Modeling Patterns of Agreement and Disagreement

Alan Agresti
Department of Statistics
University of Florida
Gainesville, Florida 32611

Abstract

This article presents a survey of ways of statistically modeling patterns of observer agreement and disagreement. Main emphasis is placed on modeling inter-observer agreement for categorical responses, both for nominal and ordinal response scales. Models discussed include (1) simple cell-probability models based on Cohen's kappa that focus on beyond-chance agreement, (2) loglinear models for square tables, such as quasi-independence and quasi-symmetry models, (3) latent class models that express the joint distribution between ratings as a mixture of clusters for homogeneous subjects, each cluster having the same "true" rating, and (4) Rasch models, which decompose subject-by-observer rating distributions using observer and subject main effects. Models can address two distinct components of agreement – strength of association between ratings, and similarity of marginal distributions of the ratings.

Key Words: association models; categorical data; inter-observer agreement; kappa; latent class models; loglinear models; marginal homogeneity; odds ratios; ordinal data; quasi independence; quasi symmetry; Rasch models; rater agreement; reliability; square contingency tables.

ACKNOWLEDGMENT

This work was partially funded by a research grant from the National Institutes of Health.

1 Introduction

Suppose several observers measure a response variable for the same set of subjects. Regardless of the nature of the response variable, different observers need not make the same response for a given subject, because of several types of "measurement error" that can occur. This is particularly true when the response variable has a subjective rating scale, as is often the case for categorical responses. Discrepancies between ratings can be attributable not only to classification errors by the observers, but also to the categories not having an objectively precise definition. Different observers may simply have different perceptions about what the categories mean. Even if there is a common perception, measurement variability can occur. For instance, repeated observations by the same observer of a given subject may exhibit variability. Issues that arise in analyzing inter-observer agreement apply also to the analysis of intra-observer agreement.

The study of observer agreement is very important in many medical applications. Measurement scales dealing with the presence of a symptom or the severity of a disease are often quite subjective, especially when the scale is ordinal. For instance, presence of a symptom might be measured using a categorical scale having labels such as (no, unlikely, somewhat likely, probable, very probable, yes); severity of disease might be measured as (none, slight, moderate, severe). Landis and Koch[1], in a highly informative survey of methods (*circa* 1975) for analysis of observer reliability studies, discussed several medical applications in which the observer may be an important source of measurement error. One of the earliest types of application in which observer variation was studied concerned interpretation of film results in chest radiography[2]. Similar issues arose in applications involving estimating the precision of measuring instruments in laboratories[3], and estimating observer reliability in psychiatric evaluation and educational testing[4,5,6]. The degree of observer agreement is also an important consideration in many non-medical contexts, such as in the study of interviewer reliability in sample surveys[7].

To illustrate methods for modeling agreement, later in the article we shall analyze Table 1, based on data presented by Landis and Koch[8]. Seven pathologists separately classified 118 slides in terms of carcinoma of the uterine cervix, using a five-point ordinal scale with categories (1) negative, (2) atypical squamous hyperplasia, (3) carcinoma in situ, (4) squamous carcinoma with early stromal invasion, and

(5) invasive carcinoma. Few observations occurred in the fifth category, and Table 1 combines that one with the fourth category. Table 1 contains data for the first two observers, A and B.

Most of the early statistical research dealing with analyzing observer agreement focused on the development of summary measures of agreement. The object was to develop a statistic that indicated whether the degree of agreement between two observers was (almost perfect, substantial, moderate, fair, slight, poor)[1,9] . The most popular measure of this type, Cohen's kappa, gives a number on a scale for which 0 indicates agreement no better than would be expected if the ratings were statistically independent, and 1 indicates perfect agreement. For Table 1, for instance, this measure equals 0.49, indicating "moderate" agreement. Discussion of such summary measures is not the primary focus of this article. Instead, we survey ways of *modeling* patterns of observer agreement and disagreement. With the modeling approach, one can more fully describe agreement. For instance, one can often (1) provide a parsimonious representation for the joint distribution of observers' ratings, (2) provide residuals that compare the frequency with which certain types of agreements or disagreements occurred compared to what would be expected with some predicted pattern, and (3) estimate conditional probabilities such as those involving ratings by one observer given ratings by other observers, those involving ratings by an observer given a "true" rating, and those involving the "true" rating given ratings by several observers.

2 Modeling Agreement on a Categorical Scale

We discuss the case in which a fixed set of d observers rate the same n subjects, giving special emphasis to $d = 2$. The measurement could be (1) binary, such as an evaluation about whether a subject has a particular disease, (2) nominal scale, such as classification into mental illness types, (3) ordinal scale, such as classifications regarding the stage or severity of a disease, or (4) interval scale, such as a serum cholesterol level. For continuous interval-scale responses, a common approach uses analysis of variance (ANOVA) models, with intraclass correlations to describe strength of agreement[1,10,11]. Dunn's article in this issue discusses this case. Our primary emphasis concerns models for categorical responses – cases (1)-(3) and case (4) with a highly discrete or grouped continuous response.

Let I denote the number of categories for a categorical response, and denote two observers by A and B. In the population of subjects of interest, let $\pi_{ij} = P(A = i, B = j)$ denote the proportion that observer A classifies in category i and observer B classifies in category j, $i = 1, ..., I, j = 1, ..., I$. Their ratings of a particular subject *agree* if they each classify the subject in the same category. In the square table $\{\pi_{ij}\}$, π_{ii} is the probability they agree that a randomly selected subject is in category i, and $\sum_i \pi_{ii}$ is the total probability of agreement. Perfect agreement occurs when $\sum_i \pi_{ii} = 1$.

Inter-observer agreement for categorical scales has two components – *distinguishability* of categories and lack of *bias.* These relate to strength of association between ratings and to similarity of their marginal distributions. A relatively large total probability of agreement requires both strong association and near marginal homogeneity, as explained next.

Darroch and McCloud[12] noted that the extent of observer agreement depends partly on how well each observer can distinguish between each pair of categories. For a pair of subjects, consider the event that each observer classifies one subject in category i and one subject in category j. The ratings are concordant if they agree on which subject is in category i and which is in category j; they are discordant if the subject that A places in category i is placed in j by B, and the one A places in category j is placed in i by B. Conditional on the event of rating the two subjects in categories i and j, the odds that the ratings are concordant rather than discordant are

$$\tau_{ij} = \pi_{ii}\pi_{jj}/\pi_{ij}\pi_{ji}, \qquad (1)$$

Distinguishability of categories increases as the association between the A and B classifications, as described by the $\{\log \tau_{ij}\}$, becomes more strongly positive. As τ_{ij} increases, the observers are more likely to agree on which subject receives each designation.

