Behavior Research Methods, Instruments, & Computers
1988, 20 (5), 513-514

Micro-CONGRU: A microcomputer program for measuring levels of overall and partial congruence among multiple observers on nominal scales

MARLEY W. WATKINS
SouthWest EdPsych Services Inc., Phoenix, Arizona
and
JOSEPH C. KUSH
Deer Valley Unified School District, Phoenix, Arizona

A wide variety of statistical procedures has been used for assessing the degree and significance of agreement among raters in assignments of objects or subjects to nominal scales. The simplest procedure, raw percent of agreement, has been described as inadequate and misleading (Spitzer, Cohen, Fleiss, & Endicott, 1967). The phi coefficient also has been shown to be inappropriate, since it reflects only the strength of the relationship and indicates nothing conclusive about agreement (McDermott, in press). Similarly, chi-square and related contingency coefficients have been demonstrated by Cohen (1960) to test null hypotheses with regard to association but not agreement, and therefore will be inflated by any departure from chance association within a data set (McDermott, in press).

The most powerful statistical technique for establishing nominal scale agreement is the kappa coefficient (κ). Originally developed by Cohen (1960) and later refined by Light (1971), κ essentially represents the normalized proportion of interrater agreement in excess of what would be expected by chance among raters. As originally developed, κ was restricted to instances where the number of raters was two, and the same two raters classified each object or subject. Fleiss (1971) developed an extension of κ for use in situations where the number of observers would be greater than two, and where there was no assumption requiring the set of observers to remain constant across all cases. Fleiss also provided formulas for measuring the response agreement among many raters for each nominal category.

Kappa has been applied widely in behavioral research. Spitzer and colleagues (American Psychiatric Association, 1980; Spitzer, Forman, & Nee, 1979) used κ to determine agreement among diagnosticians for psychiatric categories in the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-III*). McDermott & Hale (1982) applied κ to determine agreement among human and computer-generated psychoeducational diagnostic categorizations. Journal reviewers' agreement concerning manuscript publishability was analyzed via κ statistics by Watkins (1979). Other applications of κ have been discussed by Brennan and Prediger (1981) and by Fleiss (1981).

A number of useful computer programs have been written for calculating κ (Antonak, 1977; Berk & Campbell, 1976; Chan, 1987; McDermott & Watkins, 1979; Wixon, 1979). Some of these programs are limited to the two-rater case, while others operate only on large mainframe computers. The program described in this article utilizes a microcomputer to calculate both general and conditional agreement among many raters on the basis of Fleiss's (1971) computational formulas.

Language and computer. Micro-CONGRU is written in Applesoft BASIC; it requires 64K RAM under the DOS 3.3 operating system. The program is designed to run on the Apple II family of computers, including Apple II+, IIe, IIc, and IIgs, with one disk drive. A printer is optional. Programmers may be able to translate the program for operation on other microcomputers.

Input. Data may be input from the keyboard or directly from sequential text files on disk. Keyboard input is fully interactive and allows editing and review of data entered from the keyboard and the disk.

Output. Printed output includes: (1) the overall percentage of agreement among raters before chance agreement is excluded; (2) the overall coefficient (κ) of agreement; (3) the estimated variance and standard error of the overall κ; (4) the value of the unit normal deviate and the level of significance for the overall κ; (5) the conditional percentage of agreement among raters for each category prior to chance; (6) the conditional coefficients for each category; (7) the variances and standard errors for each partial κ; and (8) the unit normal deviates and significance levels for each partial κ.

Limitations. The Micro-CONGRU program will allow up to 450 cases to be assigned by 999 or fewer raters for up to a maximum of 15 nominal categories.

Program Availability. Micro-CONGRU is available on disk by sending $10.00, to cover postage and reproduction costs, to M. W. Watkins at P.O. Box 1870, Phoenix, AZ, 85001. Requests from outside the United States should include $15.00 for airmail delivery.

Address correspondence to M. W. Watkins, P.O. Box 1870, Phoenix, AZ 85001.

REFERENCES

AMERICAN PSYCHIATRIC ASSOCIATION. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington DC: Author.

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.

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.

CHAN, T. S. C. (1987). A DBASE III program that performs significance testing for the Kappa coefficient. *Behavior Research Methods, Instruments, & Computers*, **19**, 53-54.

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

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* (2nd ed.). New York: Wiley.

LIGHT, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, **76**, 365-377.

MCDERMOTT, P. A. (in press). Agreement among diagnosticians or observers: Its importance and determination. *Professional School Psychology*.

MCDERMOTT, P. A., & HALE, R. L. (1982). Validation of a systems-actuarial computer process for multidimensional classification of child psychopathology. *Journal of Clinical Psychology*, **35**, 477-486.

MCDERMOTT, P. A., & WATKINS, M. W. (1979). A program to evaluate general and conditional agreement among categorical assignments of many raters. *Behavior Research Methods & Instrumentation*, **11**, 399-400.

SPITZER, R. L., COHEN, J., FLEISS, J. L., & ENDICOTT, J. (1967). Quantification of agreement in psychiatric diagnosis. *Archives of General Psychiatry*, **17**, 83-87.

SPITZER, R. L., FORMAN, J. B., & NEE, J. (1979). DSM-III field trials: I. Initial interrater diagnostic reliability. *American Journal of Psychiatry*, **136**, 815-817.

WATKINS, M. W. (1979). Chance and interrater agreement on manuscripts. *American Psychologist*, **34**, 796-798.

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

(Revision accepted for publication June 25, 1988.)