Number 49 December 4, 1978

Citation Classics

Ebel R L. Estimation of the reliability of ratings. *Psychometrika* **16**:407-24, 1951.

When several judges rate several products, questions about the reliability of their ratings, i.e., the consistency of the ratings across judges, may arise. This paper considers several procedures for estimating that reliability, and recommends one that is generally most satisfactory. [The *Science Citation Index*® (*SCI*®) and the *Social Sciences Citation Index*™ (*SSCI*™) indicate that this paper was cited a total of 219 times in the period 1961-1977.]

Robert L. Ebel
Professor of Education and Psychology
Michigan State University
East Lansing, Michigan 48824

January 30, 1978

"As part of the program of general education launched at the University of Iowa in the early 1940's, students were required to demonstrate or develop skill in writing and speaking, among other abilities. The themes they wrote and the speeches they gave were rated by professors in the Communication Skills Program. Directors of the program were concerned that the ratings should be consistent across raters, not only in fairness to the students, but in pursuit of agreement among the professors on the elements of quality in a theme or speech. The ratings were analyzed in the Examinations Service of the University.

"My predecessor as director of that Service, the late Professor Paul Blommers, had worked out a routine for calculating the extent of agreement in the ratings, based on R. A. Fisher's intraclass correlation coefficient. There were, I found, two other formulas which appeared to be applicable. But when the three formulas were applied to the same sets of ratings they gave a somewhat discrepant result. Something was wrong, and I set out to find what it was. The discovery might just possibly put me one step closer to a firm grasp of the ideas developed by R.A. Fisher, Charles C. Peters, or Paul Horst.

"Taking some simple numerical hypothetical examples of possible ratings, I applied the three formulas and studied the results. It became apparent that the different results sometimes yielded by the formulas were due to differences among the formulas in two characteristics: (1) whether the overall within-raters variance was the arithmetic mean or the geometric mean of the separate within-raters variances, and (2) whether the between-raters variance was included or excluded in the error term.

"If examples are chosen in which the arithmetic mean is the same as the geometric mean, and if between-raters variance is always included as error, the three formulas will give identical results.

"Good mathematicians mistrust general inferences from specific numerical examples. Harold Gulliksen, reviewing an early draft of the paper, suggested that I support the conclusions I had reached from the numerical examples with a generalized algebraic derivation. Substituting persistence for brilliance, I managed to do this. Still, I do find simple numerical examples helpful in suggesting generalizations worthy of algebraic validation.

"Why has the paper been cited often? Not, I think, because many others are concerned with the discrepancy that attracted my attention Rather, it may be because the paper included a fairly simple explanation, with simple examples, of the use of analysis of variance methods in solving some frequently encountered problems of reliability estimation."