The marginal distributions $\{\pi_{i+} = \sum_j \pi_{ij}, i = 1, ..., I\}$ and $\{\pi_{+j} = \sum_i \pi_{ij}, j = 1, ..., I\}$ describe how the observers separately allocate subjects to the response categories. Bias of one observer relative to another refers to discrepancies between these marginal distributions. Bias decreases as the marginal distributions become more nearly equivalent, lack of bias meaning that $\pi_{i+} = \pi_{+i}$ for all i. We use the term "bias" here simply to refer to marginal heterogeneity for A and B, and other types of bias may exist; for instance, A and B may have identical marginal distributions, yet their common distribution may differ from that of the unknown "true" rating or that of a "standard."

Strong agreement requires both similar marginal distributions and strong positive association. For instance, one could have identical marginal distributions, yet statistical independence in the joint distribution $\{\pi_{ij}\}$ for the ratings; in that case, one observer's rating is unrelated to the other's rating. Or, one could have strong association between ratings, yet one observer's rating could be systematically different from the other's (*e.g.*, for an ordinal scale, A's rating might be consistently one category higher than B's). A difficulty in modeling agreement is that many standard models for categorical data refer to only one of the two components. For instance, loglinear models focus on the joint distribution, and hence the strength of association between ratings. Our discussion of each model will assess the extent to which it addresses both components.

Though we focus on categorical responses, we first briefly discuss an ANOVA model that helps clarify these two components of agreement. Let Y_{ij} denote the rating of subject i by observer j, $i = 1, ..., n, j = 1, ..., d$. Consider the model

$$Y_{ij} = \mu + \xi_i + \tau_j + \epsilon_{ij} \qquad (2)$$

where $\{\tau_j\}$ are observer effects and $\{\xi_i\}$ are subject effects. Suppose $\{\xi_i\}$ are treated as random effects, and $\{\xi_i\}$ and $\{\epsilon_{ij}\}$ are mutually independent random variables. For given variability among the subjects, the agreement between Y_{ij} and Y_{ik} increases as the error variability σ^2_ϵ decreases; that is, as the correlation between Y_{ij} and Y_{ik} increases. However, strong agreement also requires relatively little variability in the observer effects $\{\tau_j\}$ (*i.e.*, little inter-observer bias).

Though this article emphasizes models for agreement rather than summary descriptive measures, Section 3 combines the two approaches. It describes simple models for cell probabilities that contain the *kappa* measure of agreement as a model parameter. Section 4 describes loglinear models for joint distributions of ratings. Loglinear models pertaining to square tables, with the same categories in each dimension, have particular relevance. Examples include quasi-independence models that highlight agreement occurring beyond chance, and more general quasi-symmetry models.

Section 5 discusses Rasch models, which utilize separate rating response distributions for each subject-observer combination. They describe inter-observer bias by expressing a logit for those distributions in terms of additive observer effects and subject effects. Rasch models are related to latent class models, which are described in Section 6. Such models regard the joint distribution between ratings as a mixture of distributions, one for each latent class. Each latent class consists of homogeneous subjects, having the same "true" rating. Within each latent class, there is statistical independence between observers' ratings. This type of model focuses less on agreement between observers than on agreement between each observer and the "true" rating. Sections 3-6 discuss models both for nominal and ordinal measurement scales. Section 7 illustrates the use of these methods for the analysis of Table 1.

3 Kappa-Based Models

Bloch and Kraemer[13] noted that kappa, a measure of agreement for categorical ratings, has several versions. For nominal scales, the most popular one seems to be Cohen's[14] kappa. Its population value is defined as

$$\kappa = \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{1 - \sum \pi_{i+}\pi_{+i}} \tag{3}$$

The numerator compares the total probability of agreement to that expected if the ratings were statistically independent, referred to as "chance" agreement; the denominator is the maximum possible value of the numerator, for the given marginal probabilities.

A related kappa measure, introduced by Scott[15], replaces $\pi_{i+}\pi_{+i}$ by π_i^2, where $\pi_i = (\pi_{i+} + \pi_{+i})/2$. These versions of kappa are designed for nominal classifications. Much controversy has surrounded their use, particularly regarding their dependence on the marginal distributions[12,16,17,18]. A difficulty is that the same diagnostic process can yield different values of kappa, depending on the proportions of cases of the various types. The many generalizations of kappa include weighted versions for ordinal scales and versions for multi-observer agreement. We do not address these measures in this article, but simply note kappa's use in a simple contingency-table model for agreement and disagreement.

Suppose the observers are interchangeable, in the sense that the marginal distributions are identical, with $\pi_i = \pi_{i+} = \pi_{+i},\ i = 1, ..., I$. Then, Cohen's and Scott's measures are identical. Several authors[13,19,20] have discussed the use of a separate kappa for each category, equivalent to

$$\kappa_i = (\pi_{ii} - \pi_i^2)/[\pi_i(1 - \pi_i)] \tag{4}$$

When $I = 2, \kappa_1 = \kappa_2$. When $I > 2$, the overall κ is a weighted average of $\{\kappa_i\}$, with weight $\pi_i(1 - \pi_i)$ for κ_i. For these parameters,

$$\pi_{ii} = \pi_i^2 + \kappa_i \pi_i (1 - \pi_i),$$

and κ_i describes the degree to which agreement for category i exceeds that expected under independence of ratings.

When all $\{\kappa_i\}$ are identical, a simple model expresses the cell probabilities in terms of the common value κ, namely

$$\pi_{ii} = \pi_i^2 + \kappa\pi_i(1-\pi_i)$$

$$\pi_{ij} = \pi_i\pi_j(1-\kappa),\ i \neq j.$$

For this model, $\{\pi_{ij}\}$ satisfy *symmetry* and *quasi-independence*. That is, $\pi_{ij} = \pi_{ji}$ for all i and j, and conditional on the ratings differing, the rating by one observer is statistically independent of the rating by the other observer. When this model holds, $\kappa = 0$ is equivalent to statistical independence of the ratings and $\kappa = 1$ is equivalent to perfect agreement. The joint distribution is a weighted average of a distribution satisfying statistical independence and a distribution having perfect agreement, with weights $(1-\kappa)$ and κ. Cohen[21] used this as a model for positively associated clustering. James[22] and Agresti[23] also studied it. The model is overly simplistic for most applications, and does not have broad scope because of its requirement of identical marginal distributions. However, it does represent a case in which kappa describes both the pattern and the strength of agreement. For further discussion of kappa-based methods and other summary measures of agreement, see Fleiss[24], Fleiss and Cohen[25], Kraemer[26,27], Landis and Koch[1], Schouten[28], and the article in this issue by Kraemer.

