This research was partially supported by Research Grants GM-00038-20 and GM-70004-04 from the National Institute of General Medical Sciences and by the U. S. Bureau of the Census through Joint Statistical Agreement JSA 74-2.

A REVIEW OF STATISTICAL METHODS IN THE ANALYSIS OF DATA ARISING FROM OBSERVER RELIABILITY STUDIES

By

J. Richard Landis and Gary G. Koch

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 956

OCTOBER 1974

TABLE OF CONTENTS

A REVIEW OF STATISTICAL METHODS IN THE ANALYSIS OF DATA ARISING FROM OBSERVER RELIABILITY STUDIES

J. Richard Landis and Gary G. Koch
Department of Biostatistics
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27514

1. INTRODUCTION

For many years now, researchers in medicine and epidemiology, in psychiatric diagnosis, in psychological measurement and testing, and in sample surveys have been aware of the observer (rater or interviewer) as an important source of measurement error. Fletcher and Oldham [1964] list a bibliography on observer error and variation which includes more than 70 papers alone in the area of clinical studies, radiology, pathology and clinical chemistry, and related fields. In one of these papers, Cochrane et al. [1951] indicate that there appeared to have been a general reluctance to recognize observer error in medical judgment situations, and even more hesitation to analyze this observer variability in some rigorous, quantitative manner. However, during the period 1940-1960, researchers in many disciplines began reporting studies which indicated the importance of assessing the variation in measurements due to different observers.

One prominent area in medical practice which gave rise to increased attention to observer error was chest radiography. It soon became appar-

ent in the late 1940's that inter-observer error (the inconsistency of interpretations among different observers) and intra-observer error (the failure of an observer to be consistent with himself in independent readings of the same film) were factors in medical diagnosis that required serious consideration. Birkelo, et al. [1947] reported a study which involved four different types of films on each of 1256 persons. These films were interpreted independently by two radiologists and three chest specialists. The observer variation in film interpretation was of such large magnitude that they recommended an extensive and detailed investigation of the problem. In referring to this study, Yerushalmy [1947] addressed himself to a basic problem in diagnosis by multiple observers, viz. the inability to compare the diagnosis by each observer with a known, standard conclusion. That is, it is not known who in the sample is positive for the disease, and who is negative. To deal with this difficulty, he recommended a group opinion procedure to determine which cases are really positive.

Since dual reading in mass radiography was recommended by these studies mentioned previously, investigators continued further work in determining the effectiveness of reading films twice. Fletcher and Oldham [1949] reported marked disagreement among ten observers in diagnosing pneumoconiosis from films. Yerushalmy, et al. [1950] also found a disturbing amount of disagreement among six readers independently grading 1807 photofluorograms. In addition, Yerushalmy, et al. [1951] reported significant disagreement among observers in studying the progression of pulmonary tuberculosis in a set of 150 film pairs obtained three months apart on the same individuals.

Besides these examples of reported observer disagreement in the

reading of films for diagnostic purposes, serious inconsistencies in the reporting of signs in physical exams were cited by other workers. Comroe and Botelho [1947] found extreme variability in observers noting cyanosis in different patients with the same arterial oxygen saturation. He reported that 25% of the observers did not detect definite cyanosis even at arterial saturation levels of 70-75%, a point at which diagnosis should have been consistent. In studying the reliability of observing various symptoms related to emphysema, Fletcher [1952] concluded that the clinical diagnosis could not be made confidently by a single observer, except perhaps in the most advanced stages of the disease, because of the high level of observer disagreement. Pyke [1954] arrived at the same conclusion in an investigation of finger clubbing. In this study 16 observers were asked to examine 12 patients and indicate whether fingers were clubbed or not. The results ranged from complete agreement on some patients to total disagreement (8 yes and 8 no) on one patient. In an epidemiological study of respiratory disease among male cotton workers in England, Schilling, et al. [1955] reported disagreement between two observers in diagnosing byssinosis in 24% of the cotton workers. In addition, there was significant disagreement between the observers in their diagnosis of other clinical signs. In a similar report, Smyllie, et al. [1965] investigated the level of observer disagreement in twenty respiratory signs in patients with various diseases. Butterworth and Reppert [1960] reported the results of testing 933 physicians in auscultatory diagnosis. Fifteen "unknowns", which had been classified by cardiac specialists, were presented to each physician simultaneously by tape recording and oscilloscope. The average number of correct diagnoses was only 49%, pointing out the need for

continuing training and education in auscultation of the heart. Other papers of interest which report similar variability among observers recording physical signs include those of Cochrane, et al. [1951], Reynolds [1952], and Fletcher [1964]. This brief summary in no way purports to cover the enormous body of literature dealing with observer variability in medical diagnosis but does serve to indicate that investigations of the incidence and severity of a disease in a community do depend on the reliability of the clinicians taking histories and observing physical signs.

Since large scale, multi-center trials have become more of a prominent type of clinical research in medicine, epidemiology, and public health in recent years, the detection, quantification, and control of observer variability have similarly become increasingly more crucial. This involves consistent usage of eligibility criteria for patients in different centers, uniform protocol adherence in the separate clinics, and reliable detection of the various responses under study. Large observer disagreement in any of these phases of the collaborative research may alter the effectiveness of the study, and in some cases could completely distort the results. Examples of situations in clinical trials which require inter-observer and inter-laboratory agreement include laboratory measurements of blood values such as serum cholesterol level in lipids research, and reliable HL-A tissue typing of donor kidneys in renal transplantation.

Another area involving observer variability is the estimation of the precision of measuring instruments for industrial or laboratory usage. Grubbs [1948] considers the situation in which a test is destructive so that on any given piece of material only one test can be performed, but

it can be observed simultaneously by several instruments. The problem is to obtain separate estimates of the variances of the instrument error and of the variability of the material under study. Smith [1950] also considered the problem of determining the precision of two instruments in measuring an item where there can be no theoretically absolute values against which the instruments can be calibrated. Other papers dealing with reliability in the measuring process include Mandel [1959], Thompson [1963] and Hahn and Nelson [1970].

Since rating scales have been widely used in psychiatric research and in psychological and educational testing, the need for determining the scale reliability is well known and has been discussed in terms of test-retest reliability by Guttman [1946]. However, Overall [1968] pointed out that reliability testing has been primarily concerned with the rating instrument without adequate attention to the fact that the reliability of the scores depends on the skill of the rater or observer. Burdock, et al. [1963] indicated the importance of assessing observer variability in using an instrument developed for recording critical aspects of the behavior of mental patients observed under different situations. In an educational research application, Ebel [1951] described the situation in which the agreement among raters was considered as a criterion for validating a selection procedure for admitting graduate students into a physics research program. Also, Fleiss [1966] pointed out that much of social science research employs scores or ratings assigned to subjects by interviewers or observers. Other situations in psychiatric research in which observer agreement is considered are reported by Fleiss, et al. [1965], Sandifer, et al. [1964], and Fleiss, et al. [1972].

As mentioned previously, interviewer error is an important consideration in sample survey situations. Glasser, et al. [1970] reported the results of a telephone coincidental survey designed to determine which, if any, television program was being viewed immediately prior to the receipt of the telephone call. Since this survey involved many interviewers, possible bias and variability was an important concern. Kish [1962] showed that in sample survey situations, even when the variability attributed to interviewers was low (eg., 0.07 of the total variation), the variance of the mean can be significantly increased. Koch [1971a] developed models analogous to those of the U. S. Bureau of the Census which allow for effects which reflect the different values expected for the response of a specific population element when obtained by different interviewers.

Even though the effects of the observer in the measurement situations reviewed here must be assessed, there does not seem to be any unified body of statistical methodology developed for this complex problem. Instead, as reviewed in the following sections, researchers in these various disciplines have employed different procedures for their specific problems, and in many cases have not utilized similar developments in related areas.

2. MODELS PROPOSED FOR NUMERICAL DATA

When the scores assigned by observers are quantitative measurements or numerical data such as height, weight, or serum cholesterol level, a suitable analysis of variance model can be used to assess observer variation. The relevant theory of standard variance components models is given by Anderson and Bancroft [1952], Scheffé [1959], and Searle [1971].

These models arise in situations in which the same set of subjects is measured independently by several observers. This type of a reliability study is used to determine the reliability of individual measurements, and to assess the level of observer variability or bias.

2.1. One Characteristic; One Trial

The most elementary reliability testing situation involves each of a observers independently measuring the specified characteristic once on each of the same set of n subjects. The initial model proposed by Ebel [1951] for this design can be rewritten in standard form as

$$y_{ij} = \mu + s_i + e_{ij} \tag{2.1.1}$$

for i = 1,2,...,n and j = 1,2,...,a. Here

y_{ij} = the score assigned to the i-th subject by the j-th observer,

μ = an overall mean response,

s_i = the i-th subject effect,

e_{ij} = the residual error for the (i,j)-th observation.

Assuming that the n subjects are a random sample from a population of interest, the additional assumptions necessary to make standard tests of hypotheses are $\{s_i\}$ are NID$(0,\sigma_s^2)$, $\{e_{ij}\}$ are NID$(0,\sigma_e^2)$, where NID means normal, independent, and identically distributed. The sets $\{s_i\}$ and $\{e_{ij}\}$ are also mutually independent.

A standard measure studied in this situation is the intraclass correlation coefficient between any two scores assigned to the same subject, denoted by $\tilde{\rho}$. From these assumptions, $\tilde{\rho}$ has been shown to be

$$\tilde{\rho} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} . \tag{2.1.2}$$

Using the usual analysis of variance (ANOVA) calculations, where

$$MS_s = \text{mean square for subjects,}$$

$$\text{and } MS_e = \text{mean square for residual error,}$$

the sample estimate of $\tilde{\rho}$, given by

$$\hat{\tilde{\rho}} = \frac{MS_s - MS_e}{MS_s + (a-1)MS_e}, \tag{2.1.3}$$

is called the reliability of individual ratings. This index is indicative of the reliability of the measuring procedure or the intrinsic accuracy of the instrument being used. As such, $\hat{\tilde{\rho}}$ does not yield specific information about differences among observers, but it is of considerable importance in assessing the accuracy of the measurement process, since it reflects the proportion of total variance which is due to the inherent variance attributable to the subjects.

However, since model (2.1.1) includes the "among observers" variance in the error term, the level of variation due to the observers cannot be estimated. Both Ebel [1951] and Burdock, et al. [1963] present models in which a variance component due to the observers is considered. To maintain similarity with the previous notation, this model can be written as

$$y_{ij} = \mu + s_i + b_j + e_{ij}, \tag{2.1.4}$$

where the additional parameter is

$$b_j = \text{the j-th observer effect.}$$

All the other parameters are the same as defined for model (2.1.1). At this point two possible assumptions for the observer effects need to be considered.

CASE I -- Random Observer Effects

If the a observers are assumed to be a random sample from a large population of potential observers of interest, then the usual random effects assumptions are that $\{b_j\}$ are NID$(0,\sigma_o^2)$, and $\{b_j\}$, $\{s_i\}$, and $\{e_{ij}\}$ are mutually independent. If

MS_o = mean square for observers,

then unbiased and consistent estimates of these variance components are given by

$$\hat{\sigma}_o^2 = (MS_o - MS_e)/n, \tag{2.1.5}$$

$$\hat{\sigma}_s^2 = (MS_s - MS_e)/a, \tag{2.1.6}$$

$$\hat{\sigma}_e^2 = MS_e. \tag{2.1.7}$$

In terms of this experimental situation, these components can be interpreted as follows:

i) $\hat{\sigma}_o^2$ is the estimate of the variance due to differences in the average judgment of observers. As such, it is an estimate of observer variability or bias;

ii) $\hat{\sigma}_s^2$ is the estimate of the variance due to differences in the responses of subjects over all observers. Thus, it is an estimate of subject variability;

iii) $\hat{\sigma}_e^2$ is the estimate of the observer reliability or measurement error.

As a result, the value for $\hat{\sigma}_o^2$ reflects the extent of the disagreement of the observers in their usage of the measurement scale. Also, the value for $\hat{\sigma}_e^2$ indicates the variability that is not accounted for by the main effects of observers and subjects. This would include the possible interaction of observers and subjects which is not estimable from model (2.1.4).

If the observer effects are assumed to be random, the intraclass

correlation coefficient of (2.1.2) is now given by

$$\tilde{\rho} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_o^2 + \sigma_e^2}. \qquad (2.1.8)$$

A sample estimate of $\tilde{\rho}$ can be obtained by substituting the respective estimates from (2.1.5), (2.1.6), and (2.1.7) into (2.1.8). This formulation of $\tilde{\rho}$ in (2.1.8) shows that the observer variance component reflecting observer bias has an inverse effect on the reliability measure. That is, if observer bias is large, $\tilde{\rho}$ will necessarily be reduced.

Another indication of the relative importance of σ_o^2 in the model can be obtained by tests of hypotheses. The standard test in the variance components model is

$$H_0: \ \sigma_o^2 = 0 \quad \text{vs.} \quad H_1: \ \sigma_o^2 > 0.$$

This can be performed by using the test statistic $F = MS_o/MS_e$, which is distributed as $\mathcal{F}$ with (a-1) and (a-1)(n-1) degrees of freedom under H_0. The failure to reject H_0 indicates that the observers are not exhibiting statistically significant disagreement in their assignment of scores to subjects. However, as pointed out by Burdock, et al. [1963], instead of testing whether the observer variance component is zero, which is typically of limited concern, it is usually of much more interest to test the modified hypothesis

$$H_0: \ \sigma_o^2 \leq \theta_o \sigma_e^2 \quad \text{vs.} \quad H_1: \ \sigma_o^2 > \theta_o \sigma_e^2,$$

which tests whether σ_o^2 is less than an acceptable proportion of σ_e^2, the measurement error variance component. This reduces to the problem of determining a reasonable value for θ_o, and then forming a confidence interval for the ratio $\theta = \sigma_o^2/\sigma_e^2$.

