The Canadian Journal of Statistics
Vol. 27, No. 1, 1999, Pages 3–23
La revue canadienne de statistique

Beyond kappa: A review of interrater agreement measures*

Mousumi BANERJEE

Wayne State University School of Medicine

Michelle CAPOZZOLI, Laura McSWEENEY, and Debajyoti SINHA

University of New Hampshire

Key words and phrases: Kappa coefficient, intraclass correlation, log-linear models, nominal data, ordinal data.
AMS 1991 subject classifications: 62F03, 62G05, 62H20, 62P10.

ABSTRACT

In 1960, Cohen introduced the kappa coefficient to measure chance-corrected nominal scale agreement between two raters. Since then, numerous extensions and generalizations of this interrater agreement measure have been proposed in the literature. This paper reviews and critiques various approaches to the study of interrater agreement, for which the relevant data comprise either nominal or ordinal categorical ratings from multiple raters. It presents a comprehensive compilation of the main statistical approaches to this problem, descriptions and characterizations of the underlying models, and discussions of related statistical methodologies for estimation and confidence-interval construction. The emphasis is on various practical scenarios and designs that underlie the development of these measures, and the interrelationships between them.

RÉSUMÉ

C'est en 1960 que Cohen a proposé l'emploi du coefficient kappa comme outil de mesure de l'accord entre deux évaluateurs exprimant leur jugement au moyen d'une échelle nominale. De nombreuses généralisations de cette mesure d'accord ont été proposées depuis lors. Les auteurs jettent ici un regard critique sur nombre de ces travaux traitant du cas où l'échelle de réponse est soit nominale, soit ordinale. Les principales approches statistiques sont passées en revue, les modèles sous-jacents sont décrits et caractérisés, et les problèmes liés à l'estimation ponctuelle ou par intervalle sont abordés. L'accent est mis sur différents scénarios concrets et sur des schémas expérimentaux qui sous-tendent l'emploi de ces mesures et les relations existant entre elles.

1. INTRODUCTION

In medical and social science research, analysis of observer or interrater agreement data often provides a useful means of assessing the reliability of a rating system. The observers may be physicians who classify patients as having or not having a certain medical condition, or competing diagnostic devices that classify the extent of disease in patients

*This research was partially supported by grant R29-CA69222-02 from the National Cancer Institute to D. Sinha.

into ordinal multinomial categories. At issue in both cases is the intrinsic precision of the classification process. High measures of agreement would indicate consensus in the diagnosis and interchangeability of the measuring devices.

Rater agreement measures have been proposed under various practical situations. Some of these include scenarios where readings are recorded on a continuous scale: measurements on cardiac stroke volume, peak expiratory flow rate, etc. Under such scenarios, agreement measures such as the concordance correlation coefficient (Lin 1989, Chinchilli *et al.* 1996) are appropriate. Specifically, the concordance correlation coefficient evaluates the agreement between the two sets of readings by measuring the variation from the unit line through the origin. Our focus, however, is on agreement measures that arise when ratings are given on a nominal or ordinal *categorical* scale. Scenarios where raters give categorical ratings to subjects occur commonly in medicine; for instance, when routine diagnostic tests are used to classify patients according to the stage and severity of disease. Therefore, the topic of interrater agreement for categorical ratings is of immense importance in medicine.

Early approaches to studying interrater agreement focused on the observed proportion of agreement (Goodman and Kruskal 1954). However, this statistic does not allow for the fact that a certain amount of agreement can be expected on the basis of chance alone and could occur even if there were no systematic tendency for the raters to classify the same subjects similarly. Cohen (1960) proposed kappa as a chance-corrected measure of agreement, to discount the observed proportion of agreement by the expected level of agreement, given the observed marginal distributions of the raters' responses and the assumption that the rater reports are statistically independent. Cohen's kappa allows the marginal probabilities of success associated with the raters to differ. An alternative approach, discussed by Bloch and Kraemer (1989) and Dunn (1989), assumes that each rater may be characterized by the same underlying success rate. This approach leads to the intraclass version of the kappa statistic obtained as the usual intraclass correlation estimate calculated from a one-way analysis of variance, and is algebraically equivalent to Scott's index of agreement (Scott 1955). Approaches based on log-linear and latent-class models for studying agreement patterns have also been proposed in the literature (Tanner and Young 1985a, Agresti 1988, 1992).

Just as various approaches have evolved in studying interrater agreement, many generalizations have also been proposed to the original case of two raters using a nominal scale rating. For example, Cohen (1968) introduced a weighted version of the kappa statistic for ordinal data. Extensions to the case of more than two raters (Fleiss 1971, Light 1971, Landis and Koch 1977a, b, Davies and Fleiss 1982, Kraemer 1980), to paired-data situations (Oden 1991, Schouten 1993, Shoukri *et al.* 1995) and to the inclusion of covariate information (Graham 1995, Barlow 1996) have also been proposed.

The purpose of this paper is to explore the different approaches to the study of interrater agreement, for which the relevant data comprise either nominal or ordinal categorical ratings from multiple raters. It presents a comprehensive compilation of the main statistical approaches to this problem, descriptions and characterizations of the underlying models, as well as discussions of related statistical methodologies for estimation and confidence interval construction. The emphasis is on various practical scenarios and designs that underlie the development of these measures, and the interrelationships between them. In the next section, we review the basic agreement measures. Section 3 presents the various extensions and generalizations of these basic measures, followed by concluding remarks in Section 4.

2. BASIC AGREEMENT MEASURES

2.1. Cohen's Kappa Coefficient.

The most primitive approach to studying interrater agreement was to compute the observed proportion of cases in which the raters agreed, and let the issue rest there. This approach is clearly inadequate, since it does not adjust for the fact that a certain amount of the agreement could occur due to chance alone. Another early approach was based on the chi-square statistic computed from the cross-classification (contingency) table. Again, this approach is indefensible, since chi-square, when applied to a contingency table, measures the degree of *association*, which is not necessarily the same as agreement. The chi-square statistic is inflated quite impartially by any departure from chance association, either disagreement or agreement.

A chance-corrected measure introduced by Scott (1955), was extended by Cohen (1960) and has come to be known as Cohen's kappa. It springs from the notion that the observed cases of agreement include some cases for which the agreement was by chance alone. Cohen assumed that there were two raters, who rate n subjects into one of m mutually exclusive and exhaustive nominal categories. The raters operate independently; however, there is no restriction on the marginal distribution of the ratings for either rater. Let p_{ij} be the proportion of subjects that were placed in the i,jth cell, i.e., assigned to the ith category by the first rater and to the jth category by the second rater ($i,j = 1,\ldots,m$). Also, let $p_{i\cdot} = \sum_{j=1}^{m} p_{ij}$ denote the proportion of subjects placed in the ith row (i.e., the ith category by the first rater), and let $p_{\cdot j} = \sum_{i=1}^{m} p_{ij}$ denote the proportion of subjects placed in the jth column (i.e., the jth category by the second rater). Then, the kappa coefficient proposed by Cohen is

$$\hat{\kappa} = \frac{p_o - p_c}{1 - p_c},$$

where $p_o = \sum_{i=1}^{m} p_{ii}$ is the observed proportion of agreement and $p_c = \sum_{i=1}^{m} p_{i\cdot}p_{\cdot i}$ is the proportion of agreement expected by chance. Cohen's kappa is an extension of Scott's index in the following sense: Scott defined p_c using the underlying assumption that the distribution of proportions over the m categories for the population is known, and is equal for the two raters. Therefore, if the two raters are interchangeable, in the sense that the marginal distributions are identical, then Cohen's and Scott's measures are equivalent. To determine whether $\hat{\kappa}$ differs significantly from zero, one could use the asymptotic variance formula given by Fleiss *et al.* (1969) for the general $m \times m$ table. For large n, Fleiss *et al.*'s formula is practically equivalent to the exact variance derived by Everitt (1968) based on the central hypergeometric distribution. Under the hypothesis of only chance agreement, the estimated large-sample variance of $\hat{\kappa}$ is given by

$$\widehat{Var}_0(\hat{\kappa}) = \frac{p_c + p_c^2 - \sum_{i=1}^{m} p_{i\cdot}p_{\cdot i}(p_{i\cdot} + p_{\cdot i})}{n(1 - p_c)^2}.$$