4 Loglinear and Association Models

Kappa summarizes agreement by comparing the probability of agreement to the baseline of chance probability of agreement. It is not possible to incorporate kappa as a parameter in non-trivial loglinear models for $\{\pi_{ij}\}$. However, loglinear models can express agreement in terms of components, such as chance agreement and beyond-chance agreement. Also, they can display patterns of agreement among several observers, or compare patterns of agreement when subjects are stratified by values of a covariate.

Let $\{m_{ij} = n\pi_{ij}\}$ denote expected frequencies for ratings of n subjects by observers A and B. Chance agreement, or statistical independence of the ratings, has loglinear model representation

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B$$

For the usual Poisson or multinomial sampling models for observed cell counts $\{n_{ij}\}$, the maximum likelihood (*ML*) fitted values for this model are $\{\hat{m}_{ij} = n_{i+}n_{+j}/n\}$. Normally we would not expect this model to fit well, but its cell residuals provide information about patterns of agreement and disagreement.

Let $\{p_{ij} = n_{ij}/n\}$. The *adjusted residual*[29],

$$r_{ij} = \frac{n_{ij} - \hat{m}_{ij}}{[\hat{m}_{ij}(1 - p_{i+})(1 - p_{+j})]^{1/2}}$$

is useful because it has an asymptotic standard normal null distribution. Cells having large positive residuals give strong evidence of agreement that is greater than that expected by chance.

A useful generalization of the independence model is the quasi-independence model,

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \delta_i I(i = j) \tag{5}$$

where the indicator $I(i = j)$ equals 1 when $i = j$ and equals 0 when $i \neq j$. Conditional on disagreement by the observers, the rating by A is statistically independent of the rating by B. When $\delta_i > 0$, more agreements regarding outcome i occur than would be expected by chance. The *ML* fit is perfect on the main diagonal; that is, $\hat{m}_{ii} = n_{ii}$ for all i. This model, like other loglinear models we discuss, is easily fitted using many computer packages that have programs for loglinear models, such as *GLIM*. For details, see Bishop *et al.*[30] and Agresti[31].

Loglinear models are good vehicles for studying association between classifications. For them, odds ratios are useful measures of association, since they relate to model parameters. For studying agreement, the odds ratios $\{\tau_{ij}\}$ defined in (1) are relevant. For model (5),

$$\log \tau_{ij} = \log m_{ii} + \log m_{jj} - \log m_{ij} - \log m_{ji} = \delta_i + \delta_j$$

for all $i \neq j$. The odds that the rating by A is i rather than j is $exp(\delta_i + \delta_j)$ times as high when the rating by B is i than when it is j. When $\delta_1 = ... = \delta_I$, all these odds ratios are identical, and a simpler model holds having a single term $\delta I(i = j)$ added to the independence model. This makes the fit on the main diagonal unsaturated, satisfying $\sum \hat{m}_{ii} = \sum n_{ii}$, rather than $\{\hat{m}_{ii} = n_{ii}\}$. Several authors have proposed the use of quasi-independence loglinear models for studying agreement[30,32,33,34].

Ordinal rating scales almost always exhibit a positive association between ratings. Conditional on observer disagreement, there usually remains a tendency for high (low) ratings by A to occur with relatively high (low) ratings by B. Hence, the quasi-independence model is normally inadequate for ordinal scales. One often obtains better fitting models by partitioning beyond-chance agreement into two parts: Agreement due to a baseline association between the ratings, and perhaps other increments that reflect agreement in excess of that occurring by chance or from the baseline association.

One model of this type uses a *linear-by-linear* baseline association with ordered scores $\{u_i\}$ for the categories[35,36], namely

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \beta u_i u_j \tag{6}$$

This model tends to fit well when there is an underlying continuous rating scale such that the joint ratings have a bivariate normal distribution[37,38]. For this model

$$\log \tau_{ij} = \beta(u_j - u_i)^2$$

so category distinguishability increases as β increases and as the distance between the categories increases.

Model (6) is fitted using iterative methods, such as the Newton-Raphson method. One can use standard software such as GLIM or SAS to do this[30,38]. The model can be generalized, for instance by adding parameters representing extra agreement on the main diagonal[31,34,39], using association models having parameter scores instead of fixed scores[36,39,40], or by simultaneously describing agreement among several observers[41].

Darroch and McCloud[12] argued that models for agreement should satisfy the property of *quasi symmetry.* We next outline their approach, which seems quite reasonable. They utilized subject-by-observer-specific probability distributions. Let ϕ_{sri} denote the probability of response in category i, when observer r evaluates subject s. This stochastic approach recognizes that the same observer may classify a subject differently on separate occasions. There are two basic assumptions. First, for a given subject s, a "local independence" assumption states that classifications by separate observers are independent; for instance, the probability that observer A makes rating i and observer B makes rating j equals $\phi_{s1i}\phi_{s2j}$. This assumption seems reasonable when ratings are done "blindly" by different observers. The second assumption states that $\{\phi_{sri}\}$ satisfy the condition of no three-factor interaction; that is, ϕ_{sri} has form

$$\phi_{sri} = \alpha_{sr}\beta_{si}\gamma_{ri}.$$

This means that the signal emitted by the subjects being rated and the observer differences combine without interaction in affecting the response. Darroch and McCloud gave arguments supporting these two basic assumptions.

Suppose the subjects are randomly sampled from a population of subjects. Under local independence, the probability π_{ij} is the average of $\phi_{s1i}\phi_{s2j}$ computed for all subjects in the population. The variability among subjects in their distributions $\{\phi_{sri}\}$ results in a joint distribution $\{\pi_{ij}\}$ displaying association. Under the additional assumption of no three-factor interaction, Darroch and McCloud noted that this distribution satisfies quasi symmetry; that is, it has form

$$\pi_{ij} = a_i b_j d_{ij}$$

where $d_{ij} = d_{ji}$ for all i and j. All models discussed so far in this article have this property. In loglinear form, quasi symmetry has expression

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}, \tag{7}$$

where $\lambda_{ij} = \lambda_{ji}$ for all i and j. Though loglinear models directly address the association component of agreement and not the lack of bias component, in the quasi-symmetric case they do yield some information about bias. Under (7), the main effect parameters describe differences between the margins, marginal homogeneity being equivalent[30] to $\lambda_i^A = \lambda_i^B$ for all i.

Loglinear models treat the observers in a symmetric manner. In some applications, one classification might be a "known standard" rating, in which case asymmetric interpretations may be of greater interest. For instance, one can re-express models in terms of logits of probabilities for an observer's rating, conditional on the standard rating. Also in some applications it is important to stratify the sample into groups, according to values of relevant covariates. Then, one can check whether the agreement pattern is homogeneous across levels of the covariate. One could check whether a loglinear model fits well having identical association structure between ratings in each stratum. Stratified loglinear models are also relevant when we model the agreement between each of several observers and a standard rating, and each observer evaluates a separate sample of subjects.

