Behavior Research Methods & Instrumentation
1979, Vol. 11 (3), 399-400

A program to evaluate general and conditional agreement among categorical assignments of many raters

PAUL A. MCDERMOTT
University of Pennsylvania, Graduate School of Education
Philadelphia, Pennsylvania 19104

and

MARLEY W. WATKINS
Department of Educational Psychology and Measurements
University of Nebraska, Lincoln, Nebraska 68588

A variety of statistical procedures have been proposed for assessing the degree and significance of agreement between raters in assignments of objects or subjects to nominal scales. Foremost among the available techniques is the κ coefficient devised by Cohen (1960) and later refined by Light (1971). This statistic essentially represents the normalized proportion of interrater agreement in excess of that expected on the basis of chance or random assignments.

In recent years, a number of useful programs have been developed for computer applications of κ (Antonak, 1977; Berk & Campbell, 1976; Cicchetti, Lee, Fontana, & Dowds, 1978). However, such applications of the κ coefficient are bound by two important constraints. First, the use of κ is appropriate only when testing agreement between two raters: It is not appropriate for answering questions about the conjoint agreement among many raters. For this reason, Light (1971) developed an extension of the earlier statistic, known as κ_m, which is the coefficient of multiple-observer agreement.

The second constraint to uses of κ is found in the assumption underlying both Cohen's (1960) original statistic and Light's (1971) later extension that the respective pairs or sets of raters assigning the various objects to categories will remain identical across all cases. In applied research settings, it is frequently the circumstance that neither assumption holds because more than two independent raters may be involved in the categorical assignment of each case, and the sets of raters may vary as a function of time or convenience (as, e.g., when subjects in different levels of a treatment or educational program must be observed and rated in different settings on repeated occasions).

Fleiss (1971) developed formulas that revise and extend κ for use in situations where the number of observers may be greater than two and where there is no assumption that the sets of raters will remain constant throughout all cases. The resulting statistic may be viewed as the general or overall coefficient of agreement of many raters across all nominal categories. Fleiss also provided formulas to measure the response agreement among many raters on each specific nominal category considered. This conditional coefficient is designed to test the probability that randomly chosen raters assign any randomly selected object or subject to the identical category. Such a conditional coefficient may be applied to evaluate the integrity or viability of any given categorical value or classification.

The program described in this paper calculates both general and conditional coefficients and tests the statistical significance of agreement among many raters assigning objects to nominal scales based upon Fleiss's (1971) computational formulas.

Input. Each analysis requires two control cards and a data card deck as follows: (1) a title card, (2) a problem card to specify the number of cases being categorized, number of categories, and number of raters, and (3) a set of case cards, one card per case, specifying the number of raters choosing each category.

Output. The information provided for each analysis includes (1) an alphanumeric job title, (2) the general percentage of agreement among raters before chance agreement is excluded, (3) the value of the general coefficient of agreement, (4) the estimated variance and standard error for the general coefficient, (5) the value of the unit normal deviate and level of significance for the general coefficient, (6) the conditional percentages of agreement among raters for each category prior to the exclusion of chance, (7) the values of the conditional coefficients for each category, (8) the variances and standard errors for each conditional coefficient, and (9) the unit normal deviates and significance levels for each conditional coefficient.

Computer and Language. Written in FORTRAN IV, the program is compatible with machines in the IBM 360 series. Variables are in mnemonic form according to Fleiss's (1971) computational formulas. Input editing and output specifications are provided for user's syntactical errors.

Restrictions. Currently, the program will permit up to 1,000 cases to be assigned by 100 or fewer raters to a maximum of 25 categories.

Availability. A source listing, user's manual, and test input and output data may be obtained at no cost by writing to Paul A. McDermott, University of Pennsylvania, Graduate School of Education C1, 3700 Walnut Street, Philadelphia, Pennsylvania 19104.

REFERENCES

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

BERK, R. A., & CAMPBELL, K. A. A FORTRAN program for

0005-7878/79/030399-02$00.45/0

Cohen's kappa coefficient of observer agreement. *Behavior Research Methods & Instrumentation,* 1976, **8**, 396.

CICCHETTI, D. V., LEE, C., FONTANA, A. F., & DOWDS, B. N. *A computer program for assessing specific category rater agreement for qualitative data.* West Haven, Conn: Veterans Administration Hospital, 1978.

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

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

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

(Accepted for publication April 13, 1979.)