Assuming that $\hat{\kappa}/\sqrt{\widehat{Var}_0(\hat{\kappa})}$ follows a normal distribution, one can test the hypothesis of chance agreement by reference to the standard normal distribution. In the context of reliability studies, however, this test of hypothesis is of little interest, since generally the raters are trained to be reliable. In this case, a lower bound on kappa is more appropriate. This requires estimating the nonnull variance of $\hat{\kappa}$, for which Fleiss *et al.* provided an

approximate asymptotic expression, given by:

$$\widehat{Var}(\hat{\kappa}) = \frac{1}{n(1-p_c)^2}\left(\sum_{i=1}^{m} p_{ii}\{1-(p_{i\cdot}+p_{\cdot i})(1-\hat{\kappa})\}^2 + (1-\hat{\kappa})^2\sum_{i\neq j}^{m} p_{ij}(p_{i\cdot}+p_{\cdot j})^2 - \{\hat{\kappa}-p_c(1-\hat{\kappa})\}^2\right).$$

Cicchetti and Fleiss (1977) and Fleiss and Cicchetti (1978) have studied the accuracy of the large-sample standard error of $\hat{\kappa}$ via Monte Carlo simulations.

Landis and Koch (1977a) have characterized different ranges of values for kappa with respect to the degree of agreement they suggest. Although these original suggestions were admitted to be "clearly arbitrary", they have become incorporated into the literature as standards for the interpretation of kappa values. For most purposes, values greater than 0.75 or so may be taken to represent excellent agreement beyond chance, values below 0.40 or so may be taken to represent poor agreement beyond chance, and values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance.

Much controversy has surrounded the use and interpretation of kappa, particularly regarding its dependence on the marginal distributions. The marginal distributions describe how the raters separately allocate subjects to the response categories. "Bias" of one rater relative to another refers to discrepancies between these marginal distributions. Bias decreases as the marginal distributions become more nearly equivalent. The effect of rater bias on kappa has been investigated by Feinstein and Cicchetti (1990) and Byrt *et al.* (1993). Another factor that affects kappa is the true *prevalence* of a diagnosis, defined as the proportions of cases of the various types in the population. The same raters or diagnostic procedures can yield different values of kappa in two different populations (Feinstein and Cicchetti 1990, Byrt *et al.* 1993). In view of the above, it is important to recognize that agreement studies conducted in samples of convenience or in populations known to have a high prevalence of the diagnosis do not necessarily reflect on the agreement between the raters.

Some authors (Hutchinson 1993) deem it disadvantageous that Cohen's kappa mixes together two components of disagreement that are inherently different, namely, disagreements which occur due to bias between the raters, and disagreements which occur because the raters rank-order the subjects differently. A much-adopted solution to this is the intraclass kappa statistic (Bloch and Kraemer 1989) discussed in Section 2.3. However, Zwick (1988) points out that rather than straightway ignoring marginal disagreement or attempting to correct for it, researchers should be *studying* it to determine whether it reflects important rater differences or merely random error. Therefore, any assessment of rater agreement should routinely begin with the investigation of marginal homogeneity.

2.2. Weighted Kappa Coefficient.

Often situations arise when certain disagreements between two raters are more serious than others. For example, in an agreement study of psychiatric diagnosis in the categories personality disorder, neurosis and psychosis, a clinician would likely consider a diagnostic disagreement between neurosis and psychosis to be more serious than between neurosis and personality disorder. However, $\hat{\kappa}$ makes no such distinction, implicitly treating all disagreements equally. Cohen (1968) introduced an extension of kappa called the weighted kappa statistic ($\hat{\kappa}_w$), to measure the proportion of weighted agreement corrected

for chance. Either degree of disagreement or degree of agreement is weighted, depending on what seems natural in a given context.

The statistic $\hat{\kappa}_w$ provides for the incorporation of ratio-scaled degrees of disagreement (or agreement) to each of the cells of the $m \times m$ table of joint assignments such that disagreements of varying gravity (or agreements of varying degree) are weighted accordingly. The nonnegative weights are set prior to the collection of the data. Since the cells are scaled for degrees of disagreement (or agreement), some of them are not given *full* disagreement credit. However, $\hat{\kappa}_w$, like the unweighted $\hat{\kappa}$, is *fully* chance-corrected.

Assuming that w_{ij} represents the weight for agreement assigned to the i,jth cell ($i, j = 1,\ldots,m$), the weighted kappa statistic is given by

$$\hat{\kappa}_w = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}p_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}p_{i\cdot}p_{\cdot j}}{1 - \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}p_{i\cdot}p_{\cdot j}}.$$

Note that the unweighted kappa is a special case of $\hat{\kappa}_w$ with $w_{ij} = 1$ for $i = j$ and $w_{ij} = 0$ for $i \neq j$. If, on the other hand, the m categories form an ordinal scale, with the categories assigned the numerical values $1, 2, \ldots, m$, and if $w_{ij} = 1 - (i-j)^2/(m-1)^2$, then $\hat{\kappa}_w$ can be interpreted as an intraclass correlation coefficient for a two-way ANOVA computed under the assumption that the n subjects and the two raters are random samples from populations of subjects and raters, respectively (Fleiss and Cohen 1973).

Fleiss *et al.* (1969) derived the formula for the asymptotic variance of $\hat{\kappa}_w$, for both the null and the nonnull case. Their formula has been evaluated for its utility in significance testing and confidence-interval construction by Cicchetti and Fleiss (1977) and Fleiss and Cicchetti (1978). Based on Monte Carlo studies, the authors report that only moderate sample sizes are required to test the hypothesis that two independently derived estimates of weighted kappa are equal. However, the minimal sample size required for setting confidence limits around a single value of weighted kappa is $n = 16m^2$, which is inordinately large in most cases.

2.3. Intraclass Kappa.

Bloch and Kraemer (1989) introduced the intraclass correlation coefficient as an alternative version of Cohen's kappa, using the assumption that each rater is characterized by the same underlying marginal probability of categorization. This intraclass version of the kappa statistic is algebraically equivalent to Scott's index of agreement (Scott 1955).

The intraclass kappa was defined by Bloch and Kraemer (1989) for data consisting of blinded dichotomous ratings on each of n subjects by two fixed raters. It is assumed that the ratings on a subject are interchangeable; i.e., in the population of subjects, the two ratings for each subject have a distribution that is invariant under permutations of the raters. This means that there is no *rater bias*. Let X_{ij} denote the rating for the ith subject by the jth rater, $i = 1,\ldots,n$, $j = 1, 2$, and for each subject i, let $p_i = P(X_{ij} = 1)$ be the probability that the rating is a success. Over the population of subjects, let $\mathcal{E}\, p_i = P$, $P' = 1 - P$ and $\mathcal{V}ar\ p_i = \sigma_P^2$. The intraclass kappa is then defined as

$$\kappa_I = \frac{\sigma_P^2}{PP'}. \tag{1}$$

An estimator of the intraclass kappa can be obtained by introducing the probability model in Table 1 for the joint responses, with the kappa coefficient explicitly defined in its parametric structure. Thus, the log-likelihood function is given by

$$\ln L(P, \kappa_I | n_{11}, n_{12}, n_{21}, n_{22}) = n_{11} \ln(P^2 + \kappa_I PP') + (n_{12} + n_{21}) \ln\{PP'(1 - \kappa_I)\} + n_{22} \ln(P'^2 + \kappa_I PP').$$

TABLE 1: Underlying model for estimation of intraclass kappa.

