Interpreting Kappa in Observational Research: Baserate Matters

Cornelia Taylor Bruckner
Sonoma State University

Paul Yoder
Vanderbilt University

Abstract

Kappa (Cohen, 1960) is a popular agreement statistic used to estimate the accuracy of observers. The response of kappa to differing baserates was examined and methods for estimating the accuracy of observers presented. Results suggest that setting a single value of kappa as "minimally acceptable" (i.e., a criterion value) is not useful in ensuring adequate accuracy of observers. Instead, researchers should use the best estimate of the true baserate of the target behavior *and* the obtained kappa to estimate the accuracy of observers. Investigators can then compare the estimated accuracy of observers to a preselected criterion level. Guidelines are given for selecting a criterion accuracy level.

For investigators who code behavior from direct observation of participants, agreement between two independent observers on the occurrence and nonoccurrence of events is a critical first step to determining the precision of a behavioral definition and the accuracy of observers (Bakeman, 2000). If observers do not agree on events, training and behavioral-definition refinement is implemented to improve agreement. If observers continue to disagree, after trying to remediate disagreement, investigators often drop the contested code from their coding system or change the definition of the code repeatedly. In some cases, the code might be operationalized to the point of losing construct "validity." Retraining observers and dropping codes highlight the importance of statistics used to make decisions about the sufficiency of agreement.

In the social sciences, many researchers use the coefficient kappa (Cohen, 1960) to determine whether observer agreement is sufficient. In addition to the accuracy of observers, the magnitude of kappa is affected by the baserate of the target behavior. Therefore, the baserate should be considered when judging the sufficiency of kappa. In the current review, we describe the relation between baserate and kappa. The primary contribution of the paper is to provide a table and rationale for interpreting the adequacy of obtained kappas in the context of the estimated baserate of the target behavior.

Agreement and Accuracy

To determine functional relations between events, it is important to measure events accurately. If an observer is perfectly accurate, his or her data will coincide exactly with the true events of the coded sessions. For example, if we replaced one of our fallible observers with an infallible observer, we would know an observer's agreement with the true state of affairs, namely, their accuracy (Bakeman, Quera, McArthur, & Robinson, 1997). The fallible observer's agreement with the true occurrence of events is their sensitivity. The fallible observer's agreement with the true nonoccurrence of events is their specificity. Sensitivity and specificity, or accuracy, are ways to describe fallible observer's accuracy when an infallible record of events is available.

© American Association on Mental Retardation

Unfortunately, we cannot know the accuracy of an observer because we do not have an infallible record of the events of a coded session. Instead, we use agreement between two fallible observers to estimate accuracy. Point-by-point interobserver agreement is achieved when independent observers record the same event at the same time. If low point-by-point interobserver agreement occurs, we can assume that at least one of our observers is not accurately applying the coding system. If point-by-point interobserver agreement is high, we often assume that the two observers are observing and recording events accurately. This latter interpretation is tenuous because of rival hypotheses for high agreement, including systematic error and chance agreement (Fleiss, 1981). In the next section we describe a method for estimating agreement excluding chance agreement and, thereby, improving estimates of accuracy from point-by-point interobserver agreement.

Kappa and Chance Agreement

In his original paper on kappa, Cohen (1960) reflected, "it takes little in the way of sophistication" (p. 38) to appreciate the inadequacy of the primitive solution to quantifying observer agreement by looking at the proportion of the total events upon which observers agreed. The main flaw he identified with this "primitive" approach was the failure to separate the amount of agreement due to chance from the total amount of agreement. Cohen attempted to correct this flaw by adjusting observed agreement through subtracting chance agreement. The result was the coefficient kappa. *Kappa* is the proportion of nonchance agreement observed out of the total nonchance agreement available. The model used to calculate chance agreement in kappa is the joint probability of two independent events. This is the product of each observers' estimate of the baserates for each event in the coding system. For example, in a timed event system with seconds as the unit of measurement, if Observer 1 recorded an event in 79 out of 100 s and Observer 2 recorded it in 85 out of 100 s, the chance of both observers recording the event for any one second is the product of .79 and .85, or .67.

