DOCUMENT RESUME

ED 093 975 TM 003 831

AUTHOR Brennan, Robert L.; Light, Richard J.
TITLE Measuring Agreement When Two Observers Classify People Into Categories Not Defined in Advance.
PUB DATE May 73
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April, 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Analysis of Variance; Correlation; *Hypothesis Testing; *Rating Scales; *Reliability; *Sampling; Tests of Significance

ABSTRACT

Basic to many psychological investigations is the question of agreement between observers who independently categorize people. Several recent studies have proposed measures of agreement when a set of nominal scale categories have been pre-defined and imposed on both observers. This study, in contrast, developes a measure of agreement for settings where observers independently define their own categories. Thus, it is possible for observers to delineate different numbers of categories, with different names. Computational formulae for the mean and variance of the proposed agreement measure are given; further, a statistic with a large-sample normal distribution is suggested for testing the null hypothesis of random agreement. A computer based comparison of the large sample approximation with the exact distribution of the test statistic shows a generally good fit, even for moderate sample sizes. Finally, a worked example involving two psychologists' classifications of children illustrates the computations. (Author)

21.06

MEASURING AGREEMENT WHEN TWO OBSERVERS CLASSIFY PEOPLE INTO CATEGORIES NOT DEFINED IN ADVANCE

Robert L. Brennan
State University of New York at Stony Brook

and

Richard J. Light
Harvard University

May, 1973

MEASURING AGREEMENT WHEN TWO OBSERVERS CLASSIFY PEOPLE INTO CATEGORIES NOT DEFINED IN ADVANCE

Submitted June 10, 1973

Abstract

Basic to many psychological investigations is the question of agreement between observers who independently categorize people. Several recent studies have proposed measures of agreement when a set of nominal scale categories have been pre-defined and imposed on both observers. This study, in contrast, develops a measure of agreement for settings where observers independently define their own categories. Thus, it is possible for observers to delineate different numbers of categories, with different names. Computational formulae for the mean and variance of the proposed agreement measure are given; further, a statistic with a large-sample normal distribution is suggested for testing the null hypothesis of random agreement. A computer based comparison of the large sample approximation with the exact distribution of the test statistic shows a generally good fit, even for moderate sample sizes. Finally, a worked example involving two psychologists' classifications of children illustrates the computations.

Many variations of the problem of measuring agreement between two or more observers have been investigated by psychologists and statisticians. When measurements are taken on a variable with a continuous metric, agreement is generally expressed as a reliability or generalizability coefficient. As discussed thoroughly by Cronbach et. al. (1972) these coefficients are usually some version of the well-known intraclass correlation.

Suppose, on the other hand, that two psychologists each independently distribute N people among a set of mutually exclusive categories. When categories are specified in advance, Cohen (1960) suggested a measure of agreement Kappa for two observers who each assign N people among these categories. This measure has been extended by Cohen (1968), Everitt (1968), and Fleiss, Cohen, and Everitt (1969). It was further extended to three or more observers by Light (1971) and Fleiss (1971). For measuring agreement among several observers when each person is scored dichotomously, Fleiss (1965) has suggested procedures that are basically combinatorial. All of the suggestions have led to a useful and impressive set of procedures. All of these procedures, however, begin with the assumption that ~~all~~ several categories, each with a specific name, have been preselected, and that observers distribute people among these categories.

The problem we consider here is somewhat different. Suppose two psychologists are asked to partition a group of people into several subgroups. The specific criteria for partitioning is left up to each psychologist. Thus the two psychologists may develop different numbers of subgroups. Moreover, since no precise set of subgroups have been labeled in advance, each psychologist may use different criteria resulting in categories with different labels. This situation is illustrated in Table 1, where two psychologists, after studying a group of children, independently categorized each of 15 children into one of three subgroups based on behavior patterns.

Table 1 about here

An important question to be asked of these data is, "Do the two observers' lists agree beyond chance?" In the following sections we examine this question. We begin by developing a measure of agreement that provides a basis for testing the null hypothesis of random agreement. In order to examine the behavior of this measure, we present computational formulae for its mean and variance which are then incorporated into a large sample test statistic. Finally, after investigating the appropriateness of the test statistic for moderate sample sizes, we apply our procedure to the data in Table 1.

