THE ESTIMATION OF INTEROBSERVER AGREEMENT IN BEHAVIORAL ASSESSMENT

April A. Bryington, Darcy J. Palmer, and Marley W. Watkins
The Pennsylvania State University

Direct observation of behavior has traditionally been a core component of behavioral assessment. However, systematic observational data is not intrinsically reliable and valid. It is well known that observer accuracy and consistency can be influenced by a variety of factors. Therefore, interobserver agreement is frequently used to quantify the psychometric quality of behavioral observations. Two of the commonly used interobserver agreement indices, percentage of agreement and kappa, are reviewed. Although percentage agreement is popular due to its computational simplicity, kappa has been found to be a superior measure because it corrects for chance agreement among observers and allows for multiple observers and categories. A description of kappa and computational methods are presented.

Direct observation of behavior has traditionally been a core component of behavioral assessment (Ciminero, 1986; Tryon, 1998). Originally, it was thought unnecessary to establish the reliability and validity of direct observations of behavior since by definition direct observation is free of bias and valid. However, various aspects of methodology can confound the data and therefore lead to invalid results (Hops, Davis, & Longoria, 1995).

Kazdin (1977) reviewed research that demonstrated that observer accuracy and reliability can be influenced by variables such as knowledge that accuracy is being checked, drift from original definitions of the observed behavior, the complexity of the coding system being used, and observer expectancies combined with feedback. In addition, Wasik and Loven (1980) reported that characteristics of the recording procedures, characteristics of the observer, and characteristics unique to the observation setting are sources of inaccuracy that can jeopardize the reliability and validity of observational data. Consequently, Cone (1998) suggested that the quality of any observations of behavior must be determined regardless of the procedures used to quantify them.

INTEROBSERVER AGREEMENT

Researchers have identified procedures that can be used to measure the psychometric properties of data obtained from direct observation (Primavera, Allison, & Alfonso, 1997). The most common of these procedures is interobserver agreement (Skinner, Dittmer, & Howell, 2000). There are diverse opinions of what interobserver agreement actually measures. Hops et al. (1995) defined interobserver agreement as a measure of consistency and, therefore, as representing a form of reliability. In contrast, Alessi (1988) described interobserver agreement as an estimate of objectivity that indicates the degree to which the data reflect the behavior being observed rather than the behavior of the observer. Alessi’s definition implies that interobserver agreement is tapping into aspects of validity. Suen (1988, 1990) indicated that interobserver agreement could serve as a measure of both reliability and validity depending upon the degree to which two or more observers agree on occurrences or nonoccurrences, whether a criterion-referenced or norm-referenced orientation is used, and the ratio of random to systematic error. Although there are divergent views about what agreement actually measures, it is generally accepted that it is fundamental to sound behavioral measurement for both researchers and practitioners (Bloom, Fischer, & Orme, 1999; Hayes, Barlow, & Nelson-Gray, 1999; Hoge, 1985; Hops et al., 1995; Kazdin, 2001; Kratochwill, Sheridan, Carlson, & Lasecki, 1999; Maag, 1999; McDermott, 1988; Salvia & Ysseldyke, 2001; Suen, 1988).

Assessing Interobserver Agreement

Many different methods of calculating interobserver agreement have been proposed (Berk, 1979; Hartmann, 1977; House, House, & Campbell, 1981; Shrout, Spitzer, & Fleiss, 1987). The two most commonly cited methods are percent of agreement and kappa.

Overall Percent of Agreement

The most frequently used method for determining interobserver agreement is overall percent of agreement (Berk, 1979; Hartmann, 1977; McDermott, 1988). Percent of agreement is calculated by the benefits of using overall percent of agreement include its ease of calculation and interpretation (Hartmann). The disadvantages of percent of agreement, however, have caused many researchers to caution against its use (Berk; Birkimer & Brown, 1979; Hartmann; Hops et al., 1995; McDermott; Shrout et al., 1987; Suen & Lee, 1985; Towstopiat, 1984).