Kappa is composed of two estimated agreement quantities: observed agreement (P_o) and chance agreement (P_e). P_o, or *observed agreement,* is the proportion of the total events upon which observers agreed. P_o is what Cohen (1960) called the "primitive approach." P_e, or *chance agreement,* is the proportion of events on which agreement is expected by chance or the joint probability of two observers independently recording any event in the set. Another way to think about the model of chance used in kappa is that two observers randomly assign codes to seconds according to prior knowledge of the baserate of the behavior. Each coder then independently assigns the behavior to seconds without observing the behavior or knowing the coding definition. In this model, if two observers happen to agree about the occurrence of a behavior during a particular second, it is only due to the chance that they both randomly recorded the behavior during that second. We use the proportion of seconds in which each observer recorded the behavior, or the *baserate,* to estimate this random process. The chance of the two observers agreeing on the occurrence of a behavior in a particular second is the proportion of seconds that Observer A recorded the behavior multiplied by the proportion of seconds that Observer B recorded the behavior. Kappa is calculated using the equation $K = (P_{ov} - P_e)/(1 - P_e)$.

Other assumptions implicit in the calculation of kappa are that (a) categories are nominal, (b) data are coded in a mutually exclusive and exhaustive manner, and (c) analysis units are independent. The assumption that the categories are nominal means that each event is categorized as being the same or different from another, with no category being ordered as greater or less than another. Events are mutually exclusive if the occurrence of Event A precludes the occurrence of any other events in the set. A dataset is exhaustive if every unit that can be coded is included in the kappa table, for example, in a timed event system, every second is categorized (Agresti, 1996; Bakeman, 2000). Two events can be considered independent if the occurrence of Event A does not influence the probability of Event B (Agresti, 1996). When analysis units are not independent, there will be an inflation of P_o (observed agreement) . Researchers should be conservative when interpreting values of kappa obtained from nonindependent analysis units.

Importance of Interpreting Kappa in Two × Two Tables

The following discussion of kappa will be limited to the case of a two × two matrix. Kappa is an omnibus measure of agreement that offers a

 © American Association on Mental Retardation

single statistical summary of agreement across all categories in a code set (Cicchetti & Feinstein, 1990). For this reason, category by category agreement is required to estimate agreement for each category, retrain observers, and refine code sets (Kraemer, Periyakoil, & Noda, 2004). Judging the accuracy or validity of a single category from an omnibus multicategory kappa is similar to judging the significance of the difference between two means based on a significant *F* test including more than two means. A significant *F* test must be followed up with pairwise comparisons of means to make conclusions about the difference between two specific means. In the same way a two × two table comparing the category of interest to the collapsed sum of the other categories should be used to determine agreement on a specific category.

Any kappa table can be collapsed into a two × two table. The two observers' agreement on the target category's occurrence is represented in the upper left cell. Agreement on the target category's nonoccurrence is represented in the lower right cell. Note that the particular category that observers record can be different and still be tallied in the lower right cell as long as neither observer records the target category. This occurs because in the collapsed two × two tables, the bottom right corner cell is defined as Observer 1's other category and Observer 2's other category. Disagreements regarding the target category are represented on the upper right and lower left cells, depending on which coder recorded the target category.

Baserate Influences the Magnitude of Kappa

Effect of Unequal Baserates

When events in a mutually exclusive and exhaustive set have highly unequal simple probabilities (*baserates*), values of kappa will be lower than when baserates are equal, even when observers are highly accurate. For example, if both observers have an accuracy of 90% and the baserates of the target and other events are .1 and .9, respectively, the obtained kappa is .39. With that same accuracy, but equal baserates (i.e., 50–50), the obtained kappa is .64. Figure 1 shows variation in obtained kappas across different baserate and accuracy levels. Readers should note that as accuracy increases, kappa increases when baserate is held constant. Also, kappa increases as baserate approaches .5,

