448
http://jah.sagepub.com
hosted at
http://online.sagepub.com

Testing Measurement Reliability in Older Populations

Methods for Informed Discrimination in Instrument Selection and Application

Peter H. Van Ness, PhD, MPH
Virginia R. Towle, MPhil
Manisha Juthani-Mehta, MD
Yale University School of Medicine, New Haven, Connecticut

Objective: The authors recommend confidence intervals as measures of precision for reliability coefficients, regression modeling as supplements for such omnibus reliability statistics, and unreliability detection as a goal of reliability testing distinct from reliability inference. **Methods:** Illustrative reliability analyses are conducted on measures selected from a study of clinical features associated with urinary tract infection in older nursing home residents. **Results:** Standard methods for reliability testing (e.g., kappa coefficients) are often inappropriate for small samples, and exact methods or descriptive reliability statistics are viable alternatives. **Discussion:** Supplementation of omnibus statistics by loglinear regression modeling is especially appropriate for aging research because it facilitates tests of marginal homogeneity and comparisons of reliability results for relatively young and old subgroups. Latent class regression analysis is useful for older samples because multifactorial health conditions are often measured in multiple ways and assessment of their reliability can be integrated, granting certain assumptions, with validity assessment.

Keywords: *reliability testing; confidence intervals; loglinear models; latent class analysis; aging*

Reliability testing in clinical aging research includes comparisons of results of measurements given on separate occasions—test–retest reliability—and measurements obtained by different raters—interrater reliability. Such testing seeks to determine whether results obtained by measurement instruments will likely be replicated. Cohen's kappa coefficient is most often used for categorical measurements and the intraclass correlation

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

coefficient for continuous measurement scales (Cohen, 1960; Fisher, 1925). The objectives of this article are threefold: to recommend that confidence intervals as measures of precision accompany reliability coefficients, to indicate that regression modeling can overcome some limitations of these omnibus reliability statistics, and to distinguish between reliability inference and unreliability detection as goals of reliability testing. Clarity about these three points will allow us to address practical problems that arise in testing reliability in an illustrative study of clinical features associated with urinary tract infection in older nursing home residents.

Confidence Intervals Recommended

Even conscientiously conducted reliability studies that compare results to some minimally acceptable level of reliability often fail to include measures of precision with kappa or intraclass correlation coefficients (Gregson et al., 2000; Wolinsky, Miller, Andresen, Malmstrom, & Miller, 2005). A kappa coefficient of 0.45 alone does not provide sufficient evidence to infer that the tested measurement instrument satisfies the often-cited Landis and Koch level of greater than 0.40 for "moderate" reliability (Landis & Koch, 1977, p. 165). A 90% confidence interval (CI) with a lower bound greater than 0.4 is also required, thereby documenting that a null hypothesis is rejected for a one-sided significance level of .05. If a measurement instrument has a kappa value of 0.45 but a confidence interval the lower bound of which extends considerably below 0.40, then the next study participant randomly drawn from the same population might be measured with less than moderate reliability.

Minimum acceptable values of an intraclass correlation coefficient have been discussed. Fleiss (1986) described values from 0.40 to 0.75 as "fair to good" (p. 7); Streiner and Norman (1995) recommended values greater than 0.75 for continuous scales used in health research. (These criterial levels should be used with care and common sense; a recent article described them as "hopelessly arbitrary"; de Mast, 2007, p. 152.) Flack, Afifi, Lachenbruch, and Schouten (1988) and Walter, Eliasziw, and Donner (1998) provided sample size formulae for the kappa and intraclass correlation coefficients, respectively, so that reliability studies can be correctly powered (see Figure 1).

Authors' Note: This study was supported by Claude D. Pepper OAIC at Yale University School of Medicine (#P30AG21342). We thank Heather G. Allore for her assistance. Please address correspondence to Peter H. Van Ness, Yale University School of Medicine, Department of Internal Medicine, Program On Aging, 300 George Street, Suite 775, New Haven, CT 06511; e-mail: peter.vanness@yale.edu.

