BIOMETRICS *33*, 159–174
March 1977

The Measurement of Observer Agreement for Categorical Data

J. RICHARD LANDIS

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109 U.S.A.

GARY G. KOCH

Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27514 U.S.A.

Summary

This paper presents a general statistical methodology for the analysis of multivariate categorical data arising from observer reliability studies. The procedure essentially involves the construction of functions of the observed proportions which are directed at the extent to which the observers agree among themselves and the construction of test statistics for hypotheses involving these functions. Tests for interobserver bias are presented in terms of first-order marginal homogeneity and measures of interobserver agreement are developed as generalized kappa-type statistics. These procedures are illustrated with a clinical diagnosis example from the epidemiological literature.

1. Introduction

Researchers in many fields have become increasingly aware of the observer (rater or interviewer) as an important source of measurement error. Consequently, reliability studies are conducted in experimental or survey situations to assess the level of observer variability in the measurement procedures to be used in data acquisition. When the data arising from such studies are quantitative, tests for interobserver bias and measures of interobserver agreement are usually obtained from standard ANOVA mixed models or random effects models such as those discussed in Anderson and Bancroft [1952], Scheffé [1959], and Searle [1971]. As a result, hypothesis tests of observer effects are used to investigate interobserver bias, i.e., differences in the mean response among observers, and estimates of intraclass correlation coefficients are used to measure interobserver reliability. Modifications and extensions of these standard ANOVA models have been proposed by Grubbs [1948, 1973], Mandel [1959], Fleiss [1966], Overall [1968], and Loewenson, Bearman and Resch [1972] to evaluate the measurement error in various types of applications. Although assumptions of normality for these models may not be warranted in certain cases, the ANOVA procedures discussed in Searle [1971] and the symmetric square difference procedure in Koch [1967, 1968] still permit the estimation of the appropriate components of variance and the reliability coefficients.

On the other hand, many observer reliability studies involve categorical data in which the response variable is classified into nominal (or possibly ordinal) multinomial categories. As reviewed in Landis and Koch [1975a, 1975b], a wide variety of estimation and testing procedures have been recommended for the assessment of observer variability in these

Key Words: Observer agreement; Multivariate categorical data; Kappa statistics; Repeated measurement experiments; Weighted least squares.

cases. In this paper we propose a unified approach to the evaluation of observer agreement for categorical data by expressing the quantities which reflect the extent to which the observers agree among themselves as functions of observed proportions obtained from underlying multidimensional contingency tables. These functions are then used to produce test statistics for the relevant hypotheses concerning interobserver bias in the overall usage of the measurement scale and interobserver agreement on the classification of individual subjects. For illustrative purposes, this general methodology is developed within the context of a typical data set which resulted from an investigation of observer variability in the clinical diagnosis of multiple sclerosis.

2. A Clinical Diagnois Example

Let us consider the data arising from the diagnosis of multiple sclerosis reported in Westlund and Kurland [1953]. Among other things, the investigators were interested in comparing patient groups to study possible differences in the geographical distributions of the disease. For this purpose, a series of patients in Winnipeg, Manitoba and a separate series of patients in New Orleans, Louisiana were selected and were examined by a neurologist in their respective locations. After the completion of all the examinations, each neurologist was requested to review all the records without seeing his earlier summary and diagnosis and to classify them into one of the following diagnostic classes:

1. Certain multiple sclerosis;
2. Probable multiple sclerosis;
3. Possible multiple sclerosis (odds 50 : 50);
4. Doubtful, unlikely, or definitely not multiple sclerosis.

In order to evaluate agreement between the diagnosticians, the Winnipeg neurologist then reviewed and classified each of the New Orleans patient records, and vice versa. The data resulting from these review diagnoses are presented in Table 1.

A preliminary inspection of the Winnipeg data indicates that the Winnipeg neurologist tended to diagnose more of the patients as certain (1) or probable (2) multiple sclerosis than did his counterpart in New Orleans. As a result, they agreed on the diagnosis of only 64/149 (43 percent) of the patients. Although the differences in the overall crude distributions of the diagnoses seem to be less prominent within the New Orleans patients, the neurologists diagnosed only 33/69 (48 percent) of them into identically the same category. The statistical issues concerning these differences in diagnosis can be summarized within the framework of the following basic questions:

(1) Are there any differences between the two patient populations with respect to the overall crude distribution of the diagnoses by each of the two neurologists?
(2) Are there any differences between the overall crude distributions of the diagnoses by the two neurologists within each of the respective patient populations?
(3) Is there any neurologist $\times$ sub-population interaction in the overall crude distribution of the diagnoses?
(4) Is there any difference between the two patient populations with respect to the overall agreement of the two neurologists on the specific diagnosis of individual patients?
(5) Is the agreement of the two neurologists on the specific diagnosis of individual patients significantly different from clance agreement based on their overall crude distributions of diagnoses?

Table 1
DIAGNOSTIC CLASSIFICATION REGARDING MULTIPLE SCLEROSIS

Sub-population		Winnipeg Patients (1)					
Observer		Winnipeg Neurologist (2)					
	Diagnostic Class	1	2	3	4	Total	Proportion
New Orleans Neurologist (1)	1	38	5	0	1	44	0.295
	2	33	11	3	0	47	0.315
	3	10	14	5	6	35	0.235
	4	3	7	3	10	23	0.154
	Total	84	37	11	17	149	
	Proportion	0.564	0.248	0.074	0.114		

Sub-population		New Orleans Patients (2)					
Observer		Winnipeg Neurologist (2)					
	Diagnostic Class	1	2	3	4	Total	Proportion
New Orleans Neurologist (1)	1	5	3	0	0	8	0.116
	2	3	11	4	0	18	0.261
	3	2	13	3	4	22	0.319
	4	1	2	4	14	21	0.304
	Total	11	29	11	18	69	
	Proportion	0.159	0.420	0.159	0.261		

(6) Are there certain patterns of disagreement which may reflect significant imprecision in the diagnostic criteria?