Response type			
X_{i1}	X_{i2}	Obs. freq.	Expected probability
1	1	n_{11}	$P^2 + \kappa_I PP'$
1	0	n_{12}	$PP'(1 - \kappa_I)$
0	1	n_{21}	$PP'(1 - \kappa_I)$
0	0	n_{22}	$P'^2 + \kappa_I PP'$

The maximum-likelihood estimators $\hat{p}$ and $\hat{\kappa}_I$ for P and κ_I are obtained as

$$\hat{p} = \frac{2n_{11} + n_{12} + n_{21}}{2n},$$

and

$$\hat{\kappa}_I = \frac{4(n_{11}n_{22} - n_{12}n_{21}) - (n_{12} - n_{21})^2}{(2n_{11} + n_{12} + n_{21})(2n_{22} + n_{12} + n_{21})}, \tag{2}$$

with the estimated standard error for $\hat{\kappa}_I$ given by (Bloch and Kraemer 1989)

$$\mathrm{SE}(\hat{\kappa}_I) = \left\{ \frac{1 - \hat{\kappa}_I}{n} \left((1 - \hat{\kappa}_I)(1 - 2\hat{\kappa}_I) + \frac{\hat{\kappa}_I(2 - \hat{\kappa}_I)}{2\hat{p}(1 - \hat{p})} \right) \right\}^{\frac{1}{2}}. \tag{3}$$

The estimate $\hat{\kappa}_I$, the MLE of κ_I as defined by (1) under the above model, is identical to the estimator of an intraclass correlation coefficient for 0-1 data. If the formula for the intraclass correlation for continuous data (Snedecor and Cochran 1967) is applied to dichotomous data, then the estimate $\hat{\kappa}_I$ is obtained. Assuming $\hat{\kappa}_I$ is normally distributed with mean κ_I and standard error $\mathrm{SE}(\hat{\kappa}_I)$, the resulting $100(1 - \alpha)\%$ confidence interval is given by $\hat{\kappa}_I \pm z_{1-\alpha/2}\ \mathrm{SE}(\hat{\kappa}_I)$, where $z_{1-\alpha/2}$ is the $100(1 - \alpha)$ percentile point of the standard normal distribution. The above confidence interval has reasonable properties only in very large samples that are not typical of the sizes of most interrater agreement studies (Bloch and Kraemer 1989, Donner and Eliasziw 1992).

Bloch and Kraemer (1989) also derive a variance-stabilizing transformation for $\hat{\kappa}_I$, which provide improved accuracy for confidence-interval estimation, power calculations or formulations of tests. A third approach (Bloch and Kraemer 1989, Fleiss and Davies 1982) is based on the jackknife estimator $\hat{\kappa}_J$ of κ_I. This estimator is obtained by averaging the estimators $\hat{\kappa}_{-i}$, where $\hat{\kappa}_{-i}$ is the value of $\hat{\kappa}_I$ obtained over all subjects except the ith. Bloch and Kraemer present a large-sample variance for $\hat{\kappa}_J$ which can be used to construct confidence limits. However, the authors point out that the probability of obtaining degenerate results ($\hat{\kappa}_J$ undefined) is relatively high in smaller samples, especially as P approaches 0 or 1 or κ_I approaches 1.

For confidence-interval construction in small samples, Donner and Eliasziw (1992) propose a procedure based on a chi-square goodness-of-fit statistic. Their approach is based on equating the computed one-degree-of-freedom chi-square statistic to an appropriately selected critical value, and solving for the two roots of kappa. Using this approach, the upper ($\hat{\kappa}_U$) and lower ($\hat{\kappa}_L$) limits of a $100(1 - \alpha)\%$ confidence interval for κ_I are obtained as

$$\kappa_L = (\tfrac{1}{9}y_3^2 - \tfrac{1}{3}y_2)^{\frac{1}{2}} \left(\cos \tfrac{\theta + 2\pi}{3} + \sqrt{3} \sin \tfrac{\theta + 2\pi}{3} \right) - \tfrac{1}{3}y_3$$

$$\kappa_U = 2(\tfrac{1}{9}y_3^2 - \tfrac{1}{3}y_2)^{\frac{1}{2}} \cos \tfrac{\theta + 5\pi}{3} - \tfrac{1}{3}y_3, \qquad \pi = 3.14,$$

where

$$\theta = \arccos \frac{V}{W}, \qquad V = \tfrac{1}{27} y_3^3 - \tfrac{1}{6}(y_2 y_3 - 3y_1), \qquad W = (\tfrac{1}{9} y_3^2 - \tfrac{1}{3} y_2)^{\frac{1}{2}};$$

and

$$y_1 = \frac{\{n_{12} + n_{21} - 2n\hat{P}(1-\hat{P})\}^2 + 4n^2\hat{P}^2(1-\hat{P})^2}{4n\hat{P}^2(1-\hat{P})^2(\chi^2_{1,1-\alpha} + n)} - 1,$$

$$y_2 = \frac{(n_{12} + n_{21})^2 - 4n\hat{P}(1-\hat{P})\{1 - 4\hat{P}(1-\hat{P})\}\chi^2_{1,1-\alpha}}{4n\hat{P}^2(1-\hat{P})^2(\chi^2_{1,1-\alpha} + n)} - 1,$$

$$y_3 = \frac{n_{12} + n_{21} + \{1 - 2\hat{P}(1-\hat{P})\}\chi^2_{1,1-\alpha}}{\hat{P}(1-\hat{P})(\chi^2_{1,1-\alpha} + n)} - 1.$$

The coverage levels associated with the goodness-of-fit procedure have improved accuracy in small samples across all values of κ_I and P. Donner and Eliasziw (1992) also describe hypothesis-testing and sample-size calculations using this goodness-of-fit procedure. The above approach has been extended recently by Donner and Eliasziw (1997) to the case of three or more rating categories per subject. Their method is based on a series of nested, statistically independent inferences, each corresponding to a binary outcome variable obtained by combining a substantively relevant subset of the original categories.

2.4. Tetrachoric Correlation Coefficient.

In the health sciences, many clinically detected abnormalities which are apparently dichotomous have an underlying continuum which cannot be measured as such, for technical reasons or because of the limitations of human perceptual ability. An example is radiological assessment of pneumoconiosis, which is assessed from chest radiographs displaying a profusion of small irregular opacities. Analytic techniques commonly used for such data treat the response measure as if it were truly binary (abnormal-normal). Irwig and Groeneveld (1988) discuss several drawbacks of this approach. Firstly, it ignores the fact that ratings from two observers may differ because of threshold choice. By "threshold" we mean the value along the underlying continuum above which raters regard abnormality as present. Two raters may use different thresholds due to differences in their visual perception or decision attitude, even in the presence of criteria which attempt to define a clear boundary. Furthermore, with such data, the probability of misclassifying a case across the threshold is clearly dependent on the true value of the underlying continuous variable; the more extreme the true value (the further away from a specified threshold), the smaller the probability of misclassification. Since this is so for all the raters, their misclassification probabilities cannot be independent. Therefore, kappa-type measures (i.e., unweighted and weighted kappas, intraclass kappa) are inappropriate in such situations.

When the diagnosis is regarded as the dichotomization of an underlying continuous variable that is unidimensional with a standard normal distribution, the tetrachoric correlation coefficient (TCC) (Pearson 1901) is an obvious choice for estimating interrater agreement. Specifically, the TCC estimates the correlation between the *actual* latent (unobservable) variables characterizing the raters' probability of abnormal diagnosis, and is based on assuming bivariate normality of the raters' latent variables. Therefore, not only does the context under which TCC is appropriate differ from that for kappa-type measures, but quantitatively they estimate two different, albeit related, entities (Kraemer

1997). Several twin studies have used the TCC as a statistical measure of concordance among monozygotic and dizygotic twins, with respect to certain dichotomized traits (Corey *et al.* 1992; Kendler *et al.* 1992; Kvaerner *et al.* 1997).

