Journal of Behavioral Education, Vol. 10, No. 4, December 2000 (© 2001), pp. 205–212

Interobserver Agreement in Behavioral Research: Importance and Calculation

Marley W. Watkins, Ph.D.,[1,3] **and Miriam Pacheco, M.Ed.**[2]

Behavioral researchers have developed a sophisticated methodology to evaluate behavioral change which is dependent upon accurate measurement of behavior. Direct observation of behavior has traditionally been the mainstay of behavioral measurement. Consequently, researchers must attend to the psychometric properties, such as interobserver agreement, of observational measures to ensure reliable and valid measurement. Of the many indices of interobserver agreement, percentage of agreement is the most popular. Its use persists despite repeated admonitions and empirical evidence indicating that it is not the most psychometrically sound statistic to determine interobserver agreement due to its inability to take chance into account. Cohen's (1960) kappa has long been proposed as the more psychometrically sound statistic for assessing interobserver agreement. Kappa is described and computational methods are presented.

KEY WORDS: interobserver agreement; kappa; interrater reliability; observer agreement.

Behavioral research has historically placed great importance on the assessment of behavior and has developed a sophisticated idiographic methodology to evaluate behavioral change (Johnson & Pennypacker, 1993). Utilizing single case experimental designs, behavioral researchers attempt to determine the effect of an independent variable on a dependent variable while controlling threats to experimental validity (Gresham, 1998). Acknowledging and embracing the necessity

[1]Professor, Department of Educational and School Psychology, The Pennsylvania State University, University Park, PA.
[2]Recently completed her Master of Education degree in School Psychology at The Pennsylvania State University, University Park, PA.
[3]Correspondence should be directed to Marley W. Watkins, Department of Educational and School Psychology and Special Education, 227 CEDAR Building, The Pennsylvania State University, University Park, PA 16802; e-mail: mww10@psu.edu.

1053-0819/00/1200-0205$18.00/0

for a scientific approach, behavioral research rests on a foundation of accurate measurement (Cone, 1988).

Direct observation of behavior has traditionally been a mainstay of behavioral research (Foster, Bell-Dolan, & Burge, 1988). Although behavioral assessment using the observation of overt behavior was once considered bias free and inherently valid, it is now agreed that certain aspects of observational methodology can confound data and lead to invalid results (Hops, Davis, & Longoria, 1995). Wasik and Loven (1980) asserted that "the use of systematic behavioral observation methods, however, does not necessarily ensure reliable and valid information" (p. 211). They identified the human observer as perhaps the most easily biased variable in behavioral research and suggested that more attention be paid to the psychometric properties of observational measures. This caution has been reiterated by other behavioral researchers (Cone, 1988; Hoge, 1985; Hops, Davis, & Longoria, 1995).

INTEROBSERVER AGREEMENT

The cardinal psychometric properties which observational methods must address are reliability and validity (Hoge, 1985). As conceptualized in classical test theory, reliability and validity are central to adequate measurement (APA, AERA, & NCME, 1985). The most common method of assessing the reliability and validity of observational data is via interobserver agreement (Foster et al., 1988). The concepts of reliability and validity as applied to interobserver agreement have been interpreted somewhat differently within nomothetic and idiographic measurement paradigms. Some see interobserver agreement as a relatively simple form of reliability or consistency across observers (Hops et al., 1995) while others see it as capable of tapping aspects of both reliability and validity depending upon the ratio of random to systematic error, frame of interpretation, and magnitude of effects (Suen, 1988). All agree that interobserver agreement is the bedrock upon which sound behavioral measurement rests (Cone, 1977, 1988; Suen, 1988). McDermott (1988) insightfully articulated a consensus position when he wrote that "resolving agreement is not equivalent to discovering *truth*; rather, it affords professionals approaches to assessment that maintain enough *scientific integrity* [italics added] to serve pro temporee as best approximations to truth" (p. 239).

Assessing Interobserver Agreement

Numerous methods of assessing interobserver agreement have been developed, applied, and debated over the past two decades (Baer, 1977; Berk, 1979;

Hartmann, 1977; Landis & Koch, 1977; Langenbucher, Labouvie, & Morgenstern, 1996; Shrout, Spitzer, & Fleiss, 1987). Most prominent among these indices are percent agreement (including occurrence/nonoccurrence percent agreement) and kappa. Percent of agreement is defined as the number of agreements between observers in assigning cases or events to descriptive categories divided by the sum of both agreements and disagreements, and then multiplied by 100 to yield a percentage. Occurrence/nonoccurrence percent agreement are calculated based only upon occurrence or nonoccurrence of a category to partially correct for chance agreement. Unlike percent of agreement indices, kappa fully adjusts for chance in the calculation of observer agreement by subtracting chance from observed agreement.