Developing a Measure of Agreement

Viewing our data in the format of a two dimensional contingency table will help to clarify the definition of agreement. The raw data in Table 1 are an example of data that can always be displayed in an R × C table, as illustrated in Table 2. Notice that the first observer's categories are indexed by a_i, $i = 1,\ldots,R$, while the second observer's categories are b_j, $j = 1,\ldots,C$. Note also that this format differs from that of Cohen (1960), Light (1971), and Fleiss (1971) in that here R can be different from C, and the row categories are not necessarily the same as the column categories. In Table 2, n_{ij} represents the number of persons classified into category a_i by observer 1, and into category b_j by observer 2. Finally

$$n_{i+} = \sum_{j=1}^{C} n_{ij} \; ; \; i = 1,\ldots,R; \quad \text{and} \quad n_{+j} = \sum_{i=1}^{R} n_{ij} \; ; \; j = 1,\ldots,C.$$

Table 2 about here

How can agreement be measured from this table? The strategy is to study all possible pairs of children, and classify each pair as an agreement or disagreement in the following way. Let us focus on the particular pair of children, Adam and Bonnie. Let this pair constitute an agreement if:

a. Observer 1 classifies Adam and Bonnie into the same category, say a_i, and Observer 2 classifies Adam and Bonnie into the same category, say b_j, or

b. Observer 1 classifies Adam and Bonnie into different categories, and Observer 2 classifies Adam and Bonnie into different categories.

Any other situation constitutes a disagreement, as summarized in Table 3.

Table 3 about here

Given these definitions, the concepts of agreement and disagreement have a relatively simple interpretation in terms of the cell entries in Table 2. If the two persons in any given pair are in the same cell, or they are in neither the same row nor the same column, then that pair constitutes an "agreement". Using this idea, and remembering that any table will have $\binom{N}{2}$ possible pairs, the total number of observed agreements in any table will reduce to:

$$(1) \qquad A' = \binom{N}{2} + \sum_{i=1}^{R} \sum_{j=1}^{C} n_{ij}^{2} - \frac{1}{2}\left(\sum_{i=1}^{R} n_{i+}^{2} + \sum_{j=1}^{C} n_{+j}^{2} \right) .$$

The Expected Number of Agreements

For any two observers, we wish to examine the observed number of agreements from (1), and compare this number to the number expected from "chance" agreement. Thus, we test the null hypothesis H_0: $A = E(A')$ against the one-tailed alternative hypothesis H_1: $A > E(A')$, where A is the population parameter. Let us now turn to the development of E(A').

The expected number of agreements for any observed set of data depends upon the cell entries, n_{ij}, of a table such as Table 2. But the distribution of the n_{ij} depends upon whether the marginal totals of the table are assumed to be fixed (hypergeometric model) or variable (multinomial model). We will take the marginals to be fixed, although Kendall and Stuart (1967) point out that for large samples both assumptions lead to the same large sample distribution for cell entries.

Therefore, to find the expected number of agreements in any given table, we take the expected value of A' from (1). This becomes:

$$(2) \qquad E(A') = \binom{N}{2} - \frac{1}{2}\Big(\sum_i n_{i+}^2 + \sum_j n_{+j}^2\Big) + \sum_i \sum_j \Big[{}_{i,j}\mu'_{[2]} + {}_{i,j}\mu'_{[1]}\Big] .$$

In equation (2),

$$(3) \qquad {}_{i,j}\mu'_{[r]} = E[n_{ij}(n_{ij}-1)\ldots(n_{ij}-r+1)] = \frac{n_{i+}^{[r]}\, n_{+j}^{[r]}}{N^{[r]}}$$

where, in general, $n^{[r]} = n(n-1)\ldots(n-r+1)$. (Kendall and Stuart, 1967.)