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

Figure 1
Power Tables for the Kappa Coefficient (κ) and the Intraclass Correlation Coefficient (ICC)

Power Table for Kappa Coefficients

With Two Observations per Subject for a

Binary Variable With .40 Prevalence

Using a One-Sided Test at Alpha = 0.05

and With a κ = 0.40 Null Hypothesis*

	Observed κ			
Sample Size	0.60	0.70	0.80	0.90
20	0.22	0.40	0.66	0.94
30	0.30	0.56	0.86	0.99
50	0.44	0.79	0.98	0.99
100	0.72	0.98	0.99	0.99

Power Table for

Intraclass Correlation Coefficients

With Two Observations per Subject

Using a One-Sided Test at Alpha = 0.05

and With an ICC = 0.75 Null Hypothesis#

	Observed ICC			
Sample Size	0.80	0.85	0.90	0.95
20	0.13	0.34	0.70	0.98
30	0.16	0.45	0.85	0.99
50	0.22	0.63	0.97	0.99
100	0.35	0.88	0.99	0.99

*(Nee, 1998.) #(Hinze, 2004).

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

The practice of not reporting measures of precision for reliability test results originates, perhaps, from the irrelevance of *p* values from hypothesis tests whose null values are zero, for example, reliability assumed to be only marginally greater than chance is hardly worth the effort to assess. Measures of precision, such as confidence intervals, are not required for Cronbach's alpha coefficient because it is mathematically already the lower bound of a reliability coefficient (Cronbach, 1951). Its reporting may have set an historical precedent for reporting kappa coefficients and intraclass correlation coefficients. Whatever the origin of the practice might be, when these kappa and intraclass correlation coefficients are used for statistical inference, they should be accompanied by confidence intervals.

Omnibus Statistics

Kappa and intraclass correlation coefficients can be described as "omnibus quantities" because they summarize several dimensions of relevant data in a single number, and thus this same number can represent a plurality of dimensional configurations (van Belle, 2002, pp. 6-7; compare Figures 2a and 2b). This omnibus status makes them easy to calculate and to interpret; however, it also has limitations. Reliability studies often present results for several measurement instruments and at least implicitly claim to indicate which of the tested instruments are most reliable. Even when accompanied by suitable measures of precision, kappa coefficients and intraclass correlation coefficients for ordinal data (which are approximately equivalent to kappa values when quadratically weighted; Fleiss & Cohen, 1973) cannot adequately discriminate between the reliability of two instruments unless an unrealistic presupposition is met. This presupposition posits that the distributions of what is being measured are approximately the same for the two tables (Thompson & Walter, 1988). Even for a single table, if marginal totals (summations across specific rows or columns in a contingency table representing agreement data) vary from one rater to another, the kappa coefficient may take on different values even though the total proportion of agreement remains the same (Feinstein & Cicchetti, 1990).

In discussing the ambiguities introduced by kappa coefficients with unbalanced distributions, that is, the prevalence of the condition of interest differs for two ratings summarized in the same table, Cicchetti and Feinstein (1990) recommended supplementing reports of the coefficient with the proportion of agreement for each level of the measurement variable. (Proportions of positive and negative agreement discriminate between Figures 2a and 2b, for which the kappa coefficients are the same; they also suggest that the

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

kappa value from Figure 2c is more readily comparable to the kappa value for Figure 2b than for Figure 2a.) Although proportions of agreement are convenient and valuable, supplementing omnibus reliability statistics with relevant regression modeling techniques is a more informative and more general approach.

Reliability and Regression Modeling

In the case of the intraclass correlation coefficient, the advent of linear mixed effect models allows for its calculation from a single regression model. It can be obtained as an item in the model's correlation matrix (*SAS/STAT User's Guide,* SAS Institute, 2005). It provides the flexibility of calculating different versions of the intraclass correlation coefficient, ones for only randomly selected study participant samples, ones with only randomly selected raters, and ones with random terms for both study participants and raters. Confidence intervals and subgroup analyses can easily be calculated. Two other regression modeling techniques are especially helpful for evaluating reliability for nominal and ordinal scales.

Loglinear Regression Models

Loglinear regression models have been used to analyze rater agreement since the mid-1980s (Tanner & Young, 1985a, 1985b). Loglinear models are used instead of standard linear models because agreement data occur as discrete counts rather than on continuous scales. Counts can be treated as independent observations from a Poisson distribution. Loglinear models lend themselves to modeling agreement beyond chance because their simplest form—the independence model—assumes that the mean values of cells in a 2×2 table, $m_{ij,}$ can be estimated by the product of the table sample size, n, and the probabilities of counts occurring in a specified row, π_{i+}, and column, π_{+j}. The natural logarithm of the mean number of counts is used because this transformation makes the aforementioned multiplicative relationship additive, that is, linear in the parameters. Modifications of this model attempt to capture patterns of agreement beyond chance.