The tetrachoric correlation coefficient is obtained as the maximum-likelihood estimate for the correlation coefficient in the bivariate normal distribution, when only information in the contingency table is available (Tallis 1962, Hamdan 1970). The computation of TCC is based on an iterative process, using tables for the bivariate normal integral (Johnson and Kotz 1972). It has recently been implemented in SAS, and can be obtained through the `/plcorr` option with the `tables` statement in the `PROC FREQ` procedure.

3. EXTENSIONS AND GENERALIZATIONS

3.1. Case of Two Raters.

(a) Kappa coefficient from paired data.

Suppose two raters classify both the left and right eyes in a group of n patients for the presence or absence of a specified abnormality. Interrater agreement measures based on rating such paired body parts should allow for the positive correlation generally present between observations made on the paired organs of the same patient. It is incorrect to treat the data as if they arose from a random sample of $2n$ organs. The application of a variance formula such as that given by Fleiss *et al.* (1969) may lead to unrealistically narrow confidence intervals for kappa in this context, and spuriously high rejection rates for tests against zero. This is often countered by calculating separate kappa values for the two organs. However, this approach is again inefficient and lacks conciseness in the presentation of the results.

Oden (1991) proposed a method to estimate a pooled kappa between two raters when both raters rate the same set of pairs of eyes. His method assumes that the true left-eye and right-eye kappa values are equal and makes use of the correlated data to estimate confidence intervals for the common kappa. The pooled kappa estimator is a weighted average of the kappas for the right and left eyes, and is given by

$$\hat{\kappa}_{\text{pooled}} = \frac{(1 - \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}\rho_{i\cdot}\rho_{\cdot j})\hat{\kappa}_{\text{right}} + (1 - \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}\lambda_{i\cdot}\lambda_{\cdot j})\hat{\kappa}_{\text{left}}}{(1 - \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}\rho_{i\cdot}\rho_{\cdot j}) + (1 - \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}\lambda_{i\cdot}\lambda_{\cdot j})},$$

where

ρ_{ij} = proportion of patients whose right eye was rated i by rater 1 and j by rater 2,
λ_{ij} = proportion of patients whose left eye was rated i by rater 1 and j by rater 2,
w_{ij} = agreement weight that reflects the degree of agreement between raters 1 and 2 if they use ratings i and j respectively for the same eye,

and $\rho_{i\cdot}$, $\rho_{\cdot j}$, $\lambda_{i\cdot}$, $\lambda_{\cdot j}$ have their usual meanings. Applying the delta method, Oden obtained an approximate standard error of the pooled kappa estimator. The pooled estimator was shown to be roughly unbiased (the average bias, based on simulations, was of the order of 10^{-3}) and had better performance than either the naive two-eye estimator (which treats the data as a random sample of $2n$ eyes) or the estimator based on either single eye, in terms of correct coverage probability of the 95% confidence interval for the true kappa (Oden 1991).

Schouten (1993) presented an alternative approach in this context. He noted that existing formula for the computation of weighted kappa and its standard error (Cohen

TABLE 2: Binocular data frequencies and agreement weights.

Grader 1	Grader 2: R+L+	R+L−	R−L+	R−L−	Total
R+L+	f_{11} (1.0)	f_{12} (0.5)	f_{13} (0.5)	f_{14} (0.0)	$f_{1\cdot}$
R+L−	f_{21} (0.5)	f_{22} (1.0)	f_{23} (0.0)	f_{24} (0.5)	$f_{2\cdot}$
R−L+	f_{31} (0.5)	f_{32} (0.0)	f_{33} (1.0)	f_{34} (0.5)	$f_{3\cdot}$
R−L−	f_{41} (0.0)	f_{42} (0.5)	f_{43} (0.5)	f_{44} (1.0)	$f_{4\cdot}$
Total	$f_{\cdot 1}$	$f_{\cdot 2}$	$f_{\cdot 3}$	$f_{\cdot 4}$	n

1968, Fleiss *et al.* 1969) can be used if the observed as well as the chance agreement is averaged over the two sets of eyes and then substituted into the formula for kappa. To this end, let each eye be diagnosed normal or abnormal, and let each patient be categorized into one of the following four categories by each rater:

R+L+: abnormality is present in both eyes,
R+L−: abnormality is present in the right eye but not in the left eye,
R−L+: abnormality is present in the left eye but not in the right eye,
R−L−: abnormality is absent in both eyes.

The frequencies of the ratings can be represented as shown in Table 2.

Schouten used the weighted kappa statistic to determine an overall agreement measure. He defined the agreement weights w_{ij} to be 1.0 (complete agreement) if the raters agreed on both eyes, 0.5 (partial agreement) if the raters agreed on one eye and disagreed on the other, and 0.0 (complete disagreement) if the raters disagreed on both eyes. The agreement weights for each cell are represented in parenthesis in Table 2.

The overall agreement measure is then defined to be $\hat{\kappa}_w = (p_o - p_c)/(1 - p_c)$, where

$$p_o = \frac{\sum_{i=1}^{4} \sum_{j=1}^{4} w_{ij} f_{ij}}{n},$$

and

$$p_c = \frac{\sum_{i=1}^{4} \sum_{j=1}^{4} w_{ij} f_{i\cdot} f_{\cdot j}}{n^2},$$

and the w_{ij}'s are as defined in Table 2. Formulae for the standard error can be calculated as in Fleiss *et al.* (1969). Note that the above agreement measure can be easily extended to accommodate more than two rating categories by simply adjusting the agreement weights. Furthermore, both Oden's and Schouten's approaches can be generalized for the setting in which more than two (similar) organs are evaluated, e.g., several glands or blood vessels.

Shoukri *et al.* (1995) consider a different type of pairing situation where raters classify individuals blindly by two different rating protocols into one of two categories. The purpose is to establish the congruent validity of the two rating protocols. For example, two recent tests for routine diagnosis of paratuberculosis in cattle animals are the dot immunobinding assay (DIA) and the enzyme linked immunosorbent assay (ELISA). Comparison of the results of these two tests depends on the serum samples obtained from cattle. One then evaluates the same serum sample using both tests — a procedure that clearly creates a realistic "matching".

Let $X_i = 1$ or 0 according to whether the ith ($i = 1, 2, \ldots, n$) serum sample tested by DIA was positive or negative, and let $Y_i = 1$ or 0 denote the corresponding test status of the matched serum sample when tested by the ELISA. Let π_{kl} ($k, l = 0, 1$) denote the

where n_{1h} is the number of subjects in study h who received success ratings from both raters, n_{2h} is the number who received one success and one failure rating, n_{3h} is the number who received failure ratings from both raters, and $n_h = n_{1h}+n_{2h}+n_{3h}$. An overall measure of agreement among the studies is estimated by computing a weighted average of the individual $\hat{\kappa}_h$, yielding

$$\hat{\kappa} = \frac{\sum_{h=1}^{N} n_h \hat{P}_h(1-\hat{P}_h)\hat{\kappa}_h}{\sum_{h=1}^{N} n_h \hat{P}_h(1-\hat{P}_h)}.$$

To test $\mathcal{H}_0 : \kappa_1 = \kappa_2 = \cdots = \kappa_N$, Donner *et al.* propose a goodness-of-fit test using the statistic

$$\chi^2_G = \sum_{h=1}^{N}\sum_{l=1}^{3} \frac{\{n_{lh} - n_h\hat{\pi}_{lh}(\hat{\kappa})\}^2}{n_h\hat{\pi}_{lh}(\hat{\kappa})},$$

where $\hat{\pi}_{lh}(\hat{\kappa})$ is obtained by replacing P_h by $\hat{P}_h$ and κ_h by $\hat{\kappa}$ in $\pi_{lh}(\kappa_h)$; $l = 1,2,3$; $h = 1,2,\ldots,N$. Under the null hypothesis, χ^2_G follows an approximate chi-square distribution with $N-1$ degrees of freedom. Methods to test a variety of related hypotheses, based on the goodness-of-fit theory, are described by Donner and Klar (1996).

