Behavior Research Methods, Instruments, & Computers
1998, 30 (4), 731-734

Software Announcments

ComKappa: A Windows '95 program for calculating kappa and related statistics

Assessing observer agreement is a necessary step in establishing the reliability of systematic behavioral observations. The statistics that provide measures of observer agreement are discussed most often in terms of establishing the trustworthiness of a data set once data have been collected. However, assessing level of agreement is also useful in providing feedback to observers in training, and in the assessment of reliability drift (i.e., the tendency for agreement to decline after training is completed) at multiple timepoints throughout the data collection process (Bakeman, in press; Bakeman & Gottman, 1997).

When a coding scheme is categorical, an agreement proportion, calculated as the number of agreements divided by the total number of events, is the most basic index of agreement. Cohen (1960) noted, however, that a simple agreement proportion capitalizes on chance agreements between observers. That is, if one or both of the observers guess, a certain proportion of their guesses will be the same simply by chance. Thus, Cohen proposed the kappa coefficient as a statistic that corrects for this chance factor by adjusting the proportion of observed agreement on the basis of joint marginal distributions. Kappa is widely accepted and appears as the agreement statistic of choice in a number of general texts concerned with data analysis (Bakeman & Gottman, 1997; Siegal & Castellan, 1988). Since Cohen first proposed the kappa coefficient in 1960, a number of kappa variants and related statistics have been developed that allow researchers to assess agreement in a manner more appropriate for particular experimental designs. Weighted kappa is one variant of the original kappa that allows a researcher to incorporate partial credit or scale the level of disagreements between different codes (Cohen, 1968). Two other forms, $kappa_n$ and $kappa_b$, formulate kappa under differing models of chance. $Kappa_n$ calculates agreement corrected for chance under a model in which observers assume an even distribution of events across all codes. $Kappa_b$ assumes a chance model in which, when unsure of how to code an event, the observer will default to the code with the highest base rate. Brennan and Prediger (1981) and Umesh, Peterson, and Sauber (1989) have discussed the details of these two lesser known kappa variants. Additionally, a number of kappa-related statistics aid in the interpretation of obtained kappas. The maximum value of kappa refers to the largest value that kappa could obtain given the marginal distributions. The ratio of kappa to kappa maximum allows researchers to compare kappas obtained in sessions with differing base rates for the same code (Umesh et al., 1989). Finally, the standard error of kappa can be estimated, allowing the obtained kappa to be tested for significance against a null distribution (Bakeman & Gottman, 1997; Fleiss, Cohen, & Everitt, 1969).

Several computer programs have been developed that calculate some form of kappa (e.g., Antonak, 1977; Bakeman & Gottman, 1997; Berk & Campbell, 1976; Cicchetti, Showalter, & McCarthy, 1990; Strube, 1989; Valiquette, Lesage, Cyr, & Toupin, 1994; Wixon, 1979). However, these programs are based on older operating systems, calculate only a limited number of the kappa variants, or require the use of programming skills to update program listings and compile the code into executable format. Bakeman and Quera's (1995) GSEQ program also computes kappa but assumes sequential data as input. The present program (ComKappa) differs from similar programs in that its graphical interface interacts with the user in an easy and natural way (e.g., data may be entered into an on-screen matrix, and the program is controlled by on-screen menus, buttons, and tabs). In addition, it calculates a number of kappa and kappa-related statistics not available in other programs. The user-friendly nature of the program allows users not familiar with more sophisticated statistical packages (e.g., observers in training) to calculate kappa statistics. Furthermore, the on-screen matrix facilitates the input of data and allows users to examine the distribution of errors and identify codes that may be problematic. These two features allow less sophisticated users to measure progress during training and assess reliability drift without the constant intervention of a statistical consultant. In addition, more sophisticated users will find ComKappa useful in calculating not just kappa but weighted kappa, $kappa_n$, $kappa_b$, kappamax, and the standard error of kappa. ComKappa runs under (and requires) the Windows '95 operating system, so its conventions will be familiar to many computer users. It was written in Borland's Delphi 2, a programming system based on Borland's Object Pascal.

Input. The user enters the number of codes in the scheme and, if they are desired, labels for the codes. The program generates an on-screen matrix for the input of observed frequencies. A similar matrix allows users to define a weight matrix for the calculation of weighted kappa, if desired. Once a frequency matrix is defined, users select the kappa statistics to be calculated.

Output. Results are presented in the main window of the program. Agreement and weight matrices can be saved to or retrieved from disk files or cut and pasted to other programs via the Windows clipboard. Likewise, kappa statistics can be cut and pasted to other programs for saving or printing.