Results from loglinear agreement models overcome the shortcomings of the kappa omnibus reliability statistic in several ways. Its primary measure of association, the agreement odds ratio, is more discriminating; for example, it is less liable to give the same numerical value for different data configurations. (Note the distinct odds ratios in Figures 2a–2c; Figure 2d shows odds ratios to be invariant under transposition of both rows and columns.)

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

Figure 2
Comparison of the Kappa Coefficient (κ) and the Odds Ratio (OR) as Omnibus Statistics in Agreement Tables

a

	Rater 2		
Rater 1	35	15	50
	15	35	50
	50	50	

κ = 0.40 (90% CI* 0.25, 0.55)

OR = 5.44

proportion + agreed = 0.70 = 2a/ [N + (a – d)][#]

proportion – agreed = 0.70 = 2d/[N – (a – d)][#]

b

	45	5	50
	25	25	50
	70	30	

κ = 0.40 (90% CI 0.26, 0.54)

OR = 9.00

proportion + agreed = 0.75

proportion – agreed = 0.625

*CI = Confidence Interval.

[#](Cicchetti and Feinstein, 1990) Agreement tables are lettered consecutively by rows, from left to right and top to bottom.

(*continued*)

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

Figure 2 (continued)

c

45	15	60
15	25	40
60	40	

κ = 0.375 (90% CI* 0.22, 0.53)

OR = 5.00

proportion + agreed = 0.75

proportion – agreed = 0.625

d

25	15	40
15	45	60
40	60	

κ = 0.375 (90% CI 0.22, 0.53)

OR = 5.00

proportion + agreed = 0.625

proportion – agreed = 0.75

The agreement odds ratio for two raters can be defined in an analogous way to an odds ratio used in cohort studies. Assuming for pairs of participants that each rater classifies them in one of two categories, *i* and *j*:

$$\text{AOR} = \frac{\text{odds of rater 2 classifying participants in } i \text{ when rater 1 classifies them in } i}{\text{odds of rater 2 classifying participants in } i \text{ when rater 1 classifies them in } j} = \frac{\text{odds of concordance in category classification among raters}}{\text{odds of discordance in category classification among raters}}$$

Like the kappa and intraclass correlation coefficients, larger values of the agreement odds ratio indicate that observers are more likely to agree for

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

the given pair of categories (Agresti, 2002). (Although the agreement odds ratio is interpretable in a way analogous to traditional odds ratios, it is calculated differently.) When there are only two response categories, the agreement odds ratio is calculated from a parameter δ that represents the extent of exact agreement beyond chance. For models with ordinal response categories, a second parameter (β) can be estimated that represents beyond chance agreement because of a linear association between ratings obtained from two raters. Thus, in these models not only can agreement be decomposed into agreement because of chance and beyond chance agreement, but beyond chance agreement is further decomposed into parts attributable to exact agreement and linear association (Velema, Blettner, Restrepo, & Munoz, 1991).

Figure 3a shows an agreement table relevant to reliability testing of a measure of ease of distraction administered to an older nursing home population. It is one of several variables designed to measure changes in mental status that are thought to be clinical features of urinary tract infections in this population. The weighted (quadratic) kappa coefficient for this table is 0.34 (90% CI = 0.12, 0.57). This is not an acceptable level of reliability, and one might be interested in the nature and sources of the unreliability. One might assess the marginal homogeneity of the agreement table, that is, whether the probability of falling in any category of the row classification is equal to the probability of falling in any category of the corresponding column classification. Intuitively, it tests whether disagreements—cell counts occurring off of the left-to-right table diagonal—occur in a differential pattern that might be amenable to correction by further training, or in a more random way that might simply reflect a limitation of the measurement instrument.