Finding the Variance of A'

We now develop the exact variance for the agreement statistic. Although the formulae to follow appear tedious, they are straightforward to apply. The variance of A' can be expressed generally as:

$$(4) \qquad Var(A') = \sum_i^R \sum_j^C V(n_{ij}^2) + \sum_i^R \sum_j^C \sum_k^R \sum_\ell^C Cov(n_{ij}^2, n_{k\ell}^2) \; , \quad i \neq k \text{ and } j \neq \ell ,$$

where the subscripts k and ℓ are alternative row and column subscripts respectively. To find the first term in (4), we take:

$$(5) \qquad Var(n_{ij}^2) = E(n_{ij}^4) - [E(n_{ij}^2)]^2$$

where, in terms of the notation introduced in (3), (6a) $$E(n_{ij}^{4}) = \mu'_{i,j}[4] + 6\,\mu'_{i,j}[3] + 7\,\mu'_{i,j}[2] + \mu'_{i,j}[1]$$

and

(6b) $$E(n_{ij}^{2}) = \mu'_{i,j}[2] + \mu'_{i,j}[1] \,.$$

Finding the second term of (4) requires breaking the overall covariance into three parts. The three parts are:

(7a) $$\mathrm{Cov}(n_{ij}^{2}, n_{i\ell}^{2}) = E(n_{ij}^{2}, n_{i\ell}^{2}) - E(n_{ij}^{2})E(n_{i\ell}^{2}) \quad \text{where} \quad j \neq \ell \;;$$

(7b) $$\mathrm{Cov}(n_{ij}^{2}, n_{kj}^{2}) = E(n_{ij}^{2}, n_{kj}^{2}) - E(n_{ij}^{2})E(n_{kj}^{2}) \quad \text{where} \quad i \neq k \;; \text{ and}$$

(7c) $$\mathrm{Cov}(n_{ij}^{2}, n_{k\ell}^{2}) = E(n_{ij}^{2}, n_{k\ell}^{2}) - E(n_{ij}^{2})E(n_{k\ell}^{2}) \quad \text{where} \quad i \neq k \text{ and } j \neq \ell.$$

The factors in the last term in (7a, b, and c) can be found using the format given in (6b) above. The computing formulae for the first terms in (7a, b, and c) are given in Table 4.

Table 4 about here

An Approximate Test of Significance

The null hypothesis of random agreement can be tested using the statistic $Z_{A'}$:

(8) $$Z_{A'} = \frac{A' - E(A')}{\sqrt{\mathrm{Var}(A')}}$$

which for large R, C, and N has a standard normal distribution.

An Empirical Investigation of the Normal Approximation

In theory, the approximation Z_A given in (8) assumes large R, C, and N. However, this type of normal approximation is frequently used with very modest sample sizes. We therefore undertook an empirical investigation to examine the validity of our approximation for several small tables. Specifically, we chose eight 3 × 3 tables, one each with an N of 15, 30, and 51, and five with N's of 42. The results are presented in Table 5.

Table 5 about here

For each of the eight cases in Table 5, we generated by computer the exact probability distribution of the test statistic A'. Then, using the results of (2) and (4) above, we found the 0.05 and 0.01 cutoff values for A' in terms of the normal approximation proposed in (8). Finally, we found what proportion of the area under the exact distribution fell in the tail beyond these cutoff points. These results are given in Table 5.

To illustrate how Table 5 was constructed, let us focus on the first row. Here we have a table where each of the three rows and each of the three columns has a marginal total of 5. The next three columns of Table 5 reference the normal approximation at the 0.05 level of significance. Since $E(A') = 62.14$ and $V(A') = 20.41$, and $Z_\alpha = 1.65$ at $\alpha = 0.05$, the estimated cutoff point for A' from the normal approximation is $62.14 + 1.65\sqrt{20.41} = 69.55$. Since the exact distribution of A' is discrete, we choose the next highest value in the exact distribution. This value, 71.00, cuts off the top 0.064 of the exact distribution of A'. Similarly, for the 0.01 level of significance, the normal approximation cutoff value of A' equals 72.66, which corresponds to a value of 75.00 in the exact distribution. The tail area in the exact distribution beyond 75.00 is 0.016.

Three conclusions emerge from Table 5. First, for moderately small sample sizes, and for tables of small dimensionality (i.e., R = C = 3), the normal approximation to the distribution of A' is consistently quite good. Second, for tables with constant R = C = 3, increasing the sample size N has no dramatic effect on the quality of the normal approximation. This confirms the results expected from asymptotic normal theory (Wilks, 1962), which are that in contingency tables such as ours, the validity of the normal approximation is more affected by R and C than by N. Third, we investigated the effects of asymmetry in the marginal totals for the five tables with N = 42. These results are given in cases 3 through 7, which indicate that the quality of the normal approximation is essentially unaffected by different degrees of skewness in the marginals. Overall, then, for any but the smallest tables, the normal approximation should provide a reasonable guide for a test of the null hypothesis of random agreement.