As stated in Koch *et al.* [1977], questions (1)–(3) are directly analogous to the hypotheses of "no whole-plot effects," "no split-plot effects," and "no whole-plot × split-plot interaction" in standard split-plot experiments. In this context, question (1) addresses differences among the sub-populations, question (2) involves the issue of interobserver bias, and question (3) is concerned with the observer × sub-population interaction. Thus, the first-order marginal distributions of response for each of the neurologists within each sub-population contain the relevant information for dealing with these questions. In

contrast to overall crude differences, questions (4)–(6) are addressed at interobserver agreement on a subject-to-subject basis; and, as such they are directly analogous to hypotheses concerning intraclass correlation coefficients in random effects models. Hence, certain functions of the diagonal cells of various subtables are used to provide information for estimating and testing the significance of agreement on the classification of individual subjects.

In the following sections a general methodology for answering these questions is developed in terms of specific hypotheses. These procedures are then illustrated with an analysis of the data in Table 1.

3. *Methodology*

Let $i = 1, 2, \cdots, s$ index a set of sub-populations from which random samples have been selected. Suppose that the same response variable is measured separately by each of d observers using an L-point scale. Let the $r = L^d$ response profiles be indexed by a vector subscript $\mathbf{j} = (j_1, j_2, \cdots, j_d)$, where $j_g = 1, 2, \cdots, L$ for $g = 1, 2, \cdots, d$. Furthermore, let $\pi_{i\mathbf{j}} = \pi_{ij_1,j_2,\cdots,j_d}$ represent the joint probability of response profile $\mathbf{j}$ for randomly selected subjects from the ith sub-population. Then let the first-order marginal probability

$$\phi_{igk} = \sum_{\mathbf{j} \text{ with } j_g=k} \cdots \sum \pi_{ij_1,j_2,\cdots,j_d} \quad \text{for} \quad \begin{matrix} i = 1, 2, \cdots, s \\ g = 1, 2, \cdots, d \\ k = 1, 2, \cdots, L \end{matrix} \tag{3.1}$$

represent the probability of the kth response category for the gth observer in the ith sub-population.

3.1 Hypotheses Involving Marginal Distributions

Hypotheses directed at the questions of differences among sub-populations and inter-observer bias involve distributions of the response profiles and can be expressed in terms of constraints on the first-order marginal probabilities $\{\phi_{igk}\}$. As a result, the specific hypotheses associated with questions (1)–(3) are directly analogous to H_{SM}, H_{CM}, and H_{AM} outlined in Koch *et al.* [1977] in expressions (2.4), (2.5), and (2.9), respectively. In particular, the d observers correspond to the d conditions, and thus the hypothesis of *first order marginal symmetry* (*homogeneity*) addresses the issue of interobserver bias. These hypotheses can also be expressed in terms of constraints on *mean score* functions associated with each observer such as the $\{\eta_{ig}\}$ summary indexes specified in (2.14) in Koch *et al.* [1977]. Further discussion of hypothoses involving marginal distributions within the context of observer agreement studies is given in Landis [1975].

3.2 Hypotheses Involving Generalized Kappa-Type Measures

Whereas the previous hypotheses concerning differences among sub-populations and interobserver bias involved only the first-order marginal probabilities, hypotheses directed at the extent to which observers agree among themselves on the classification of individual subjects must be formulated in terms of the internal elements of the table. For example, the estimate of the crude proportion of agreement between two observers is simply the sum of the observed proportions on the main diagonal of the corresponding two-way table. In addition, if partial credit is permitted for certain types of disagreement, an estimate of the weighted proportion of agreement will involve the weighted inclusion of the off-diagonal cells.

As reviewed in Landis and Koch [1975a, 1975b], numerous measures of observer agreement have been proposed for categorical data, e.g., Goodman and Kruskal [1954], Cohen [1960, 1968], Fleiss [1971], Light [1971], and Cicchetti [1972]. Most of these quantities are of the form

$$\kappa = \frac{\pi_0 - \pi_e}{1 - \pi_e}, \tag{3.2}$$

where π_0 is an observational probability of agreement and π_e is a hypothetical expected probability of agreement under an appropriate set of baseline constraints such as total independence of observer classifications. Ranging from $[-\pi_e/(1 - \pi_e)]$ to $+1$, κ indicates the extent to which the observational probability of agreement is in excess of the probability of agreement hypothetically expected under the baseline constraints. Furthermore, as shown in Fleiss and Cohen [1973] and Fleiss [1975], κ is directly analogous to the intraclass correlation coefficient obtained from ANOVA models for quantitative measurements and can be used as a measure of the reliability of multiple determinations on the same subjects.