CASE II -- Fixed Observer Effects

Other reliability studies may involve situations in which the a observers are a fixed set which will be used in later experiments, and are the only observers of interest. In this case, the assumptions of the model given in (2.1.4) need to be revised to those of the usual mixed model analysis of variance. Here we have

$$y_{ij} = \mu + s_i + \beta_j + e_{ij} \qquad (2.1.9)$$

for $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,a$, where β_j are fixed constants under the constraint

$$\sum_{j=1}^{a} \beta_j = 0.$$

In this case, $\{\beta_j\}$ are the fixed observer effects. All the other assumptions given for model (2.1.4) apply here also. Under this mixed model the measure of reliability of ratings, $\tilde{\rho}$, is the same as given in (2.1.2), and the test for observer bias is given by

$$H_0: \ \beta_j = 0, \ j = 1,2,\ldots,a,$$

which can be tested by using the same F statistic cited for the random effects model.

2.2. One Characteristic; r Trials

In the situations discussed in section 2.1 each observer measured the specified characteristic exactly one time on each subject. As a result, none of the proposed models provides estimates for "within-observer" variance or interaction between observers and subjects. However, Burdock, et al. [1963] also consider the situation in which a observers each independently make r independent measurements of the same specified characteristic on each of the same n subjects.

CASE I -- Random Observer Effects

The random effects model can then be written as

$$y_{ijk} = \mu + s_i + b_j + (sb)_{ij} + e_{ijk} \qquad (2.2.1)$$

for i = 1,2,...,n; j = 1,2,...,a; and k = 1,2,...,r. The additional parameter in this model compared to that of (2.1.4) is

$(sb)_{ij}$ = the interaction of the j-th observer and the i-th subject.

Again, the random effects assumptions are that $\{s_i\}$ are NID$(0,\sigma_s^2)$, $\{b_j\}$ are NID$(0,\sigma_o^2)$, $\{(sb)_{ij}\}$ are NID$(0,\sigma_{os}^2)$ and $\{e_{ijk}\}$ are NID$(0,\sigma_e^2)$. Also, these sets of parameters are mutually independent. From the expected mean squares resulting from the usual ANOVA calculations, it can be demonstrated that unbiased and consistent estimates of these variance components are given by

$$\hat{\sigma}_o^2 = (MS_o - MS_{os})/nr, \qquad (2.2.2)$$

$$\hat{\sigma}_s^2 = (MS_s - MS_{os})/ar, \qquad (2.2.3)$$

$$\hat{\sigma}_{os}^2 = (MS_{os} - MS_e)/r, \qquad (2.2.4)$$

$$\hat{\sigma}_e^2 = MS_e. \qquad (2.2.5)$$

The interpretations given previously in section 2.1 for these components still apply here. In addition, the interaction effect $(sb)_{ij}$ measures the extent to which observer j departs from his usual pattern when measuring subject i. As a result, $\hat{\sigma}_{os}^2$ reflects how observers vary in their overall rating of the same subject. The statistical significance of this effect in the model can be tested by

$$H_0: \sigma_{os}^2 = 0 \quad \text{vs.} \quad H_1: \sigma_{os}^2 > 0$$

using the test statistic $F = MS_{os}/MS_e$, which is distributed as $\mathcal{F}$ with

$(a-1)(n-1)$ and $(r-1)$an degrees of freedom under H_0.

Under these assumptions, the intraclass correlation coefficient becomes

$$\tilde{\rho} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_o^2 + \sigma_{os}^2 + \sigma_e^2}, \qquad (2.2.6)$$

with the same interpretation of reliability already given in section 2.1.

CASE II -- Fixed Observer Effects

If the observer effects are assumed to be fixed, the usual ANOVA mixed model is

$$y_{ijk} = \mu + s_i + \beta_j + (s\beta)_{ij} + e_{ijk}, \qquad (2.2.7)$$

where the β_j are fixed constants under the constraint

$$\sum_{j=1}^{a} \beta_j = 0,$$

and for each $i = 1,2,\ldots,n$, $(s\beta)_{ij}$ are fixed constants under the constraint

$$\sum_{j=1}^{a} (s\beta)_{ij} = 0.$$

The assumptions on the other parameters are the same as those given for (2.2.1). In the mixed model situation, an intraclass correlation coefficient may be expressed by

$$\tilde{\rho} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_{os}^2 + \sigma_e^2}, \qquad (2.2.8)$$

as the observer effects are assumed fixed.

An example using an ANOVA mixed model is given by Bearman, et al. [1964] in a study of variability in tuberculin test reading. In this study the measurement of the transverse diameter of the indurated

area was recorded in millimeters for each reading. Each reaction was read four times on each subject by each of four observers in a completely blind fashion. Even though all the F ratios are reported to be significant at p = .0005, the dominant source of variation was shown to be attributable to the subjects. From the fixed observers assumption, the estimate of the reliability coefficient using (2.2.8) is given by $\hat{\tilde{\rho}} = 0.786$, or by making the assumptions for the random effects model the estimate is $\hat{\tilde{\rho}} = 0.724$ using (2.2.6). In either case, even though the observer variability or bias is significant, the reliability is approximately 75%.

2.3. Modifications of ANOVA Models

To this point, all the proposed models have assumed that the variance component due to random measurement error, σ_e^2, is constant across all observers. In many cases this may be an unwarranted assumption. Overall [1968] discusses the problem of estimating a separate measurement error variance component for each observer using an experimental design in which such estimates cannot be computed directly from the usual ANOVA calculations. He specifically considers the situation in which n subjects are randomly assigned to k treatments, and are subsequently rated independently by two judges. For each rater, the total score variability, $\sigma_{x_j}^2$, is assumed to be

$$\sigma_{x_j}^2 = \sigma_{\alpha_j}^2 + \sigma_{w_j}^2, \tag{2.3.1}$$

where

$$\sigma_{w_j}^2 = \sigma_{s_j}^2 + \sigma_{e_j}^2, \; j = 1,2 . \tag{2.3.2}$$

Here

$\sigma_{\alpha_j}^2$ = the variance component of the treatment effects for the j-th observer,

$\sigma_{s_j}^2$ = the variance component of the subject effects for the j-th observer,

$\sigma_{e_j}^2$ = the measurement error variance component for the j-th observer.

By performing a separate analysis of variance on the measurements recorded by each rater, the usual estimates of $\sigma_{w_j}^2$ and $\sigma_{\alpha_j}^2$ can be obtained by equating

$$\hat{\sigma}_{w_j}^2 = MS_{w_j}, \tag{2.3.3}$$

$$\text{and } \hat{\sigma}_{\alpha_j}^2 = (MS_{A_j} - MS_{w_j})/m, \tag{2.3.4}$$

assuming $\underline{m}$ subjects within each treatment group. Now given that each rater is unaware of the treatment group of each subject, Overall assumes that

$$\sigma_s^2 = c\sigma_{\alpha_.}^2. \tag{2.3.5}$$

This assumption simply requires that of the total subject differences recorded by different raters, the proportion of variance due to treatment effects is constant. In addition, the inter-rater correlation is shown to be

$$r_{12} = \frac{\sigma_{s_1}\sigma_{s_2}}{\sigma_{w_1}\sigma_{w_2}} = \frac{c\,\sigma_{\alpha_1}\sigma_{\alpha_2}}{\sigma_{w_1}\sigma_{w_2}}, \tag{2.3.6}$$

by including assumption (2.3.5). By using the usual product moment correlation coefficient as the estimate of r_{12}, the solution for c is obtained by

$$\hat{c} = \frac{\hat{r}_{12}\,\hat{\sigma}_{w_1}\,\hat{\sigma}_{w_2}}{\hat{\sigma}_{\alpha_1}\,\hat{\sigma}_{\alpha_2}}\,, \tag{2.3.7}$$

substituting the estimates of (2.3.3) and (2.3.4). Thus, using (2.3.2), the separate estimates of measurement error components are given by

$$\hat{\sigma}_{e_j}^{\,2} = \hat{\sigma}_{w_j}^{\,2} - \hat{c}\,\hat{\sigma}_{\alpha_j}^{\,2}\,. \tag{2.3.8}$$

In addition, separate reliability measures for each rater using the intraclass correlation coefficient estimates are given by

$$r_{ii} = \frac{c\sigma_{\alpha_i}^{\,2}}{c\sigma_{\alpha_i}^{\,2} + \sigma_{e_i}^{\,2}}\,,\ i = 1,2. \tag{2.3.9}$$

Using the previously defined estimates, these are given by

$$\hat{r}_{11} = \hat{r}_{12}\,\frac{\hat{\sigma}_{w_2}\,\hat{\sigma}_{\alpha_1}}{\hat{\sigma}_{w_1}\,\hat{\sigma}_{\alpha_2}}$$

and (2.3.10)

$$\hat{r}_{22} = \hat{r}_{12}\,\frac{\hat{\sigma}_{w_1}\,\hat{\sigma}_{\alpha_2}}{\hat{\sigma}_{w_2}\,\hat{\sigma}_{\alpha_1}}\,.$$

The variance components models which include parameters to reflect observer variability are similar, and in some cases are identical to measurement error models developed for estimating precision of instrumentation in laboratories and industrial plants. In an early paper in this area, Grubbs [1948] attempts to separate the variance of the products being measured from the variance of the measurement instruments. In the case with two instruments, I_1 and I_2, he proposes the model

$$y_{ij} = x_i + e_{ij} \tag{2.3.11}$$

for i = 1,2,...,n items each being measured once by each instrument, indexed by j = 1,2. Here x_i is considered to be the unknown true value of the i-th item and e_{ij} is the measurement error using the j-th instrument. The assumptions are that the $\{x_i\}$ are mutually independent of the $\{e_{ij}\}$, and that the $\{e_{i1}\}$ are mutually independent of the $\{e_{i'2}\}, \forall i,i'$. As a result, the variance terms can be written as

$$\text{Var}\{y_{i1}\} = \sigma_x^2 + \sigma_{e_1}^2, \tag{2.3.12}$$

and

$$\text{Var}\{y_{i2}\} = \sigma_x^2 + \sigma_{e_2}^2. \tag{2.3.13}$$

This formulation allows for a separate variance component, $\sigma_{e_j}^2$, for the measurement error of each instrument, and for σ_x^2, the variance of the products being measured. From the independence assumptions, the covariance term is

$$\text{Cov}\{y_{i1},y_{i2}\} = \sigma_x^2. \tag{2.3.14}$$

By defining

$$S_{12} = \frac{1}{n-1} \sum_{i=1}^{n} (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2), \tag{2.3.15}$$

and

$$S_{jj} = \frac{1}{n-1} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2, \; j = 1,2, \tag{2.3.16}$$

unbiased estimates of these variance components can be obtained by setting

$$\hat{\sigma}_{e_1}^2 = S_{11} - S_{12}, \tag{2.3.17}$$

$$\hat{\sigma}_{e_2}^2 = S_{22} - S_{12}, \tag{2.3.18}$$

$$\hat{\sigma}_x^2 = S_{12}. \tag{2.3.19}$$

In addition, by making normality assumptions for the model of (2.3.11), the variances of the estimated components are given in terms of the unknown variance components.

These results are then extended to the case with $\underline{N}$ instruments in a straightforward manner by Grubbs. The analogous model for this situation is

$$y_{ij} = x_i + e_{ij} \tag{2.3.20}$$

for $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,N$ instruments. Under the assumptions of independent errors, the variances and covariances are

$$\text{Var}\{y_{ij}\} = \sigma_x^2 + \sigma_{e_j}^2 \quad , \; j = 1,2,\ldots,N, \tag{2.3.21}$$

and

$$\text{Cov}\{y_{ij}, y_{ij'}\} = \sigma_x^2 \quad , \; j \neq j'. \tag{2.3.22}$$

By a similar procedure, he shows that an unbiased estimate of $\sigma_{e_1}^2$ is given by

$$\sigma_{e_1}^2 = S_{11} - \frac{2}{N-1} \sum_{j=2}^{N} S_{1j} + \frac{2}{(N-1)(N-2)} \sum_{2 \leq j < k}^{N} S_{jk}. \tag{2.3.23}$$

The other measurement error components are estimated in an analogous way by an obvious choice of the proper subscripts in (2.3.23). In addition, the product variability is estimated by the sample average of the covariances as

$$\hat{\sigma}_x^2 = \frac{2}{N(N-1)} \sum_{1 \leq i < j}^{N} S_{ij}. \tag{2.3.24}$$

In terms of our previous notation, these models developed by Grubbs [1948] can be rewritten as

$$y_{ij} = \mu + s_i + \beta_j + e_{ij} \qquad (2.3.25)$$

for i = 1,2,...,n and j = 1,2,...,a instruments. In this case,

y_{ij} = the value of the i-th item using the j-th instrument,

μ = an overall mean response,

s_i = the i-th item effect,

β_j = the bias of the j-th instrument,

e_{ij} = the residual error for the (i,j)-th observation.