Availability. The program is available for downloading through the WWW at http://www.gsu.edu/~psyrab/bakeman.html. Alternatively, the program can be obtained by sending an IBM-formatted disk and stamped return mailer to the first author.

REFERENCES

Antonak, R. F. (1977). A computer program to compute measures of response agreement for nominal scale data obtained from two judges. *Behavior Research Methods & Instrumentation*, **9**, 553.

Bakeman, R. (in press). Behavioral observation and coding. In H. T. Reis & C. K. Judd (Eds.), *Handbook of research methods in social psychology*. New York: Cambridge University Press.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). New York: Cambridge University Press.

Bakeman, R., & Quera, V. (1995). *Analyzing interaction: Sequential analysis with SDIS and GSEQ*. New York: Cambridge University Press.

Berk, R. A., & Campbell, K. L. (1976). A FORTRAN program for Cohen's kappa coefficient of observer agreement. *Behavior Research Methods & Instrumentation*, **8**, 396.

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

Cicchetti, D. V., & Showalter, D., & McCarthy, P. (1990). A computer program for calculating subject-by-subject kappa or weighted kappa coefficients. *Educational & Psychological Measurement*, **50**, 153-158.

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

Cohen, J. A. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreements or partial credit. *Psychological Bulletin*, **70**, 213-220.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, **72**, 323-327.

Siegal, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Strube, M. J. (1989). A general program for the calculation of the kappa coefficient. *Behavior Research Methods, Instruments, & Computers*, **21**, 643-644.

Umesh, U. N., Peterson, R. A., & Sauber, M. H. (1989). Interjudge agreement and the maximum value of kappa. *Educational & Psychological Measurement*, **49**, 835-850.

Valiquette, C. A. M., Lesage, A. D., Cyr, M., & Toupin, J. (1994). Computing Cohen's kappa coefficients using SPSS MATRIX. *Behavior Research Methods, Instruments, & Computers*, **26**, 60-61.

Wixon, D. R. (1979). Cohen's kappa coefficient of observer agreement: A BASIC program for minicomputers. *Behavior Research Methods & Instrumentation*, **11**, 602.

Byron F. Robinson
Emory University, Atlanta, Georgia

and

Roger Bakeman
Georgia State University, Atlanta, Georgia

(Manuscript received January 15, 1997;
revision accepted for publication May 21, 1997.)

Automated syndromic profile and discriminant classification analyses for the Adjustment Scales for Children and Adolescents

The Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993) is an objective behavior rating instrument designed to be completed by a student's classroom teacher, for use with all noninstitutionalized youth ages 5–17 years. The ASCA is composed of 156 behavioral descriptions in 29 specific situations in which teachers may observe students' behaviors rather than a symptom or problem checklist. Of the 156 items, 96 are scorable and based on factor analyses, singularly assigned to one of six core syndromes (Attention-Deficit/Hyperactive, Solitary Aggressive–Provocative, Solitary Aggressive–Impulsive, Oppositional Defiant, Diffident, and Avoidant) or two supplementary syndromes (Delinquent and Lethargic/Hypoactive). Raw scores are converted to normalized *T* scores that are based on the nationally representative standardization sample.

The Syndromic Profile Interpretation approach was developed to facilitate differential diagnosis and classification using the ASCA. McDermott and Weiss (1995) produced a cluster analysis for the 1,400 students in the ASCA standardization sample that yielded 22 syndromic profile types (14 major and 8 clinical subtypes) using the six core syndromes. Profiles varied with regard to distinct behavior and severity as well as by prevalence for age, gender, socioeconomic status, cognitive ability, achievement, handicapping condition, and physical problems (McDermott, 1994; McDermott & Weiss, 1995). Detailed descriptions of the process of comparing the normalized *T* score profile for a selected syndromic profile type with the obtained *T* score profile for a youth in question with the use of the generalized distance score (GDS) were also presented by these authors. The GDS is a measure of profile similarity (dissimilarity) that is obtained by examining deviations of a youth's *T* scores from the average *T* scores for a specified group (profile type). Calculation of GDSs is fully explained and an example is provided in the ASCA manual (McDermott, 1994). The youth's profile is classified as most similar to the ASCA profile type that results in the *smallest* GDS. Characteristics associated with each of the 22 syndromic profile types are also presented in the ASCA manual (McDermott, 1994) as well as in McDermott and Weiss (1995).

The Discriminant Classification Interpretation method is based on the discriminant function equations obtained from a validity study showing that the ASCA differentiated "normal" from "socially/emotionally disturbed" youths. Although the quadratic discriminant function equations were slightly more accurate (by an average of about 2%), the linear method is presented in the ASCA