Several kappa-type measures of interobserver agreement can be formulated to investigate selected patterns of disagreement simultaneously by choosing corresponding sets of weights which reflect the role of each response category in a given agreement index. For example, a set of weights can be chosen so that the resulting agreement measure indicates the combined performance of all the observers, such as majority or consensus agreement, or sets of weights can be directed at subsets of observers, such as all possible pairwise agreement measures. Alternatively, these weights can be chosen so that the associated kappa measures indicate the increments in agreement which result by successively combining relevant categories of the response variable. Such kappa measures are said to be in a hierarachical relationship with each other. Thus, in general, let $w_{1\mathbf{j}}, w_{2\mathbf{j}}, \cdots, w_{u\mathbf{j}}$ be u sets of weights assigned to the response profiles indexed by $\mathbf{j} = (j_1, j_2, \cdots, j_d)$. Moreover, let $0 \leq w_{h\mathbf{j}} \leq 1$ for $h = 1, 2, \cdots, u$ over all $\mathbf{j}$, so that the resulting estimates are interpretable as probabilities of agreement. Then the observational probability of agreement associated with the hth set of weights in the ith sub-population is the weighted sum

$$\lambda_{ih} = \sum_{\mathbf{j}} \cdots \sum_{\mathbf{j}} w_{h\mathbf{j}}\pi_{i\mathbf{j}} \quad \text{for} \quad \begin{matrix} i = 1, 2, \cdots, s \\ h = 1, 2, \cdots, u. \end{matrix} \tag{3.3}$$

Correspondingly, the expected proportion of agreement associated with (3.3) is the weighted sum

$$\gamma_{ih} = \sum_{\mathbf{j}} \cdots \sum_{\mathbf{j}} w_{h\mathbf{j}}\pi_{i\mathbf{j}}^{(e)} \quad \text{for} \quad \begin{matrix} i = 1, 2, \cdots, s \\ h = 1, 2, \cdots, u, \end{matrix} \tag{3.4}$$

where $\pi_{i\mathbf{j}}^{(e)}$ represents the joint hypothetical expected probability of response profile $\mathbf{j}$ for randomly selected subjects from the ith sub-population.

These expected probabilities are determined by the choice of a particular set of baseline constraints assumed for the response profiles. For this purpose, let $\underline{E} = \{\underline{E}_1, \underline{E}_2, \cdots\}$ represent such underlying constraints on the marginal probabilities $\{\phi_{ig\mathbf{k}}\}$ of (3.1). In this context, the following sets of constraints are of interest in creating interobserver agreement measures:

(i) Under the assumption of total independence among the response variables from the d observers, the $\{\pi_{i\mathbf{j}}^{(e)}\}$ satisfy

$$\underline{E}_1 : \pi_{ij_1j_2\cdots j_d}^{(e)} = \phi_{i1j_1}\phi_{i2j_2}\cdots\phi_{idj_d} = \prod_{k=1}^{d}\phi_{ikj_k} \quad \text{for} \quad i = 1, 2, \cdots, s. \tag{3.5}$$

(ii) Under the assumption of "no interobserver bias" the hypothesis of first-order marginal homogeneity (H_{CM} in Koch *et al.* [1977]) holds. In this situation, let the common probability of classification into the kth category be

$$\psi_{ik} = \phi_{i1k} = \phi_{i2k} = \cdots = \phi_{idk} \tag{3.6}$$

for $i = 1, 2, \cdots, s$ and $k = 1, 2, \cdots, L$. Then under the baseline constraints of total independence and marginal homogeneity the $\{\pi_{ij}^{(e)}\}$ satisfy

$$\underline{E}_2 : \pi_{ij_1j_2\cdots j_d}^{(e)} = \psi_{ij_1}\psi_{ij_2}\cdots\psi_{ij_d} = \prod_{g=1}^{d}\psi_{ij_g} \quad \text{for} \quad i = 1, 2, \cdots, s. \tag{3.7}$$

Consequently, a generalized kappa-type measure of agreement directly analogous to (3.2) can be formulated by

$$\kappa_{ih} = \frac{\lambda_{ih} - \gamma_{ih}}{1 - \gamma_{ih}} \quad \text{for} \quad \begin{matrix} i = 1, 2, \cdots, s \\ h = 1, 2, \cdots, u, \end{matrix} \tag{3.8}$$

under a set of specified constraints in $\underline{E}$. Here κ_{ih} represents an agreement measure among the d observers in the ith sub-population with respect to the hth set of weights.

Within this framework, the specific hypotheses associated with questions (4)–(6) can now be formulated as follows:

(4) If there are no differences among the s sub-populations with respect to the measures of overall specific agreement among the d observers under $\underline{E}_z$, then the $\{\kappa_{ih}\}$ satisfy the hypothesis

$$H_{SA|\underline{E}_z} : \kappa_{1h} = \kappa_{2h} = \cdots = \kappa_{sh} \quad \text{for} \quad h = 1, 2, \cdots, u, \tag{3.9}$$

where SA denotes sub-population agreement.

(5) If the level of observed agreement is equal to that expected under $\underline{E}_z$, then the $\{\kappa_{ih}\}$ satisfy the hypothesis

$$H_{NA|\underline{E}_z} : \kappa_{ih} = 0 \quad \text{for} \quad \begin{matrix} i = 1, 2, \cdots, s \\ h = 1, 2, \cdots, u, \end{matrix} \tag{3.10}$$

where NA denotes no agreement.