Percent of Agreement

The most popular index for describing interobserver agreement is percent of agreement (Berk, 1979; McDermott, 1988). Its popularity is illustrated by a review of six recent issues of the *Journal of Behavioral Education* (Volume 7, Nos. 1–4 and Volume 8, Nos. 1–2) which found that 24 out of 27 articles (88%) employed percent of agreement to assess observational agreement.

Although percent agreement has benefits, mainly computational simplicity (Hops et al., 1995) and apparent ease of interpretation (Baer, 1977), it has many noteworthy faults. Primary is that percent of agreement values represent total agreement between observers without any consideration given to the operation of chance. As a result, there is no way of determining how an obtained percent of agreement compares to a level that could have resulted from random assignment alone. Thus, percent of agreement statistics tend to inflate the degree of perceived observer agreement, especially for frequently occurring behaviors, making it potentially misleading (Berk, 1979).

Figure 1 illustrates, with simulated data, how percent of agreement can result in an inflated estimate when the agreement between observers that might be due to chance is ignored. Suen and Lee (1985) confirmed this inflation with actual data. They obtained information from studies that used percent of agreement to report interobserver agreement. Out of the 50 randomly selected studies they analyzed, between one-fourth and three-fourths of the reported observations would have been judged as unreliable against a lenient kappa criterion (i.e., 0.60). Between one-half and three-fourths of the observations would have been judged unreliable against a more stringent kappa criterion (i.e., 0.75). From these results, Suen and Lee (1985) concluded that percent of agreement had seriously undermined the reliability of past observations and its application could no longer be justified.

This is not an isolated conclusion. Percent of agreement indices have long been considered weak measures of interobserver agreement (Dunn & Everitt, 1995; Fleiss, 1981; Foster et al., 1988; Hartmann, 1977; Hops et al., 1995; Langenbucher

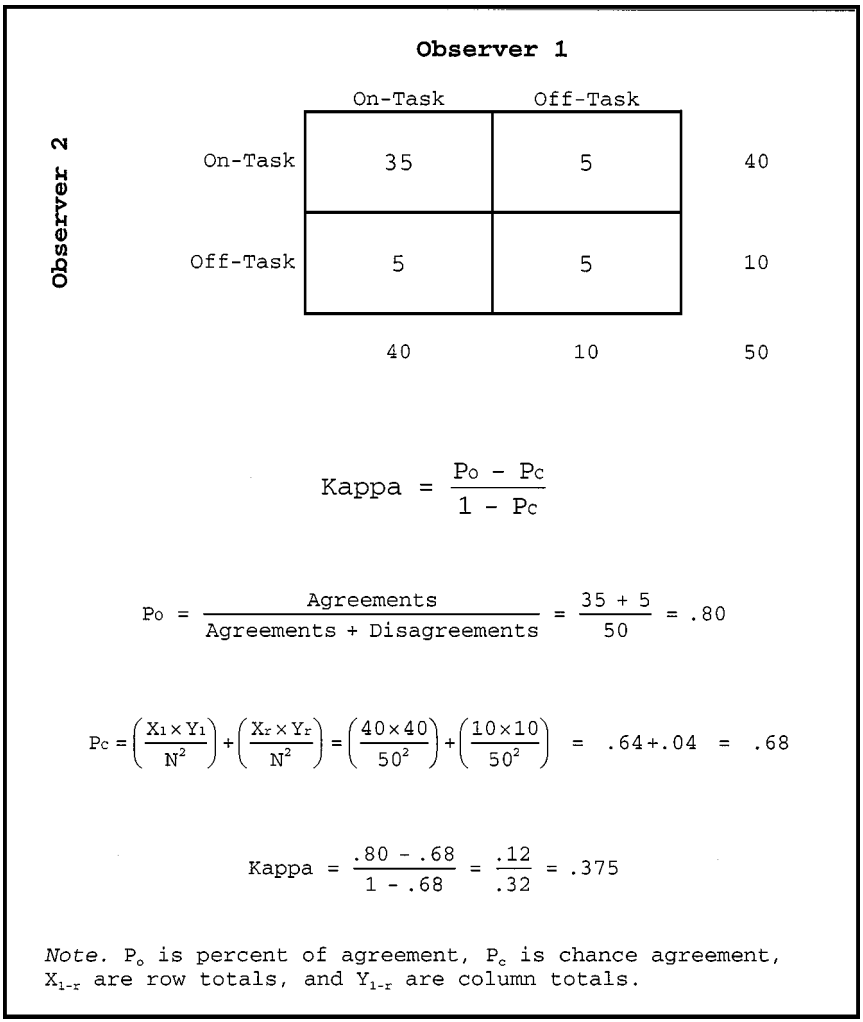

Fig. 1. Algorithm and sample calculation for Kappa for two observers who rate 50 cases into two mutually exclusive categories.

