EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1979, 39

A COMPUTER PROGRAM FOR MEASURING LEVELS OF OVERALL AND PARTIAL CONGRUENCE AMONG MULTIPLE OBSERVERS ON NOMINAL SCALES

MARLEY W. WATKINS
University of Nebraska-Lincoln

PAUL A. McDERMOTT
University of Pennsylvania

A computer program entitled Program CONGRU is presented which analyzes the overall conjoint agreement among many observers for their classifications on categorical scales. A special feature assesses the significance of observers' congruence in assignments to each nominal category considered.

OF the many descriptive and inferential indices of the amount of agreement among observers in their assigning objects or subjects to nominal categories, the most widely practiced standard has been to offer statements of the percentage of concurrent agreement among observers. As indicated in the review by Hartmann (1977), this approach has been especially evident in the behavioral sciences in which it has been held that some degree of reliability must be afforded in rendering classifications of characteristics or qualities, particularly those qualities being attributed to human beings. Unfortunately though, as Yelton, Wildman, and Erickson (1977) and McDermott (1979) have pointed out, indices specifying percentages of interobserver agreement in nominal scale classification are very often inappropriately applied. They are misleading both for the researcher and for the practitioner.

In an alternative approach Cohen (1960) essentially sought to control for the unpropitious elements inherent in statements of simple percentages of agreement. He recognized that in any classification situation a certain amount of agreement among observers would be found by sole reason of chance events. Thus, any statement of interobserver agreement must reflect not only how much agreement is

evident but also, more importantly, how much agreement is in evidence beyond that which would be expected by chance alone. For this purpose, Cohen (1960) defined the statistic κ as

$$\kappa = \frac{\begin{array}{c}\text{Proportion of Simple} \\ \text{Agreement Between Observers}\end{array} - \begin{array}{c}\text{Proportion of Agreement} \\ \text{Expected by Chance}\end{array}}{1 - \begin{array}{c}\text{Proportion of Agreement} \\ \text{Expected by Chance}\end{array}}, \quad (1)$$

where κ represents the normalized proportion (or percentage with decimal removed) of agreement between observers after chance agreement has been excluded. In its general application, the use of κ is attendant upon a number of assumptions: (a) the categories must be mutually exclusive, exhaustive of possible alternatives and, of course, nominal; (b) the cases (of objects or subjects) classified are independent; and (c) the observers operate independently in their classification.

Several conceptually sound approaches and corresponding programs for computer analysis have been designed (Cicchetti, Aivano, and Vitale, 1977; Cicchetti, Lee, Fontana, and Dowds, 1978) which utilize the κ statistic in assessing levels of agreement for qualitative data. However, these techniques, although very useful in the context of certain applied and research settings, are contingent upon additional assumptions which are difficult to meet in many investigations or which, in other situations, are not generalizable to the inquiries being addressed. In the first instance, the available techniques are restricted to the case of agreement between *two* observers. This restriction is true even though some programs (e.g., Cicchetti et al., 1978) permit the user to test all possible pairwise comparisons among observers. However, much research is directed toward the question of interobserver agreement among *many* (i.e., more than two) observers conjointly. This concern with interobserver agreement is necessary, for example, when several persons have been called upon to validate a categorical rating device or when a number of specialists have been asked to render diagnoses for the same subject. The question of fundamental moment is clearly whether the set of observers is as a group in agreement, or equivalently, whether the observers are congruent in their categorizations. Moreover, it may be necessary to determine whether the degree of congruence among observers varies as a function of the nominal categories to which they are assigning subjects. Such queries demand a statistic of multiple observer agreement.

Field research may further constrain the investigator by making it impossible to retain identical pairs or groups of observers for the rating of all subjects. The maintenance of consistent sets of observers is a clear assumption of standard applications of κ. Yet, what might

the researcher do to contend with the typical circumstance wherein categorical ratings must be provided by asystematically or randomly varied sets of observers? Such a dilemma is encountered, for example, whenever students' classroom behaviors must be observed and classified by teachers who remain inconstant from classroom to classroom and from grade level to grade level.