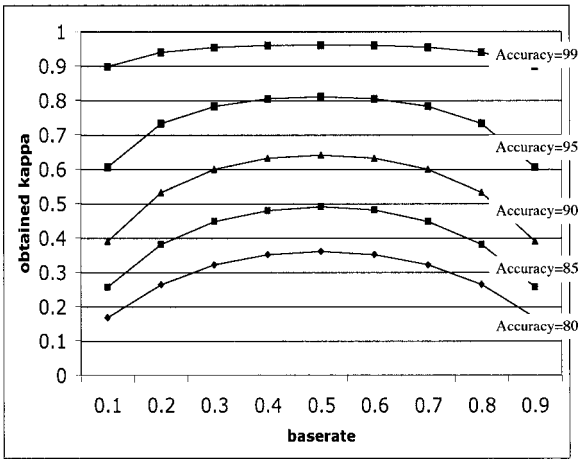

Figure 1. The variation of kappa across baserates and accuracy. The best estimate of the baserate is represented on the horizontal axis. The value of kappa is represented on the vertical axis. The lines are differentiated by the observer accuracy that they represent, with the bottom line representing accuracy of .80 and the top line representing accuracy of .99.

when accuracy is held constant. (For a three-dimensional expansion of Figure 1, see Spitznagel and Helzer, 1985.) The math used to produce these examples is discussed later in this paper. At this point, readers are asked to accept the math for the sake of making the point that baserate of the target behaviors influences the kappa even when accuracy with the true state of affairs remains constant. In our example, kappas of .39 and .64 represent the same amount of observer accuracy but different distribution of events across categories. If we use a criterion value for minimally acceptable kappa (e.g., .60), we would conclude that a kappa of .39 is inadequate and the kappa of .64 is adequate. This would be incorrect because the observed difference in the magnitude of kappa is due to the baserates of the events and not the accuracy of observers.

Kraemer (1979) derived the mathematical relationship between baserate, observer accuracy, and kappa. This phenomenon has been reported in several different fields (see Grove, Andreasen, McDonald-Scott, Keller, & Shapiro, 1981, for an applied review). The danger of setting criterion values for kappa has been repeated by biostatisticians (Feinstein & Cicchetti, 1990), psychiatrists (Grove et al., 1981), psychometricians (Brennan & Prediger, 1981), and psychologists (Bakeman et al., 1997; Uebersax, n.d.). However, it is rarely heeded

© American Association on Mental Retardation

(e.g., Horner et al., 2005). There are two possible solutions: (a) use a different statistic to quantify agreement that is not affected by baserate or (b) consider the baserate and the obtained kappa to estimate accuracy, which is then compared to a criterion accuracy value.

Alternative Indices to Kappa

Three statistics with reduced sensitivity to baserate have been evaluated to determine their utility as a replacement for kappa. These include the odds ratio, Yule's Y, and the ϕ coefficient (Cicchetti & Feinstein, 1990; Maxwell, 1977; Spitznagel & Helzer, 1985). These coefficients can be divided into two classes: marginal independent/nonchance corrected and marginally dependent/chance corrected (Cicchetti & Feinstein, 1990; Maxwell, 1977). The first class is described as marginally independent because only the cell values of the agreement table are used in calculation. The second class is described as marginally dependent because the totals of the rows and columns (i.e., marginals) of the agreement table are used in calculation. Because chance is often modeled using the product of the marginals, coefficients that include these marginal values are described as having overt corrections for chance.