This is directly analogous to the mixed model given in (2.1.9). However, the more general assumption on the error components is that $\{e_{ij}\}$ are NID$(0,\sigma_{e_j}^2)$, for i = 1,2,...,n. As a result, the variance-covariance structure for this model can be displayed in an $\underline{a}$ x $\underline{a}$ matrix $\underset{\sim}{V}$ as

$$\underset{\sim}{V} = \text{diag}(\sigma_{e_j}^2)\ \underset{\sim}{I}_{\underline{a}} + \sigma_x^2\ \underset{\sim}{J}_{\underline{a}}$$

$$= \begin{bmatrix} \sigma_{e_1}^2+\sigma_x^2 & \sigma_x^2 & \cdots & \sigma_x^2 \\ \sigma_x^2 & \sigma_{e_2}^2+\sigma_x^2 & \cdots & \sigma_x^2 \\ \vdots & & \cdots & \\ \sigma_x^2 & \sigma_x^2 & \cdots & \sigma_{e_a}^2+\sigma_x^2 \end{bmatrix}, \qquad (2.3.26)$$

where $\underset{\sim}{I}_{\underline{a}}$ is the $\underline{a}$ x $\underline{a}$ identity matrix and $\underset{\sim}{J}_{\underline{a}}$ is an $\underline{a}$ x $\underline{a}$ matrix with all elements equal to 1. This demonstrates that all the covariances among the $\underline{a}$ instruments are σ_x^2, and thus indicates the rationale for the choice of the particular estimates of (2.3.23) and (2.3.24).

Smith [1950] considers the same basic problem of estimating the precision of measuring instruments. By assuming that the measurement scale of one instrument is linearly related to the scale of the other instrument, he develops estimates of these parameters in a manner

analogous to that of Grubbs [1948]. Other papers related to this type of precision estimation include those of Thompson [1963] and Hahn and Nelson [1970].

In a major paper describing the measuring process, Mandel [1959] discusses the nature of experimental error by means of a highly abstract formulation of the topic. Then in a testing situation involving n replicates of the same b materials being tested in a different laboratories he attempts to put the problem into a larger framework than the usual ANOVA model. By transforming the data, so that within-cell standard deviations across materials are relatively constant, he proposes the model

$$y_{ijk} = \mu_i + \beta_i(\xi_j - c) + \lambda_{ij} + \varepsilon_{ijk}, \qquad (2.3.27)$$

where,

μ_i = the overall mean in the i-th laboratory,

β_i = the slope for the linear effect in the i-th laboratory,

ξ_j = the mean across laboratories for material j,

c = the grand overall mean,

λ_{ij} = the interaction of laboratory i with material j,

ε_{ijk} = the within-cell random measurement error.

Note that this model assumes a linear trend among laboratories. By equating the appropriate expected mean squares from modified analysis of variance calculations, he demonstrates how the variance components for this model can be estimated.

In a recent series of reliability studies involving the scoring of the presence and severity of cerebrovascular atherosclerosis, Loewenson, et al. [1972] applied a modified mixed model ANOVA given by

$$y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ij} + \delta_{ijk}, \qquad (2.3.28)$$

where,

μ = overall mean,

τ_i = i-th coder effect,

β_j = j-th specimen effect,

ε_{ij} = experimental error, distributed $N(0,\sigma_\varepsilon^2)$,

δ_{ijk} = sampling error, distributed $N(0,\sigma_\delta^2)$.

From this framework, they estimated the difference in mean scores of pairs of coders, the experimental variability, the variability for repeated scoring, and separate variances for each of the coders.

2.4. Response Error Models in Sample Surveys

In the previous sections the proposed models were developed under the assumptions of numerical data, and in most cases, normal distributions. However, in many situations, particularly in estimation contexts, no distributional assumptions are required to permit the calculations of various model effects. In fact, the data need not necessarily be numerical. Much of the work in response error models in sample survey data involves the estimation of components of variation on 0-1 type data. Hansen, et al. [1964] present a model for the measurement of attribute data and the related measurement errors. In this model the total variance is partitioned into three components defined as response variance, sampling variance, and the interaction variance. The simple response variance, σ_{dG}^2, is defined as "the basic trial-to-trial variability in response, averaged over all the elements in the population." In situations in which the interaction variance can be assumed to be zero, the total variance can be written as

$$\sigma^2_{p_{tG}} = \sigma_{dG}^2 + \sigma_{eG}^2, \tag{2.4.1}$$

where, for simplicity, the simple random sampling variance is denoted by σ_{eG}^2. This result is then used to motivate an index of unreliability or inconsistency of classification as

$$I_{dG} = \frac{\sigma_{dG}^2}{\sigma_{p_{tG}}^2}. \tag{2.4.2}$$

As such, this index reflects the proportion of the total variance attributable to the simple response variance.

The relationship between the index of inconsistency, I_{dG}, and the reliability index or intraclass correlation coefficient of the previous ANOVA models in sections 2.2 and 2.3 can be demonstrated in the following manner. Let

Y_{jt} = the value assigned to the j-th element of the population on the t-th trial

in the sample survey framework. Then I_{dG} of (2.4.2) can be written as

$$I_{dG} = \frac{E\{Y_{jt} - \bar{Y}_{j\cdot}\}^2}{E\{Y_{jt} - \bar{Y}_{\cdot\cdot}\}^2} \tag{2.4.3}$$

By writing the usual ANOVA model of (2.1.4) for this situation as

$$Y_{jt} = \mu + s_j + b_t + e_{jt} \tag{2.4.4}$$

for $j = 1,2,\ldots,n$ elements in the sample and $t = 1,2,\ldots,a$ conceptually similar trials, the index of inconsistency is given by

$$I_{dG} = \frac{\sigma_t^2 + \sigma_e^2}{\sigma_s^2 + \sigma_t^2 + \sigma_e^2}, \tag{2.4.5}$$

where the trial effects $\{b_j\}$ are assumed to have variance σ_t^2. This is precisely the complement of the intraclass correlation coefficient for this model, since

$$\tilde{\rho} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_t^2 + \sigma_e^2} \quad . \tag{2.4.6}$$

As a result, the index of inconsistency can be written as

$$I_{dG} = 1 - \tilde{\rho} \tag{2.4.7}$$

If, instead of assuming that the sample is taken repeatedly over several trials, the separate trials represent different observers determining the specified attribute on the same sample of individuals, then the response error model index of inconsistency, I_{dG}, reflects the level of unreliability or inconsistency of the different observers.

In a study of interviewer variance for attitudinal variables, Kish [1962] also uses the intraclass correlation coefficient to determine the proportion of the total variance that can be attributed to the usage of several interviewers. He uses a variance components model directly analogous to that of (2.1.1) to obtain

$$\rho = \frac{s_a^2}{s_a^2 + s_b^2} \quad , \tag{2.4.8}$$

where s_a^2 corresponds to the measurement error component, σ_e^2, in the previous model of (2.1.1). As such, this is precisely $1 - \tilde{\rho}$ from (2.1.2), and thus reflects the level of unreliability or inconsistency of the interviewers. He then demonstrates the effect of interviewer variance on the sample mean by showing that

$$\text{Var}\{\bar{y}\} = \frac{s_a^2 + s_b^2}{n}\,[1 + \rho(n/a - 1)]. \qquad (2.4.9)$$

2.5. Related Multivariate Extensions

All the models to this point have been univariate, and hence restricted to the situation in which only one variable or attribute is measured on each subject or observation. However, since many situations involve the concurrent measurement of several variables, it is of considerable importance to be able to assess the level of observer variability in a multivariate reliability study.

Gleser, et al. [1965] and Maxwell and Pilliner [1968] consider the situation in which k judges independently measure each of m items on each of n persons. They present the usual three-factor, completely balanced, random effects ANOVA model to evaluate the reliability of these psychological measurements. This produces estimates of three main effects -- subjects, observers, and items, three two-way interactions, and residual error. By evaluating the relative size of the appropriate variance component estimates, the effect of the observers can be determined. However, this model does not account for the repeated measures or "split-plot" nature of the problem. Instead, it assumes a random sample within each of the subject x observer x item cells.

In a multivariate treatment of this problem, Fleiss [1966] considers a reliability study in which each of k observers assigns scores on p tests to each of a sample of subjects. He proposes a model which is a direct extension of the mixed model of (2.1.9). In this univariate model, under the normality assumption, the 1 x k vector $\mathbf{Y}' = (Y_1, \ldots, Y_k)$, where

Y_j = the score from the j-th observer,

is assumed to be distributed k-variate normal, viz. $MN_k(\underset{\sim}{\mu}, \underset{\sim}{\Sigma})$, where $\underset{\sim}{\mu}' = (\mu + \beta_1, \ldots, \mu + \beta_k)$ and $\underset{\sim}{\Sigma} = \sigma_e^2 \underset{\sim}{I_k} + \sigma_s^2 \underset{\sim}{J_k}$, using the notation of sections 2.2 and 2.3. Fleiss [1966] extends this model to the multivariate case by writing the observation on the i-th subject as

$$\underset{1 \times pk}{\underset{\sim}{Y}_i} = (\underset{\sim}{Y}_i^{(1)}, \underset{\sim}{Y}_i^{(2)}, \ldots, \underset{\sim}{Y}_i^{(k)}), \tag{2.5.1}$$

where $\underset{\sim}{Y}_i^{(j)}$ are the p component scores assigned by observer j. The multivariate model is given by

$$\underset{1 \times p}{\underset{\sim}{Y}_i^{(k)}} = \underset{\sim}{\mu}^{(k)} + \underset{\sim}{b}_i + \underset{\sim}{e}_i^{(k)}, \tag{2.5.2}$$

where

$\underset{\sim}{\mu}^{(k)}$ = the vector of mean scores assigned by the k-th observer,

$\underset{\sim}{b}_i$ = the random vector of effects for the i-th subject,

$\underset{\sim}{e}_i^{(k)}$ = the random vector of residual errors.

He then assumes that

$$\underset{\sim}{b}_i \sim MN_p(\underset{\sim}{0}, \underset{p \times p}{\underset{\sim}{B}}) \tag{2.5.3}$$

and

$$\underset{\sim}{e}_i^{(k)} \sim MN_p(\underset{\sim}{0}, \underset{p \times p}{\underset{\sim}{A} - \underset{\sim}{B}}). \tag{2.5.4}$$

From this, the variance-covariance structure for the 1 x pk vector of observations $\underset{\sim}{X}_i$ is given by

$$\underset{pk \times pk}{\underset{\sim}{\Sigma}_x} = (\underset{\sim}{A} - \underset{\sim}{B}) \otimes \underset{\sim}{I_k} + \underset{\sim}{B} \otimes \underset{\sim}{J_k}$$

$$= \begin{bmatrix} \underset{\sim}{A} & \underset{\sim}{B} & \cdots & \underset{\sim}{B} \\ \underset{\sim}{B} & \underset{\sim}{A} & \cdots & \underset{\sim}{B} \\ \vdots & & & \vdots \\ \underset{\sim}{B} & \underset{\sim}{B} & \cdots & \underset{\sim}{A} \end{bmatrix}, \tag{2.5.5}$$

where $\otimes$ is the Kronecker product for matrices.

Under these assumptions, he derives the likelihood ratio multivariate test criterion for testing H_0: $\underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)} = \ldots = \underset{\sim}{\mu}^{(k)}$, to assess the degree of observer agreement. In addition, as a single measure of reliability among the observers, he recommends

$$R_p = \frac{1}{p} \operatorname{trace}\{\underset{\sim\sim}{B}\underset{\sim\sim}{A}^{-1}\}. \tag{2.5.6}$$

In the special case when both $\underset{\sim}{A}$ and $\underset{\sim}{B}$ are diagonal, R_p is the arithmetic mean of the p univariate intraclass correlation coefficients $\tilde{\rho}$ of (2.1.2). Finally, an approximate $(1 - \alpha)100\%$ confidence interval for R_p is given by appealing to the appropriate χ^2 distributions.

In the sample survey context, Koch [1971a] presents a multivariate extension to the response error model discussed in section 2.4. This is essentially the p-variate analogue to the model in which the total variance is partitioned into the response variance, sampling variance, and the interaction variance. In another extension, Koch [1971b] develops a multivariate response error model with components of variation attributed to interviewers. In these multivariate response error models it is also of interest to determine the proportion of the total variance attributable to the simple response variance. In another paper, Bershad [1969] proposes an index of inconsistency for the multivariate classification scheme which is a direct extension of the index I_{dG} given in (2.4.3). Let

Y_{kjt} = the value of the k-th classification assigned to the j-th individual on the t-th trial,

and $\underset{\sim}{Y}_{jt} = (Y_{1jt},\ldots,Y_{Ljt})$, assuming that there are $k = 1,2,\ldots,L$ characteristics. Then he defines

$$I_{\underset{\sim}{L}} = \frac{E|\underset{\sim}{Y}_{jt} - \underset{\sim}{Y}_{j\cdot}|^2}{E|\underset{\sim}{Y}_{jt} - \underset{\sim}{Y}_{\cdot\cdot}|^2}. \tag{2.5.7}$$

By writing

$$E(Y_{kjt} - Y_{kj\cdot})^2 = \sigma^2_{R(k)} \tag{2.5.8}$$

and

$$E(Y_{kjt} - Y_{k\cdot\cdot})^2 = \sigma^2_{(k)}, \tag{2.5.9}$$

the index of inconsistency for the $\underline{L}$-variate case can be written as

$$I_{\underline{L}} = \frac{E\{\sum_{k=1}^{L} (Y_{kjt} - Y_{kj\cdot})^2\}}{E\{\sum_{k=1}^{L} (Y_{kjt} - Y_{k\cdot\cdot})^2\}} = \frac{\sum_{k=1}^{L} \sigma^2_{R(k)}}{\sum_{k=1}^{L} \sigma^2_{(k)}}. \tag{2.5.10}$$

Finally, by letting

$$I_{2(k)} = \frac{\sigma^2_{R(k)}}{\sigma^2_{(k)}}, \tag{2.5.11}$$

the overall index of inconsistency can be written as

$$I_{\underline{L}} = \frac{\sum_{k=1}^{L} I_{2}(k)\, \sigma^2_{(k)}}{\sum_{k=1}^{L} \sigma^2_{(k)}}, \tag{2.5.12}$$

which is a weighted average of the univariate indices of inconsistency, with weights determined by the total variance associated with each variable.