Donner *et al.* (1996) also discuss another method of testing $\mathcal{H}_0 : \kappa_1 = \kappa_2 = \cdots = \kappa_N$ using a large-sample variance approach. The estimated large-sample variance of $\hat{\kappa}_h$ (Bloch and Kraemer 1989, Fleiss and Davies 1982) is given by

$$\widehat{Var}\ \hat{\kappa}_h = \frac{1-\hat{\kappa}_h}{n_h}\left((1-\hat{\kappa}_h)(1-2\hat{\kappa}_h) + \frac{\hat{\kappa}_h(2-\hat{\kappa}_h)}{2\hat{P}_h(1-\hat{P}_h)}\right).$$

Letting $\hat{W}_h = 1/\widehat{Var}\ \hat{\kappa}_h$ and $\tilde{\kappa} = (\sum_{h=1}^{N} \hat{W}_h\hat{\kappa}_h)/(\sum_{h=1}^{N} \hat{W}_h)$, an approximate test of $\mathcal{H}_0$ is obtained by referring $\chi^2_V = \sum_{h=1}^{N} \hat{W}_h(\hat{\kappa}_h - \tilde{\kappa})^2$ to tables of the chi-square distribution with $N-1$ degrees of freedom.

The statistic χ^2_V is undefined if $\hat{\kappa}_h = 1$ for any h. Unfortunately, this event can occur with fairly high frequency in samples of small to moderate size. In contrast, the goodness-of-fit test statistic, χ^2_G, can be calculated except in the extreme boundary case of $\hat{\kappa}_h = 1$ for all $h = 1,2,\ldots,N$, when a formal test of significance has no practical value. Neither test statistic can be computed when $\hat{P}_h = 0$ or 1 for any h, since then $\hat{\kappa}_h$ is undefined. Based on a Monte Carlo study, the authors found that the two statistics have similar properties for large samples ($n_h > 100$ for all h). In this case differences in power tend to be negligible except in the case of unequal π_h's or very unequal n_h's, where χ^2_G tends to have a small but consistent advantage over χ^2_V. For smaller sample sizes, clearly the goodness-of-fit statistic χ^2_G is preferable.

3.2. Case of Multiple Raters: Generalizations of Kappa.

Fleiss (1971) proposed a generalization of Cohen's kappa statistic to the measurement of agreement among a constant number of raters (say, K). Each of the n subjects are rated by K (> 2) raters independently into one of m mutually exclusive and exhaustive nominal categories. This formulation applies to the case of *different* sets of raters (i.e., random ratings) for each subject. The motivating example is a study in which each of 30 patients was rated by 6 psychiatrists (selected randomly from a total pool of 43 psychiatrists) into one of five categories.

Let K_{ij} be the number of raters who assigned the ith subject to the jth category, $i = 1,\ldots,n, j = 1,\ldots,m$, and define

$$p_j = \frac{1}{Kn}\sum_{i=1}^{n} K_{ij}.$$

Note that p_j is the proportion of all assignments which were to the jth category. The chance-corrected measure of overall agreement proposed by Fleiss (1971) is given by

$$\hat{\kappa} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} K_{ij}^2 - Kn\{1 + (K-1)\sum_{j=1}^{m} p_j^2\}}{nK(K-1)(1 - \sum_{j=1}^{m} p_j^2)}.$$

Under the null hypothesis of no agreement beyond chance, the K assignments on each subject are multinomial variables with probabilities $p_1,\ldots,p_m$. Using this, Fleiss (1971) obtained an approximate asymptotic variance of $\hat{\kappa}$ under the hypothesis of no agreement beyond chance:

$$\mathcal{V}ar_0\,\hat{\kappa} = \frac{2}{nK(K-1)} \times \frac{\sum_{j=1}^{m} p_j^2 - (2K-3)(\sum_{j=1}^{m} p_j^2)^2 + 2(K-2)\sum_{j=1}^{m} p_j^3}{(1 - \sum_{j=1}^{m} p_j^2)^2}.$$

In addition to the $\hat{\kappa}$-statistic for measuring overall agreement, Fleiss (1971) also proposed a statistic to measure the extent of agreement in assigning a subject to a particular category. A measure of the beyond-chance agreement in assignment to category j is given by

$$\hat{\kappa}_j = \frac{\sum_{i=1}^{n} K_{ij}^2 - Knp_j\{1 + (K-1)p_j\}}{nK(K-1)p_j(1-p_j)}.$$

It is easy to see that $\hat{\kappa}$, the measure of overall agreement, is a weighted average of the $\hat{\kappa}_j$'s, with the corresponding weights $p_j(1-p_j)$. Under the null hypothesis of no agreement beyond chance, the approximate asymptotic variance of $\hat{\kappa}_j$ is

$$\mathcal{V}ar_0\,\hat{\kappa}_j = \frac{\{1 + 2(K-1)p_j\}^2 + 2(K-1)p_j(1-p_j)}{nK(K-1)^2 p_j(1-p_j)}.$$

In a different, unbalanced setting (subjects rated by different numbers of raters), Landis and Koch (1977b) associated Fleiss's $\hat{\kappa}$ with the intraclass correlation coefficient computed for a one-way random-effects ANOVA model with the single factor corresponding to the (random) subjects. Davies and Fleiss (1982) demonstrated this equivalence for a two-way balanced layout. Specifically, the authors proposed a kappa-like statistic for a set of multinomial random variables arrayed in a two-way (subject by rater) layout, and showed that this statistic can be obtained either via chance-correction of the average proportion of pairwise agreement, or via an analysis of variance for a two-way layout. In contrast to Fleiss's (1971) formulation, theirs applies to the case of a *common* set of raters who judge all subjects. Applications include the case where each of the *same* set of several physicians classifies each of a sample of patients into one of several mutually exclusive and exhaustive categories, or the case where blood samples are classified by each of the same set of several assay methods.

Kraemer (1980) considered the issue of different numbers of ratings per subject. She also relaxed the conditions of mutually exclusive categories by allowing a subject to be classified into more than one category by the same rater. For example, a subject could

be classified in category A or category B equally (A/B) or in category A primarily and category B secondarily (AB). Although both of these categorizations involve both A and B, they are treated differently. The extension of the kappa coefficient in this scenario ($\hat{\kappa}_o$, say) is derived by regarding an observation on a subject as a rank ordering of the m categories. In the example above, a rating of A/B would impose a rank of 1.5 on the categories A and B and a rank of $\frac{1}{2}(m+3)$ on the other $m-2$ categories. Using the Spearman rank correlation coefficient, $\hat{\kappa}_o$ is defined to be $\hat{\kappa}_o = (r_I - r_T)/(1 - r_T)$, where r_I is the unweighted average of the intrasubject correlation coefficients and r_T is the average Spearman rank correlation coefficient among all pairs of observations in the sample. Equivalently, this kappa statistic can be computed using an analysis of variance on the ranks.

For most forms of kappa in the multiple-raters case, only the asymptotic variance, under the null hypothesis of no beyond-chance agreement, is known. Davies and Fleiss (1982) discuss some interesting applications where the hypothesis that the population kappa equals zero might be of interest. Specifically, in estimating familial aggregation of certain psychiatric disorders, where several family members are asked to report independently on other members of their family, failure to reject the null hypothesis might suggest that reliance on the reports from available relatives for information about unavailable relatives might produce totally unreliable data.

3.3. *Modelling Patterns of Agreement.*

(a) Log-linear models.