A Worked Example

In this concluding section, we illustrate the computations for measuring and testing agreement using the raw data from the children in Table 1. We begin by placing the raw data into the 3 × 3 contingency table given in Table 6. Notice that all the marginal totals n_{i+} and n_{+j} are 5, which follows from the data in Table 1.

Table 6 about here

Computing A' from (1):

$$A' = 105 + (16 + 0 + 1 + \ldots + 1) - 75 = 75 \ .$$

Computing E(A') from (2) requires us to find first ${}_{i,j}\mu'[1]$ and ${}_{i,j}\mu'[2]$:

$${}_{i,j}\mu'[1] = \frac{5 \cdot 5}{15} = 1.667 \quad \text{for all } i,j;$$

and

$${}_{i,j}\mu'[2] = \frac{5 \cdot 4 \cdot 5 \cdot 4}{15 \cdot 14} = 1.905 \quad \text{for all } i,j \ .$$

Now, finding E(A'):

$$E(A') = 105 - 75 + 9(1.667 + 1.905) = 62.143 \ .$$

We now perform the slightly lengther calculation of the variance of A' from (4). First, we need to find ${}_{i,j}\mu'[3]$ and ${}_{i,j}\mu'[4]$.

$${}_{i,j}\mu'_{[3]} = \frac{5\cdot4\cdot3\cdot5\cdot4\cdot3}{15\cdot14\cdot13} = 1.319 \qquad \text{for all } i,j \;;$$

and

$${}_{i,j}\mu'_{[4]} = \frac{5\cdot4\cdot3\cdot2\cdot5\cdot4\cdot3\cdot2}{15\cdot14\cdot13\cdot12} = 0.440 \qquad \text{for all } i,j \;.$$

Then, finding other needed terms:

$$E(n_{ij}^4) = 0.440 + 6(1.319) + 7(1.905) + 1.667 = 23.352$$

$$E(n_{ij}^2) = (1.667 + 1.905) = 3.571$$

$$\mathrm{Var}(n_{ij}^2) = 23.352 - 3.571 = 10.597$$

$$\sum_i \sum_j \mathrm{Var}(n_{ij}^2) = 9(10.597) = 95.369$$

From the first two computational formulae in Table 4:

$$E(n_{ij}^2, n_{i\ell}^2) = E(n_{ij}^2, n_{kj}^2) = \frac{5\cdot4\cdot5\cdot4}{15\cdot14} + \frac{5\cdot4\cdot3\cdot5\cdot5\cdot4}{15\cdot14\cdot13}$$

$$+ \frac{5\cdot4\cdot3\cdot5\cdot4\cdot5}{15\cdot14\cdot13} + \frac{5\cdot4\cdot3\cdot2\cdot5\cdot4\cdot5\cdot4}{15\cdot14\cdot13\cdot12} = 8.242 \;;$$

and therefore, from formulae (7a) and (7b):

$$\mathrm{Cov}(n_{ij}^2, n_{i\ell}^2) = \mathrm{Cov}(n_{ij}^2, n_{kj}^2) = 8.242 - (3.571)^2 = -4.513 \;.$$

Further, from the third formula in Table 4:

$$E(n_{ij}^2, n_{k\ell}^2) = \frac{5\cdot5\cdot5\cdot5}{15\cdot14} + \frac{5\cdot5\cdot5\cdot4\cdot5\cdot4}{15\cdot14\cdot13} + \frac{5\cdot4\cdot5\cdot4\cdot5\cdot5}{15\cdot14\cdot13}$$

$$+ \frac{5\cdot4\cdot5\cdot4\cdot5\cdot4\cdot5\cdot4}{15\cdot14\cdot13\cdot12} = 15.186 ;$$

and from (7c):

$$\mathrm{Cov}(n_{ij}^2, n_{k\ell}^2) = 15.186 - (3.571)^2 = 2.431 .$$

So, summing up all the covariance terms,

$$\sum_i \sum_j \sum_k \sum_\ell \mathrm{Gov}(n_{ij}^2, n_{k\ell}^2) = 36(-4.513) + 36(2.431) = -74.961 .$$