Two-by-two tables can be tested for marginal homogeneity using a McNemar test for symmetry (McNemar, 1947). Rejection of the null hypothesis of symmetry in this case implies rejection of a null hypothesis of marginal homogeneity, which indicates that differential disagreement occurs to an extent statistically significant at some specified level, usually .05. This approach to testing marginal homogeneity is applicable to 2×2 tables but not for larger square tables. Loglinear regression techniques provide a more general way to test for marginal homogeneity. A loglinear model can be fit that assumes that counts occurring off of the main diagonal of a square contingency table are symmetrically distributed. A likelihood ratio chi-square statistic measures model goodness of fit. With only a slight modification a second loglinear model can be fit that relaxes the symmetry assumption to allow for marginal heterogeneity (Darroch & McCloud, 1986). Comparison of the likelihood ratio chi-square from this quasi-symmetry model and the aforementioned symmetry model allows for a statistical test of a null hypothesis

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

of marginal homogeneity. Rejection of this null hypothesis indicates that differential disagreement between raters occurs in a way that is statistically significant. For Figure 3a the symmetry model yields a likelihood ratio chi-square statistic of 4.53 with 3 degrees of freedom (*df*), and the quasisymmetry model yields values of 0.19 with 1 *df*. Hence, upon subtracting the latter values from the former, a chi-square test of 4.34 for 2 *df* has a p value of .114. In some cases, especially for small sample sizes, results will be questionable because of poor-fitting models; this topic is addressed subsequently.

A second issue that a clinical researcher might want to investigate for the ease of distraction measure is whether the reliability of this instrument differs for two subgroups. Often of interest are possible differences in reliability when the younger versus the older portions of the cohort are compared. Figures 3b1 and 3b2 represent agreement data for two age-related subgroups. The weighted kappa for the younger half (65-86) is 0.27 (90% CI = 0.00, 0.53), and it is 0.47 (90% CI = 0.17, 0.77) for the older half (87+). A test for the equality of the two kappa coefficients fails to reject the null hypothesis of equality, $\chi^2(1) = 0.68$, $p = .411$. Two factors require that this test for equality be interpreted with caution. First, the marginal distributions of the two age-related agreement tables are different, so comparison of the two coefficients is problematic. Second, the small sample size makes the test underpowered. Geriatric researchers are also often interested in knowing whether reliability differs for proxy responses versus older study participant responses. A similar subgroup analysis would be insightful for investigating this issue.

An advantage of the loglinear regression approach is that it can incorporate a binary covariate into the model that allows for statistical inferences as to whether the reliability of an instrument differs for two groups (Graham & Jackson, 2000). It allows for the calculation of agreement odds ratios for pairs of levels in the measurement scale and thereby avoids some of the ambiguity in comparing kappa coefficients of differently distributed agreement tables.

Latent Class Agreement Regression Models

The dependence of the kappa coefficient on imbalances in marginal distributions of agreement data motivates its supplementation with additional information that can be provided effectively by loglinear regression models. A second consideration motivates supplementation of omnibus reliability statistics with latent class regression models. Testing a measurement instrument for reliability pragmatically implicates that it successfully measures what it is intended to measure; it implicates validity understood as diagnostic

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

Figure 3
Subgroup Decomposition by Age of a 3 × 3 Agreement Table for a Measure of Ease of Distraction ($N = 30$)

a

Rater 2

Rater 1

8	6	1
2	9	3
0	1	0

Table 3b1: Age ≤ 86

3	4	1
0	4	2
0	1	0

Table 3b2: Age > 86

5	2	0
2	5	1
0	0	0

accuracy. An instrument that consistently misses its mark is little redeemed by the consistency of its errors. Validity is difficult to test statistically. Latent class models provide some insight.

Latent variable regression models (also describable as finite mixture models) differ from traditional regression models by containing parameters that describe unobserved variables. When modeling rater agreement, they model the joint distribution of ratings as a mixture of distributions for levels of a latent variable. They effectively relax the traditional assumption that the same probability model holds for the entirety of the data set being analyzed. In clinical reliability analyses, the latent variable might be disease severity such that the rating scale posits certain disease thresholds that correspond to rating levels. These thresholds might be understood to mark points on continuum of disease severity (latent trait models) or to specify transitions between

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

homogenous stages of disease progression (latent class models). The simplest case of a latent class model posits two classes of a health condition—its presence and absence (Uebersax, 1992; Uebersax & Grove, 1990).

Measurement error in this context is relative, and its assessment is based on an important assumption and a key data requirement. It is assumed that (a) if two ratings disagree, one is correct and the other incorrect, and (b) if a plurality of ratings gives the same result, this result is correct. These assumptions allow some assessment of validity in the absence of a definitive criterion but obviously require that there be data from at least three, and preferably more, raters. Some latent variable models permit inferences about rating sensitivity, specificity, and the area under a Receiver Operating Characteristic curve. These model results have the advantage of being easily interpretable in a clinical context and readily comparable to other relevant information.