Rather than summarizing agreement by a single number, Tanner and Young (1985a) model the structure of the agreement in the data. They consider log-linear models to express agreement in terms of components, such as chance agreement and beyond-chance agreement. Using the log-linear modelling approach, one can display patterns of agreement among several observers, or compare patterns of agreement when subjects are stratified by values of a covariate. Assuming that there are n subjects who are rated by the same K raters ($K \geq 2$) into m *nominal* categories, Tanner and Young express chance agreement, or statistical independence of the ratings, using the following log-linear model representation:

$$\log v_{ij\ldots l} = u + u_i^{R_1} + u_j^{R_2} + \cdots + u_l^{R_K}, \qquad i,j,\ldots,l = 1,\ldots,m, \tag{4}$$

where $v_{ij\ldots l}$ is the expected cell count in the $ij\ldots l$th cell of the joint K-dimensional cross-classification of the ratings, u is the overall effect, $u_c^{R_k}$ is the effect due to categorization by the kth rater in the cth category ($k = 1,\ldots,K$; $c = 1,\ldots,m$), and $\sum_{i=1}^{m} u_i^{R_1} = \cdots = \sum_{l=1}^{m} u_l^{R_K} = 0$. A useful generalization of the independence model incorporates agreement beyond chance in the following fashion:

$$\log v_{ij\ldots l} = u + u_i^{R_1} + u_j^{R_2} + \cdots + u_l^{R_K} + \delta_{ij\ldots l}. \tag{5}$$

The additional term $\delta_{ij\ldots l}$ represents agreement beyond chance for the $ij\ldots l$th cell. To test a given hypothesis concerning the agreement structure, the parameters corresponding to the agreement component $\delta_{ij\ldots l}$ are assigned to specific cells or groups of cells in the contingency table. The term $\delta_{ij\ldots l}$ can be defined according to what type of agreement pattern is being investigated. For example, to investigate homogeneous agreement among $K = 2$ raters, one would define δ_{ij} to be equal to a common δ when $i = j$, and 0 when

$i \neq j$. On the other hand, to investigate a possibly nonhomogeneous pattern of agreement (i.e., differential agreement by response category), one would consider $\delta_{ij} = \delta_i I(i = j)$, $i,j = 1,\ldots,m$, where the indicator $I(i = j)$ equals 1 when $i = j$, and 0 when $i \neq j$. For the general scenario of $K > 2$ raters, this approach addresses higher-order agreement as well as pairwise agreement (Tanner and Young 1985a). The parameters then describe conditional agreement: for instance, the agreement between two raters for fixed ratings by the other raters.

As Agresti (1992) points out for the case of two raters, the odds that the ratings are concordant rather than discordant can be related to parameters in the log-linear model representation. This makes log-linear models good vehicles for studying agreement. Furthermore, under this representation, the models for independence, homogeneous agreement and nonhomogeneous agreement form a nested sequence of models. Therefore, using the partitioning property of the likelihood-ratio chi-square statistic, one could examine the improvement in fit given the introduction of a set of parameters. Specifically, a comparison of the likelihood-ratio chi-square statistics for the model of independence and the model of homogeneous agreement can be used to assess whether, in fact, there is any beyond-chance agreement. Similarly, assuming that the model for nonhomogeneous agreement is correct, a comparison of the associated likelihood-ratio chi-square statistic to that corresponding to the model for homogeneous agreement can be used to test whether the agreement is uniform. These models can be easily fitted using available software such as GLIM or SAS. Under the log-linear modelling approach, the use of odds ratios offer an alternative to characterizing interobserver agreement by a kappa statistic, or intraclass correlation coefficient. The latter approach has important advantages too. For example, one advantage of the intraclass correlation approach is that this parameter is bounded, with specific values within its range representing relatively well-understood levels of agreement. However, a major disadvantage of this approach is the loss of information from summarizing the table by a single number. In this respect, the log-linear modelling approach is better, since it allows one to *model* the *structure* of agreement, rather than simply describing it with a single summary measure.

Log-linear models treat the raters in a symmetric manner. In some applications, one rater might be a gold-standard device, in which case asymmetric interpretations may be of greater interest. In such situations, one can express the models (4) and (5) in terms of logits of probabilities for a rater's response, conditional on the standard rating (Tanner and Young 1985a).

Graham (1995) extended Tanner and Young's approach to accommodate one or more categorical covariates in assessing agreement pattern between two raters. The baseline for studying covariate effects is taken as the conditional independence model, thus allowing covariate effects on agreement to be studied independently of each other and of covariate effects on the marginal observer distributions. For example, the baseline model for two raters and a categorical covariate X is given by

$$\log v_{ijx} = u + u_i^{R_1} + u_j^{R_2} + u_x^X + u_{ix}^{R_1X} + u_{jx}^{R_2X}, \qquad i,j = 1,\ldots,m,$$

where $u_i^{R_1}$, $u_j^{R_2}$ are as defined in Equation (4), u_x^X is the effect of the xth level of the covariate X, and $u_{ix}^{R_kX}$ is the effect of the partial association between the kth rater ($k = 1,2$) and the covariate X. Given the level of the covariate X, the above model assumes independence between the two raters' reports. Graham uses this conditional independence model as the baseline from which to gauge the strength of agreement. The beyond-chance agreement is modelled as follows:

$$\log v_{ijx} = u + u_i^{R_1} + u_j^{R_2} + u_x^X + u_{ix}^{R_1X} + u_{jx}^{R_2X} + \delta^{R_1R_2} I(i = j) + \delta_x^{R_1R_2X} I(i = j),$$

where $\delta^{R_1R_2}I(i = j)$ represents overall beyond-chance agreement, and $\delta_x^{R_1R_2X}I(i = j)$ represents additional chance-corrected agreement associated with the xth level of the covariate X. Likelihood-ratio goodness-of-fit statistics can be used to compare the full model, which is saturated in the case of a single covariate and a binary response, with a reduced model which assumes $\delta_x^{R_1R_2X} = 0$ for all x, that is, the magnitude of chance-corrected agreement is the same at all levels of the covariate (e.g., the magnitude of chance-corrected agreement is the same for men and women). Inferences can also be based on the covariate agreement parameter estimates and their estimated asymptotic standard errors.

(b) Latent-class models.

Several authors have proposed latent-class models to investigate interrater agreement (Aickin 1990, Uebersax and Grove 1990, Agresti 1992). Latent-class models express the joint distribution of ratings as a mixture of distributions for classes of an unobserved (latent) variable. Each distribution in the mixture applies to a cluster of subjects representing a separate class of a categorical latent variable, those subjects being homogeneous in some sense. Following Agresti (1992), we describe a basic latent-class model for interrater agreement data. This approach treats both the observed scale and the latent variable as discrete.

Suppose there are three raters, namely, A, B and C, who rate each of n subjects into m nominal categories. The latent-class model assumes that there is an unobserved categorical scale X, with L categories, such that subjects in each category of X are homogeneous. Because of this homogeneity, given the level of X, the joint ratings of A, B and C are assumed to be statistically independent. This is referred to as *local independence*. For a randomly selected subject, let π_{ijlk} denote the probability of ratings (i,j,l) by raters (A,B,C) and categorization in class k of X. Furthermore, let v_{ijlk} denote the expected frequencies for the A-B-C-X cross-classification. The observed data then constitute a three-way marginal table of an unobserved four-way table. The latent-class model corresponding to log-linear model (AX, BX, CX) is the nonlinear model having form

$$\log v_{ijl+} = u + u_i^A + u_j^B + u_l^C + \log \sum_k \exp(u_k^X + u_{ik}^{AX} + u_{jk}^{BX} + u_{lk}^{CX}).$$

One can use the fit of the model to estimate conditional probabilities of obtaining various ratings by the raters, given the latent class. One can also estimate probabilities of membership in various latent classes, conditional on a particular pattern of observed ratings, and use these to make predictions about the latent class to which a particular subject belongs. In that sense, latent class models focus less on agreement between the raters than on the agreement of each rater with the "true" rating. This is useful information if the latent classes truly correspond to the actual classification categories. But, of course, one never knows whether that is the case. It seems, therefore, that a combination of log-linear and latent-class modelling should be a useful strategy for studying agreement.