When $d > 2$ observers rate the same sample, one approach constructs a loglinear model that applies to the joint d-dimensional cross-classification of the ratings[34]. It addresses higher-order agreement as well as pairwise agreement. The parameters in such models describe conditional agreement – for instance, the agreement between two observers for fixed ratings by the other observers. This may be somewhat unnatural, since the assessment of agreement between two observers depends on how many and which other observers are included. An alternative approach fits models simultaneously to two-way marginal tables of the multi-dimensional table. One can then consider whether patterns of agreement are homogeneous for different pairs of observers. Such models are not easy to fit, since the likelihood refers to the interior cells in the table rather than the two-way marginal tables, and multi-observer tables are often very large and sparse. One approach obtains consistent estimates by treating the separate two-way tables as independent layers of a three-way table, and uses a jackknife to obtain estimated standard errors of parameter estimates that reflect the actual dependence[41].

Other applications of the loglinear approach include using other types of quasi-symmetry models[37,42,43], other types of diagonals-parameter models[44,45], other types of ordinal models[46,47,48] that model ordinal-scale disagreement, and models based directly on assumed underlying normal distributions[49,50].

5 Rasch Models

We next discuss a model that is specified directly in terms of the separate subject-by-observer rating distributions. It yields comparisons of marginal distributions of the observers' ratings for those distributions. We first consider the case of $I = 2$ categories, for which the model is the well-known *Rasch model*[51], having form

$$\log(\phi_{sr1}/\phi_{sr2}) = \alpha + \gamma_s + \beta_r. \tag{8}$$

That is, it assumes no three-factor interaction for $\{\phi_{sri}\}$. This model naturally addresses bias among observers. For instance, for each subject, the odds of a rating by observer A in category 1 are $exp(\beta_A - \beta_B)$ times the odds for observer B.

In fitting the Rasch model, for a given subject s one assumes "local independence." As $n \to \infty$, ordinary ML estimators of the observer effects $\{\beta_r\}$ are inconsistent[52], because of the concomitant increase in the number of subject parameters. Consistent estimates result from conditioning on sufficient statistics for $\{\gamma_s\}$. We noted previously that the assumptions of no three-factor interaction and local independence generate the quasi-symmetry model for the averaging of joint distributions over subjects (*e.g.*, for $\{\pi_{ij}\}$ when $d = 2$). Not surprisingly, the Rasch model relates to the quasi-symmetry model for the 2^d table that cross classifies responses for the d observers. It follows[53] that ML estimates of $\{\lambda_1^A - \lambda_2^A,\ \lambda_1^B - \lambda_2^B, ...\}$ in the quasi-symmetry model are also conditional ML estimates of $\{\beta_r\}$ in model (8).

When $I > 2$, a multinomial logit generalization of the Rasch model relates to the quasi-symmetry model for the I^d cross classification of the subjects' responses. We mention here a special case of a generalization presented by Andersen[52], applicable for ordinal responses. It has the adjacent-categories logit representation

$$\log(\phi_{sri}/\phi_{s,r,i+1}) = \alpha_i + \gamma_s + \beta_r, \tag{9}$$

holding simultaneously for $i = 1, ..., I-1$. For each subject, the odds of observer A making response i instead of response $i+1$ are $exp(\beta_A - \beta_B)$ times the odds for observer B, and this holds uniformly in i.

One can obtain conditional ML estimates of $\{\beta_r\}$ in (9) by ordinary ML fitting of a corresponding quasi-symmetric loglinear model[54]. To illustrate, for two observers one derives the loglinear model by expressing π_{ij} as the average of $\phi_{s1i}\phi_{s2j}$ with respect to some unknown distribution for the subject effects. The resulting loglinear model has form

$$\log m_{ij} = \mu + \lambda_i + \lambda_j + \beta_A x_i + \beta_B x_j + \rho_t \tag{10}$$

where $\{x_h = h\}$ and $t = i + j$. The observer parameters β_A and β_B are the same in 9 and 10, and the $\{\rho_t\}$ in 10 are nuisance parameters that relate to $\{\gamma_s\}$ in 9. The sufficient statistics for β_A and β_B are the first-order marginal means $\sum ip_{i+}$ and $\sum jp_{+j}$. For model (10), conditional on the "total score" $t = i + j$, the responses are independent. Thus, local independence applies to groups of subjects who are homogeneous in terms of having the same total score. This is the idea behind latent class modeling, which is the subject of the next section. One can regard the Rasch model as a latent class model with as many classes as there are different total scores.

6 Latent Class Models

Latent class models express the joint distribution of ratings as a mixture of distributions for classes of an unobserved (latent) variable. Each latent class consists of homogeneous subjects, such that local independence holds among the ratings. Since Goodman's[55] development of ML procedures for estimating parameters in latent class models, such models have been used for a variety of applications; see, for instance, Haberman[56] and, for a more elementary introduction, McCutcheon[57]. This section summarizes a basic latent class model[58] , applied to describe inter-rater agreement. This approach treats both the observed scale and the latent variable as discrete.

We illustrate latent class models for three observers, denoted by A, B, and C. The data to be analyzed are sample cell counts $\{n_{hij}\}$ specifying the numbers of occurrences for the I^3 possible combinations of ratings by the three observers. The latent class model assumes there is an unobserved categorical scale

X, with L categories, such that subjects in each category of X are homogeneous. Because of this homogeneity, the joint ratings of A, B, and C are assumed to be statistically independent, given the level of X. Subjects occurring in the same latent category might be ones having the same "true" rating. For instance, when $I = 2$ (say, *positive* and *negative* possible ratings), the 2^d joint distribution for the d ratings may be a mixture of two distributions − statistical independence among observers for subjects whose true rating is positive, and statistical independence among observers for subjects whose true rating is negative.

For a randomly selected subject, let π_{hijk} denote the probability of ratings (h, i, j) by observers (A, B, C), and categorization in class k of X. For a sample of size n, let $\{m_{hijk} = n\pi_{hijk}\}$ denote expected frequencies for the A-B-C-X cross-classification. The observed data $\{n_{hij}\}$ are a three-way marginal table of an unobserved four-way table. The distribution $\{\pi_{hijk}\}$ satisfies the loglinear model having as sufficient statistics the marginal configurations represented by the notation (AX, BX, CX). The latent class model corresponding to loglinear model (AX, BX, CX) is the nonlinear model having form

$$\log m_{hij+} = \mu + \lambda_h^A + \lambda_i^B + \lambda_j^C + \log[\sum_k exp(\lambda_k^X + \lambda_{hk}^{AX} + \lambda_{ik}^{BX} + \lambda_{jk}^{CX})]$$

Strong associations between each observer and X can induce strong marginal associations between pairs of observers. One can use the fit of the model to estimate conditional probabilities of obtaining various ratings by the observers, given the latent class. When $L = I$ and one interprets the latent classes as the same as the observed scale, estimates of probabilities $\{P(A = i \mid X = i), P(B = i \mid X = i), P(C = i \mid X = i)\}$ are of interest. 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. A practical problem with latent class models is the large number of parameters. To achieve identifiability it is sometimes necessary to reduce the parameter space by setting certain parameters equal to zero or introducing equality constraints[55,59].