Finally, we find from (4) the variance of A':

$$\mathrm{Var}(A') = 95.369 - 74.961 = 20.408 .$$

These results permit us to test the null hypothesis of random agreement for our example. Using the test in (8):

$$Z_{A'} = \frac{75.000-62.143}{\sqrt{20.408}} = 2.846 .$$

Since the computed value of $Z_{A'}$ far exceeds the critical value of 1.65 for the 0.05 level of significance, we reject the null hypothesis for the data in Table 6, and conclude that agreement beyond chance exists between the two psychologists.

Generalizability of the Agreement Statistic

While the development in this paper has focused entirely on the problem of measuring agreement between two observers who categorize N people, the statistic A' has substantially broader application. We suggest four such points [redacted] here.

First, while our worked example involved placing people (young children) into categories, the procedure is applicable for any phenomena that can be nominally scaled. Thus, for example, two psychologists may have a list of N behaviors to be classified according to some unspecified set of psychological disorders. Alternatively, two reading specialists may be asked to classify a set of N words into subgroups according to common student misconceptions in meaning. Notice that in these two examples it was not people but rather phenomena that were being categorized.

Second, although in our worked example each dimension of the table represented one observer, there is no reason why the categorizations and the question of their agreement could not emerge from two groups working independently.

Third, a recurring issue in applied research involves comparing data from different studies (Light and Smith, 1971). For example, suppose two states independently arrive at different classification schemes for the same set of job titles. If one is interested in the extent to which these two schemes are consistent, or "agree", the measure A' and its test statistic are applicable.

Fourth, while in our worked example $R = C = 3$, it will frequently occur that $R \neq C$. Therefore, the formulas for measuring agreement are general, and not restricted to square tables.

Table 1

Data on Two Psychologists' Categorization of 15 Children

Psychologist 1

Subgroups

Primarily Interested in Athletics		Primarily Interested in Popularity		Primarily Interested in Scholarship	
Adam	(A)	Francis	(F)	Kathy	(K)
Bonnie	(B)	George	(G)	Larry	(L)
Claire	(C)	Harold	(H)	Michael	(M)
David	(D)	Ira	(I)	Nora	(N)
Edward	(E)	Jennifer	(J)	Oscar	(O)

Psychologist 2

Subgroups

Primarily Interested in Popularity		Primarily Interested in Athletics		No Clear Interests	
Adam	(A)	George	(G)	Edward	(E)
Bonnie	(B)	Kathy	(K)	Harold	(H)
Claire	(C)	Larry	(L)	Ira	(I)
David	(D)	Michael	(M)	Jennifer	(J)
Francis	(F)	Nora	(N)	Oscar	(O)

Table 2

Generalized R×C Contingency Table Format for Agreement Problem

Observer 2

		b_1	b_2	b_j	...	b_C	
	a_1	n_{11}	n_{12}	n_{1j}	...	n_{1C}	n_{1+}
	a_2	n_{21}	n_{22}	n_{2j}	...	n_{2C}	n_{2+}
Observer 1	a_i	n_{i1}	n_{i2}	n_{ij}	...	n_{iC}	n_{i+}
		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	a_R	n_{R1}	n_{R2}	n_{Rj}	...	n_{RC}	n_{R+}
		n_{+1}	n_{+2}	n_{+j}	...	n_{+C}	N

Note--that Using this format:

a) Rows and/or columns may be permuted with no loss of information.

b) It is not true that cells on the main diagonal represent agreement and cells off the main diagonal represent disagreement.

Table 3

Definitions of Agreement and Disagreement

		Observer 2	
	Classification of Any Pair	Same Category	Different Categories
Observer 1	Same Category	Agreement	Disagreement
	Different Categories	Disagreement	Agreement

Table 4

Computational Formulae for First Part of Covariance Terms in 7a, 7b, and 7c.