To fit latent-class models, one can use data augmentation techniques, such as the EM algorithm. The E (expectation) step of the algorithm approximates counts in the complete A-B-C-X table using the observed A-B-C counts and the working conditional distribution of X, given the observed ratings. The M (maximization) step treats those approximate counts as data in the standard iterative reweighted least-squares algorithm for fitting log-linear models. Alternatively, one could adopt for the entire analysis a scoring algorithm for fitting nonlinear models or a similar method for fitting loglinear models with missing data (Haberman 1988).

3.4. Agreement Measures for Ordinal Data.

Medical diagnoses often involve responses taken on an ordinal scale, many of which are fairly subjective. For example, in classifying $p21^{WAF1}$ gene expression in human breast cancer cells, pathologists use the following ordered categories: (1) no expression, (2) moderate expression, and (3) overexpression. For such scales, disagreement between raters may arise partly because of differing perceptions about the meanings of the category labels and partly because of factors such as interrater variability. As Maclure and Willett (1987) point out, with ordinal data, an intermediate category will often be subject to more misclassification than an extreme category because there are two directions in which to err away from the extremes. Therefore, a modification of kappa which accounts for severity of discordance or size of discrepancy is better suited for ordinal data. The weighted kappa statistic (Cohen 1968, Landis and Koch 1977a) offers such a modification. However, since the magnitude of the weighted kappa is greatly influenced by the relative magnitudes of the weights, some standardization in usage of the weights should be employed (Maclure and Willett 1987). Within each off-diagonal band of the $m \times m$ cross-classification table, the cell indexes differ by $1, 2, \ldots,$ or $m-1$ units, and the magnitude of disagreement in the ratings corresponds to the degree of disparity between indexes. An intuitively appealing standard usage, therefore, is to take the disagreement weight w_{ij}, $i,j = 1, \ldots, m$, to be proportional to the distance (or its square) between the two points i and j on the ordinal scale. With the above choice of weights, the weighted kappa statistic reduces to a standard intraclass correlation coefficient (Fleiss and Cohen 1973).

O'Connell and Dobson (1984) propose a general class of chance-corrected agreement measures suitable for ordinal data and K (> 2) raters with distinct marginal distributions. Specifically, these authors propose subject-specific measures of agreement, which allow identification of subjects who are especially easy or difficult to classify. The overall agreement measure based on the whole group of n subjects is an average of the subject-specific measures, and includes the weighted kappa statistic (Cohen 1968) as a special case. A modification of O'Connell and Dobson's overall agreement measure has also been suggested (Posner *et al.* 1990). This allows weighting of subjects to reflect a stratified sampling scheme.

Log-linear models for ordinal scale ratings have been proposed in the literature from the perspectives of modelling *disagreement* (Tanner and Young 1985b) as well as *agreement* (Agresti 1988) patterns. In ordinal scale ratings, magnitudes as well as directions of the disagreements of ratings are important. Therefore, the primary advantage of the log-linear framework over statistics like weighted kappa is that it provides a natural way of modelling "how" the chance-corrected frequencies differ across the off-diagonal bands of the cross-classification table. For example, is there a systematic direction bias in one of the raters? Tanner and Young's formulation (1985b) considers the independence model as the baseline for chance correction, and the authors incorporate an "additional" component for the off-diagonal cells of the $m \times m$ cross-classification table to model disagreement.

Agresti (1988) argues that ordinal scale ratings almost always exhibit a positive association between the ratings. Conditional on the ratings not being identical, there is still a tendency for high (low) ratings by one rater to be accompanied by high (low) ratings by another rater. Therefore, to model agreement between ordinal scale ratings it is inappropriate to simply take the independence model as the baseline. For two raters using the same ordered categories, Agresti (1988) proposes a model of *agreement plus linear-by-linear association.* This approach specifically combines Tanner and Young's

(1985a) model and the uniform association model (Goodman 1979) for bivariate cross-classifications of ordinal variables. This model partitions overall agreement into three parts: chance agreement (what would occur even if the classifications were independent), agreement due to a baseline association between the ratings, and an increment that reflects agreement in excess of that occurring simply from chance agreement or from the baseline association. It can be represented as

$$\log v_{ij} = u + u_i^{R_1} + u_j^{R_2} + \beta\lambda_i\lambda_j + \delta_{ij}, \tag{6}$$

where

$$\delta_{ij} = \begin{cases} \delta, & i = j, \\ 0, & \text{otherwise}, \end{cases}$$

$\lambda_1 < \cdots < \lambda_m$ are fixed scores assigned to the response categories, and the u's and the v's are as defined in Equation (4) of Section 3.3. The model (6) is a special case of the quasisymmetry model, and has simple interpretations through odds ratios.

Latent-class models that utilize the ordinality of ordered categories (Bartholomew 1983) have also been applied to studies of rater agreement. One such model of this type also treats the unobserved variable X as ordinal, and assumes a linear-by-linear association between each classification and X (Agresti and Lang 1993), using scores for the observed scale as well as for the latent classes. Another approach is to posit an underlying continuous variable (Qu *et al.* 1992, 1995). Instead of assuming a fixed set of classes for which local independence applies, one could assume local independence at each level of a continuous latent variable. Williamson and Manatunga (1997) extended Qu *et al.*'s (1995) latent-variable models to analyze ordinal-scale ratings (with m categories) arising from n subjects who are being assessed by K raters (e.g., physicians) using R different rating methods (e.g., medical diagnostic tools). Overall agreement and subject-level agreement between the raters are estimated based on the marginal and association parameters, using the generalized estimating equations approach (Liang and Zeger 1986). This method allows for subject- and/or rater-specific covariates to be included in the model.

4. CONCLUSION

The literature on interrater agreement analyses has grown extensively over the last decade. In this paper, we have presented a comprehensive survey of the various methods for the study of interrater agreement when the response variable is nominal or ordinal categorical in nature. Our focus was on various practical scenarios and designs that underlie the development of these methods, and the interrelationships between them.

The version of the kappa statistic selected should depend on the population model underlying the study in question. If each rater uses the same underlying marginal distribution of ratings, then the intraclass kappa is appropriate. When the assumption of a common marginal distribution across raters within a study is not tenable, methods using Cohen's kappa are more appropriate.

Most of the methods we discussed are designed to quantify variance attributable to the rating process. In that sense, they focus on how the ratings characterize the raters. Agreement is assessed at multiple levels: firstly, at the overall level; secondly, whether certain individual raters vary appreciably from an established gold-standard norm of rating; and, thirdly, whether there is nonhomogeneous agreement between different groups of raters (e.g., rater groups that differ in training and/or experience). A different context arises when the primary focus is on how the ratings characterize the subjects

(Kraemer 1992). For example, a patient given a diagnosis carrying serious cost and risk consequences often seeks a second (or third or fourth) diagnostic opinion, for even the most expert and careful physician using the best of medical facilities can go wrong. How many such opinions suffice to *guarantee* the diagnosis? When one obtains all the multiple opinions, what rule (of consensus) should be used to yield the best decision? In such contexts, subject-specific agreement measures can provide valuable information.

REFERENCES

Agresti, A. (1988). A model for agreement between ratings on an ordinal scale. *Biometrics*, 44, 539–548.

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statist. Methods Med. Res.*, 1, 201–218.

Agresti, A., and Lang, J.B. (1993). Quasi-symmetric latent class models, with application to rater agreement. *Biometrics*, 49, 131–139.

Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive model, and its relation to Cohen's kappa. *Biometrics*, 46, 293–302.

Barlow, W. (1996). Measurement of interrater agreement with adjustment for covariates. *Biometrics*, 52, 695–702.

Barlow, W., Lai, M.Y., and Azen, S.P. (1991). A comparison of methods for calculating a stratified kappa. *Statist. Med.*, 10, 1465–1472.

Bartholomew, D.J. (1983). Latent variable models for ordered categorical data. *J. Econometrics*, 22, 229–243.