(6) In some cases the weights for the kappa measures are chosen to be in a hierarchical relationship with each other in order to investigate specific disagreement patterns. In these situations, if the extent of disagreement is the same for the categories combined by the $(h + 1)$-st set of weights as for those combined by the hth set, then the $\{\kappa_{ih}\}$ satisfy the hypothesis

$$H_{HA|\underline{E}z} : \kappa_{i,h+1} = \kappa_{i,h} \quad \text{for} \quad i = 1, 2, \cdots, s, \tag{3.11}$$

where HA denotes hierarchical agreement.

In order to maintain consistent nomenclature when describing the relative strength of agreement associated with kappa statistics, the following labels will be assigned to the corresponding ranges of kappa:

Kappa Statistic	*Strength of Agreement*
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

Although these divisions are clearly arbitrary, they do provide useful "benchmarks" for the discussion of the specific example in Table 1.

3.3 Estimation and Hypothesis Testing

Test statistics for the hypotheses considered in the previous sections as well as estimators for corresponding model parameters can be obtained by using the general approach for the analysis of multivariate categorical data proposed by Grizzle, Starmer and Koch [1969] (hereafter abbreviated GSK) as outlined in Appendix 1 in Koch *et al.* [1977]. The hypotheses in Section 3.1 involving constraints on the first-order marginal probabilities can be tested by expressing the estimates of the $\{\phi_{i\vartheta k}\}$ or the $\{\eta_{i\vartheta}\}$ as linear functions of the type given in Appendix 1 (A.14) in Koch *et al.* [1977]. These particular matrix expressions have already been discussed in considerable detail in Koch and Reinfurt [1971] and Koch *et al.* [1977], and thus they will not be elaborated here. Otherwise, their specific construction for these hypotheses in observer agreement studies is documented in Landis [1975].

In contrast to the linear functions which pertain to the hypotheses in Section 3.1, all the hypotheses involving generalized kappa-type measures require the expression of the ratio estimates of the $\{\kappa_{ih}\}$ as compounded logarithmic-exponential-linear functions of the observed proportions as formulated in Appendix 1 (A.20) in Koch *et al.* [1977]. As a result, the test statistics for the hypotheses in Section 3.2 can also be generated by the corresponding expression given in Appendix 1 (A.11) in Koch *et al.* [1977].

4. Analysis of Multiple Sclerosis Data

This section is concerned with the analysis of the multiple sclerosis data in Table 1 with primary emphasis given to illustrating the methodology in Section 3. Tests of significance are used in a descriptive context to identify important sources of variation as opposed to a rigorous inferential context; and thus issues pertaining to multiple comparisons are ignored here. These, however, can be handled by the Scheffé type procedures given in Grizzle, Starmer and Koch [1969]. The design for this example involves $s = 2$ sub-populations, $d = 2$ observers, and $L = 4$ response categories. Thus, there are $r = L^d = 16$ possible multivariate response profiles within each of the sub-populations.

4.1 Marginal Homogeneity Tests

The functions required to test the hypotheses involving marginal distributions can be generated in the formulation of (A.14) in Appendix 1 in Koch *et al.* [1977] with the function vector $\mathbf{F}' = (\mathbf{F}_1', \mathbf{F}_2')$ where

$$\begin{aligned} \mathbf{F}_1' &= (0.295, 0.315, 0.235, 0.564, 0.248, 0.074) \\ \mathbf{F}_2' &= (0.116, 0.261, 0.319, 0.159, 0.420, 0.159), \end{aligned} \tag{4.1}$$

Table 2
HIERARCHICAL WEIGHTS FOR AGREEMENT MEASURES

Weights		$w_{1\underset{\sim}{j}}$				$w_{2\underset{\sim}{j}}$				$w_{3\underset{\sim}{j}}$				$w_{4\underset{\sim}{j}}$			
Observer		2				2				2				2			
Diagnostic Class		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Observer 1	1	1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0
	2	0	1	0	0	1	1	0	0	1	1	0	0	1	1	1	0
	3	0	0	1	0	0	0	1	0	0	0	1	1	0	1	1	1
	4	0	0	0	1	0	0	0	1	0	0	1	1	0	0	1	1

which contain the marginal proportions for diagnostic classes "1," "2" and "3" for the two observers within the two sub-populations. The test statistic for H_{SM} is $Q_C = 46.37$ with d.f. = 6, which implies that there are significant ($\alpha = 0.01$) differences in the distributions of the observed response profiles between the Winnipeg and New Orleans patients. The tests of this hypothesis within each of the observers also indicate statistically significant ($\alpha = 0.01$) differences between the two sub-populations, although the Winnipeg neurologist represents the more dominant component. Similarly, the test statistic for H_{CM} is $Q_C = 69.01$ with d.f. = 6, which implies that there are significant ($\alpha = 0.01$) differences in the response profiles between the two neurologists within each sub-population. Moreover, the dominant component of these observer differences is within the Winnipeg patient group. These results suggest that significant interobserver bias exists between the two neurologists in their overall usage of the diagnostic classification scale. In addition, the goodness-of-fit statistic for testing the interaction hypothesis H_{AM} is $Q = 14.09$ with d.f. = 3. This significant ($\alpha = 0.01$) observer $\times$ sub-population interaction is consistent with the result that the observer differences are more substantial in the Winnipeg patient group ($Q_C = 58.47$) than in the New Orleans patient group ($Q_C = 10.54$).

Table 3
DESCRIPTION OF HIERARCHICAL WEIGHTS

Set of Weights	Disagreement Permitted for Agreement Statistic
1	None; requires perfect agreement.
2	Certain (1) with Probable (2).
3	Certain (1) with Probable (2); Possible (3) with Doubtful (4).
4	Certain (1) with Probable (2); Probable (2) with Possible (3); Possible (3) with Doubtful (4).