This basic latent class model treats the rating scale as nominal. When the rating scale is ordinal and when (for some fixed L) model (AX, BX, CX) fits well, it can be beneficial to fit special cases of that model that utilize the ordinality. Such models are more parsimonious and have simpler interpretations for associations between the observed and latent variables. One model that treats X and the observed scale as ordinal assumes a linear-by-linear association between each observer and X. The model uses scores $\{u_h\}$ for the observed scale and scores $\{x_k\}$ for the latent classes, and assumes that

$$\log m_{hij+} = \mu + \lambda_h^A + \lambda_i^B + \lambda_j^C + \log[\sum_k exp(\lambda_k^X + \beta^{AX} u_h x_k + \beta^{BX} u_i x_k + \beta^{CX} u_j x_k)] \qquad (11)$$

More generally, one could replace one or both sets of scores by parameters, yielding latent class versions of a log-multiplicative model introduced by Goodman[36]. For examples of various latent class approaches for ordinal variables, see Agresti and Lang[60], Clogg[59], and Hagenaars[61].

When the rating scale is truly nominal with distinct categories (*e.g.*, positive, negative), it is not unreasonable to expect a latent class model to hold, with the same number of latent classes; that is, one expects that if subjects are relatively homogeneous within each true state, the observers' judgments will be approximately independent conditional on that state. However, when there are gradations of the symptom studied, such as when the measurement scale is ordinal with a subjective scale, this is not so plausible[17]. Instead, it is often natural to posit an underlying continuous variable. 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. Models of this type are called *latent trait* models. Uebersax[62] described the use of such models for agreement analyses. Each observer is assumed to have a set of threshholds that divide the continuum into I intervals. A logistic distribution determines the probability a subject is classified into each category, at each latent trait level. Each such level corresponds to a distinct point on the continuum, but the perceived level differs from it because of various sources of measurement error. Fitting the models produces estimates of the proportions in various latent classes, a dispersion parameter that reflects measurement error, levels on the continuum for the latent classes, and the observer threshholds. For examples of related latent trait models, see Andrich[63], Bartholomew[64,65], and Rost[66,67].

Another possible extension assumes quasi symmetry for the distribution among the observed classifications implied by a latent class model. This results in a simplification by which associations are identical between each observer and the latent variable[60]. When $I = 2$, such latent class models are special cases of the Rasch model[68].

To fit latent class models, one can assume that $\{n_{hij}\}$ have independent Poisson(m_{hij}) distributions, and apply the EM algorithm[55]. 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 loglinear models. Alternatively, one could adapt for the entire analysis a scoring algorithm for fitting nonlinear models[69] or a similar method for fitting loglinear models with missing data[70]. Software is readily available for fitting latent class models – for instance, MLSSA[57,71], LAT[56], and NEWTON[70].

Authors who have used latent class models or adaptations of them to model inter-rater agreement include Agresti and Lang[60], Aickin[72], Clogg[73], Dillon and Mulani[74], Espeland and Handelman[75], Uebersax[62,76], and Uebersax and Grove[77]. A danger with the latent class approach, as in related methods such as factor analysis models, is the temptation to interpret latent classes too literally. For instance, it is tempting to treat a rating of i given that the subject falls in latent class i as necessarily being a "correct" classification. One should realize the tentative nature of the latent classes and be careful not to make the error of reification[78].

An advantage of the latent class model approach using loglinear models, compared to summarizing association by the kappa measure, is that parameters and their estimates are identical whether the sampling scheme is Poisson, multinomial, or independent multinomial. For instance, in some studies samples of subjects may be chosen to give a good spread across categories. Sampling relatively more or fewer observations from some latent classes has no impact on the value of the association parameters relating the latent factor to each observer. If a model holds and the sampling proportions are changed, the model still holds with the same values for the association parameters. This is not the case for kappa. Its value is highly dependent on the allocation across categories of the latent variable or across categories of any of the observed classifications.

7 Example

To illustrate methods for modeling agreement, we now analyze Table 1. The sample version of Cohen's kappa for Table 1 is 0.493, with estimated standard error 0.057. The difference between observed and chance agreement is about 50% of the maximum possible difference. If we assume identical margins, using $(p_{i+} + p_{+i})/2$ to estimate π_i, we obtain estimates of component-wise kappa values (4) of $\{0.781$, 0.247, 0.402, $0.435\}$. To investigate the pattern of agreement more fully, we fitted several models to Table 1. Table 2 reports likelihood-ratio statistics (G^2) for testing their fit. For some models, at least one sufficient statistic takes its maximum or minimum possible value for given values of other sufficient statistics, in which case ML estimates do not exist unless a small constant is added to certain empty cells. The lack of existence is purely a consequence of dealing with the log scale for model parameters, and it is standard to add a very small constant to the empty cells or to all cells so that the estimates exist but their values are not overly smoothed (Agresti[31], Sec. 7.7). The results we quote are based on adding 10^{-8} to each cell.

First we consider the loglinear modeling approach. The model of statistical independence of the ratings is simply the ordinary latent class model with $L = 1$. It fits poorly ($G^2 = 118.0, df = 9$), as one would expect. Adjusted residuals for the model, reported in Table 3, show that agreement for each category is greater than expected by chance, especially for the first category. Similarly, disagreements tend to occur less than would be expected by chance, though the evidence of this tends to be weaker for categories closer together. The most marked disagreements are with observer B choosing category 3 and observer A instead choosing category 2 or 4. The quasi-independence model fits much better ($G^2 = 13.2, df = 5$) than independence, but still gives an inadequate fit. Thus, simpler models that make this assumption, such as the kappa-based model of Section 3, will be inadequate. The quasi-symmetry model fits well ($G^2 = 1.0, df = 3$), though, so it is worth investigating more parsimonious models that have this property.

The quasi-independence model has an adjusted residual of 3.31 in cell (4,3) and 2.62 in cell (2,1) and some relatively large negative residuals in cells in which one rating is high and the other is low. Conditional on the ratings by A and B differing, high (low) ratings by A tend to correspond to high (low) ratings by B. The residual ordinal association that remains after fitting the quasi-independence model is explained well by a linear-by-linear term ($G^2 = 1.1, df = 4$). A simpler model having a single diagonal parameter (δ) and a linear-by-linear term also fits well ($G^2 = 4.8, df = 7$). This model has form

$$\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \beta u_i u_j + \delta I(i = j) \qquad (12)$$

with $\{u_i = i\}$. Table 4 displays its ML fit. Agresti[39] gave examples of the use of SAS and GLIM to fit this model.