2.6. Estimation Procedures for Variance Components

As shown in the previous sections, the models for assessing observer variability result in random effects or mixed model ANOVA estimation problems. In the situation with balanced data and constant measurement error variance across observers, the usual ANOVA estimators are fairly easy to compute in the random effects model. However, in the more complex univariate mixed models, and certainly in the multivariate models, the estimation of the relevant variance components becomes a difficult task.

In the univariate case, the ANOVA method of estimation involves equating mean squares to their expectations, and solving for the respective variance components. In the balanced case, this involves using the usual calculations for the fixed effects ANOVA table discussed in Scheffé [1959] and Anderson and Bancroft [1952]. In addition, Searle [1971] presents least squares solutions to the non-orthogonal case when the data are not balanced.

Another method of estimation involves maximum likelihood which is discussed by Hartley and Rao [1967]. More recently, Hemmerle and Hartley [1973] presented an iterative solution to the maximum likelihood equations in the ANOVA mixed model. As is true for all maximum likelihood solutions, these estimates depend on distributional assumptions, which may be difficult to validate in the observer variability context.

A quite general procedure has been developed by Koch [1967, 1968] which does not require the ANOVA sums of squares as do most of the other standard procedures. These estimates are unbiased and consistent, and can be calculated regardless of whether or not the data are balanced. This method takes advantage of the fact that the squared difference of any two observations is some function of the variance components in the

model. Unbiased estimates of these components can be obtained by forming the appropriate normalization of these symmetric squared differences across all relevant pairs of observations.

The multivariate situation becomes extremely complex because of the inter-relationships of the variance components across variables. Cole [1969] has shown the solution to the estimation of various components when the variance-covariance structure of the variables follows certain patterns, such as the one shown in (2.3.26) resulting from different measurement error variances for each observer. However, all these procedures require such assumptions as multivariate normality, which may not be warranted in many cases.

3. METHODS PROPOSED FOR DICHOTOMOUS DATA

There are many situations in which a dichotomous judgment is the finest discrimination that is of interest or that can be determined. Examples include the decision as to whether or not a drug is better than a placebo, the establishment of the presence or absence of lung involvement from a radiograph, or the determination of the presence or absence of a psychological trait. By assigning the values of 0 or 1 to the absence or presence of the property, respectively, one could then appeal to the variance components methodology discussed previously in section 2. However, even though the intraclass correlation coefficient of reliability is directly analogous to certain of the measures of agreement developed for categorical data, in some cases, the assumptions of normality are usually not warranted. As a result, the specific methodology developed for dichotomous data will be explored in this section.

3.1. Measures of Association

In the situation in which $a = 2$ observers make independent dichotomous judgments on n subjects, the resulting data can be classified in 2 x 2 tables as observed frequencies in Table 3.1, and as observed proportions in Table 3.2.

Table 3.1 Observed Frequencies of Dichotomous Scores Assigned to n Subjects by Two Observers

		Observer 2		
		1	0	Total
Observer 1	1	n_{11}	n_{12}	$n_{1\cdot}$
	0	n_{21}	n_{22}	$n_{2\cdot}$
	Total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

Table 3.2 Observed Proportions of Dichotomous Scores Assigned to n Subjects by Two Observers

		Observer 2		
		1	0	Total
Observer 1	1	p_{11}	p_{12}	$p_{1\cdot}$
	0	p_{21}	p_{22}	$p_{2\cdot}$
	Total	$p_{\cdot 1}$	$p_{\cdot 2}$	1

Various measures of association have been developed specifically for the 2 x 2 contingency table shown in Table 3.1. Kendall and Stuart [1961] reported the Q statistic, proposed by Yule, which can be expressed as

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} . \qquad (3.1.1)$$

Q ranges from -1 to +1, and has the following properties:

$$Q = \begin{cases} +1, & \text{if } n_{12}n_{21} = 0, \text{ i.e., } n_{12} \text{ or } n_{21} = 0 \\ 0, & \text{if } n_{11}n_{22} = n_{12}n_{21}, \text{ i.e., the ratings of the observers are independent} \\ -1, & \text{if } n_{11}n_{22} = 0, \text{ i.e., } n_{11} \text{ or } n_{22} = 0. \end{cases}$$

They also discussed the V or ϕ coefficient denoted by

$$\phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\{n_{1\cdot}n_{\cdot 1}n_{2\cdot}n_{\cdot 2}\}^{1/2}} . \qquad (3.1.2)$$

In this case, $\phi^2 = \chi^2/n$, where χ^2 is the usual Pearson chi-square statistic for the 2 x 2 table. ϕ also ranges between -1 and +1, and has the following properties:

$$\phi = \begin{cases} +1, & \text{if } n_{12} = n_{21} = 0, \text{ i.e., perfect agreement between the two observers} \\ 0, & \text{if } n_{11}n_{22} = n_{12}n_{21}, \text{ i.e., the ratings of the two observers are independent} \\ -1, & \text{if } n_{11} = n_{22} = 0, \text{ i.e., complete disagreement between the two observers.} \end{cases}$$

In addition, it can be shown that ϕ is directly analogous to Kendall's Tau-a coefficient of concordance on 0-1 data.

It should be pointed out here that both of the statistics of (3.1.1) and (3.1.2) are measures of <u>association</u> rather than <u>agreement</u>. That is, they tend to measure a type of correlation between the scores assigned by the two observers instead of producing a direct measure of the level of agreement. However, ϕ does have the property of ranging between complete disagreement (-1) and complete agreement (+1) of the observers.

3.2. Measures of Agreement

The most elementary index,

$$p = p_{11} + p_{22}, \qquad (3.2.1)$$

based on the proportion of all subjects on whom the two observers agree, has been called the "index of crude agreement" by Rogot and Goldberg [1966]. This index essentially assigns equal weight to agreement on the absence of the property, as well as to agreement on the presence of the property.

However, if the property is judged to be absent more often than present, then indices which ignore p_{22} may be of interest. One such measure due to Dice [1945] is

$$S_D = \frac{p_{11}}{(p_{1\cdot} + p_{\cdot 1})/2} \,. \tag{3.2.2}$$

The index S_D reflects the conditional probability of agreement on the presence of the property given the average of the proportions judged to be present by the separate observers. By writing

$$S_{D'} = \frac{p_{22}}{(p_{2\cdot} + p_{\cdot 2})/2} \,, \tag{3.2.3}$$

Rogot and Goldberg [1966] proposed the arithmetic mean of S_D and $S_{D'}$ as a measure of agreement denoted by

$$A_2 = \frac{p_{11}}{p_{1\cdot} + p_{\cdot 1}} + \frac{p_{22}}{p_{2\cdot} + p_{\cdot 2}} \,. \tag{3.2.4}$$

As such, A_2 ranges from 0, when there is complete disagreement, to +1, when there is complete agreement. In addition, they also proposed another measure of agreement based on conditional probabilities, defined as

$$A_1 = \frac{1}{4}\left\{\frac{p_{11}}{p_{1\cdot}} + \frac{p_{11}}{p_{\cdot 1}} + \frac{p_{22}}{p_{2\cdot}} + \frac{p_{22}}{p_{\cdot 2}}\right\} \,. \tag{3.2.5}$$

A_1 also ranges from complete disagreement (0) to complete agreement (+1).

3.3. Chance Corrected Measures of Agreement

None of the indices of agreement cited in Section 3.2 have accounted for the level of agreement expected by chance alone. However, for statistics such as p, S_D, A_1, and A_2 defined previously, suitable corrections for chance agreement can be derived. Using the notation of Fleiss [1973], let I_o denote the observed value of the index, and let I_e denote the value of the same index expected on the basis of chance alone. Then $I_o - I_e$ represents the excess agreement beyond chance, and $1 - I_e$ denotes the maximum possible excess agreement beyond chance. Then define

$$M(I) = \frac{I_o - I_e}{1 - I_e}, \qquad (3.3.1)$$

as a standardized measure of agreement adjusting for the agreement due to chance. Note that

$$M(I) = \begin{cases} +1, & \text{if both observers are in complete agreement} \\ 0, & \text{if the level of agreement is equal to the expected agreement due to chance} \\ <0, & \text{if observed agreement is less than the expected agreement due to chance.} \end{cases}$$

In the special case of $I_e = 1/2$, M(I) has a minimum of -1.

Fleiss [1973] reports that Scott [1955] and Cohen [1960] appear to be the first to recommend such measures as M(I). Both chose I_o to be the crude index of agreement, $p_{11} + p_{22}$ of (3.2.1). By assuming that the two judges had identical underlying rates, Scott [1955] set

$$I_e = \bar{p}^2 + \bar{q}^2,$$

where $\bar{p} = \frac{1}{2}(p_{1\cdot} + p_{\cdot 1})$, and $\bar{q} = 1 - \bar{p}$. So appealing to (3.3.1), this corrected crude index becomes

$$\pi = \frac{4(p_{11}p_{22} - p_{12}p_{21}) - (p_{12} - p_{21})^2}{(p_{1\cdot} + p_{\cdot 1})(p_{2\cdot} + p_{\cdot 2})} . \qquad (3.3.2)$$

Under the assumption of no bias, the intraclass correlation coefficient r* derived by Fleiss [1965], is equivalent to π in the case with 2 observers. However, Cohen [1960] calculated the expected values for each cell under the usual assumption of independence, and used

$$I_e = p_{1\cdot}p_{\cdot 1} + p_{2\cdot}p_{\cdot 2}.$$

This index, denoted by κ (kappa), is given by (3.3.1) as

$$\begin{aligned} \kappa &= M(p) \\ &= \frac{2(p_{11}p_{22} - p_{12}p_{21})}{p_{1\cdot}p_{\cdot 2} + p_{\cdot 1}p_{2\cdot}} . \end{aligned} \qquad (3.3.3)$$

By using the chance corrected index of (3.3.1), the measures of agreement given previously can be unified, in most cases. Fleiss [1973] shows that by letting $I_e = 2p_{1\cdot}p_{\cdot 1}/(p_{1\cdot} + p_{\cdot 1})$ for the index S_D of (3.2.2), that $M(S_D) = \kappa$ of (3.3.3). In a similar manner, for the index A_2 of (3.2.4), with

$$I_e = \frac{p_{1\cdot}p_{\cdot 1}}{p_{1\cdot} + p_{\cdot 1}} + \frac{p_{2\cdot}p_{\cdot 2}}{p_{2\cdot} + p_{\cdot 2}} ,$$

$M(A_2)$ is again equivalent to κ of (3.3.3). This demonstrates that by using the chance corrected index M(I) of (3.3.1), three of the measures of agreement have been shown to be equivalent. However, this result does not hold for A_1 defined in (3.2.5). In this case, since $I_e = 1/2$, as shown by Rogot and Goldberg [1966], Fleiss [1973] demonstrated that

$$M(A_1) = \frac{(p_{11}p_{22} - p_{12}p_{21})(p_{1\cdot}p_{2\cdot} + p_{\cdot 1}p_{\cdot 2})}{2p_{1\cdot}p_{2\cdot}p_{\cdot 1}p_{\cdot 2}} \qquad (3.3.4)$$

$$= \frac{p_{11}p_{22} - p_{12}p_{21}}{\text{H.M.}} ,$$

where H.M. is the harmonic mean of the two marginal variances $p_{1\cdot}p_{2\cdot}$ and $p_{\cdot 1}p_{\cdot 2}$. One additional property of $M(A_1)$ is that it assumes the value -1 when there is complete disagreement between the observers. The κ statistic of (3.3.3), on the other hand, is equal to -1 only when, in addition, $p_{1\cdot} = p_{\cdot 1} = 1/2$. Fleiss [1973] also points out that the ϕ coefficient of (3.1.2) can be written as

$$\phi = \frac{p_{11}p_{22} - p_{12}p_{21}}{\text{G.M.}} , \qquad (3.3.5)$$

where G.M. is the geometric mean of the two marginal variances. Another similar measure, proposed by Maxwell and Pilliner [1968], was derived from psychometric theory, and can be written as

$$r_{11} = \frac{p_{11}p_{22} - p_{12}p_{21}}{\text{A.M.}} , \qquad (3.3.6)$$

where A.M. is the arithmetic mean of the two marginal variances. This demonstrates that the three measures $M(A_1)$, ϕ, and r_{11} differ only in the type of average variance used in the denominator. However, Fleiss [1973] indicates that only r_{11} of (3.3.6) is interpretable as an intra-class correlation coefficient.

3.4. One Characteristic; a Observers

The situation involving $a > 2$ observers independently rating n subjects on a dichotomous scale has been considered by Armitage, *et al*.

[1966], Fleiss [1965] and Bennett [1972]. The general goal is to create some overall measure of the extent to which observers agree or disagree in their recording of the same sign S for the same subjects. In order to utilize comparable notation, the data can be displayed in Table 3.3 similar to that of Bennett [1972].