A useful application of latent class modeling addresses the multifactorial nature of many health conditions among older persons. For instance, ease of distraction is not the only dimension of a change in mental status that might be relevant to diagnosing a urinary tract infection. Others are measures of altered perception, disorganized speech, restlessness, lethargy, and daily mental variability. The latent classes of change and no change in mental status are identified as a function of the covariances among the six variables (Lanza, Lemmon, Schafer, & Collins, 2007; see Table 1). The lethargy and daily mental variability variables are least sensitive, with the lethargy variable also having the worst specificity. (Confidence intervals should likewise accompany measures of sensitivity and specificity; Ely et al., 2001.) This suggests that the lethargy variable is poorly measuring the change in mental status that one intends to measure with the other variables and might best be deleted from the group in study analyses. Note that the ease of distraction has fairly strong sensitivity and specificity results despite its apparently limited interrater reliability. Using regression techniques like latent class analysis provides additional perspectives on measurement instruments and makes possible informed discrimination in instrument selection and/or correction.

Reliability Inference and Unreliability Detection

In formal reliability studies in which inferences are drawn, statistics like the kappa and intraclass correlation coefficients should be accompanied by confidence intervals. Evaluation of measurement reliability, however, is often undertaken for more practical purposes such as detecting unreliability in instrument administration that might be corrected by further training

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

Table 1
Results of a Latent Class Analysis of Six Change of Mental Status Variables

	Diagnostic Accuracy Estimates	
Variable	Sensitivity (95% CI)	Specificity (95% CI)
Ease of distraction	0.78 (0.61, 0.95)	0.87 (0.77, 0.97)
Altered perception	0.73 (0.55, 0.92)	0.87 (0.77, 0.97)
Disorganized speech	0.88 (0.75, 1.00)	0.98 (0.94, 1.00)
Restlessness	0.71 (0.52, 0.90)	0.86 (0.75, 0.97)
Lethargy	0.63 (0.43, 0.83)	0.77 (0.64, 0.90)
Daily mental variability	0.64 (0.44, 0.84)	0.88 (0.78, 0.98)

Note: $N = 62$. For latent class prevalence estimates, percentage of no change = 0.65, and percentage of change = 0.35. For both latent class prevalence estimates and diagnostic accuracy estimates, the likelihood ratio chi-square statistic = 57.73 with 50 *df* ($p = .211$), so there is a failure to reject the null hypothesis of goodness of fit. CI = confidence interval.

or scale modification. In clinical aging research, small sample sizes are often used for such practical purposes, rendering statistically significant results unlikely. What is especially important in these circumstances is to avoid bias introduced by small sample sizes. Exact versions of kappa coefficients are available as are exact tests of marginal homogeneity (*StatXact User's Guide;* Cytel Statistical Software & Services, 2006). (For data in Figure 3a, exact methods yield substantively similar analytic results as reported previously here.) Unbalanced distributions can be especially pronounced in small agreement tables and thereby generate kappa coefficients that are hard to interpret. Alternatives to the kappa might be sought that are less influenced by such imbalances and so more easily interpreted (Brennan & Prediger, 1981; Munoz & Bangdiwala, 1997).

Finally, descriptive reliability statistics might have to suffice for small samples. Using the percentage of overall agreement, Byrt and colleagues propose for 2 × 2 tables a "prevalence-adjusted and bias-adjusted kappa" that is equal to 2 times the overall percentage of agreement, minus 1 (Byrt, Bishop, & Carlin, 1993). They also propose a bias index for such tables that provides insight comparable to a test for marginal homogeneity and a prevalence index that integrates information from percentages of positive and negative agreement. These descriptive statistics sometimes yield results that concur with the kappa coefficient (Figure 4a) and in other cases suggest different reliability results (Figure 4b). Use of such simple descriptive statistics is preferable to inferential methods in circumstances for which

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

Figure 4
Descriptive Reliability Statistics for Two Urine-Related Measures ($N = 20$)

a: Change in Odor

5	1
1	13

PABAK = 0.80 = [(2*(a + d)/N)–1][#]

Bias Index = 0.00 = (b – c)/N[#]

Prevalence Index = –0.40 = (a – d)/N[#]

Kappa = 0.76 (90% CI* 0.50, 1.00)

b: Change in Incontinence

0	2
1	17

PABAK = 0.70

Bias Index = 0.05

Prevalence Index = –0.85

Kappa = –0.07 (90% CI–0.16, 0.01)

#(Byrt, Bishop, and Carlin, 1993)Agreement tables are lettered consecutively by rows, from left to right and top to bottom.

