CC/NUMBER 3
JANUARY 20, 1986

This Week's Citation Classic®

Cohen J. A coefficient of agreement for nominal scales.
Educ. Psychol. Meas. 20:37-46, 1960.
[New York University, NY]

The kappa coefficient, the proportion of agreement corrected for chance between two judges assigning cases to a set of *k* categories, is offered as a measure of reliability. Its limits, maximum and minimum, and sampling characteristics are given. [The *Social Sciences Citation Index*® (*SSCI*®) indicates that this paper has been cited in over 810 publications since 1960, making it the most-cited paper published in this journal.]

Jacob Cohen
Department of Psychology
New York University
New York, NY 10003

November 18, 1985

The nature of psychological measurement being what it is, it was altogether fitting that it spawned around the turn of the century a theory of mental tests that centered around errors of measurement and reliability. Among its many fruits was the demonstration of how a test's reliability constrained its validity, i.e., its ability to correlate with other variables. But the theory was for tests that yielded numeric *scores*.

During my first postdoctoral decade, I was in a clinical setting where an important form of measurement was, in J.P. Stevens's classic scheme, nominal scaling (the assignment of units to qualitative categories, as in psychiatric diagnosis). Other examples include the classification of botanical specimens and referees' disposition decisions for manuscript submissions. The fact that this form of measurement is not quantitative does not prevent the same issues of reliability from arising. It is intuitively evident that poor interjudge agreement, say in diagnosis, will limit the possible degree of association between diagnosis and anything else.

This being fairly obvious, it was standard practice back then to report the reliability of such nominal scales as the percent agreement between pairs of judges. Thus, two psychiatrists independently making a schizophrenic-nonschizophrenic distinction on outpatient clinic admissions might report 82 percent agreement, which sounds pretty good.

But is it? Assume for a moment that instead of carefully interviewing every admission, each psychiatrist classifies 10 percent of the admissions as schizophrenic, but does so blindly, i.e., *completely at random*. Then the expectation is that they will jointly "diagnose" .10 X .10 = .01 of the sample as schizophrenic and .90 X .90 = .81 as nonschizophrenic, a total of .82 "agreement," obviously a purely chance result. This is no more impressive than an ESP demonstration of correctly calling the results of coin tosses blindfolded 50 percent of the time!

The example is a bit extreme, but the principle it illustrates is unexceptionable: the proportion or percent of observed agreement between two judges assigning cases to a set of *k* mutually exclusive, exhaustive categories inevitably contains an expected proportion that is entirely attributable to chance.

Thus was *kappa* born. It is simply the proportion of agreement corrected for chance. For both the diagnosis and ESP examples, *kappa* equals zero, as it should. If the psychiatrists had agreed on 95 percent of the cases, the *kappa* would be (.95 - .82)/(1 -.82) = .72.

I later extended the concept of *kappa* to weighted *kappa* (to provide for partial credit) and to weighted chi square. *Kappa* has become something of a cottage industry among psychometricians, who have produced dozens of articles on *kappa* over the years. It has been extended to multiple judges, generalized and specialized in various ways, and integrated into formal measurement theories. Its statistical properties have been thoroughly studied both analytically and through Monte Carlo studies.[1-4] It has been variously programmed for mainframe and personal computers, and has found its way into standard textbooks.

The main reason for its heavy citation, however, has been its application in many fields, but chiefly, harking back to its origins, in psychiatric diagnosis.

1. **Light R J.** Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychol. Bull.* 76:365-77, 1971. (Cited 80 times.)
2. **Kraemer H C.** Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* 44:461-72, 1979.
3. **Cohen J.** Weighted kappa: nominal scale agreement with provision for scale and disagreement or partial credit. *Psychol. Bull.* 70:213-20, 1968. (Cited 275 times.)
4. ---------- Weighted chi-square: an extension of the kappa method. *Educ. Psychol. Meas.* 32:61-74, 1972.