Table 3.3 Data Resulting from n Subjects Independently Rated by a Observers on a Dichotomous Scale

		Observers					
		1	2	3	...	a	Total
Subjects	1	x_{11}	x_{12}	x_{13}	...	x_{1a}	r_1
	2	x_{21}	x_{22}	x_{23}	...	x_{2a}	r_2
	3	x_{31}	x_{32}	x_{33}	...	x_{3a}	r_3
	.	.	.	.		.	.
	.	.	.	.		.	.
	.	.	.	.		.	.
	n	x_{n1}	x_{n2}	x_{n3}	...	x_{na}	r_n
	Total	y_1	y_2	y_3	...	y_a	y

Here,

$$x_{ij} = \begin{cases} 1, & \text{if the i-th subject is determined to be positive for S by observer j} \\ 0, & \text{otherwise,} \end{cases}$$

r_ℓ = 0,1,..., or a is the number of positive findings for the ℓ-th subject,

y_j is the number of subjects found positive by observer j.

Let

$$\tau_\ell = \frac{1}{a} r_\ell, \quad \ell = 1,2,...,n,$$

be the average number of positive findings for the ℓ-th subject. From this we note that if

$$\tau_\ell > 1/2, \; r_\ell \text{ observers agree with the majority,}$$

and if

$$\tau_\ell < 1/2, \; a - r_\ell \text{ observers agree with the majority.}$$

Also, let

$$\bar{\tau} = \frac{1}{n} \sum_{\ell=1}^{n} \tau_\ell = \frac{1}{na} \sum_{\ell=1}^{n} r_\ell \tag{3.4.1}$$

be the overall average proportion of positive findings for all subjects.

In the framework of this notation, the majority agreement index (M.A.I.) of Armitage, et al. [1966] can be expressed as

$$a_\ell = |2\tau_\ell - 1|, \quad \ell = 1,2,\ldots,n. \tag{3.4.2}$$

Thus, the M.A.I. index a_ℓ assumes the value +1 if all observers agree on the presence or absence of the sign on the ℓ-th subject, and it assumes the value 0 if the observers are evenly divided. A small peculiarity of a_ℓ is that it can take on the value 0 only if a is even; otherwise, its minimum value is 1/a. As a summary statistic across subjects, Armitage, et al. [1966] propose

$$\bar{a} = \frac{1}{n} \sum_{\ell=1}^{n} a_\ell, \tag{3.4.3}$$

which is called the mean majority agreement index. In the special case when $a = 2$ observers, the measure $\bar{a}$ of (3.4.3) reduces to $p = p_{11} + p_{22}$ of (3.2.1).

Since there are $a > 2$ observers, we have $\binom{a}{2}$ possible pairs of

observers, of which $r_\ell(a - r_\ell)$ disagree about the ℓ-th subject. Armitage, et al. [1966] proposed a pair disagreement index (P.D.I.) defined by

$$d_\ell = \frac{2r_\ell(a - r_\ell)}{a(a - 1)} \quad , \quad \ell = 1,2,\ldots,n. \tag{3.4.4}$$

As such, the P.D.I. index d_ℓ assumes the value 0 if all observers agree on the presence or absence of the sign on the ℓ-th subject, and it assumes the value $a/2(a-1)$ if the observers are evenly divided. They also proposed

$$\bar{d} = \frac{1}{n} \sum_{\ell=1}^{n} d_\ell, \tag{3.4.5}$$

which is called the mean pair disagreement index. In the special case when $a = 2$ observers, the measure $\bar{d}$ of (3.4.5) reduces to $1-p$, where $p = p_{11} + p_{22}$ of (3.2.1).

The third measure of observer agreement proposed by Armitage, et al. [1966] is based on the variation of τ_ℓ from subject to subject. They define the standard deviation agreement index (S.D.A.I.) as

$$s_a = \left\{ \frac{\sum_{\ell=1}^{n} (r_\ell - \bar{r})^2}{n - 1} \right\}^{1/2} . \tag{3.4.6}$$

Note that an increase in the S.D.A.I. index s_a reflects an increase in the variability of the r_ℓ's, and therefore an increase in observer agreement. As long as not all the subjects are found to be either positive or negative by all the observers, the S.D.A.I. assumes the value 0 when there is complete disagreement among the observers. Fleiss [1973] has shown that when $a = 2$ observers, the S.D.A.I. index becomes

$$s_2 = \{\frac{n}{n-1}[(p_{11} + p_{22}) - (p_{11} - p_{22})^2]\}^{1/2}. \qquad (3.4.7)$$

In this situation, s_2 ranges from 0 to a maximum value of $\sqrt{\frac{n}{n-1}}$, only when $p_{11} + p_{22} = 1$ and $p_{11} = p_{22} = 1/2$. This feature makes s_2 unattractive as a general usage agreement index.

Armitage, et al. [1966] indicate that one advantage of the mean M.A.I. and the mean P.D.I. is that they can also be evaluated separately for each observer. As such, the mean M.A.I. for one observer is the proportion of his observations which agree with the majority opinion on the same subject, and the mean P.D.I. for one observer is the proportion of disagreements among all the paired comparisons involving that observer. In that paper, and also in Smyllie, et al. [1965] they illustrate these indices with data obtained from observers reporting physical signs in respiratory disease.

3.5. Intraclass Correlation Coefficients

Even though the data in the dichotomous case are limited to the values of 1 and 0, it is of interest to compare these measures of agreement with the reliability coefficient $\tilde{\rho}$ defined in Section 2 for various ANOVA models. Fleiss [1973] has shown that in the case with $a = 2$ observers, the usual ANOVA table can be written as shown in Table 3.4, using the notation from Table 3.2.

If we consider the most elementary ANOVA model given in (2.1.1), the sums of squares for observers (0) will be included in the sums of squares due to residual error. In this case, the relevant mean squares are given by

$$MS_s = S/(n-1),$$

$$\text{and} \quad MS_e = (0 + E)/n.$$

Table 3.4 ANOVA Table for Dichotomous Data (a = 2 Observers)

SOURCE	D.F.	SUMS OF SQUARES
Observers	1	$O = \frac{n}{2}(p_{12} - p_{21})^2$
Subjects	n-1	$S = \frac{n}{2}[(p_{11} + p_{22}) - (p_{11} - p_{22})^2]$
Residual Error	n-1	$E = \frac{n}{2}[(p_{12} + p_{21}) - (p_{12} - p_{21})^2]$
Total	2n-1	$\frac{n}{2}[1 - (p_{11} - p_{22})^2]$

Using the sample estimate of $\tilde{\rho}$ given in (2.1.3), and assuming that n is sufficiently large to ignore $n/(n-1)$, we obtain

$$R_1 = \hat{\tilde{\rho}}_1 = \frac{S - (O + E)}{S + (O + E)} = \frac{4(p_{11}p_{22} - p_{12}p_{21}) - (p_{12} - p_{21})^2}{(p_{1\cdot} + p_{\cdot 1})(p_{2\cdot} + p_{\cdot 2})}, \quad (3.5.1)$$

which is identical to Scott's π (3.3.2) and Fleiss's r*, as presented in Fleiss [1965]. Thus, by failing to partition out the observer effects, we note that the reliability coefficient R_1 is equivalent to the π coefficient of (3.3.2), which resulted by assuming that the two judges had identical underlying rates. This aspect of R_1, and equivalently of π, makes it less attractive as an agreement measure than the kappa statistic which does allow for observer bias.

If we now consider the ANOVA model of (2.1.4), which does include observer effects, two measures of reliability result, depending on whether the observer effects are assumed to be fixed or random. In the fixed effects case, the mean squares are given by

$$MS_s = S/(n - 1),$$

$$\text{and } MS_e = E/(n - 1),$$

and hence, the intraclass correlation coefficient of (2.1.2) is estimated by

$$R_2 = \hat{\tilde{\rho}}_2 = \frac{S - E}{S + E} = \frac{2(p_{11}p_{22} - p_{12}p_{21})}{p_{1\cdot}p_{2\cdot} + p_{\cdot 1}p_{\cdot 2}}, \qquad (3.5.2)$$

which is equivalent to Maxwell and Pilliner's r_{11} statistic of (3.3.6). If, instead, we assume <u>random effects</u> for the observers, the mean squares are

$$MS_o = 0,$$

$$MS_s = S/(n-1),$$

$$MS_e = E/(n-1).$$

Again, if we can assume that $n/(n-1)$ is essentially equivalent to 1, the reliability coefficient of (2.1.8) is estimated by

$$R_3 = \hat{\tilde{\rho}}_3 = \frac{S - E}{S + E + 2(O)}. \qquad (3.5.3)$$

Fleiss, <u>et al</u>. [1973] and Krippendorff [1970] have both shown that R_3 of (3.5.3) is identically equivalent to κ of (3.3.3). These last two results can be summarized as follows:

<u>FIXED OBSERVERS</u>	<u>RANDOM OBSERVERS</u>
$\hat{\tilde{\rho}}_2 = R_2 = \frac{2(p_{11}p_{22}-p_{12}p_{21})}{(p_{1\cdot}p_{2\cdot}+p_{\cdot 1}p_{\cdot 2})} = r_{11}$	$\hat{\tilde{\rho}}_3 = R_3 = \frac{2(p_{11}p_{22}-p_{12}p_{21})}{(p_{1\cdot}p_{\cdot 2}+p_{\cdot 1}p_{2\cdot})} = \kappa.$

This demonstrates that observer bias, <u>i.e.</u>, the failure of $p_{1\cdot}$ and $p_{\cdot 1}$ to be equal, forces R_3 to be less than R_2, and as a result,

$$\kappa \leq r_{11}. \qquad (3.5.4)$$

These results tend to make κ and r_{11} the most attractive as measures of agreement in the dichotomous situation with 2 observers, since they are corrected for chance agreement, and have a direct interpretation as an intraclass correlation coefficient, depending on whether the observer effects are fixed or random.

3.6. Hypothesis Tests

To test hypotheses of observer differences or bias, the dichotomous data for the $\underline{a}$ observer case from Table 3.3 can be handled by considering an $\underline{s}_o$ x $\underline{a}$ matrix $\underset{\sim}{X}$, where $s_o = 2^{\underline{a}}$ possible realizations on any subject. Then, by using the notation of Bennett [1972a], the observed and expected frequencies of these $\underline{s}_o$ patterns can be summarized in vector form by $\underset{\sim}{n}' = (n_1, n_2, \ldots, n_{s_o})$, and $\underset{\sim}{np}' = (np_1, np_2, \ldots, np_{s_o})$, where

$$\sum_{i=1}^{s_o} n_i = n, \text{ and } \sum_{i=1}^{s_o} p_i = 1.$$

In addition, by writing the column proportions of Table 3.3 in vector form as

$$\underset{\sim}{y} = \frac{1}{n}(y_1, y_2, \ldots, y_a),$$

with expected value

$$\underset{\sim}{\pi} = (\pi_1, \pi_2, \ldots, \pi_a),$$

the hypothesis concerning differences between $\underline{m}$ ($\leq$a) of the observers is

$$H_0: \ \pi_1 = \pi_2 = \ldots = \pi_m. \tag{3.6.1}$$

Bennett [1972a] indicated that this can be tested equivalently using the observed proportions

$$\underset{\sim}{\hat{p}} = (\hat{p}_1, \ldots, \hat{p}_{s_o}),$$

and writing H_0: $\underset{\sim}{B}\ \underset{\sim}{p} = \underset{\sim}{0}$, where $\underset{\sim}{B}$ is the appropriate $(\underline{m}-1) \times \underline{s}_o$ contrast matrix. Bennett [1967] derived approximate multinomial maximum likelihood (M.M.L.) estimates for this hypothesis providing an asymptotically χ^2 goodness of fit statistic with m-1 d.f. In the special case of m=2 observers, this is shown to be identical to McNemar's test for correlated proportions. In a subsequent paper, Bennett [1968] showed that the minimum modified Neyman χ_1^2 statistic for this hypothesis can be expressed as

$$\chi_1^2 = \frac{\chi^2}{1 - \chi^2/n}, \tag{3.6.2}$$

where χ^2 is based on the M.M.L. estimates. Hence, these two tests are asymptotically equivalent. These developments were used later in Bennett [1972b] to test for the equality of several diagnostic procedures.

Fleiss [1965] proposed Cochran's Q statistic for the observer bias test. Using the notation of Table 3.3, this statistic is given by

$$\begin{aligned} Q &= \frac{a(a-1)\sum_{j=1}^{a}(y_j - \bar{y})^2}{\sum_{\ell=1}^{n} r_\ell(a - r_\ell)} \\ &= \frac{2\sum_{j=1}^{a}(y_j - \bar{y})^2}{n\bar{d}}, \ \bar{d} > 0, \end{aligned} \tag{3.6.2}$$

where

$$\sum_{j=1}^{a} y_j/a = \bar{y},$$

and $\bar{d}$ is the mean pair disagreement index of Armitage, et al. [1966]. This demonstrates that as the disagreement index $\bar{d}$ decreases, the Q statistic increases. As such, this statistic depends both on agreement

and bias. Q has an approximate χ^2 distribution with (a-1) d.f. under (3.6.1). In addition, under the assumption of independence among judgments on a subject, Fleiss [1965] stated that observer differences can be tested by

$$\chi^2 = \frac{\sum_{j=1}^{a} (y_j - \bar{y})^2}{\bar{y} - \bar{y}^2/n}, \tag{3.6.3}$$

which has an approximate χ^2 distribution with (a-1) d.f. By writing

$$\frac{\chi^2}{Q} = \frac{n\bar{d}}{2(\bar{y} - \bar{y}^2/n)} = \frac{\bar{d}/2}{p_o - p_o^2}, \tag{3.6.4}$$

where $p_o = \bar{y}/n$, we observe that χ^2/Q is an index of inconsistency, as $p_o - p_o^2$ is the estimated variance of y_j under H_0 of (3.6.1). Here

$$\chi^2/Q = \begin{cases} 0, & \text{if } \bar{d} = 0, \text{ i.e., no observer disagreement} \\ \dfrac{a}{4(a-1)(p_o - p_o^2)}, & \text{if observers are evenly divided.} \end{cases}$$

Fleiss [1965] shows that

$$r^* = 1 - \chi^2/Q \tag{3.6.5}$$

is the common intraclass correlation between all pairs of judgments on the same subjects.