4.2 Hierarchical Kappa-Type Measures of Agreement

Specific patterns of disagreement between the neurologists on the diagnostic classification of individual subjects can be investigated by selecting a hierarchy of weights which successively combine adjoining categories of diagnosis in order to create potentially less stringent reliability measures. For example, the four sets of weights in Table 2 can be used to investigate the sources of imprecise diagnostic criteria. As indicated in Table 3, these weights are chosen so that specific disagreement patterns are successively tolerated in the corresponding estimates of agreement. In particular, w_{1j} represents the set of weights which generate the kappa measure of perfect agreement proposed in Cohen [1960].

The sequence of hierarchical kappa-type statistics within each of the two patient populations associated with the weights given in Table 2 can be expressed in the formulation (A.20) in Appendix 1 in Koch *et al.* [1977] under the baseline constraints of total independence $\underset{\sim}{E}_1$ in (3.5) by letting

$$\underset{24\times 32}{\mathbf{A}_1} = \begin{bmatrix} 1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&1&1&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&1&1&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1 \\ 1&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0 \\ 0&1&0&0&0&1&0&0&0&1&0&0&0&1&0&0 \\ 0&0&1&0&0&0&1&0&0&0&1&0&0&0&1&0 \\ 0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&1 \\ 1&0&0&0&0&1&0&0&0&0&1&0&0&0&0&1 \\ 1&1&0&0&1&1&0&0&0&0&1&0&0&0&0&1 \\ 1&1&0&0&1&1&0&0&0&0&1&1&0&0&1&1 \\ 1&1&0&0&1&1&1&0&0&1&1&1&0&0&1&1 \end{bmatrix} \otimes \mathbf{I}_2 \; ; \tag{4.2}$$

$$\underset{40\times 24}{\mathbf{A}_2} = \begin{bmatrix} 1&0&0&0&1&0&0&0&0&0&0&0 \\ 1&0&0&0&0&1&0&0&0&0&0&0 \\ 1&0&0&0&0&0&1&0&0&0&0&0 \\ 1&0&0&0&0&0&0&1&0&0&0&0 \\ 0&1&0&0&1&0&0&0&0&0&0&0 \\ 0&1&0&0&0&1&0&0&0&0&0&0 \\ 0&1&0&0&0&0&1&0&0&0&0&0 \\ 0&1&0&0&0&0&0&1&0&0&0&0 \\ 0&0&1&0&1&0&0&0&0&0&0&0 \\ 0&0&1&0&0&1&0&0&0&0&0&0 \end{bmatrix} \otimes \mathbf{I}_2 \; ; \tag{4.3}$$

$$\underset{40\times24}{\mathbf{A}_2} = \begin{bmatrix} 0\,0\,1\,0 & 0\,0\,1\,0 & 0\,0\,0\,0 \\ 0\,0\,1\,0 & 0\,0\,0\,1 & 0\,0\,0\,0 \\ 0\,0\,0\,1 & 1\,0\,0\,0 & 0\,0\,0\,0 \\ 0\,0\,0\,1 & 0\,1\,0\,0 & 0\,0\,0\,0 \\ 0\,0\,0\,1 & 0\,0\,1\,0 & 0\,0\,0\,0 \\ 0\,0\,0\,1 & 0\,0\,0\,1 & 0\,0\,0\,0 \\ 0\,0\,0\,0 & 0\,0\,0\,0 & 1\,0\,0\,0 \\ 0\,0\,0\,0 & 0\,0\,0\,0 & 0\,1\,0\,0 \\ 0\,0\,0\,0 & 0\,0\,0\,0 & 0\,0\,1\,0 \\ 0\,0\,0\,0 & 0\,0\,0\,0 & 0\,0\,0\,1 \end{bmatrix} \otimes \mathbf{I}_2 \; ; \qquad (4.3)\ \text{Cont.}$$

$$\underset{16\times40}{\mathbf{A}_3} = \begin{bmatrix} -1 & 0\,0\,0 & 0 & -1 & 0\,0 & 0 & 0 & -1 & 0 & 0\,0 & 0 & -1 & 1\,0\,0\,0 \\ -1 & -1\,0\,0 & -1 & -1 & 0\,0 & 0 & 0 & -1 & 0 & 0\,0 & 0 & -1 & 0\,1\,0\,0 \\ -1 & -1\,0\,0 & -1 & -1 & 0\,0 & 0 & 0 & -1 & -1 & 0\,0 & -1 & -1 & 0\,0\,1\,0 \\ -1 & -1\,0\,0 & -1 & -1 & -1\,0 & 0 & -1 & -1 & -1 & 0\,0 & -1 & -1 & 0\,0\,0\,1 \\ 0 & 1\,1\,1 & 1 & 0 & 1\,1 & 1 & 1 & 0 & 1 & 1\,1 & 1 & 0 & 0\,0\,0\,0 \\ 0 & 0\,1\,1 & 0 & 0 & 1\,1 & 1 & 1 & 0 & 1 & 1\,1 & 1 & 0 & 0\,0\,0\,0 \\ 0 & 0\,1\,1 & 0 & 0 & 1\,1 & 1 & 0 & 0 & 1 & 1\,1 & 0 & 0 & 0\,0\,0\,0 \\ 0 & 0\,1\,1 & 0 & 0 & 0\,1 & 1 & 0 & 0 & 0 & 1\,1 & 0 & 0 & 0\,0\,0\,0 \end{bmatrix} \otimes \mathbf{I}_2 \; ; \qquad (4.4)$$

$$\underset{8\times16}{\mathbf{A}_4} = [\mathbf{I}_4 - \mathbf{I}_4] \otimes \mathbf{I}_2 \; ; \qquad (4.5)$$