The most significant problem with percent of agreement is its failure to take into account agreement that may be due to chance (House et al., 1981). As McDermott (1988) pointed out, when using percent of agreement “there exists no means of determining whether obtained agreement is effectively beyond what might be produced by completely naive observers or by the toss of dice” (p. 229). Not only does percent of agreement fail to control for chance, it is also influenced by the frequency of behaviors being observed. A researcher may obtain a level of percentage agreement that he or she feels is adequate, when in reality, it may be inflated due to chance or the high frequency of the behavior being observed (Towstopiat, 1984). Figure 1 illustrates this potential inflation with data from House, Farber, and Nier (1983).

Suen and Lee (1985) provided empirical evidence that disregarding chance can lead to inflated levels of agreement. They randomly selected 12 studies that reported percentage agreement. From these studies, they chose a simple random sample of 50 observation points and found that between one-fourth and three-fourths of the observations would have been determined to be unreliable against a lenient chance-corrected criterion. Between one-half and three-fourths of the observations would have been judged unreliable against a more stringent chance-corrected criterion. Suen and Lee concluded that percent of agreement has seriously undermined the reliability of past observations and that “its continued use can no longer be justified” (p. 232).

Occurrence and Nonoccurrence Percent of Agreement

The failure of overall percent of agreement to take chance into account can be partially corrected by using percent of agreement only on the occurrence or nonoccurrence of the target behavior rather than the overall level of agreement. If the occurrence of the target behavior is the focus of interest then percent of agreement on occurrence of the target behavior may be appropriate. Conversely, if agreement on nonoccurrence is most important then percent of agreement on nonoccurrence of the target behavior can be used. These indices indicate the percentage of time in which two or more observers agree that a target behavior either occurred or did not occur.

The benefits of percent agreement on occurrence or nonoccurrence are simplicity of calculation and partial resistance to the distorting effects of chance. However, they do not completely control for chance (Hopkins & Herman, 1977) and they can potentially produce incongruent indices of agreement. Like overall percent of agreement, percent agreement on occurrence or nonoccurrence is only applicable when two observers are monitoring a dichotomous target behavior (Primavera et al., 1997).

Kappa Coefficient of Agreement

Kappa (*k*; Cohen, 1960) has become the preferred index for measuring interobserver agreement (Hops et al., 1995). For example, Primavera et al. (1997) highly recommended kappa “when data are dichotomous or nominal” (p. 64) while Langenbucher, Labouvie, and Morgenstern (1996) suggested that kappa “should be the default measure” (p. 1287) when assessing diagnostic agreement in psychiatry. Kappa has also been favored for determining observer agreement in medicine (Everitt, 1994).

Strengths of kappa

One of kappa’s strengths is its ability to correct for chance agreement across two or more nominal categories. Another is its known sampling distribution that allows for the construction of confidence intervals and tests of statistical significance (Cohen, 1960). An original limitation of kappa was that it could only be used with two observers and the same two observers had to rate every observation. This was corrected by Fleiss (1971) who extended kappa to be used in situations in which there are a constant number of raters, but the raters do not necessarily have to be the same across observations. Fleiss's $k_{m,}$ (the subscript $_{m}$ signifying *k* for multiple observers) automatically reduces to *k* when there are only two observers for all observations.

Another beneficial characteristic of kappa is that it allows for generalizability across different experimental conditions. Foster and Cone (1986) pointed out that chance agreement changes as the base rate or prevalence of behavior changes. Because percent of agreement does not correct for chance, it is differentially inflated in situations with different rates of behavior, hindering comparison across conditions. Kappa, however, allows for standardized comparisons by statistically removing chance.