The odds ratio, described as a measure of agreement by Cicchetti and Feinstein (1990) and Yule's Y, described by Spitznagel and Helzer (1985), are marginally independent omnibus measures of association that do not have overt corrections for chance but should take on values of no association when agreement is due only to chance. The odds ratio can range from 0 to ∞, with values of 1 indicating no association (Agresti, 1996). A problem with scaling occurs because a null value of 1 translates poorly to audiences accustomed to interpreting values of kappa or agreement proportions where a value of 1 indicates perfect agreement. Also, the odds ratio cannot be calculated when either off-diagonal cell is 0 and may take on inappropriately low values when either of the diagonal cells is 0 (Cicchetti & Feinstein, 1990). Yule's Y can range from −1.0 to 1.0, with a value of 0 indicating no association (Yule, 1912). The scale of Yule's Y is difficult to interpret because the numerator and denominator are square-root transformations that have no intuitive meaning (Cicchetti & Feinstein, 1990). It turns out that Yule's Y actually *is* dependent on baserate when the baserate is greater than 50% (Spitznagel & Helzer, 1985). Both indices are designed to measure association between two different variables. That is, they were not designed to compute agreement between two observers on the same variable. Therefore, the meaning of these statistics as agreement indices is not transparent. Problems with scaling, calculation, interpretation, and some dependency on baserates make supposedly marginally independent indices an unlikely replacement for kappa.

The ϕ coefficient, described as a measure of observer agreement by Cohen (1960) and Maxwell (1977), is the correlation between two dichotomous variables (Agresti, 1996). The ϕ coefficient is marginally dependent with an overt correction for chance (Maxwell, 1977). The ϕ coefficient has been shown to be identical to kappa in some cases and, when different, is only slightly larger than kappa (Cicchetti & Feinstein, 1990). Also, the ϕ coefficient is smaller than kappa, when baserates are unequal and cannot reach perfect correlation values of 1.0, or −1.0, when marginals are not proportional (Cohen, 1960). As with the marginally independent indices, the ϕ coefficient was designed as a measure of correlation between two different variables, and its meaning as a measure of agreement between two observers on one variable is not well-understood. Because the ϕ coefficient is often equivalent to kappa, difficult to interpret, and sensitive to baserates, it is unlikely to be used as a replacement for kappa.

Considering Baserate When Selecting a Criterion Kappa

Because no replacement for kappa has been shown to be sufficient, we used existing statistical theory to develop an easy method for simultaneously considering baserate and obtained kappa to estimate accuracy. We say "estimate" accuracy because the baserate and kappa are obtained from a sample of time and raters and may be different than the true values. This solution to judging the sufficiency of an obtained kappa is innovative and practical. It is innovative because it suggests that we use the obtained baserates and kappa to estimate accuracy of the observers. No other proposed methods have attempted to provide what investigators really want to know: the accuracy of their observers. Direct estimates of accuracy have never been used to judge agreement statistics, although folk knowledge of agreement is that it estimates accuracy. This solution is practical be-

 © American Association on Mental Retardation

cause it provides a method for estimating the accuracy of observers from information we already obtain when we compute point-by-point agreement from a mutually exclusive and exhaustive coding of an observation session: an obtained kappa and the two observers' estimates of the baserate of the target behavior. This will allow practitioners to apply the information from complicated probability calculations to their data using a handheld calculator and the provided table.

Using probability theory, investigators can calculate the accuracy of observers from the estimated baserate of the behavior and the obtained kappa, assuming that accuracy of both coders are equal, and sensitivity and specificity are equal. Bakeman et al. (1997) provided an excellent discussion of the statistical theory behind the calculation of cells in an observer agreement table. For the readers' convenience, we have included a summary of the theory presented by Bakeman et al. in the appendix of this article. Using the formula for the expected value of kappa, we created a table depicting the relationship between kappa, observer accuracy, and baserate of the target category in a two × two table when sensitivity and specificity are equal and raters are equally fallible (Table 1). To calculate obtained values of kappa for I × J tables and/or observers with different sensitivity and specificity, we suggest use of the program FalliObs (Bakeman et al., 1997), which is available from www.gsu.edu/psychology/bakeman.