For the data in Table 1, these estimates are given by

$$\mathbf{F} = \begin{bmatrix} \hat{\kappa}_{11} \\ \hat{\kappa}_{12} \\ \hat{\kappa}_{13} \\ \hat{\kappa}_{14} \\ \hat{\kappa}_{21} \\ \hat{\kappa}_{22} \\ \hat{\kappa}_{23} \\ \hat{\kappa}_{24} \end{bmatrix} = \begin{bmatrix} 0.208 \\ 0.328 \\ 0.408 \\ 0.596 \\ 0.297 \\ 0.332 \\ 0.386 \\ 0.789 \end{bmatrix}, \qquad (4.6)$$

where $\hat{\kappa}_{ih}$ is the estimate of the agreement measure in the ith sub-population associated with the hth set of weights shown in Table 2. In addition, the estimated covariance matrix of $\mathbf{F}$ is given by

$$\mathbf{V_F} = \begin{bmatrix} 0.2546 & 0.2122 & 0.1868 & 0.1442 & & & & \\ 0.2122 & 0.4005 & 0.3862 & 0.2912 & & \mathbf{0} & & \\ 0.1868 & 0.3862 & 0.5200 & 0.3832 & & & & \\ 0.1442 & 0.2912 & 0.3832 & 0.5700 & & & & \\ & & & & 0.6163 & 0.5582 & 0.5046 & 0.2185 \\ & \mathbf{0} & & & 0.5582 & 0.6879 & 0.6544 & 0.3010 \\ & & & & 0.5046 & 0.6544 & 1.0030 & 0.4147 \\ & & & & 0.2185 & 0.3010 & 0.4147 & 0.7720 \end{bmatrix} \times 10^{-2}. \quad (4.7)$$

The test statistics for the hierarchical hypotheses in (3.11) are displayed in Table 4. These results indicate that all increases in successive agreement measures within the Winnipeg patient group are significant ($\alpha = 0.05$); but for the New Orleans patient group, the only significant ($\alpha = 0.05$) increase in agreement pertained to the final set of weights. Thus, the neurologists are exhibiting significant disagreement between diagnoses (1,2), (2,3) and (3,4) in the Winnipeg group and significant disagreement between diagnoses (2,3) in the New Orleans group, as evidenced by the inflated frequencies in these off-diagonal cells in Table 1. On the other hand, the estimates in (4.6) suggest that the hierarchical

Table 4
STATISTICAL TESTS FOR HIERARCHICAL HYPOTHESES

Hypothesis	D.F.	Q_C
Combined Patient Groups		
$\kappa_{12} = \kappa_{11}$; $\kappa_{22} = \kappa_{21}$	2	6.89*
$\kappa_{13} = \kappa_{12}$; $\kappa_{23} = \kappa_{22}$	2	5.15
$\kappa_{14} = \kappa_{13}$; $\kappa_{24} = \kappa_{23}$	2	28.13**
Winnipeg Patients (1)		
$\kappa_{12} = \kappa_{11}$	1	6.20**
$\kappa_{13} = \kappa_{12}$	1	4.38*
$\kappa_{14} = \kappa_{13}$	1	10.96**
New Orleans Patients (2)		
$\kappa_{22} = \kappa_{21}$	1	0.69
$\kappa_{23} = \kappa_{22}$	1	0.76
$\kappa_{24} = \kappa_{23}$	1	17.17**

* means significant at $\alpha = 0.05$;
** means significant at $\alpha = 0.01$.

kappa measures within both patient groups exhibit the same increasing trend. Since the estimated variances of the kappa statistics are much larger for the New Orleans patient group (due to the smaller sample size), the agreement patterns may indeed be essentially the same in both patient groups.

If the two neurologists are indeed exhibiting the same agreement patterns with respect to the weights given in Table 2 within the two groups of patients, then under (3.5) the $\{\kappa_{ih}\}$ satisfy the following hypotheses from (3.9)

$$H_{SA|\underline{E}_1} : \kappa_{1h} = \kappa_{2h} \quad \text{for} \quad h = 1, 2, 3, 4. \tag{4.8}$$

Test statistics for these hypotheses both individually and jointly are presented in Table 5.

The results in Tables 4 and 5 suggest that a reduced model can be used to combine parameters which are essentially equivalent. For this purpose, the agreement statistics in (4.6) can be modeled by

$$\mathbf{E}_A\{\mathbf{F}\} = \mathbf{X}\beta = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \kappa_1 \\ \kappa_2 \\ \kappa_3 \\ \kappa_4 \\ \kappa_5 \end{bmatrix}, \tag{4.9}$$

where "$\mathbf{E}_A$" denotes "asymptotic expectation." For this model, the goodness-of-fit statistic is $Q = 2.27$ with d.f. $= 3$. Thus, this reduced model provides a satisfactory characterization of the variation among these agreement measures. Specific test statistics for the corresponding hypotheses in (3.10) and (3.11) pertaining to the model $\mathbf{X}$ in (4.9) are given in Table 6. These results suggest that all the parameters are significantly ($\alpha = 0.01$) different from zero, and moreover, are significantly ($\alpha = 0.05$) different from each other. Furthermore, by reducing the model to these smoothed estimates, the marginally significant ($\alpha = 0.10$) difference between κ_{14} and κ_{24} in Table 5 is now significant ($\alpha = 0.05$) for the