*CI = Confidence Interval

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

the latter are not applicable. This point emphasizes that the goal of reliability testing, being the effective selection and application of measurement instruments, should be pursued by different methods as circumstances require.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687-699.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence, and kappa. *Journal of Clinical Epidemiology, 46*, 423-429.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43*, 551-558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cytel Statistical Software & Services. (2006). *StatXact user's guide, version 7*. Cambridge, MA: Author.

Darroch, J. N., & McCloud, P. I. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics, 28*, 371-388.

de Mast, J. (2007). Agreement and kappa-type indices. *American Statistician, 61*, 148-153.

Ely, E. W., Inouye, S. K., Bernard, G. R., Gordon, S., Francis, J., May, L., et al. (2001). Delirium in mechanically ventilated patients: Validity and reliability of the Confusion Asessment Method for the Intensive Care Unit (CAM-ICU). *Journal of the American Medical Association, 286*, 2703-2710.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*, 543-549.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver and Boyd.

Flack, V. F., Afifi, A. A., Lachenbruch, P. A., & Schouten, H. J. A. (1988). Sample size determinations for the two rater kappa statistic. *Psychometrika, 53*, 321-325.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York: Wiley.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613-619.

Graham, P., & Jackson, R. (2000). A comparison of primary and proxy respondent reports of habitual physical activity, using kappa statistics and log-linear models. *Journal of Epidemiology and Biostatistics, 5*, 255-265.

Gregson, J. M., Leathley, M. J., Moore, A. P., Smith, T. L., Sharma, A. K., & Watkins, C. L. (2000). Reliability of measurements of muscle tone and muscle power in stroke patients. *Age and Ageing, 29*, 223-228.

Hinze, J. L. (2004). *PASS 2005 user's guide*. Kaysville, UT: Number Cruncher Statistical Systems.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016

Lanza, S. T., Lemmon, D., Schafer, J. L., & Collins, L. M. (2007). *PROC LCA & PROC LTA User's Guide Version 1.1.3*. University Park: Methodology Center, Pennsylvania State University.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*, 153-157.

Munoz, S. R., & Bangdiwala, S. I. (1997). Interpretation of kappa and B statistics measures of agreement. *Journal of Applied Statistics, 24*, 105-111.

Nee, J. C. (1998). *EasyStat 3.04 guide*. New York: New York Psychiatric Institute.

SAS Institute. (2005). *SAS/STAT user's guide, version 9.1.3*. Cary, NC: Author.

Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use* (2nd ed.). New York: Oxford University Press.

Tanner, M. A., & Young, M. A. (1985a). Modeling agreement among raters. *Journal of the American Statistical Association, 80*, 175-180.

Tanner, M. A., & Young, M. A. (1985b). Modeling ordinal scale disagreement. *Psychological Bulletin, 98*, 408-415.

Thompson, W. D., & Walter, S. D. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology, 41*, 949-958.

Uebersax, J. S. (1992). Modeling approaches for the analysis of observer agreement. *Investigative Radiology, 27*, 738-743.

Uebersax, J. S., & Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine, 9*, 559-572.

van Belle, G. (2002). *Statistical rules of thumb*. Hoboken, NJ: Wiley.

Velema, J. P., Blettner, M., Restrepo, M., & Munoz, N. (1991). The evaluation of agreement by means of log-linear models: Proxy interviews on reproductive history among floriculture workers in Columbia. *Epidemiology, 2*, 107-115.

Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine, 17*, 101-110.

Wolinsky, F. D., Miller, D. K., Andresen, E. M., Malmstrom, T. K., & Miller, J. P. (2005). Reproducibility of physical performance and physiologic assessments. *Journal of Aging and Health, 17*, 111-124.

Downloaded from jah.sagepub.com at PENNSYLVANIA STATE UNIV on September 18, 2016