3.7. Multivariate Dichotomous Responses

Bennett [1972a] considered the case in which k clinicians each observe the presence or absence of c correlated signs $S_1, S_2, \ldots, S_c$ on a sample of n patients. For simplicity, he first presented a test for the case when $k = 2$ and $c = 2$. In this situation, there are four possible realizations or response patterns for S_1 and S_2, viz. (1,1), (1,0),

(0,1), and (0,0), reflecting the determinations of the two observers. This gives rise to a 4 x 4 contingency table, such as Table 3.5, with frequencies $\{n_{\ell m}\}$ and corresponding probabilities $\{\pi_{\ell m}\}$;

$$\sum_{\ell=1}^{4}\sum_{m=1}^{4}\pi_{\ell m} = 1.$$

Then the hypothesis of no observer differences, or no observer bias in

Table 3.5 Dichotomous Responses by Two Observers on Two Signs

S_1 \ S_2	(1,1)	(1,0)	(0,1)	(0,0)	TOTAL
(1,1)	n_{11}	n_{12}	n_{13}	n_{14}	n_{10}
(1,0)	n_{21}	n_{22}	n_{23}	n_{24}	n_{20}
(0,1)	n_{31}	n_{32}	n_{33}	n_{34}	n_{30}
(0,0)	n_{41}	n_{42}	n_{43}	n_{44}	n_{40}
TOTAL	n_{01}	n_{02}	n_{03}	n_{04}	n

detecting positive findings for S_1 and S_2 can be tested by H_0: $\underset{\sim}{B}\,\underset{\sim}{\pi} = \underset{\sim}{0}$, where

$$\underset{\sim}{B} = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \end{bmatrix},$$

and

$$\underset{\sim}{\pi}' = (\pi_{11},\pi_{12},\pi_{13},\pi_{14},\pi_{21},\pi_{22},\pi_{23},\pi_{24},\pi_{31},\pi_{32},\pi_{33},\pi_{34},\pi_{41},\pi_{42},\pi_{43},\pi_{44}).$$

Using the maximum likelihood estimates (M.M.L.), the χ^2 goodness-of-fit tests suggested by Bennett [1967] is given by

$$\chi^2 = \sum_{\ell=1}^{4}\sum_{m=1}^{4}\frac{(n_{\ell m} - n\hat{\pi}_{\ell m})^2}{n\hat{\pi}_{\ell m}}, \qquad (3.7.1)$$

which is asymptotically chi-square with 2 degrees of freedom (d.f.) under H_0. He then showed how these results could be extended directly to a c-dimensional table. The general hypothesis of no observer bias can be written as H_0: $\pi_1^{(t)} = \pi_2^{(t)} = \ldots = \pi_m^{(t)}$, t = 1,2,...,c, for c signs and m observers. The appropriate test statistic is a chi-square goodness of fit statistic with c(m-1) d.f.

In addition, Bennett [1972a] proposed that the various indices due to Armitage, et al. [1966] could be extended to the situation with c dichotomous signs. He illustrated this by considering the average proportion of positive findings for each of the c signs.

4. METHODS PROPOSED FOR NOMINAL AND ORDINAL DATA

Many research situations give rise to categorical data determined by nominal scales, e.g., states of mental health such as normal, neurotic, and psychotic, or ordinal scales such as mild, moderate, and severe stages of a disease. We have already considered the special case of a dichotomous scale in Section 3. In each of these instances the observer has a fixed set of categories to which he can classify each subject. As a result, techniques of analyzing categorical data or contingency table data can be applied to the observer variability problem.

Since each observer uses the same scale, the data can be cross-classified into a k^a contingency table, where k is the number of categories on the rating scale, and a is the number of observers. The case with a = 2 observers gives rise to the two-dimensional k x k contingency table shown in Table 4.1. Here observers 1 and 2 independently classify n subjects on a k-point scale. If the observers agree on the classification of a subject into category j, the result will be recorded in the

total n_{jj} on the main diagonal. If both observers agree on each subject, the main diagonal elements are the only non-zero elements in the table. Thus, n_{ij} is the total number of subjects classified into category i by observer 1 and into category j by observer 2.

Table 4.1 Classification on k-Point Scale by Two Observers

Observer 1 \ Observer 2	1	2	...	k	TOTAL
1	n_{11}	n_{12}	...	n_{1k}	n_{10}
2	n_{21}	n_{22}	...	n_{2k}	n_{20}
.	.	.	...	.	.
.	.	.	...	.	.
.	.	.	...	.	.
k	n_{k1}	n_{k2}	...	n_{kk}	n_{k0}
TOTAL	n_{01}	n_{02}		n_{0k}	n

4.1. Measures of Association

In the k x k contingency table of 4.1, we can calculate the usual Pearson chi-square statistic,

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{k} n\left(n_{ij} - \frac{n_{i0}n_{0j}}{n}\right)^2 / n_{i0}n_{0j} .$$

Kendall and Stuart [1961] give a coefficient of contingency due to Pearson denoted by

$$P = \left\{ \frac{\chi^2}{n + \chi^2} \right\}^{1/2} . \qquad (4.1.1)$$

As such, $0 \leq P \leq 1$, and P attains 0 under complete independence. How-

ever, P has limitations, since, in general, P does not have the same upper limit for different tables, and therefore fails to be comparable in different situations. Because of the upper limit problem and the lack of direct interpretation in the observer agreement question, we find that P is quite limited for our interests. It only measures association between the ratings, and not necessarily agreement.

To remedy these undesirable properties of P given in (4.1.1), Kendall and Stuart [1961] report that Tschuprow proposed an alternative function of χ^2 which is given by

$$T = \left\{\frac{\chi^2}{n(k-1)}\right\}^{1/2}, \tag{4.1.2}$$

for the k x k table. In the case of perfect agreement between the observers, T will attain the value of +1. We can also note that for k = 2, i.e., the dichotomous data case, that $T^2 = \chi^2/n$, which is precisely ϕ^2, the coefficient of (3.1.2).

4.2. Measures of Agreement Between Two Raters

Goodman and Kruskal [1954, 1959] point out that the usual tests for association, such as given in Section 4.1, do not seem appropriate as measures of observer agreement. They discuss a measure based upon optimal prediction of order given by

$$\lambda_r = \frac{\sum_a p_{aa} - \frac{1}{2}(p_{M\cdot} + p_{\cdot M})}{1 - \frac{1}{2}(p_{M\cdot} + p_{\cdot M})}, \tag{4.2.1}$$

where $p_{aa} = n_{aa}/n$ and $p_{M\cdot}$ and $p_{\cdot M}$ are the two marginal proportions corresponding to a hypothesized modal class. As defined here, λ_r takes on values from -1 when all the diagonal elements are zero and the modal probability, $p_{M\cdot} + p_{\cdot M}$, is 1, to +1 when both observers are in complete

agreement.

If we write $p_{ij} = n_{ij}/n$, from the results in Table 4.1, the most elementary index of agreement between the observers can be expressed as

$$p_o = \sum_{i=1}^{k} p_{ii}, \tag{4.2.2}$$

which has been called the index of crude agreement by Rogot and Goldberg [1966]. However, this is clearly inadequate, since a certain level of agreement is expected by chance alone. Cohen [1960] proposed a coefficient of agreement for nominal scales denoted by

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{4.2.3}$$

where

$$p_e = \sum_{i=1}^{k} p_{i0}p_{0i},$$

the level of agreement calculated under the assumption of complete independence between ratings by the two observers. This is directly analogous to the chance corrected measures of agreement discussed in Section 3.3 for dichotomous data. As defined here, κ (kappa) is simply the proportion of agreement after chance agreement is removed from consideration. κ has the same properties as the general index M(I) presented in (3.3.1), where I_o is the sum of the proportions on the main diagonal of Table 4.1.

In a later paper, Cohen [1968] introduced a modified form of kappa which makes provision for scaled disagreement or partial credit. Since disagreements between the observers in certain cells may be more important than in other cells, he defines a <u>weighted kappa</u> statistic by

$$\kappa_W = \frac{p^*_o - p^*_e}{1 - p^*_e}, \tag{4.2.4}$$

where

$$p^*_o = \sum_i \sum_j w_{ij} p_{ij},$$

and

$$p^*_e = \sum_i \sum_j w_{ij} p_{i0} p_{0j}.$$

Since w_{ij} is the weight assigned to the (i,j)-th cell, p^*_o is simply the weighted observed proportion of agreement, and p^*_e is the weighted proportion of agreement expected by chance alone. From this, we note that the original kappa statistic of (4.2.3) is a special case of weighted kappa with weights $w_{ij} = 1$ for $i = j$, and $w_{ij} = 0$, $i \neq j$.

Everitt [1968] derived the means and variances of both kappa and weighted kappa, but as pointed out by Fleiss, et al. [1969], the standard errors given by Cohen [1960, 1968] and Everitt [1968] are incorrect, since they were based on the assumption of fixed marginals, rather than under the single constraint of a fixed number of subjects. Fleiss, et al. [1969] give the large sample standard error of kappa and weighted kappa, in general, and for the hypothesis that κ or $\kappa_W = 0$. Applications of these statistics are also given by Spitzer, et al. [1967] and Feldman, et al. [1972].

Fleiss and Cohen [1973] have shown that weighted kappa is essentially equivalent to the intraclass correlation coefficient under a specific weight matrix. If we let v_{ij} be the disagreement weight for a subject placed in categories i and j by observers 1 and 2, respectively, then the mean observed proportion of disagreement can be written as

$$\bar{D}_o = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} v_{ij}, \qquad (4.2.5)$$

with expected mean proportion of disagreement under independence given by

$$\bar{D}_e = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{i0} p_{0j} v_{ij}. \tag{4.2.6}$$

In terms of these disagreement proportions, weighted kappa of (4.2.4) can also be expressed as

$$\kappa_W = \frac{\bar{D}_e - \bar{D}_o}{\bar{D}_e} . \tag{4.2.7}$$

Again, we note that kappa of (4.2.3) is a special case of κ_W when

$$v_{ij} = \begin{cases} 0, & i = j \\ 1, & i \neq j. \end{cases}$$

Cohen [1968] has shown that under observed marginal symmetry and disagreement weights $v_{ij} = (i - j)^2$, weighted kappa is precisely equal to the product-moment correlation coefficient on the integer-valued categories. Moreover, without any restrictions on the margins, if the two-way random effects model of (2.1.4) is assumed for the data scored as 1,2,...,k by each of the 2 observers, Fleiss and Cohen [1973] showed that

$$\kappa_W = \frac{SS_s - SS_e}{SS_s + 2SS_o + SS_e} , \tag{4.2.8}$$

which is an estimator of

$$\tilde{\rho}' = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_o^2 + \sigma_e^2 + \frac{1}{n-1}(\sigma_o^2 + \sigma_e^2)} . \tag{4.2.9}$$

From this it is clear that for a large number of subjects, $\tilde{\rho}'$ is essentially equivalent to the intraclass correlation coefficient $\tilde{\rho}$ of (2.1.8). However, it should be observed that the weights $v_{ij} = (i - j)^2$ are assuming an ordinal type of scale with a squared difference penalty on the diagreements. For nominal scales, the weights need to be chosen in

the context of the substantive issues of the problem.

Even though the kappa and weighted kappa statistics were originally devised for nominal scale data, they can quite readily be applied to the situation with ordinal scale data by assigning appropriate weights to each of the off-diagonal cells to reflect the degree of disagreement. If, in addition, it is desirable to have p_o^* interpretable as the proportion of agreement allowing for partial agreement, the weights need to be selected so that $w_{ij} = 1$, for $i = j$, and $0 \leq w_{ij} \leq 1$, for $i \neq j$. One such selection of weights recommended by Cicchetti [1972, 1973] is given by

$$w_{ij} = \frac{(k - 1) - (i - j)}{k - 1} . \qquad (4.2.10)$$

As an example, the weights for a 5-point ordinal scale are given in Table 4.2. Cicchetti [1972, 1973] proposes a measure of agreement

Table 4.2 Example of Weights for Ordinal Data

		Observer 2				
	w_{ij}	1	2	3	4	5
	1	1	3/4	1/2	1/4	0
	2	3/4	1	3/4	1/2	1/4
Observer 1	3	1/2	3/4	1	3/4	1/2
	4	1/4	1/2	3/4	1	3/4
	5	0	1/4	1/2	3/4	1

denoted by

$$C = \frac{p_o^* - p_e^*}{s.e.(p_o^*)} , \qquad (4.2.11)$$

where the numerator of C is identical with that of K_W of (4.2.4) with weights given by (4.2.10), and s.e.(p_o*) is the standard error of the weighted proportion p_o*. As such, C is treated as a standard normal deviate for significance testing.