Table 5
STATISTICAL TESTS BETWEEN PATIENT SUB-POPULATIONS

Hypothesis	D.F.	Q_C
$\kappa_{1h} = \kappa_{2h}$ for h = 1,2,3,4.	4	7.15
$\kappa_{11} = \kappa_{21}$	1	0.90
$\kappa_{12} = \kappa_{22}$	1	0.00
$\kappa_{13} = \kappa_{23}$	1	0.03
$\kappa_{14} = \kappa_{24}$	1	2.77

Table 6
STATISTICAL TESTS FOR MODEL **X**

Hypothesis	D.F.	Q_C	Hypothesis	D.F.	Q_C
$\kappa_2 = \kappa_1$	1	5.40*	$\kappa_1 = 0$	1	31.05**
$\kappa_3 = \kappa_2$	1	4.92*	$\kappa_2 = 0$	1	40.71**
$\kappa_4 = \kappa_3$	1	12.33**	$\kappa_3 = 0$	1	45.49**
$\kappa_5 = \kappa_4$	1	4.88*	$\kappa_4 = 0$	1	72.44**
			$\kappa_5 = 0$	1	94.97**

* means significant at $\alpha = 0.05$;
** means significant at $\alpha = 0.01$.

comparison of κ_4 and κ_5 in this final model. Finally, the predicted values for the $\{\kappa_{ih}\}$ based on the fitted model (4.9) are displayed in Table 7 together with their corresponding estimated standard errors.

Thus, these results suggest that the diagnostic criteria are not very distinct with respect to their usage by these two neurologists. In addition to bias at the macro stage, i.e., considering only the overall marginal proportions, these observers exhibited significant disagreement at the micro state, i.e., considering each individual subject, in specifying a diagnosis. Only with respect to the relatively relaxed criterion corresponding to the fourth set of weights do the kappa statistics indicate a "moderate" to "substantial" level of interobserver reliability.

5. Discussion

In some applications, one may also be interested in a set of weights which assign varying degrees of partial credit to the off-diagonal cells depending on the extent of the disagreement, rather than successively combining adjoining categories as shown in Table 2. For

Table 7
SMOOTHED ESTIMATES OF AGREEMENT UNDER MODEL **X**

Sub-population		1		2	
Weights	Agreement Statistic	Estimate Under $\underset{\sim}{X}$	Estimated Standard Error	Estimate Under $\underset{\sim}{X}$	Estimated Standard Error
$w_{1\underset{\sim}{j}}$	κ_{i1}	0.236	0.042	0.236	0.042
$w_{2\underset{\sim}{j}}$	κ_{i2}	0.311	0.049	0.311	0.049
$w_{3\underset{\sim}{j}}$	κ_{i3}	0.383	0.057	0.383	0.057
$w_{4\underset{\sim}{j}}$	κ_{i4}	0.579	0.068	0.790	0.081

Table 8
ALTERNATIVE WEIGHTS FOR OVERALL AGREEMENT MEASURES

Weights		w_{1j}				w_{2j}			
Observer		2				2			
	Diagnostic Class	1	2	3	4	1	2	3	4
Observer 1	1	1	0	0	0	1	½	¼	0
	2	0	1	0	0	½	1	½	¼
	3	0	0	1	0	¼	½	1	½
	4	0	0	0	1	0	¼	½	1

example, the weights w_{2j} in Table 8 are directly analogous to those discussed in Cohen [1968], Fleiss, Cohen and Everitt [1969] and Cicchetti [1972], which were used to generate weighted kappa and C statistics. For the data in Table 1, these estimates are given by

$$\mathbf{F} = \begin{bmatrix} \hat{\kappa}_{11} \\ \hat{\kappa}_{12} \\ \hat{\kappa}_{21} \\ \hat{\kappa}_{22} \end{bmatrix} = \begin{bmatrix} 0.208 \\ 0.315 \\ 0.297 \\ 0.407 \end{bmatrix}, \tag{5.1}$$

where the $\{\hat{\kappa}_{i1}\}$ estimate the perfect agreement kappa measure and the $\{\hat{\kappa}_{i2}\}$ estimate the partial-credit weighted kappa agreement measure between the two neurologists in the two patient populations. A more extensive analysis of these data under the weights in Table 8 is given in Landis [1975] and Landis *et al.* [1976].

Although the methodology for the assessment of observer agreement developed in this paper is quite general, these procedures have been illustrated with an example involving only two observers. However, for situations in which either the number of observers d or the number of response categories L is moderately large, the number of possible multivariate response profiles $r = L^d$ becomes extremely large. Consequently, the matrices required to implement the GSK procedures directly may be outside the scope of computational feasibility. In addition, for each of the s sub-populations many of the r possible response profiles will not necessarily be observed in the respective samples so that corresponding cell frequencies are zero. Thus, in such cases, specialized computing procedures are required to obtain the estimates of the pertinent functions.