Using Table 1

Observer accuracy can be estimated from Table 1. To estimate accuracy the user must have two pieces of information: (a) the obtained kappa and (b) the average of the two observers' estimates of the baserate. Both of these pieces of information can be calculated from the two × two agreement table described earlier in the paper. The obtained kappa can be calculated by hand or using available software (Robinson & Bakeman, 1998; Tapp & Walden, 1993). The average of the two observers' estimates of the baserate is the proportion of seconds that Observer 1 recorded the behavior plus the proportion of seconds that Observer 2 recorded the behavior divided by 2. This calculation provides a better estimate of the baserate than either observer's estimate alone. However, the accuracy of this estimate increases as the number of raters or observation sessions increase in length. The user first locates the estimated baserate of the behavior in the far left column of Table 1. There is a row of obtained kappas associated with that value of the baserate. The user then moves across the row of obtained kappas to locate the obtained kappa associated with the estimated baserate. Then the user moves up the column containing their obtained kappa to determine the accuracy represented by that column. The accuracy represented by that column is the estimated accuracy for that obtained kappa and baserate. If the obtained kappa or estimated baserate is not present in the table, then estimated accuracy can be computed using interpolation. For example, if the baserate of the behavior was .7 and the obtained kappa was .65, we would estimate the accuracy as greater than .9 and less than .95.

Assumptions of Table 1

Table 1 is limited to two × two tables and assumes that sensitivity and specificity are equal

Table 1. Expected Values of Kappa Given the Best Estimate of Baserate and Accuracy of Observers

Best est. of true baserate	Accuracy .8	Accuracy .85	Accuracy .9	Accuracy .95	Accuracy .99
.1	.168	.257	.39	.605	.897
.2	.264	.381	.532	.732	.939
.3	.321	.447	.599	.782	.953
.4	.351	.479	.631	.804	.959
.5	.36	.49	.64	.81	.96
.6	.351	.48	.631	.804	.959
.7	.321	.447	.599	.782	.953
.8	.265	.381	.532	.732	.939
.9	.168	.257	.39	.605	.897

© American Association on Mental Retardation

Table 2. Two by Two Table Representing Observer Agreement on Engagement for Participant 1

		Observer 2		
		Seconds engaged	Seconds not engaged or other	Sum of rows
Observer 1	Seconds engaged	588	36	624
	Seconds not engaged or other	76	359	435
	Sum of columns	664	395	1059

and that observers are equally fallible. Simulations have shown that when these assumptions are reasonably met, Table 1 is a good estimator of actual observer accuracy (Bakeman et al., 1997). An applied example of using the baserate by accuracy table is presented next to aid readers in applying the logic.

Applied Example of Interpolating Observer Accuracy From Kappa and Baserate

Two examples of agreement matrices are provided in Tables 2 and 3. In each, two independent observers used a timed event-coding system, with seconds as the unit of measurement, to indicate whether children with autism spectrum disorders were engaged, unengaged, or other during a 20-min parent–child interaction session. To aid in agreement estimation for engagement, we collapsed the categories of unengaged and other and calculated kappa for a two × two table with agreement or disagreement by two observers being classified into the engaged or the collapsed category (Table 2). The best estimate of baserate of engagement for Participant 1 was .61 (624 + 664)/(2 × 1059). The observed kappa was .78. The Baserate × Accuracy table (Table 1) would indicate an observed kappa of .78 and estimated baserate of .61 corresponds with estimated observer accuracy greater than .90. Table 3 provides another example with another participant's session. In this case, the best estimate of the baserate of engagement and the collapsed category for Participant 2 was .94 and .06, respectively. The baserates of the two categories are extremely different. In this case, the observed kappa is .47. Using Table 1, we estimated the accuracy of judges to be greater than .90.

In either case, most researchers would consider .9 or .95 accuracy with the true state of affairs adequate. If we had judged the adequacy of agreement using a criterion kappa value of .6 (i.e., that suggested by Odom et al., 2004), we would have rejected the data for Participant 2 as being inaccurate when, in fact, the accuracy of observers was almost as good as that represented by Participant 1's kappa of .78. The use of an arbitrary criterion kappa would have led to unnecessary retraining of observers, costing the project staff time, or altering the coding system with potential reduction in the construct validity of the coded data.