et al., 1996; McDermott, 1988; Shrout, et al., 1987). Berk (1979) described percent of observer agreement as "clearly inadequate for estimating reliability" (p. 461) and Everitt (1994) stated that "it is not an adequate index" (p. 27) of interobserver agreement. At best, percent of agreement might be helpful for teachers who need simple, easily calculated methods for use in the classroom (Wasik & Loven, 1980).

Kappa Coefficient of Agreement

Kappa (Cohen, 1960) was designed to rectify the major weaknesses of percent of agreement indices. Kappa's strength lies in its ability to take into account chance agreement between two or more observers. A second strength of kappa lies in its known sampling distribution. Kappa (k) was later expanded by Fleiss (1971) to allow generalization to situations using multiple observers who need not necessarily be the same throughout the study. Fleiss' k_m reduces to the original Cohen's k when the same two observers are used for all comparisons.

The kappa coefficient has become one of the preferred statistical methods for calculating the degree and significance of agreement between observers in their assignment of objects or subjects to nominal categories (McDermott, 1988). Kappa $= (P_o - P_c) \div (1 - P_c)$ where P_o is the proportion of observed agreement and P_c is the proportion of agreement expected by chance. Conceptually, coefficient kappa is the proportion of agreement over and above what would be expected by chance (Cohen, 1960). In other words, it is the proportion of the total amount of agreement not explained by chance for which the observers accounted. Therefore, kappa more accurately reflects, with less ambiguity, the reliability of the data.

Interpretation of kappa is simple and intuitive since it takes the form of the common correlation coefficient. That is, it ranges from -1.00 to $+1.00$. However, values less than zero (i.e., indicating observer agreement less than chance, hence disagreement) are often considered of no practical interest. Interpretative guidelines have been provided by Fleiss (1981) and Cicchetti (1994). Values of less than .40 are poor, values of .40 to .60 suggest fair agreement, values of .60 to .75 represent good agreement, and values greater than .75 indicate excellent agreement. Because of its ability to account for chance, a kappa coefficient of $+1.00$ can correctly be interpreted as perfect agreement between observers. Stated another way, the observers account for 100% of the remaining proportion of agreement not explained by chance. A kappa coefficient of zero indicates that the observers' ratings are no more precise than what could be attainable by random assignment. A kappa coefficient of .70 indicates that the observers account for 70% of the agreement over and above what would be expected by chance.

Another desirable characteristic of kappa is its comparability across experiments and conditions. Since kappa coefficients are corrected for chance, they can readily be compared to different experimental conditions even if the frequency of behavior changes across conditions (Ciminero, Calhoun, & Adams, 1986). This attribute is not possible with percent of agreement because of its differential inflation caused by disparate levels of chance for certain populations due to natural base rates or varying frequencies of behavior.

Although generally accepted as a superior interobserver agreement metric, kappa has limitations which may affect its usefulness. The most obvious limitation

is that kappa can only be computed with nominal scale data. However, this is generally only a theoretical problem since most interobserver comparisons involve nominal categorizations. Another potential limitation of kappa is that it is impossible to calculate when both observers record the target response as occurring at rates of either 0% or 100%. This is because chance agreement will then equal 100%, causing the denominator to resolve to zero (Ciminero et al., 1986). This extreme example illustrates an inherent theoretical difficulty with kappa: the highest possible value kappa can take for a given data set is determined by the marginal distributions (Cohen, 1960). Because chance is calculated by finding the joint probabilities of the marginals, kappa can only equal +1.00 if the marginals are identical (e.g., see Cohen, 1960). Although a limitation, it is more theoretical than actual because observers' agreement at rates of 0% or 100% can be practically resolved to represent "no" or "perfect agreement," respectively.

In addition, low base rates of a target response or classification within a given population may affect the computation of kappa (Shrout et al., 1987). To deal with this situation, Nelson and Cicchetti (1995) recommended that investigators ensure that there are at least ten cases in each cell to ensure maximal accuracy of the kappa agreement index. Nevertheless, after comparing agreement indices, Langenbucher et al. (1996) concluded that "*k* should be the default measure in most situations" (p. 1287).