In a study reported by Chen, et al. [1961] of paired death certificate data to see if more spouses and/or siblings than expected tend to die of cancer of the same detailed site, the data were displayed in a square contingency table similar to Table 4.1. The problem reduced to comparing the sum of the matched pairs for each site with the expected value of this sum. In the appendix, Mantel and Crittenden show that a chi-square test of the form

$$\chi^2 = \frac{\{|\sum n_{ii} - E\sum n_{ii}| - 0.5\}^2}{\mathrm{Var}\{\sum n_{ii}\}}, \qquad (4.2.12)$$

with 1 d.f., is a more powerful test for correlation than the usual chi-square test for the entire contingency table. In the observer variability framework, we note that this test is comparable to the kappa statistic in that it is based directly on the diagonal elements.

In another study reported by Spiers and Quade [1970], the values on the main diagonal of a square contingency table were of interest. In their model, the expected value for the (i,j)-th cell was considered to be a weighted average of the expected value under independence, and the expected value with the diagonals inflated to the greatest possible extent. Using the method of minimum χ^2, estimates of these weights were derived, and then a test of independence was performed.

Much work has been done recently in the area of incomplete contingency tables, e.g., Goodman [1968] and Mantel [1970]. Briefly, an incom-

plete contingency table is one in which certain cells have an a-priori probability of being zero. As a result, these techniques are developed to analyze tables with certain cells deleted. In particular, if we are interested in examining only those instances in which two observers disagree, we would have a table similar to Table 4.3, assuming that the observers are using a 5-point scale. Then by methods of iterative proportional fitting of maximum likelihood estimation, a test of quasi-

Table 4.3 Incomplete Data Table

		Observer 2				
		1	2	3	4	5
Observer 1	1	---	n_{12}	n_{13}	n_{14}	n_{15}
	2	n_{21}	---	n_{23}	n_{24}	n_{25}
	3	n_{31}	n_{32}	---	n_{34}	n_{35}
	4	n_{41}	n_{42}	n_{43}	---	n_{45}
	5	n_{51}	n_{52}	n_{53}	n_{54}	---

independence can be performed. This will reflect the extent to which the ratings by the two observers are associated on the off-diagonal (disagreement) cells.

Light [1971] has recommended a modified chi-square statistic that is sensitive to the pattern of agreement on the diagonal of the k x k table for two observers. Using the usual expected values based on independence, and combining all the off-diagonal cells, his statistic A_p is given by

$$A_p = \sum_{i=1}^{k} \frac{\left[n_{ii} - \frac{n_{io}n_{oi}}{n}\right]^2}{\frac{n_{io}n_{oi}}{n}} + \frac{1}{\sum_{i\neq j}^{k}\sum^{k} \frac{n_{io}n_{oj}}{n}} \left[\sum_{i\neq j}^{k}\sum^{k} n_{ij} - \sum_{i\neq j}^{k}\sum^{k} \frac{n_{io}n_{oj}}{n}\right]^2, \quad (4.2.13)$$

which is distributed asymptotically as chi-square with k degrees of freedom under the hypothesis of independence. It is of interest here, that κ of (4.2.3) may be essentially zero, while A_p may be large and significantly different from zero. However, if A_p is near zero, κ will be necessarily near zero. As such, A_p will reflect deviations from the expected pattern on the diagonal, while κ reflects the overall level of agreement.

4.3. Measures of Agreement Among Many Raters

When $a > 2$ observers independently classify n subjects on a k-point scale, overall measures of agreement are much more complex, and as a result, are not as numerous as those reviewed in Section 4.2. Cartwright [1956] suggested a measure of reliability when there are two or more judges using a nominal scale classification procedure. The agreement measure for a given subject, m_j, is the number of pairs of judges agreeing on the assignment. As such, the maximum value for m_j is $\binom{a}{2}$ for a judges. The coefficient, alpha, can be written as

$$A = \frac{1}{n\binom{a}{2}} \sum_{j=1}^{n} m_j. \quad (4.3.1)$$

This statistic is directly analogous to the pair agreement index recommended by Armitage, et al. [1966] as the complement of $\bar{d}$ of (3.4.5) for dichotomous data. When $a = 2$ judges, A is identical to the crude index of agreement, p_o, given in (4.2.2).

This alpha statistic is somewhat restrictive, however, because it assumes that each of the $\underline{k}$ classes on the nominal scale is equiprobable. Also, there is no provision for partial agreement similar to that of weighted kappa of (4.2.4). Under these special conditions, Cartwright [1956] developed the sampling distribution and probability values of A for selected numbers of judges and scale size. Finally, he recommended an analysis of variance of ranks, where the $\underline{k}$ treatment groups are determined by the $\underline{k}$-point scale, as a measure of multi-judge discrimination. This test reflects the relative variation in scale usage, and is of interest, since a high level of agreement could result from poor discrimination among the subjects.

Fleiss [1971] considered an extension of kappa in the situation when more than two raters independently classify each subject. If there are $\underline{a}$ ratings for each subject on a $\underline{k}$-point scale, then

$$p_j = \frac{1}{na} \sum_{i=1}^{n} n_{ij}, \tag{4.3.2}$$

where n_{ij} is the number of assignments of the i-th subject to the j-th category, is the overall proportion of assignments to the j-th category. As a measure of agreement on the i-th subject, he recommended

$$P_i = \frac{1}{a(a-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1), \tag{4.3.3}$$

which is simply the proportion of agreeing pairs out of all possible pairs of ratings. The overall measure of agreement is then given by

$$\bar{P} = \frac{1}{na(a-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1), \tag{4.3.4}$$

which is identical to A, the alpha statistic proposed by Cartwright [1956]. By assuming that assignments are made purely at random to category j with probabilities p_j, Fleiss [1971] proposed the expected mean proportion of agreement due to chance as

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2, \tag{4.3.5}$$

giving rise to the kappa statistic,

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} n_{ij}^2 - n\,a[1 + (a-1)\sum_{j=1}^{k} p_j^2]}{n\,a(a-1)(1 - \sum_{j=1}^{k} p_j^2)}. \tag{4.3.6}$$

He then derived a large-sample standard error of κ under certain restrictive assumptions.

Fleiss [1971] also proposed a measure of agreement on a particular category as the conditional probability that the second assignment is to category j, given the first assignment was to category j by two randomly selected raters. This statistic is given by

$$\bar{P}_j = \frac{\sum_{i=1}^{n} n_{ij}(n_{ij} - 1)}{\sum_{i=1}^{n} n_{ij}(a - 1)}, \tag{4.3.7}$$

which under the hypothesis of only chance agreement is expected to be p_j. This give the κ statistic,

$$\kappa_j = \frac{\bar{P}_j - p_j}{1 - p_j} = \frac{\sum_{i=1}^{n} n_{ij}^2 - n\,a p_j[1 + (a-1)p_j]}{n\,a(a-1)\,p_j q_j}, \tag{4.3.8}$$

where $q_j = 1 - p_j$. From this, we note that κ of (4.3.6) can be expressed

as

$$\kappa = \frac{\sum_{j=1}^{k} p_j q_j \kappa_j}{\sum_{j=1}^{k} p_j q_j}, \qquad (4.3.9)$$

a weighted average of the κ_j statistics of (4.3.8).

In general, the kappa statistic has been written as an observed proportion of agreement, p_o, corrected for an expected proportion of agreement due to chance, p_e, as expressed in (4.2.3). Alternatively, if $d_o = 1 - p_o$ and $d_e = 1 - p_e$ represent the respective proportions of disagreement, then kappa can be expressed as

$$\kappa = 1 - \frac{d_o}{d_e}. \qquad (4.3.10)$$

Light [1971] developed extensions of kappa of the form of (4.3.10) for situations when there are more than two raters. For $\underline{a} = 3$ observers, let $n_{ij\ell}$ denote the number of subjects assigned to the i-th category by the 1^{st} observer, to the j-th category by the 2^{nd} observer, and to the ℓ-th category by the 3^{rd} observer. He recommended a multiple observer agreement statistic, κ_3, written as

$$\kappa_3 = 1 - \frac{(A/n^5)}{(B/n^6)}, \qquad (4.3.11)$$

where

n = the number of subjects being rated independently by the three observers,

A = the overall proportion of disagreements,

B = the expected proportion of disagreements.

The values for A and B are calculated, given the three fixed margins of

the table, by

$$A = \sum_{i\neq j} n_{ijo} \sum_{i\neq \ell} n_{ioo} n_{oo\ell} \sum_{j\neq \ell} n_{ojo} n_{oo\ell} + \sum_{i\neq \ell} n_{io\ell} \sum_{i\neq j} n_{ioo} n_{ojo} \sum_{j\neq \ell} n_{ojo} n_{oo\ell}$$

$$+ \sum_{j\neq \ell} n_{oj\ell} \sum_{i\neq j} n_{ioo} n_{ojo} \sum_{i\neq \ell} n_{ioo} n_{oo\ell},$$

and

$$B = 3 \sum_{i\neq j} n_{ioo} n_{ojo} \sum_{i\neq \ell} n_{ioo} n_{oo\ell} \sum_{j\neq \ell} n_{ojo} n_{oo\ell}, \tag{4.3.12}$$

where the zero subscript denotes the sum over that index. An approximate large sample estimate of the variance of κ_3 is given by

$$\widehat{\text{Var}}(\kappa_3) = \frac{p_o(1 - p_o)}{n(1 - p_e)^2}, \tag{4.3.13}$$

where $p_o = 1 - \frac{A}{n^5}$ and $p_e = 1 - \frac{B}{n^6}$. He also proposed a conditional agreement index similar to that of Fleiss [1971] given in (4.3.8) by modifying the above values of A and B to reflect disagreements in fixed levels of the three-way table.

As discussed by Light [1971], all of the κ-type statistics assume nothing about a "true" or "correct" classification of the subjects, but rather an internal consistency of the selected observers. In addition, all the observers are given equal weight, resulting in κ reflecting an "overall group agreement" of the $\underline{a}$ observers. However, in some situations one of the observers may be the standard whose classification of the subjects is considered to be "correct." In this case, Light [1971] recommended a test of the joint agreement of the $\underline{a}$ observers with the standard. Let $n_{ij}^{(p)}$ be the number of subjects classified into the i-th

category by the standard and into the j-th category by the p-th observer, p = 1,...,a. Then he proposed the test statistic,

$$G = \frac{t_a - E\{t_a\}}{(\xi t_a)^{1/2}}, \qquad (4.3.14)$$

which is asymptotically standard normal, where

$$t_a = \sum_{p=1}^{a} \sum_{i=1}^{k} n_{ii}^{(p)},$$

$$E\{t_a\} = \frac{1}{n} \sum_{i=1}^{k} [n_{io} (\sum_{p=1}^{a} n_{oi}^{(p)})],$$

and ξt_a is the estimated variance of t_a.

Robinson [1957] discussed the statistical measurement of agreement in terms of the intraclass correlation coefficient for ordinal valued data. He defined a measure of agreement in the k-observer case as

$$A = 1 - D/D_{max}, \qquad (4.3.15)$$

where D is the observed measure of disagreement, and D_{max} is the maximum possible level of disagreement. In the ANOVA framework, D is the sum of squares due to observers, and D_{max} is the total sum of squares. He noted that the relationship between the measure of agreement A and the intraclass correlation coefficient $\tilde{\rho}$ can be expressed as

$$A = \frac{a-1}{a} \tilde{\rho} + \frac{1}{a}, \qquad (4.3.16)$$

where a is the number of observers. As such, A has a distinct advantage as a measure of agreement over $\tilde{\rho}$, since A ranges from 0 to 1 regardless of the number of observers, whereas $\tilde{\rho}$ has a lower bound of $-1/(a-1)$.

4.4. Observer Bias in Multinomial Classification

Although many of the proposed measures of agreement reviewed in the previous sections have utilized the marginal distribution to correct for chance agreement, none of these measures are directly aimed at the problem of observer bias. Krishnaswami and Rajeshwar [1968] have obtained expressions for the maximum positive bias and the minimum negative bias for the situation in which two inspectors classify each of n items into one of m classes. For the particular case of $m = 3$, lower and upper bounds for the proportion estimates have been derived based on the magnitudes of the probabilities of misclassification. This allows the construction of confidence intervals for each of the proportions.

5. SUMMARY AND PROPOSED UNIFICATION OF PROCEDURES

When the data arising from observer reliability studies are numerical, the methods of estimation and hypothesis testing are usually selected from the ANOVA-type models discussed in Section 2. Even though assumptions of normality may not be warranted in certain cases, some of these procedures still permit the estimation of the appropriate components of variance and reliability coefficients to assess the level of observer variability.

As demonstrated in Sections 3 and 4, a wide variety of procedures have been proposed to assess observer agreement when the data are categorical. In the situation when r observers classify the same subjects on the same k-point scale, the resulting data can be classified in a k^r contingency table. A unified approach to the analysis of such multidimensional contingency tables has been introduced by Grizzle, et al.

[1969]. This GSK procedure is essentially a linear models formulation using weighted least squares as a computational device to generate linearized minimum modified chi-square test statistics for the hypotheses to be tested. In a later paper, Koch and Reinfurt [1971] indicated how the GSK method can be applied to categorical data arising from mixed models. The experimental situation corresponding to such a model involves exposing each of n randomly selected subjects to a factor with r levels, (in our case, r observers), and classifying each of the r responses into k categories, (in our case, the k-point nominal/ordinal scale). In the two-way table, the test of no observer bias is essentially the test of homogeneity of the two margins. Bennett [1967] proposed a linearized maximum likelihood chi-square test for this hypothesis in the dichotomous data case, but a test can be proposed to permit nominal or ordinal data scales by appealing to the mixed model methodology in Koch and Reinfurt [1971]. For r observers, this becomes the test for homogeneity of the r first order margins. For multivariate categorical responses, a test of observer bias across variables is a natural extension of this procedure.