Model (12) has estimates $\hat{\beta} = 1.32$ ($ase = 0.42$) and $\hat{\delta} = 0.84$ ($ase = 0.43$). For this model, $\log \tau_{ij} = (j-i)^2 + 2\delta$. An interpretation is that for $i = 1, 2, 3, 4$, the odds that the estimated diagnosis of pathologist A is $i+1$ rather than i is $exp(\hat{\beta}+2\hat{\delta}) = 20.1$ times higher when the diagnosis of pathologist B is $i+1$ rather than when it is i. In this sense, category distinguishability for these observers seems strong. The special case of model (12) satisfying $\lambda_i^A = \lambda_i^B$ for all i also satisfies marginal homogeneity, but fits poorly ($G^2 = 76.6, df = 10$). Whatever lack of agreement exists in Table 1 seems due to bias moreso than to category indistinguishability. If the raters could calibrate themselves to achieve marginal homogeneity, then this model would simplify to complete symmetry, with a relatively high proportion of observations on the main diagonal.

We next describe agreement using latent class models. The ordinary latent class model (treating classes as nominal) fits poorly when $L = 2$, and that model is saturated when $L \geq 3$. However, latent class models with $L{\times}L$ terms still apply when $L \geq 3$ and (using equal-interval scores) fit well, particularly when $L = 4$ ($G^2 = 3.6, df = 4$). Models with $L = 4$ are natural to apply, as one might hope that the latent classes approximate well the categories of the observed classification. The estimates of the linear-by-linear associations between the observers and the latent rating are similar for the two observers, and a simpler model satisfying quasi symmetry that sets them equal also fits well ($G^2 = 4.6, df = 5$). Table 4 also displays the fit of this "uniform" $L{\times}L$ latent class model. The estimated common association parameter (using unit-spaced scores) is $\hat{\beta} = 4.54$. The estimated odds that an observer selects category $j+1$ instead of j are $exp(4.54) = 93$ times higher for subjects in latent category $k+1$ than for subjects in category k. There is a strong association between each rating and the latent rating.

For this latent class model, the fitted proportions in the four latent classes are {0.175, 0.161, 0.520, 0.144}. For each observer, Table 5 shows the estimated observer response probabilities in each latent class. For a given latent class, each observer is most likely to make response in the same category. The only near exception is latent class 4 for observer B, in which response 3 is almost as likely as response 4, which helps explain the positive residual in cell (4,3) for the independence model. To the extent that the latent classes are indicative of "true" ratings, classification errors are most likely in categories 2 and 3 for A and 2 and 4 for B. Similarly, one can use the fit to calculate estimated values of $\{P(X = k \mid A = i, B = j)\}$. For instance, when $A = 4$ and $B = 3$, the estimated values are {0.000, 0.000, 0.598, 0.402}.

For examples of analyses of multi-rater agreement using an expanded version of Table 1 having five additional pathologists, see Cox *et al.*[79], Landis and Koch[8], Becker and Agresti[41] and Agresti and Lang[60].

8 Commentary

This article has surveyed several ways of modeling agreement and disagreement for categorical response scales. Though we have presented the models as ways of describing inter-observer agreement, clearly they are just as valid for modeling intra-observer agreement. For instance, one might be interested in modeling the extent to which ratings agree when an expert observer uses two different measuring instruments to analyze the same sample of subjects, or when an observer rates the same subjects with the same instrument at two separate occasions.

It seems dangerous to make judgments about the "best" way to model agreement and disagreement. Work in this area is at an early stage of development, and much more will be done by the end of the century. However, we can make a few basic comparisons of the types of methods we have discussed. First, we believe that the trend toward developing models is a good one. Model-based approaches yield additional and more precise information than that provided by summary measures of agreement. Of the models, the main choices seem to be between variance-components models that lead to intraclass correlations which (in the categorical case) are kappa-type measures, loglinear and related association models, and latent class and more general latent structure models.

The loglinear and latent class models have the advantage of yielding fitted cell probabilities. Thus one can test goodness of fit and make predictions about the behavior of observers under certain situations. Loglinear models are simple but, so far at least, a single model fails to provide information about both association and observer bias. Latent class models give somewhat different information. They focus less on agreement between the observers than the agreement of each observer with the "true" rating. For instance, one can predict the probability an observer makes a particular response for subjects in a particular latent class, or predict the probability that a subject belongs to some latent class for a given set of observer responses. 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. With latent class models we simply obtain information about probabilistic connections between the true ratings and the observed ratings that *could* generate data such as observed. Currently it seems useful to assess agreement from a combination of loglinear and latent class modeling, with the latter perhaps replaced by a latent trait model if the ratings are ordinal or can be visualized as having an underlying continuous response.

The modeling of patterns of agreement and disagreement should continue to yield interesting and challenging research problems for some time to come. Our survey has focused on “fixed panel” designs, whereby the same observers rate each subject. Relatively little work refers to “varying panel” designs, in which one randomly selects a separate panel of raters for each subject[77,80]. Similarly, not much attention has been paid to modeling agreement on a categorical scale when some observers have replicate ratings. Also, there are many challenges in modeling multi-rater agreement. For such problems, data are usually sparse, with potentially much missing data. In future work, it may be beneficial to incorporate recent research on (1) robust methods using quasi likelihood and estimating equations[81,82] to handle complex dependencies in large sparse tables, (2) incomplete data methods[83] to handle missing data problems, and (3) random effects in categorical data models[84] to treat observers as a sample from a population of observers.

Alternative types of models not discussed in this review are likely to receive more attention as methods of analyzing agreement. In particular, all models we discussed treat the observers as fixed, rather than random. Clearly, in many applications one would prefer to treat the observers as a sample from some population of observers, and obtain information about the degree of agreement for a typical pair of raters in this population. In addition, it might be of interest to compare observers of different types, such as observers from different medical specialties or levels of experience. Mixed models incorporating fixed effects for types and random effects for observers may then be useful. These are important problems for future research. Other types of models that have not been discussed here include signal detection theory methods[27,85] and correspondence analysis and correlation models[37,86]. The first of these has close connections to latent trait models[76], and the second to loglinear and association models with scores for response categories. Whether having this abundance of models will produce ultimate convergence of statisticians to a favorite approach for analyzing agreement, or whether it will simply produce greater fragmentation in types of analyses, remains to be seen.

BIBLIOGRAPHY

1. Landis, J.R., and G.G. Koch. A review of statistical methods in the analysis of data arising from observer reliability studies, Parts I, II. *Statistica Neerlandica* 1975, *29*: 101-23, 151-61.