Light (1971) has developed the statistic κ_m, referred to in this paper as the multiple observer congruence statistic. This statistic represents an extention of κ to the case in which many observers categorize subjects and in which the overall conjoint agreement among observers is assessed. Fleiss (1971) has described the special case of κ_m in which the researcher may assume that the set of multiple observers does not remain constant throughout all cases.

Purpose

Within this context the purpose of this paper was to describe a computer program designated as Program CONGRU that provides multiple observer congruence statistics based primarily on Fleiss's computational formulae. In addition to supplying κ_m values for overall congruence, the program calculates for each nominal category considered, the partial κ_j coefficient. Partial κ_j's are congruency statements based upon the conditional probability that a randomly chosen observer's assignment of a given subject to category j will coincide with another randomly selected observer's assignment of the same subject. Provision of these indices allows the user to examine the degree of congruence among observers relative to the various categories considered.

Input

The job deck for each analysis consists of two control cards and a case card deck which are arranged sequentially as follows:

Title card

An alphanumeric job title may be punched anywhere in columns 1–80.

Problem card

All numbers must be right adjusted.
Columns 1–4 = number of cases.
5–6 = number of categories.
7–9 = number of observers.

Case card deck

One card must be punched for each case. Each case card must include an entry for each of the categories indicated in columns 5–6 of the Problem Card. Non-selected categories are represented as blanks. All numbers must be right adjusted.

Columns 1–3– = number of observers selecting 1st category.
4–6 = number of observers selecting 2nd category.
7–9 = number of observers selecting 3rd category.
. .
. .
. .
73–75 = number of observers selecting 25th category.

Output

The program provides the following information for each analysis:

1. *Job title* as specified by control card.
2. *Overall degree of multiple observer congruence* for all categories:
   a. Percentage of agreement among observers, i.e., $\bar{P}$, not corrected for chance.
   b. Multiple observer congruence statistic, κ_m.
   c. Estimated variance and standard error of κ_m.
   d. Value of z and level of significance for κ_m.
3. *Conditional degree of multiple observer congruence* for each separate category where $\kappa_m \geq 0$:
   a. Percentage of partial agreement among observers on each category, i.e., $\bar{P}_j$, not corrected for chance.
   b. Partial coefficient of multiple observer congruence on each category, i.e., $\kappa_1, \kappa_2 \cdots \kappa_j$.
   c. Estimated variances and standard errors for respective κ_j's.
   d. Critical values of z and significance levels for partial coefficients.

Capabilities and Limitations

Program CONGRU is written in FORTRAN IV for processing by computers in the IBM 360 series. It is fully documented with variables in mnemonic form corresponding to Fleiss's (1971) computational formulas. Input editing and output specifications are provided for user's syntactical errors. At present, the program handles a maximum of 1000 cases being assigned to as many as 25 categories by 100 or fewer observers.

Availability

A listing of the CONGRU source program, a copy of this paper, and a complete set of sample input and output data are available, without charge, from Dr. Paul A. McDermott, University of Pennsylvania, Graduate School of Education, 3700 Walnut Street, Philadelphia, Pennsylvania 19104.

REFERENCES

Cicchetti, D. V., Aivano, S. L., and Vitale, J. Computer programs for assessing rater agreement and rater bias for qualitative data. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1977, 37, 195–201.

Cicchetti, D. V., Lee, C., Fontana, A. F., and 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 coefficient of agreement for nominal scales. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1960, 20, 37–46.

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

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.

McDermott, P. A. The grouping and classification of school children: Actuarial and legal principles. In B. D. Sales and M. R. Novick (Eds.), *Perspectives in law and psychology: Volume III. Testing and evaluation.* New York: Plenum, 1979.

Yelton, A. R., Wildman, B. G., and Erickson, M. T. A probability-based formula for calculating interobserver agreement. *Journal of Applied Behavior Analysis,* 1977, 10, 127–131.