Bloch, D.A., and Kraemer, H.C. (1989). 2 × 2 kappa coefficients: Measures of agreement or association. *Biometrics*, 45, 269–287.

Byrt, T., Bishop, J., and Carlin, J.B. (1993). Bias, prevalence and kappa. *J. Clin. Epidemiol.*, 46, 423–429.

Chinchilli, V.M., Martel, J.K., Kumanyika, S., and Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics*, 52, 341–353.

Cicchetti, D.V., and Fleiss, J.L. (1977). Comparison of the null distributions of weighted kappa and the *C* ordinal statistic. *Appl. Psych. Meas.*, 1, 195–201.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Edu. and Psych. Meas.*, 20, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psych. Bull.*, 70, 213–220.

Corey, L.A., Berg, K., Solaas, M.H., and Nance, W.E. (1992). The epidemiology of pregnancy complications and outcome in a Norwegian twin population. *Obstetrics and Gynecol.*, 80, 989–994.

Davies, M., and Fleiss, J.L. (1982). Measuring agreement for multinomial data. *Biometrics*, 38, 1047–1051.

Donner, A., and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: Confidence interval construction, significance-testing and sample size estimation. *Statist. Med.*, 11, 1511–1519.

Donner, A., and Eliasziw, M. (1997). A hierarchical approach to inferences concerning interobserver agreement for multinomial data. *Statist. Med.*, 16, 1097–1106.

Donner, A. and Klar, N. (1996). The statistical analysis of kappa statistics in multiple samples. *J. Clin. Epidemiol.*, 49, 1053–1058.

Donner, A., Eliasziw, M., and Klar, N. (1996). Testing homogeneity of kappa statistics. *Biometrics*, 52, 176–183.

Dunn, G. (1989). *Design and Analysis of Reliability Studies*. Oxford Univ. Press, New York.

Everitt, B.S. (1968). Moments of the statistics kappa and weighted kappa. *British J. Math. Statist. Psych.*, 21, 97–103.

Feinstein, A.R., and Cicchetti, D.V. (1990). High agreement but low kappa I: The problems of two paradoxes. *J. Clin. Epidemiol.*, 43, 543–548.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psych. Bull.*, 76, 378–382.

Fleiss, J.L. (1973). *Statistical Methods for Rates and Proportions*. Wiley, New York, 144–147.

Fleiss, J.L., and Cicchetti, D.V. (1978). Inference about weighted kappa in the non-null case. *Appl. Psych. Meas.*, 2, 113–117.

Fleiss, J.L., and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. and Psych. Meas.*, 33, 613–619.

Fleiss, J.L., and Davies, M. (1982). Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. *Amer. J. Epidemiol.*, 115, 841–845.

Fleiss, J.L., Cohen, J., and Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. *Psych. Bull.*, 72, 323–327.

Goodman, L.A. (1979). Simple models for the analysis of association in cross classifications having ordered categories. *J. Amer. Statist. Assoc.*, 74, 537–552.

Goodman, L.A., and Kruskal, W.H. (1954). Measures of association for cross classifications. *J. Amer. Statist. Assoc.*, 49, 732–764.

Graham, P. (1995). Modelling covariate effects in observer agreement studies: The case of nominal scale agreement. *Statist. Med.*, 14, 299–310.

Haberman, S.J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociol. Methodol.*, 18, 193–211.

Hamdan, M.A. (1970). The equivalence of tetrachoric and maximum likelihood estimates of ρ in 2×2 tables. *Biometrika*, 57, 212–215.

Hutchinson, T.P. (1993). Kappa muddles together two sources of disagreement: Tetrachoric correlation is preferable. *Res. Nursing and Health*, 16, 313–315.

Irwig, L.M., and Groeneveld, H.T. (1988). Exposure-response relationship for a dichotomized response when the continuous underlying variable is not measured. *Statist. Med.*, 7, 955–964.

Johnson, N.L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* Wiley, New York, 117–122.

Kendler, K.S., Neale, M.C., Kessler, R.C., Heath, A.C. and Eaves, L.J. (1992). Familial influences on the clinical characteristics of major depression: A twin study. *Acta Psychiatrica Scand.*, 86, 371–378.

Kraemer, H.C. (1980). Extension of the kappa coefficient. *Biometrics*, 36, 207–216.

Kraemer, H.C. (1992). How many raters? Toward the most reliable diagnostic consensus. *Statist. Med.*, 11, 317–331.

Kraemer, H.C. (1997). What is the "right" statistical measure of twin concordance (or diagnostic reliability and validity)? *Arch. Gen. Psychiatry*, 54, 1121–1124.

Kvaerner, K.J., Tambs, K., Harris, J.R., and Magnus, P. (1997). Distribution and heritability of recurrent ear infections. *Ann. Otol. Rhinol. and Laryngol.*, 106, 624–632.

Landis, R.J., and Koch, G.G. (1977a). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.

Landis, R.J., and Koch, G.G. (1977b). A one-way components of variance model for categorical data. *Biometrics*, 33, 671–679.

Liang, K.Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.

Light, R.J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psych. Bull.*, 76, 365–377.

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255–268.

Maclure, M., and Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *Amer. J. Epidemiol.*, 126, 161–169.

O'Connell, D.L., and Dobson, A.J. (1984). General observer-agreement measures on individual subjects and groups of subjects. *Biometrics*, 40, 973–983.

Oden, N.L. (1991). Estimating kappa from binocular data. *Statist. Med.*, 10, 1303–1311.

Pearson, K. (1901). Mathematical contribution to the theory of evolution VII: On the correlation of characters not quantitatively measurable. *Philos. Trans. Roy. Soc. Ser. A*, 195, 1–47.

Posner, K.L., Sampson, P.D., Caplan, R.A., Ward, R.J., and Cheney, F.W. (1990). Measuring interrater reliability among multiple raters: An example of methods for nominal data. *Statist. Med.*, 9, 1103–1115.

Qu, Y., Williams, G.W., Beck, G.J., and Medendorp, S.V. (1992). Latent variable models for clustered dichotomous data with multiple subclusters. *Biometrics*, 48, 1095–1102.

Qu, Y., Piedmonte, M.R., and Medendorp, S.V. (1995). Latent variable models for clustered ordinal data. *Biometrics*, 51, 268–275.

Schouten, H.J.A. (1993). Estimating kappa from binocular data and comparing marginal probabilities. *Statist. Med.*, 12, 2207–2217.

Scott, W.A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quart.*, 19, 321–325.

Shoukri, M.M., Martin, S.W., and Mian, I.U.H. (1995). Maximum likelihood estimation of the kappa coefficient from models of matched binary responses. *Statist. Med.*, 14, 83–99.

Snedecor, G.W., and Cochran, W.G. (1967). *Statistical Methods.* Sixth Edition. Iowa State Univ. Press, Ames.

Tallis, G.M. (1962). The maximum likelihood estimation of correlations from contingency tables. *Biometrics*, 18, 342–353.

Tanner, M.A., and Young, M.A. (1985a). Modeling agreement among raters. *J. Amer. Statist. Assoc.*, 80, 175–180.

Tanner, M.A., and Young, M.A. (1985b). Modeling ordinal scale agreement. *Psych. Bull.*, 98, 408–415.

Uebersax, J.S., and Grove, W.M. (1990). Latent class analysis of diagnostic agreement. *Statist. Med.*, 9, 559–572.

Williamson, J.M., and Manatunga, A.K. (1997). Assessing interrater agreement from dependent data. *Biometrics*, 53, 707–714.

Zwick, R. (1988). Another look at interrater agreement. *Psych. Bull.*, 103, 374–378.

Received 6 May 1997
Revised 10 October 1997
Accepted 3 December 1997

Center for Healthcare Effectiveness Research
Wayne State University School of Medicine
Detroit, Michigan 48201
U.S.A.
e-mail: banerjee@kci.wayne.edu

Department of Mathematics
University of New Hampshire
Durham, New Hampshire 03824
U.S.A.