One alternative approach for handling such very large contingency tables in which most of the observed cell frequencies are zero is discussed in Landis and Koch [1977]. In this regard, the same estimators which would need to be obtained from the conceptual multidimensional contingency table can be generated by first forming appropriate indicator variables of the raw data from each subject and then computing the across-subject arithmetic means. Subsequent to these preliminary steps, the usual matrix operations discussed in Appendix 1 in Koch *et al.* [1977] can then be applied to these indicator variable means to determine the required measures of observer agreement. These alternative computations involving raw data, as well as the extended GSK procedures summarized in Appendix 1 in Koch *et al.* [1977] can all be performed by a recently developed computer program (GENCAT) discussed in Landis, *et al.* [1976].

Acknowledgments

This research was partially supported by Research Grants GM-00038-20 and GM-70004-05 from the National Institute of General Medical Sciences and by the U. S. Bureau of the Census through Joint Statistical Agreements JSA 74-2 and JSA 75-2. The authors would like to thank the referees for their helpful comments on an earlier draft of this paper. In addition, the authors are grateful to Ms. Rebecca Wesson and Ms. Lynn Wilkinson for their conscientious typing of previous drafts of this paper, and to Ms. Linda L. Blakley and Ms. Connie Massey for their efficient typing of the final version of this manuscript.

La Mesure de la Concordance Entre Observations pour des Donnees en Categories

Résumé

L'article expose une méthodologie statistique générale pour l'analyse de données multivariates en catégories provenant d'études de fiabilité d'observateurs. La procédure fait principalement appel à la construction de fonctions des proportions observées traduisant la concordance des observateurs entre eux et à la construction de statistiques de tests pour des hypothèses impliquant ces fonctions. On présente des tests pour des biais entre observateurs en fonction de l'homogénéité marginale du premier ordre et on construit des mesures de concordance entre observateurs comme des statistiques généralisant celles du type kappa. On illustre ces procédures avec un exemple de diagnostic clinique provenant de la littérature épidémiologique.

References

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

Bhapkar, V. P. [1966]. A note on the equivalence of two test criteria for hypotheses in categorical data. *Journal of the American Statistical Association 61,* 228–235.

Bhapkar, V. P. [1968]. On the analysis of contingency tables with a quantitative response. *Biometrics 24,* 329–338.

Bhapkar, V. P. and Koch, G. G. [1968a]. Hypotheses of "no interaction" in multidimensional contingency tables. *Technometrics 10,* 107–123.

Bhapkar, V. P. and Koch, G. G. [1968b]. On the hypotheses of "no interaction" in contingency tables. *Biometrics 24,* 567–594.

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

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

Cohen, J. [1968]. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin 70,* 213–220.

Fleiss, J. L. [1966]. Assessing the accuracy of multivariate observations. *Journal of the American Statistical Association 61,* 403–412.

Fleiss, J. L., Cohen, J. and Everitt, B. S. [1969]. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin 72,* 323–337.

Fleiss, J. L. [1971]. Measuring nominal scale agreement among many raters. *Psychological Bulletin 76,* 378–382.

Fleiss, J. L. and Cohen, J. [1973]. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement 33,* 613–619.

Fleiss, J. L. [1975]. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics 31,* 651–659.

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

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. *Journal of the American Statistical Association 43,* 243–264.

Grubbs, F. E. [1973]. Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics 15,* 53–66.

Goodman, L. A. and Kruskal, W. H. [1954]. Measures of association for cross classification. *Journal of the American Statistical Association 49,* 732–764.

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. and Reinfurt, D. W. [1971]. The analysis of categorical data from mixed models. *Biometrics 27,* 157–173.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., Jr. and Lehnen, R. G. [1977]. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics 33,* 133–158.

Landis, J. R. [1975]. A general methodology for the measurement of observer agreement when the data are categorical. Ph.D. Dissertation, University of North Carolina, Institute of Statistics Mimeo Series No. 1022.

Landis, J. R. and Koch, G. G. [1975a]. A review of statistical methods in the analysis of data arising from observer reliability studies (Part I). *Statistica Neerlandica 29,* 101–123.

Landis, J. R. and Koch, G. G. [1975b]. A review of statistical methods in the analysis of data arising from observer reliability studies (Part II). *Statistica Neerlandica 29,* 151–161.

Landis, J. R. and Koch, G. G. [1977]. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Accepted for publication in *Biometrics.*

Landis, J. R., Stanish, W. M., Freeman, J. L. and Koch, G. G. [1976]. A computer program for the generalized chi-square analysis of categorical data using weighted least squares (GENCAT). University of Michigan Biostatistics Technical Report No. 8. Accepted for publication in *Computer Programs in Biomedicine.*

Light, R. J. [1971]. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin 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.

Neyman, J. [1949]. Contribution to the theory of the χ^2 test. *Proceedings of the Berkeley Symposium on mathematical statistics and probability,* Berkeley and Los Angeles, University of California Press, 239–272.

Overall, J. E. [1968]. Estimating individual rater reliabilities from analysis of treatment effects. *Educational and Psychological Measurement 28,* 255–264.

Scheffé, H. [1959]. *The Analysis of Variance.* Wiley, New York.

Searle, S. R. [1971]. *Linear Models.* Wiley, New York.

Wald, A. [1943]. Tests of statistical hypotheses concerning general parameters when the number of observations is large. *Transactions of the American Mathematical Society 54,* 426–482.

Westlund, K. B. and Kurland, L. T. [1953]. Studies on multiple sclerosis in Winnipeg. Manitoba and New Orleans, Louisiana. *American Journal of Hygiene 57,* 380–396.

Received April 1975, *Revised November* 1975