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 inter-observer 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 inter-observer agreement must reflect not only how much agreement is
This concern with interobserver agreement is necessary, for example, such queries demand a statistic of multiple observer agreement. 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 $\kappa$. Yet, what might the researcher do to contend with the typical circumstance wherein categorical ratings must be provided by systematically 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’s (1971) has developed the statistic $\kappa_m$, referred to in this paper as the multiple observer congruence statistic. This statistic represents an extension of $\kappa$ 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 $\kappa_m$ 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 $\kappa_m$ values for overall congruence, the program calculates for each nominal category considered, the partial $\kappa_j$ coefficient. Partial $\kappa_j$'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. |
| Columns | 5–6 | number of categories. |
| Columns | 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.
10–12 = ...
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., $P$, not corrected for chance.
   b. Multiple observer congruence statistic, $k_m$.
   c. Estimated variance and standard error of $k_m$.
   d. Value of $z$ and level of significance for $k_m$.
3. Conditional degree of multiple observer congruence for each separate category where $k_m \geq 0$:
   a. Percentage of partial agreement among observers on each category, i.e., $P_j$, not corrected for chance.
   b. Partial coefficient of multiple observer congruence on each category, i.e., $k_j$, $k_j \cdots k_j$.
   c. Estimated variances and standard errors for respective $k_j$’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