Calculation of kappa. Probably the greatest practical limitation of kappa is its perceived computational complexity (Hops et al., 1995). This perception is not without merit in relation to percent of agreement indices. Therefore, this paper presents two methods to simplify the calculation of kappa. The first method is appropriate for the case of two observers and is easily computed by hand. An algorithm and sample calculation are provided in Figure 1.

The second method is more complex and therefore must be automated with a computer. It is based upon the Fleiss (1971) extension of kappa to the case of multiple observers, where the observers do not have to remain constant throughout the study. This computer program, entitled *MacKappa* (Watkins, 1998), calculates partial kappa coefficients to allow the investigator to verify agreement on a category by category basis as well as by the overall weighted average across categories. It also provides sampling distribution data to allow the researcher to ascertain the statistical significance of general and partial kappa coefficients.

MacKappa is a FutureBASIC program which operates on Macintosh computers under Mac OS versions 8 and 9. Data is input via a tab delimited text file. *MacKappa* will conduct analyses with 2-999 observers, 2-999 cases, and 2-25 categories. *MacKappa* is available from the first author upon receipt of an initialized Macintosh disk accompanied by appropriate first class return postage. Alternatively, *MacKappa* can be downloaded without charge from http://espse.ed.psu.edu/spsy/Watkins/SPSY-Watkins.ssi.

SUMMARY

Percent of agreement is the most common index of interobserver agreement, but it is seriously flawed due to its inability to account for chance and its noncomparability across categories and studies. Kappa has been demonstrated to be a superior index of interobserver agreement and should be the default measure in behavioral research. Computational complexity may have made it difficult for behavioral researchers to use. This is no longer a salient problem with the computational template and computer program presented in the current paper.

REFERENCES

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Baer, D. M. (1977). Reviewer's comment: Just because it's reliable doesn't mean that you can use it. *Journal of Applied Behavior Analysis, 10,* 117–119.

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency, 83,* 460–472.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290.

Ciminero, A. R., Calhoun, K. S., & Adams, H. E. (Eds.). (1986). *Handbook of behavioral assessment* (2nd ed.). New York: Wiley.

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

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy, 8,* 411–426.

Cone, J. D. (1988). Psychometric considerations and the multiple models of behavioral assessment. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (3rd Edition). NY: Pergamon.

Dunn, G., & Everitt, B. (1995). *Clinical biostatistics: An introduction to evidence-based medicine.* London: Edward Arnold.

Everitt, B. S. (1994). *Statistical methods for medical investigations* (2nd Edition). NY: Halsted Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378–382.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. NY: Wiley.

Foster, S. L., Bell-Dolan, D. J., & Burge, D. A. (1988). Behavioral observation. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (3rd Edition). NY: Pergamon.

Gresham, F. M. (1998). Designs for evaluating behavior change. In T. S. Watson & F. M. Gresham (Eds.), *Handbook of child behavior therapy.* NY: Plenum.

Hartmann, D. P. (1977, Spring). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, 10,* 103–116.

Hoge, R. D. (1985). The validity of direct observation measures of pupil classroom behavior. *Review of Educational Research, 55,* 469–483.

Hops, H., Davis, B., & Longoria, N. (1995). Methodological issues in direct observation: Illustrations with the living in familial environments (LIFE) coding system. *Journal of Clinical Child Psychology, 24,* 193–203.

Johnson, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of human behavioral research* (2nd Edition). Hillsdale, NJ: Erlbaum.

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

Langenbucher, J., Labouvie, E., & Morgenstern, J. (1996). Methodological developments: Measuring diagnostic agreement. *Journal of Consulting and Clinical Psychology*, *64*, 1285–1289.

McDermott, P. A. (1988). Agreement among diagnosticians or observers: Its importance and determination. *Professional School Psychology*, *3*, 225–240.

Nelson, L. D., & Cicchetti, D. V. (1995). Assessment of emotional functioning in brain-impaired individuals. *Psychological Assessment*, *7*, 404–413.

Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Comment: Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry*, *44*, 172–178.

Suen, H. K. (1988). Agreement, reliability, accuracy, and validity: Toward a clarification. *Behavioral Assessment*, *10*, 343–366.

Suen, H. K., & Lee, P. S. (1985). Effects of the use of percentage agreement on behavioral observation reliabilities: A reassessment. *Journal of Psychopathology and Behavioral Assessment*, *7*, 221–234.

Wasik, B. H., & Loven, M. D. (1980). Classroom observational data: Sources of inaccuracy and proposed solutions. *Behavioral Assessment*, *2*, 211–227.

Watkins, M. W. (1988). *MacKappa* [Computer software]. Pennsylvania State University: Author.