Discussion

Kappa is a popular agreement statistic used to estimate the accuracy of observers. A best estimate of the baserate of the target behavior is necessary for an accurate interpretation of kappa. Using a priori criterion values for kappa may result in unnecessary code refinement and coder retraining or abandoning research questions. Instead of using

Table 3. Two by Two Table Representing Observer Agreement on Engagement for Participant 2

		Observer 2		
		Seconds engaged	Seconds not engaged or other	Sum of rows
Observer 1	Seconds engaged	1091	8	1099
	Seconds not engaged or other	65	36	101
	Sum of columns	1156	44	1200

 © American Association on Mental Retardation

criterion kappa values, researchers can (a) create two × two agreement tables for each target category, (b) compute the observed kappa, (c) compute the best estimate of true baserate of the target category, and (d) interpolate the estimated accuracy of coders with the true state of affairs using Table 1 (Bakeman et al., 1997).

The estimate of the baserate of the target behavior improves with increases in the length of the observation session and increases in the number of raters. Because our estimate of accuracy depends on our estimate of baserate, our estimate of accuracy also improves with increases in the length of the observation session and the number of raters. Increasing the length of the observational session is analogous to increasing the number of individuals in a sample and improves our estimate of baserate by decreasing sampling error (Glass & Hopkins, 1996). Increasing the number of raters increases the number of independent estimates of the baserate. The greater the number of independent estimates of baserate used to compute the mean baserate, the more likely that mean represents the true baserate of the session (Bruckner, Yoder, & McWilliams, 2005). For these reasons, estimated baserate, and thus estimated accuracy, can be improved by increasing the length of sessions and adding more raters.

In case readers believe that the quality of data will be higher if kappas below a criterion such as .6 are rejected, a brief discussion of the balance between reliability and construct validity will be presented. For example, a code is developed to measure the construct of object play in young children. The code is then operationally defined as the duration of the session that a child uses an object in a way that shows understanding of the object's unique properties (Lifter, 2000). After training on this coding system, observers obtain a kappa that is below the criterion of .6, but each coder has above 90% accuracy. Assume the investigator decides that redefining the code is necessary. The new code definition is the number of times the child touches any object. This modification to the coding system increases the duration of the object play code to closer to .5 and moves kappa above .6. The price paid for the higher kappa is a loss of validity. Unfortunately, the new code does not even have *face* validity with the construct of differentiated object play. Now, because of what we believed to be a more rigorous scientific decision to reject all kappas below .6, we are unable to confirm research questions about differentiated object play. The challenge for the field is to change the way that agreement statistics are judged when papers are reviewed for publication. Using a criterion value to judge the sufficiency of agreement gives an unfair advantage to sets of behaviors with equal probability and unfairly discounts the value of datasets on behaviors with unequal probabilities. The field should instead judge the adequacy of accuracy estimates given the obtained kappa and estimated baserate of the target behavior. This would involve increasing awareness about the baserate problem with kappa.

In summary, we have attempted to convince readers to estimate the accuracy of observers from obtained kappa and baserate estimates. The adequacy of the estimated accuracy has not yet been determined. As a beginning point, we might consider .90 as a reasonable accuracy level. However, this would not be reasonable for variables that are extraordinarily difficult to code. As in all judgments about adequacy of reliability, the real standard is relative to the other attempts to measure the same construct or behavior and to the expected effect size involving the target variable.

References

Agresti, A. (1996). *An introduction to categorical data analysis.* New York: Wiley.

Bakeman, R. (2000). Behavioral observation and coding. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social personality psychology* (pp. 138–159). New York: Cambridge University Press.

Bakeman, R., Quera, V., McArthur, D., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods, 2,* 357–370.

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

Bruckner, C. T., Yoder, P. J., & McWilliams, R. A. (2006). Generalizability and decision studies: An example using conversational language samples. *Journal of Early Intervention, 28,* 139–153.

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

© American Association on Mental Retardation

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

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

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology*. Boston: Allyn & Bacon.

Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis: Theory and practice. *Archives of General Psychiatry, 38,* 408–413.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71,* 165–179.