Limitations of kappa

Although kappa's benefits have caused many to suggest that it is the most desirable index to use when calculating interobserver agreement, it also has several limitations that should be considered. One constraint of kappa is that it can only be used with nominal scale data. Because most interobserver comparisons involve nominal categorization, this is generally not a problem. A second possible limitation is that kappa is impossible to calculate when both observers report that the behavior occurred 100% of the time or not at all. When this occurs, chance agreement will equal 100% and the denominator of the kappa equation will resolve to zero (Foster & Cone, 1986). However, this is more of a theoretical problem than a practical one. If observers agree 100% of the time, it can be seen as perfect agreement.

Another possible limitation of the kappa coefficient is that it tends to decrease when there are low base rates of the observed behavior (Shrout et al., 1987). To alleviate this problem, Nelson and Cicchetti (1995) suggested that researchers ensure that there are at least ten occurrences of the behavior in the sample being observed. This will minimize the effect of interobserver disagreement in cases of low frequency behaviors. Similarly, the magnitude of kappa can be influenced by the relative balance of agreements and disagreements. However, Cicchetti and Feinstein (1990) pointed out that this tendency serves a legitimate scientific purpose.

Interpretation of kappa

Kappa indicates the proportion of agreement above and beyond what would be expected by chance (Cohen, 1960) and takes the form of a simple correlation coefficient that is relatively easy to interpret. Possible values range from +1.00, which indicates perfect agreement, through 0.00, which reflects chance agreement, down to a theoretical -1.00, which signifies perfect disagreement. Values less than zero are usually of no practical interest because they represent agreement that is less than would be expected by chance (Cohen). Because kappa adjusts for chance agreement, less stringent guidelines are generally applied than those used in simple percent of agreement. Cicchetti (1994) provided a summary of interpretive guidelines for kappa. Specifically, values below 0.40 indicate poor clinical significance; values between 0.40 and 0.59 indicate fair clinical significance; values between 0.60 and 0.74 indicate good clinical significance; and values between 0.75 and 1.00 indicate excellent clinical significance. Because kappa accounts for chance, a coefficient of +1.00 can be interpreted correctly as indicating perfect agreement between observers. In this case, the observers would have accounted for 100% of the agreement that was not explained by chance. If a coefficient of zero is obtained, it indicates that the observers' ratings are no more precise than what could be attained by random assignment. A kappa coefficient of 0.80 indicates that the observers have accounted for 80% of the agreement over and above what would be expected by chance.

Calculation of kappa

Conceptually, kappa is defined as: The greatest deterrent to the use of kappa may be its perceived difficulty of computation when compared to simple percent agreement (Hops et al., 1995). 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. A REALbasic computer program, entitled *Chi-Square Analysis* (Watkins, 2002), is also available for the case of two observers. Both Macintosh and Windows versions can be downloaded without charge from http://espse.ed.psu.edu/spsy/Watkins/SPSY-Watkins.ssi.

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 that operates on Macintosh computers under Mac OS 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* can be downloaded without charge from http://espse.ed.psu.edu/spsy/Watkins/SPSY-Watkins.ssi.

SUMMARY

The calculation of interobserver agreement is essential for establishing the psychometric properties of observational data. Although percentage agreement is the most commonly used agreement index, its limitations have led researchers to recommend kappa as a more desirable index of interobserver agreement. Difficult computation may have deterred its common use in the past; however, this is no longer a salient problem with the computational guide and computer programs presented in the current paper.

REFERENCES

Alessi, G. (1988). Direct observation methods for emotional/behavioral problems. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Conceptual foundations and practical applications* (pp. 14-75). New York: Guilford Press.

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

Birkimer, J. C., & Brown, J. H. (1979). Back to basics: Percentage agreement measures are adequate, but there are easier ways. *Journal of Applied Behavior Analysis, 12,* 535-543.

Bloom, M., Fischer, J., & Orme, J. G. (1999). *Evaluating practice: Guidelines for the accountable professional* (3rd ed.). Boston: Allyn and Bacon.

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

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

Ciminero, A. R. (1986). Behavioral assessment: An overview. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (2nd ed., pp. 3-11). New York: John Wiley & Sons.