In addition, by forming logarithms of ratios of observed to expected proportions as discussed in Forthofer and Koch [1973], the GSK procedure can be used to test the significance of measures of observer agreement. This involves functions of the diagonals of the table to obtain the proportions used in the κ and κ_W statistics discussed in previous sections. As a result, the asymptotic standard errors of these statistics are immediately available as matrix products, avoiding the difficulties with complex, and in some cases, incorrect standard errors. (See Cohen [1960, 1968], Everitt [1968], Fleiss, et al. [1969], and Light [1971].) By

smoothing the table under such hypotheses as marginal homogeneity and diagonal symmetry, in addition to complete independence, several agreement statistics, along with their estimated variance-covariance structure, can be estimated and tested.

Further work is now underway to develop a unified set of procedures to analyze categorical data arising from observer reliability studies. It is hoped that this methodology will be useful in determining the extent and importance of the different sources of variation in the data, one of which can be attributed to the observers.

REFERENCES

Anderson, R. L. and Bancroft, T. A. [1952]. Statistical Theory in Research. McGraw Hill, New York.

Armitage, P., Blendis, L. M. and Smyllie, H. C. [1966]. "The measurement of observer disagreement in the recording of signs." J. R. Statist. Soc. A 129, 98-109.

Bearman, J. E., Kleinman, H., Glyer, V. V. and La Croix, O. M. [1964]. "A study of variability in tuberculin test reading." Amer. Rev. Resp. Dis. 90, 913-919.

Bennett, B. M. [1967]. "Tests of hypotheses concerning matched samples." J. R. Statist. Soc. B 29, 468-474.

Bennett, B. M. [1968]. "Note of χ^2 tests for matched samples." J. R. Statist. Soc. B 30, 368-370.

Bennett, B. M. [1972a]. "Measures for clinicians' disagreements over signs." Biometrics 28, 607-612.

Bennett, B. M. [1972b]. "On comparisons of sensitivity, specificity and predictive value of a number of diagnostic procedures." Biometrics 28, 793-800.

Bershad, M. A. [1969]. "The index of inconsistency for an L-fold classification system, $L \geq 2$." U. S. Bureau of the Census, Technical Notes No. 2, 1-3.

Birkelo, C. C., Chamberlain, W. E., Phelps, P. S., Schools, P. E., Zacks, D. and Yerushalmy, J. [1947]. "Tuberculosis case finding: comparison of effectiveness of various roentgenographic and photofluorographic methods." J. Amer. Med. Assn. 133, 359-366.

Burdock, E. I., Fleiss, J. L. and Hardesty, A. S. [1963]. "A new view of inter-observer agreement." Personnel Psychol. 16, 373-384.

Cartwright, D. S. [1956]. "A rapid non-parametric estimate of multijudge reliability." Psychometrika 21, 17-29.

Chen, W. Y., Crittenden, L. B., Mantel, N. and Cameron, W. R. [1961]. "Site distribution of cancer deaths in husband-wife and sibling pairs." J. Nat. Cancer Inst. 27, 875-892.

Cicchetti, D. V. [1972]. "A new measure of agreement between rank-ordered variables." Proceedings, 80th Annual Convention, APA, 17-18.

Cicchetti, D. V. and Allison, T. [1973]. "Assessing the reliability of scoring EEG sleep records: an improved method." Proceedings and Journal of the Electro-Physiological Technologists' Association 20, 92-102.

Cochrane, A. L., Chapman, P. J. and Oldham, P. D. [1951]. "Observers' errors in taking medical histories." Lancet I, 1007-1009.

Cochrane, A. L. and Garland, L. H. [1952]. "Observer error in the interpretation of chest films." Lancet II, 505-509.

Cohen, J. [1960]. "A coefficient of agreement for nominal scales." Educ. and Psychol. Meas. 20, 37-46.

Cohen, J. [1968]. "Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit." Psychol. Bull. 70, 213-220.

Cohen, J. [1972]. "Weighted chi square: an extension of the kappa method." Educ. Psychol. Meas. 32, 61-74.

Cole, J. W. L. [1969]. "Multivariate analysis of variance using patterned covariance matrices." Institute of Statistics Mimeo Series, No. 640.

Comroe, J. H., Jr. and Botelho, S. [1947]. "The unreliability of cyanosis in the recognition of arterial anoxemia." Am. J. Med. Sci. 214, 1-6.

Dice, L. R. [1945]. "Measures of the amount of ecologic association between species." Ecology 26, 297-302.

Ebel, R. L. [1951]. "Estimation of the reliability of ratings." Psychometrika 16, 407-424.

Everitt, B. S. [1968]. "Moments of the statistics kappa and weighted kappa." Brit. J. Math. and Statist. Psychol. 21, 97-103.

Feldman, S., Kelin, D. F. and Honigfeld, G. [1972]. "The reliability of a decision tree technique applied to psychiatric diagnosis." Biometrics 28, 831-840.

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

Fleiss, J. L., Spitzer, R. L. and Burdock, E. I. [1965]. "Estimating accuracy of judgment using recorded interviews." Arch. Gen. Psychiat. 12, 562-567.

Fleiss, J. L. [1966]. "Assessing the accuracy of multivariate observations." J. Amer. Statist. Assn. 61, 403-412.

Fleiss, J. L., Cohen, J. and Everitt, B. S. [1969]. "Large sample standard errors of kappa and weighted kappa." Psychol. Bull. 72, 323-327.

Fleiss, J. L. [1971]. "Measuring nominal scale agreement among many raters." Psychol Bull. 76, 378-382.

Fleiss, J. L., Spitzer, R. L., Endicott, J. and Cohen, J. [1972]. "Quantification of agreement in multiple psychiatric diagnosis." Arch. Gen. Psychiat. 26, 168-171.

Fleiss, J. L. and Cohen, J. [1973]. "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability." Educ. Psychol. Meas. 33, 613-619.

Fleiss, J. L. [1973]. "Measuring agreement between two judges on the presence or absence of a trait." Paper presented at Joint Meetings of American Statistical Association, New York City.

Fletcher, C. M. and Oldham, P. D. [1949]. "Problem of consistent radiological diagnosis in coalminers' pneumoconiosis." Brit. J. Industr. Med. 6, 168-183.

Fletcher, C. M. [1952]. "The clinical diagnosis of pulmonary emphysema -- an experimental study." Proc. Roy. Soc. Med. 45, 577-584.

Fletcher, C. M. [1964]. "The problem of observer variation in medical diagnosis with special reference to chest diseases." Method Inform. Med. 3, 98-103.

Fletcher, C. M. and Oldham, P. D. [1964]. "Diagnosis in group research." Chapter 2 of Medical Surveys and Clinical Trials, ed. L. J. Witts. 2nd ed. Oxford Univ. Press, London.

Forthofer, R. N. and Koch, G. G. [1973]. "An analysis for compounded functions of categorical data." Biometrics 29, 143-157.

Glasser, G. J., Metzger, G. A. and Miaoulis, G. [1970]. "Measurement and control of interviewer variability." Proceeding of the Business and Economics Section, ASA. 314-319.

Gleser, G. C., Cronbach, L. J. and Rajaratnam, N. [1965]. "Generalizing of scores influenced by multiple sources of variance." Psychometrika 30, 395-418.

Goodman, L. A. and Kruskal, W. H. [1954]. "Measures of association for cross classification." J. Amer. Statist. Assn. 49, 732-764.

Goodman, L. A. and Kruskal, W. H. [1959]. "Measures of association for cross classification. II: Further discussion and references." J. Amer. Statist. Assn. 54, 123-163.

Goodman, L. A. [1968]. "The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries." J. Amer. Statist. Assn. 63, 1091-1131.

Grizzle, J. E., Starmer, C. F. and Koch, G. G. [1969]. "Analysis of categorical data by linear models." Biometrics 25, 489-504.

Grubbs, F. E. [1948]. "On estimating precision of measuring instruments and product variability." J. Amer. Statist. Assoc. 43, 243-264.

Guttman, L. [1946]. "The test-retest reliability of qualitative data." Psychometrika 11, 81-95.

Hahn, G. J. and Nelson, W. [1970]. "A problem in the statistical comparison of measuring devices." Technometrics 12, 95-102.

Hansen, M. H., Hurwitz, W. N. and Pritzker, L. [1964]. "The estimation and interpretation of gross differences and the simple response variance." Contributions to statistics presented to Professor P. S. Mahalanobis on the occasion of his 70th birthday. Pergamon Press, Calcutta.

Hartley, H. O. and Rao, J. N. K. [1967]. "Maximum-likelihood estimation for the mixed analysis of variance model." Biometrics 54, 93-108.

Hemmerle, W. J. and Hartley, H. O. [1973]. "Computing maximum likelihood estimates for the mixed A.O.V. model using the W transformation." Technometrics 15, 819-831.

Kendall, M. G. and Stuart, A. [1961]. The Advanced Theory of Statistics. Vol. 2. Hafner Pub. Co., New York.

Kish, L. [1962]. "Studies of interviewer variance for attitudinal variables." J. Amer. Statist. Assoc. 57, 92-115.

Koch, G. G. [1967]. "A general approach to the estimation of variance components." Technometrics 9, 93-118.

Koch, G. G. [1968]. "Some further remarks concerning 'A general approach to the estimation of variance components.'" Technometrics 10, 551-558.

Koch, G. G. [1971a]. "A response error model for a simple interviewer structure situation." Technical Report No. 4, Project SV-618, Research Triangle Institute.

Koch, G. G. [1971b]. "An alternative approach to the formulation of response error models." Technical Report No. 1, Project SV-618, Research Triangle Institute.

Koch, G. G. and Reinfurt, D. W. [1971]. "The analysis of categorical data from mixed models." Biometrics 27, 157-173.

Krippendorff, K. [1970]. "Bivariate agreement coefficients for reliability of data." E. F. Borgatta, ed. Sociological Methodology, 1970. Jossey-Bass, San Francisco.

Krishnaswami, P. and Rajeshwar, N. [1968]. "Bias in multinomial classification." J. Amer. Statist. Assn. 63, 298-303.

Light, R. J. [1971]. "Measures of response agreement for qualitative data: some generalizations and alternatives." Psychol. Bull. 76, 365-377.

Loewenson, R. B., Bearman, J. E., and Resch, J. A. [1972]. "Reliability of measurements for studies of cardiovascular atherosclerosis." Biometrics 28, 557-569.

Mandel, J. [1959]. "The measuring process." Technometrics 1, 251-267.

Mantel, N. [1970]. "Incomplete contingency tables." Biometrics 26, 291-304.

Maxwell, A. E. and Pilliner, A. E. G. [1968]. "Deriving coefficients of reliability and agreement for ratings." Brit. J. Math. Statist. Psychol. 21, 105-116.

Overall, J. E. [1968]. "Estimating individual rater reliabilities from analysis of treatment effects." Educ. and Psychol. Meas. 28, 255-264.

Pyke, D. A. [1954]. "Finger clubbing: validity as a physical sign." Lancet II, 352-354.

Reynolds, W. E. [1952]. "Research in public health." Paper presented at the 1951 (28th) annual conference of the Milbank Memorial Fund. Milbank Memorial Fund, New York.

Robinson, W. S. [1957]. "The statistical measurement of agreement." American Sociological Review 22, 17-25.

Rogot, E. and Goldberg, I. D. [1966]. "A proposed index for measuring agreement in test-retest studies." J. Chron. Dis. 19, 991-1006.

Sandifer, M. G., Jr., Pettus, C., and Quade, D. [1964]. "A study of psychiatric diagnosis." The Journal of Nervous and Mental Disease 139, 350-356.

Scheffé, H. [1959]. The Analysis of Variance. John Wiley & Sons, New York.

Schilling, R. S. F., Hughes, J. P. W. and Dingwall-Fordyce, I. [1955]. "Disagreement between observers in an epidemiological study of respiratory disease." Brit. Med. J. I, 65-68.

Scott, W. A. [1955]. "Reliability of content analysis: the case of nominal scale coding." Public Opinion Quart. 19, 321-325.

Searle, S. R. [1971]. Linear Models. John Wiley & Sons, New York.

Smith, H. F. [1950]. "Estimating precision of measuring instruments." J. Amer. Statist. Assoc. 45, 447-451.

Smyllie, H. C., Blendis, L. M. and Armitage, P. [1965]. "The observer disagreement in physical signs of the respiratory system." Lancet II, 412-413.

Spiers, P. S. and Quade, D. [1970]. "On the question of an infectious process in the origin of childhood leukemia." Biometrics 26, 723-737.

Spitzer, R. L., Cohen, J., Fleiss, J. L. and Endicott, J. [1967]. "Quantification of agreement in psychiatric diagnosis: a new approach." Arch. Gen. Psychiat. 17, 83-87.

Thompson, W. A., Jr. [1963]. "Precision of simultaneous measurement procedures." J. Amer. Statist. Assoc. 58, 474-479.

Yerushalmy, J. [1947]. "Statistical problems in assessing methods of medical diagnosis with special reference to X-ray techniques." Publ. Hlth. Rep. (Wash.) 62, 1432-1449.

Yerushalmy, J., Harkness, J. T., Cope, J. H. and Kennedy, B. R. [1950]. "The role of dual reading in mass radiography." Amer. Rev. Tuberc. 61, 443-464.

Yerushalmy, J., Garland, L. H., Harkness, J. T. and Zwerling, H. B. [1951]. "Evaluation of role of serial chest roentgenograms in estimating progress of disease in patients with pulmonary tuberculosis." Amer. Rev. Tuberc. 64, 225-248.