Kraemer, H. C. (1979). Ramifications of a population model for k as a coefficient of reliability. *Psychometrika, 44,* 461–472.

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2004). Kappa coefficients in medical research. In R. D. D'Agostino (Ed.), *Tutorials in biostatistics I: Statistical methods in clinical studies*. New York: Wiley.

Lifter, K. (2001). Linking assessment to intervention for children with developmental disabilities or at-risk for developmental delay: The developmental play assessment (DPA) instrument. In K. Gitlin-Weiner, A. Sandgrund, & C. Schafer (Eds.), *Play diagnosis and assessment* (pp. 228–260). New York: Wiley.

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry, 130,* 79–83.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. D., Thompson, B., & Harris, K. (2004, Fall). Quality indicators for research in special education and guidelines for evidence-based practices: Executive summary. *Newsletter of the Division for Research: Council for Exceptional Children*.

Robinson, B. F., & Bakeman, R. (1998). ComKappa: A Windows 95 program for calculating kappa and related statistics. *Behavior Research Methods, Instruments, and Computers, 30,* 731–732.

Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the baserate problem in the kappa statistic. *Archives of General Psychiatry, 42,* 725–728.

Tapp, J., & Walden, T. (1993). Procoder: A professional tape control coding and analysis system for behavioral research using video tape. *Behavior Research Methods, Instruments, and Computers, 25,* 53–56.

Uebersax, J. (n.d.). Statistical methods for rater agreement. *Kappa coefficients*. Retrieved July 14, 2005, from http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm

Yule, G. U. (1912). On methods of measuring association between two attributes. *Journal of the Royal Statistical Society, 75,* 579–642.

Received 8/22/05, accepted 1/30/06.
Editor-in-charge: William E. MacLean, Jr.

Requests for reprints should be sent to Cornelia Taylor Bruckner, Desired Results Access, CIHS/Sonoma State University, 311 Professional Center Dr., Rohnert Park, CA 94928-2152. E-mail: corneliabruckner@sonoma.edu

Appendix A

Method for Estimating Accuracy from Agreement

Accuracy can be represented as a conditional probability matrix for each observer, where Row 1 Column 1 is the probability that the observer would code the event a 1 when in truth it was a 1; Row 1 Column 2 is the probability that the observer would code an event 2 when in truth it was a 1; Row 2 Column 1 is the probability that an observer would code an event 1 when in truth it was a 2; Row 2 Column 2 is the probability that the observer would code an event a 2 when in truth it was a 2 (Bakeman et al., 1997). A familiar context for this type of conditional probability matrix is the calculation of sensitivity and specificity. If we say that the presence of a disorder is Event 1 and the absence of a disorder is Event 2, the sensitivity of a test would be Row 1 Column 1 and specificity would be Row 2 Column 2. If the accuracy of an observer is perfect, Row 1 Column 1 and Row 2 Column 2 would both equal 1, meaning every time Event 1 occurred, it would be recorded by the observer (sensitivity) and every time Event 1 did not occur it was not recorded (specificity).

 © American Association on Mental Retardation

Row 1 Column 1 in an agreement matrix is the frequency or proportion of intervals in which both observers recorded Event 1. This is distinct from the accuracy matrix where Row 1 Column 1 is the probability that a single coder would code an event given its occurrence. To calculate this from two observer accuracy matrices, we need to sum all the ways that the observers could agree on an Event. This includes both observers coding an Event 1 when it is in truth a 1 but it also includes both observers coding it a 1 when it is in truth a 2. The proportion of intervals in which both observers recorded a 1 when it was in truth a 1 is the product of the sensitivity of Observer 1, the sensitivity of Observer 2, and the baserate of Event 1. The proportion of intervals in which Observer 1 and Observer 2 would record an Event 1 when it was a 2 would be the product of Row 2 Column 1 for both observers and the baserate of Event 2. This is repeated for all four cells in the agreement matrix (see Bakeman et al., 1997, for definitions, equation, and extension to more than 2 codes).

© American Association on Mental Retardation