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

Cone, J. D. (1998). Psychometric considerations: Concepts, contents, and methods. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (4th ed., pp. 22-46). Boston: Allyn & Bacon.

Everitt, B. S. (1994). *Statistical methods in medical investigations* (2nd ed.). London: Edward Arnold.

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

Foster, S. L., & Cone, J.D. (1986). Design and use of direct observation procedures. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (2nd ed., pp. 253-324). New York: John Wiley & Sons.

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

Hayes, S. C., Barlow, D. H., & Nelson-Gray, R. O. (1999). *The scientist-practitioner: Research and accountability in the age of managed care.* Boston: Allyn and Bacon.

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

Hopkins, B. L., & Hermann, J. A. (1977). Evaluating interobserver reliability of interval data. *Journal of Applied Behavior Analysis, 10,* 121-126.

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.

House, A. E., Farber, J. W., & Nier, L. L. (1983). Differences in computational accuracy and speed of calculation between three measures of interobserver agreement. *Child Study Journal, 13*, 195-201.

House, A. E., House, B. J., & Campbell, M. B. (1981). Measures of interobserver agreement: Calculation formulas and distribution effects. *Journal of Behavioral Assessment, 3*, 37-57.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABC's of reliability. *Journal of Applied Behavior Analysis, 10*, 141-150.

Kazdin, A. E. (2001). *Behavior modification in applied settings* (6th ed.). Belmont, CA: Wadsworth/Thomson Learning.

Kratochwill, T. R., Sheridan, S. M., Carlson, J., & Lasecki, K. L. (1999). Advances in behavioral assessment. In C. R. Reynolds & T. R. Gutkin (Eds.), *The handbook of school psychology* (3rd ed.) (pp. 350-382). New York: Wiley.

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

Maag, J. W. (1999). *Behavior management: From theoretical implications to practical applications*. San Diego, CA: Singular Publishing.

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.

Primavera, L. H., Allison, D. B., & Alfonso, V. C. (1997). Measurement of dependent variables. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 41-91). Mahwah, NJ: Erlbaum.

Salvia, J., & Ysseldyke, J. E. (2001). *Assessment* (8th ed.). Boston: Houghton Mifflin.

Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry, 44*, 172-177.

Skinner, C. H., Dittmer, K. I., & Howell, L. A. (2000). Direct observation in school settings: Theoretical issues. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Theory, research, and clinical foundations* (2nd ed., pp. 19-45). New York: Guilford Press.

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

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Erlbaum.

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

Towstopiat, O. (1984). A review of reliability procedures for measuring observer agreement. *Contemporary Educational Psychology, 9*, 333-352.

Tryon, W. W. (1998). Behavioral observation. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (4th ed., pp. 79-103). Boston: Allyn & Bacon.

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.

Watkins, M. W. (2002). *Chi-Square Analysis* [Computer software]. Pennsylvania State University: Author.

		Observer 1		
		Positive	Negative	
Observer 2	Positive	2	10	12
	Negative	3	105	108
		5	115	120

$$\text{Kappa} = \frac{P_o - P_c}{1 - P_c}$$

$$P_o = \frac{\text{Agreements}}{\text{Agreements} + \text{Disagreements}} = \frac{2 + 105}{120} = .892$$

$$P_c = \left(\frac{X_1 \times Y_1}{N^2}\right) + \left(\frac{X_r \times Y_r}{N^2}\right) = \left(\frac{12 \times 5}{120^2}\right) + \left(\frac{108 \times 115}{120^2}\right) = .004 + .863 = .86$$

$$\text{Kappa} = \frac{.892 - .867}{1 - .867} = \frac{.025}{.133} = .188$$

Note. P_o is percent of agreement, P_c is chance agreement, X_{1-r} are row totals, and Y_{1-r} are column totals.

Figure 1. Algorithm and sample calculation for Kappa for two observers who rate 120 cases into two mutually exclusive categories.