2. Fletcher, C. M., and P. D. Oldham. Problem of consistent radiological diagnosis in coalminers' pneumoconiosis. *Brit. J. Industr. Med.* 1949, *6*: 168-83.

3. Grubbs, F. E. On estimating precision of measuring instruments and product variability. *Journal of the American Statistical Association* 1948, *43*: 243-64.

4. Overall, J. Estimating individual rater reliabilities from analysis of treatment effects. *Educational and Psychological Measurement* 1968, *28*: 255-64.

5. Burdock, E., J. L. Fleiss, and A. Hardesty. A new view of inter-observer agreement. *Personnel Psychol.* 1963, *16*: 373-484.

6. Fleiss, J.L., R. Spitzer, and E. Burdock. Estimating accuracy of judgment using recorded interviews. *Archives of General Psychiatry* 1965, *12*: 562-7.

7. Hansen, M. H., W. Hurwitz, and M. Bershad. Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute* 1961, *38*: 359-74.

8. Landis, J.R., and G.G. Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977, *33:* 363-74.

9. Landis, J.R., and G.G. Koch. The measurement of observer agreement for categorical data. *Biometrics* 1977, *33*: 159-74.

10. Bartko, J.J. The intraclass correlation coefficient as a measure of reliability *Psychological Reports* 1966, *19*: 3-11.

11. Shrout, P.E., and J.L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 1979, *86*: 420-28.

12. Darroch, J.N. and P.I. McCloud. Category distinguishability and observer agreement. *Australian Journal of Statistics* 1986, *28*: 371-88.

13. Bloch, D.A., and H.C. Kraemer. 2 × 2 kappa coefficients: Measures of agreement or association. *Biometrics* 1989, *45*: 269-87.

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

15. Scott, W.A. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly* 1955, *19*: 321-25.

16. Spitznagel, E. L., and J. E. Helzer. A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry* 1985, *42*: 725-28.

17. Uebersax, J.S. Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin* 1987, *101*: 140-146.

18. Zwick, R. Another look at inter-rater agreement. *Psychological Bulletin* 1988, *103*: 374-78.

19. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971, *76*: 378-82.

20. Kraemer, H.C. Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* 1979, *44*: 461-72.

21. Cohen, J. E. The distribution of the chi-squared statistic under clustered sampling from contingency tables. *Journal of the American Statistical Association* 1976, *71*: 665-70.

22. James, I.R. Analysis of nonagreements among multiple raters. *Biometrics* 1983, *39*: 651-57.

23. Agresti, A. An agreement model with kappa as parameter. *Statistics & Probability Letters* 1989, *7*: 271-3.

24. Fleiss, J.L. *Statistical Methods for Rates and Proportions*, 2nd ed. New York: Wiley, 1981.

25. Fleiss, J.L., and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 1973, *33*: 613-19.

26. Kraemer, H. C. Extensions of the kappa coefficient. *Biometrics* 1980, *36*: 207-16.

27. Kraemer, H.C. Assessment of 2 × 2 associations: Generalization of signal-detection methodology. *American Statistician* 1988, *42*: 37-49.

28. Schouten, H.J.A. Measuring pairwise interobserver agreement when all subjects are judged by the same observers. *Statistica Neerlandica* 1982, *36*:229 45-61.

29. Haberman, S. J. The analysis of residuals in cross-classification tables. *Biometrics* 1973, *29*: 205-20.

30. Bishop, Y., S. Fienberg, and P. Holland. *Discrete Multivariate Analysis.* Cambridge, MA: MIT Press, 1975.

31. Agresti, A. *Categorical Data Analysis.* New York: Wiley, 1990.

32. Bellini, P. and G. Lovison. Square concordance tables in health data analysis. *Rivista di Statistica*

Applicata 1988, *21*: 443-58.

33. Bergan, J. R. Measuring observer agreement using the quasi-independence concept. *Journal of Educational Measurement* 1980, *17*: 59-69.

34. Tanner, M.A., and M.A. Young. Modeling agreement among raters. *Journal of the American Statistical Association* 1985, *80*: 175-80.

35. Haberman, S. J. Log-linear models for frequency tables with ordered classifications. *Biometrics* 1974, *36*: 589-600.

36. Goodman, L. A. Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association* 1979, *74*: 537-52.

37. Goodman, L. A. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Annals of Statistics* 1985, *13*: 10-69.

38. Becker, M. P. Using association models to analyse agreement data: Two examples, *Statistics in Medicine* 1989, *8*: 1199-1207.

39. Agresti, A. A model for agreement between ratings on an ordinal scale. *Biometrics* 1988, *44*: 539-48.

40. Becker, M. P. On the bivariate normal distribution and association models for ordinal categorical data. *Statistics & Probability Letters* 1989, *8*: 435-40.

41. Becker, M. P., and A. Agresti. Loglinear modeling of pairwise interobserver agreement on a categorical scale. *Statistics in Medicine* 1991, *10*: to appear.

42. Becker, M. P. Quasisymmetric models for the analysis of square contingency tables. *Journal of the Royal Statistical Society B* 1990, *52*: 369-78.

43. Sobel, M. E. Some models for the multiway contingency table with a one-to-one correspondence among categories. *Sociological Methodology* 1988, *18*: 165-211.

44. Goodman, L. A. Some multiplicative models for the analysis of cross-classified data. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1* 1972, pp. 649-96. Berkeley: Univ. of California Press.

45. Goodman, L. A. Multiplicative models for square contingency tables with ordered categories. *Biometrika* 1979, *66*: 413-18.

46. Hauser, R.M., and M.P. Massagli. Some models of agreement and disagreement in repeated measures of occupation. *Demography* 1983, *20*: 449-60.

47. Hout, M., O.D. Duncan, and M.E. Sobel. Association and heterogeneity: Structural models of similarities and differences. *Sociological Methodology* 1987, *17*: 145-84.

48. Tanner, M.A., and M.A. Young. Modeling ordinal scale disagreement. *Psychological Bulletin* 1985, *98*: 408-15.

49. Jørgensen, B. Estimation of interobserver variation for ordinal rating scales. *Lecture Notes in Statistics: Generalized Linear Models* 1985, *32*: 93-104.

50. Kjaersgaard-Andersen, P., F. Christensen, S. A. Schmidt, N. W. Rasmussen, and B. Jørgensen. A new method of estimation of interobserver variation and its application to the radiological assessment of osteoarthrosis in hip joints. *Statistics in Medicine* 1988, *7*: 639-47.

51. Rasch, G. On general laws and the meaning of measurement in psychology. pp. 321-33 in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability* 1961,, vol. 4, ed. J. Neyman. Berkeley: Univ. of California Press.