For (7a),

$$E(n_{ij}^{2}, n_{i\ell}^{2}) = \frac{1}{N^{[2]}}[n_{i+}^{[2]} n_{+j}^{[1]} n_{+\ell}^{[1]}] + \frac{1}{N^{[3]}}[n_{i+}^{[3]} n_{+j}^{[1]} n_{+\ell}^{[2]} + n_{i+}^{[3]} n_{+j}^{[2]} n_{+\ell}^{[1]}] + \frac{1}{N^{[4]}}[n_{i+}^{[4]} n_{+j}^{[2]} n_{+\ell}^{[2]}].$$

For (7b),

$$E(n_{ij}^{2}, n_{kj}^{2}) = \frac{1}{N^{[2]}}[n_{+j}^{[2]} n_{i+}^{[1]} n_{k+}^{[1]}] + \frac{1}{N^{[3]}}[n_{+j}^{[3]} n_{i+}^{[2]} n_{k+}^{[1]} + n_{+j}^{[3]} n_{i+}^{[1]} n_{k+}^{[2]}] + \frac{1}{N^{[4]}}[n_{+j}^{[4]} n_{i+}^{[2]} n_{k+}^{[2]}]$$

For (7c),

$$E(n_{ij}^{2}, n_{k\ell}^{2}) = \frac{1}{N^{[2]}}[n_{i+}^{[1]} n_{+j}^{[1]} n_{k+}^{[1]} n_{+\ell}^{[1]}] + \frac{1}{N^{[3]}}[n_{i+}^{[1]} n_{+j}^{[1]} n_{k+}^{[2]} n_{+\ell}^{[2]} + n_{i+}^{[2]} n_{+j}^{[2]} n_{k+}^{[1]} n_{+\ell}^{[1]}]$$

$$+ \frac{1}{N^{[4]}}[n_{i+}^{[2]} n_{+j}^{[2]} n_{k+}^{[2]} n_{+\ell}^{[2]}].$$

Table 5

Comparison of Normal Approximation for A' to Exact Distribution of A'

				Cutoff Value of A' for $\alpha = 0.05$			Cutoff Value of A' for $\alpha = 0.01$		
		Marginals							
Case Number	N	Row	Column	Exact Distribution	Normal Approximation	$\alpha_{A'}$	Exact Distribution	Normal Approximation	$\alpha_{A'}$
1	15	5,5,5	5,5,5	71.00	69.57	.064	75.00	72.66	.016
2	30	10,10,10	10,10,10	267.00	264.01	.067	273.00	270.32	.028
3	42	14,14,14	14,14,14	513.00	509.56	.065	519.00	518.44	.035
4	42	14,14,14	2,5,35	378.00	376.26	.085	386.00	380.39	.014
5	42	10,14,18	7,12,23	484.00	482.86	.073	494.00	492.24	.029
6	42	4,10,28	7,12,23	468.00	467.29	.086	484.00	483.41	.027
7	42	10,14,18	10,14,18	499.00	498.56	.072	509.00	507.50	.030
8	51	17,17,17	17,17,17	747.00	746.21	.078	759.00	757.03	.031

Table 6

Data from Table 1 Reformulated in a Contingency Table Format to Measure Agreement

		Observer 2			
		Popularity	Athletics	None	
Observer 1	Athletics	A,B,C,D (4)	(0)	E (1)	5
	Popularity	F (1)	G (1)	H,I,J (3)	5
	Scholarship	(0)	K,L,M,N (4)	O (1)	5
		5	5	5	15

References

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

Cohen, J., "Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin*, 1968, *70*, 213-220.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N., *The Dependability of Behavioral Measurements*, New York: Wiley, 1972.

Everitt, B. S., "Moments of the statistics K and weighted Kappa," *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 97-103.

Fleiss, J. L., "Estimating the accuracy of dichotomous judgments," *Psychometrika*, 1965, *30*, 469-479.

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

Fleiss, J. L., Cohen, J., and Everitt, B. S., "Large sample standard errors of Kappa and weighted Kappa," *Psychological Bulletin*, 1969, *72*, 323-327.

Kendall, M. G., and Stuart, A., *The Advanced Theory of Statistics*, 2nd edition, Volume II, Hafner, 1967.

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

Light, R. J. and Smith, P. V., "Accumulating evidence: Strategies for resolving contradictions among different research studies," Harvard Educational Review, 1971, 41, 429-471.

Wilks, S. S., Mathematical Statistics, New York: Wiley, 1962.