52. Andersen, E. B. Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology* 1973, *26*: 31-44.

53. Tjur, T. A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics* 1982, *9*: 23-30.

54. Agresti, A. Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonals parameters. Unpublished manuscript 1991.

55. Goodman, L. A. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974, *61*: 215-31.

56. Haberman, S. J. *Analysis of Qualitative Data*, Vol. 2. New York: Academic Press, 1979.

57. McCutcheon, A. L. *Latent Class Analysis.* Beverly Hills: Sage Publications, 1987. 58. Lazarsfeld, P.

F., and N. W. Henry. *Latent Structure Analysis.* Boston: Houghton Mifflin, 1968.

59. Clogg, C. C. Some latent structure models for the analysis of Likert-type data. *Social Science Research* 1979, *8*: 287-301.

60. Agresti, A., and J. B. Lang. Quasi-symmetric latent class models, with application to rater agreement. *Biometrics* 1992, *48*: to appear.

61. Hagenaars, J. A. *Categorical Longitudinal Data: Log-linear Panel, Trend, and Cohort Analysis.* Newbury Park, CA: Sage, 1990, p.143. .

62. Uebersax, J. S. Latent class agreement analysis with ordered rating categories. Unpublished

manuscript, 1991.

63. Andrich, D. A rating formulation for ordered response categories. *Psychometrika* 1978, *43*: 561-73.

64. Bartholomew, D. J. Latent variable models for ordered categorical data. *Journal of Econometrics* 1983, *22*: 229-43.

65. Bartholomew, D. J. Latent variable models and factor analysis. London: Griffin, 1987.

66. Rost, J. Rating scale analysis with latent class models. *Psychometrika* 1988, *53*: 327-48.

67. Rost, J. A latent class model for rating data. *Psychometrika* 1985, *50*: 37-49.

68. Lindsay, B., C. C. Clogg, and J. Grego. Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association* 1991, *86*: 96-107.

69. Palmgren, J., and A. Ekholm. Exponential family non-linear models for categorical data with errors of observation. *Applied Stochastic Models and Data Analysis* 1987, *3*: 111-24.

70. Haberman, S. J. A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology* 1988, *18*: 193-211.

71. Clogg, C. C. Unrestricted and restricted maximum likelihood latent structure analysis: A manual for users. Population Issues Research Center, Working Paper 1977-09, Pennsylvania State Univ.

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

73. Clogg, C. C. New developments in latent structure analysis. In D. M. Jackson and E. F. Borgatta, eds., *Factor Analysis and Measurement in Social Research*, 1980, pp. 215-46. Beverly Hills, CA: Sage.

74. Dillon, W. R., and N. Mulani. A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research*. 1984, *19*: 438-58.

75. Espeland, M. A., and S. L. Handelman. Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics* 1989, *45*: 587-99.

76. Uebersax, J. S. Quantitative methods for the analysis of observer agreement: Toward a unifying model. Unpublished manuscript, 1991.

77. Uebersax, J. S., and W. M. Grove. Latent class analysis of diagnostic agreement. *Statistics in Medicine* 1990, *9*: 559-72.

78. Gould, S. J. *The Mismeasure of Man*. New York: W. W. Norton, 1981, p.250.

79. Cox, M.A.A., P. Przepiora, and R.L. Plackett. Multivariate contingency tables with ordinal data, *Utilitas Mathematica* 1982, *21A*: 29-42.

80. Uebersax, J. S. Validity inferences from interobserver agreement. *Psychological Bulletin* 1988, *104*: 405-16.

81. Liang, K. Y., and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika* 1986, *73*: 13-22.

82. Zhao, L. P., L. Wilkens, S. Lipsitz, J. Hankin, and L. Kolonel. A regression method for agreement measures between paired continuous or discrete responses. Unpublished manuscript, 1991.

83. Conaway, M. R. The analysis of repeated categorical measurements subject to nonignorable nonresponse. Unpublished manuscript, 1990.

84. Conaway, M. R. A random effects model for binary data. *Biometrics* 1990, *46*: 317-28.

85. Tosteson, A.N.A., and C.B. Begg. A general regression methodology for ROC curve estimation. *Medical Decision Making* 1988, *8*: 204-15.

86. Goodman, L. A. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review* 1986, *54*: 243-309.

Table 1. Diagnoses of carcinoma for pathologists A and B.

	B			
A	1	2	3	4
1	22	2	2	0
2	5	7	14	0
3	0	2	36	0
4	0	1	17	10

Source: Based on data in Landis and Koch (1977b)

Table 2. Likelihood-ratio statistics for models fitted to Table 1

No. Latent Classes	Model	Likelihood-Ratio Statistic	DF
1	Independence	118.0	9
	$L\times L$ association	8.8	8
	$L\times L$ association + diagonal	4.8	7
	Quasi Independence	13.2	5
	Quasi $L\times L$ association	1.1	4
	Quasi Symmetry	1.0	3
2	$L\times L$ LC	32.8	6
	Ordinary LC	31.9	2
3	$L\times L$ LC	8.7	5
4	Uniform $L\times L$ LC	4.6	5
	$L\times L$ LC	3.6	4

Table 3. Diagnoses of carcinoma, with adjusted residuals in parentheses for independence model.

		B		
A	1	2	3	4
1	22	2	2	0
	(8.49)	(-0.47)	(-5.95)	(-1.76)
2	5	7	14	0
	(-0.50)	(3.20)	(-0.54)	(-1.76)
3	0	2	36	0
	(-4.08)	(-1.22)	(5.51)	(-2.28)
4	0	1	17	10
	(-3.30)	(-1.32)	(0.28)	(5.93)

Table 4. Fitted values for loglinear model and for latent class model with $L = 4$ latent classes.

	B			
A	1	2	3	4
1	22	2	2	0
	(22.0)[a]	(2.0)	(1.9)	(0.0)
	(21.9)[b]	(3.2)	(0.9)	(0.0)
2	5	7	14	0
	(4.6)	(8.4)	(12.8)	(0.2)
	(5.0)	(6.9)	(14.0)	(0.2)
3	0	2	36	0
	(0.4)	(1.2)	(35.6)	(0.8)
	(0.2)	(1.5)	(35.6)	(0.8)
4	0	1	17	10
	(0.0)	(0.4)	(18.6)	(9.0)
	(0.0)	(0.4)	(18.5)	(9.1)

[a] – loglinear model

[b] – latent class model

Table 5. Estimated Response Probabilities by Latent Class

Observer	Latent Class	Response 1	2	3	4
	1	.977	.023	.000	.000
A	2	.304	.674	.021	.000
	3	.001	.207	.603	.189
	4	.000	.000	.033	.967
	1	.985	.015	.000	.000
B	2	.351	.502	.147	.000
	3	.000	.035	.953	.012
	4	.000	.000	.455	.545