WORKING PAPER SERIES

WORKING PAPER NO 7, 2008

Swedish Business School at Örebro

An Overview of Methods in the Analysis of Dependent Ordered Categorical Data: Assumptions and Implications

By

Hans Högberg
hans.hogberg@lg.se
Centre for Research and Development
Uppsala University and County
Council of Gävleborg
Sweden

Elisabeth Svensson
elisabeth.svensson@oru.se
Department of Statistics at
Swedish Business School
Örebro University
Sweden

http://www.oru.se/esi/wps
SE-701 82 Örebro
Sweden

ISSN 1403-0586

An Overview of Methods in the Analysis of Dependent Ordered Categorical Data: Assumptions and Implications

Hans Högberg
hans.hogberg@lg.se
Centre for Research and Development
Uppsala University and County
Council of Gävleborg
Sweden

Elisabeth Svensson
elisabeth.svensson@oru.se
Department of Statistics at
Swedish Business School
Örebro University
Sweden

Abstract

Subjective assessments of pain, quality of life, ability etc. measured by rating scales and questionnaires are common in clinical research. The resulting responses are categorical with an ordered structure and the statistical methods must take account of this type of data structure. In this paper we give an overview of methods for analysis of dependent ordered categorical data and a comparison of standard models and measures with nonparametric augmented rank measures proposed by Svensson. We focus on assumptions and issues behind model specifications and data as well as implications of the methods. First we summarise some fundamental models for categorical data and two main approaches for repeated ordinal data; marginal and cluster-specific models. We then describe models and measures for application in agreement studies and finally give a summary of the approach of Svensson. The paper concludes with a summary of important aspects.

JEL classification: C14
Keywords: Dependent ordinal data, GEE, GLMM, logit, modelling

Introduction

More and more frequently assessments of subjective phenomena such as pain, mood, functioning, quality of life, quality of care, ability etc. occur in clinical research. Stevens [1] characterised the properties of data in terms of nominal, ordinal, interval and ratio scales. This characterisation is widely adopted although it is by no means complete. Measurements in physical, biological and medical sciences obtained from laboratory instruments are standardised and often on the ratio scale. Contrary, subjective assessments and judgements are qualitative, and the data are not standardised. The data are merely a categorization of individuals in different classes. The categorisation indicates either just a classification into a category different from another (nominal), or a classification into a category indicating a relative order (ordinal). The subjective assessments are often measured by some rating scale or questionnaire. The rating scales often results in responses that are ordinal with a discrete or a continuous range of possible values.

The aim of clinical research is often to evaluate changes in responses to treatment or medical care. Then the study design is longitudinal and the same individuals are evaluated at two or more occasions. Observations are taken independently but data are related within individual repetitions. Other research questions in clinical research may concern the quality of rating scales. Validity and reliability may be evaluated by intra or inter rater designs to assess agreement. Raters judge independently the same individuals repeatedly and the response is dependent within repetitions. Thus, in many studies researchers are faced with dependent ordered categorical data and the statistical methods must take account of this type of data structure.

Overviews and survey articles of statistical analysis of dependent ordered categorical data appear occasionally in the scientific literature. Often there is either a very broad scope or a more specific focus in the articles, e.g. overviews of ordinal categorical data or surveys of modelling pattern of agreement [2-8]. This paper aims at an overview of methods for analysis of dependent ordered categorical data and to compare standard methods and measures with the measures of Svensson´s approach [9-13]. The focus will be at the assumptions behind model specifications and data, the usefulness for descriptions and inferences, and implications. First we will briefly describe some of the fundamental models for categorical

data and then describe two approaches for clustered, e.g. repeated, ordinal data. We then turn to log linear models. Descriptions of models and summary measures for application to order consistency and agreement precede an introduction to Svensson's methods, and the paper concludes with a summary of important aspects and conclusions.

Methods and estimation

Basic asymmetric models for categorical data

The basic model for nominal data is the *baseline category model,* which in its logit form looks like

$$\log\left[\frac{P(Y=i)}{P(Y=I)}\right] = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x} \qquad \text{, for categories i=1,…, I-1.}$$

One category is defined as a baseline category and the other categories are paired off with this category. The model simultaneously estimates the effects. The effects may vary with the paired comparisons. A severe limitation is that the model formulation not recognizes the ordered structure in the response variable.

For ordinal data the most common and useful model is the *cumulative logit model* which may be written as

$$\text{logit}\left[P(Y \le i)\right] = \log\left[\frac{P(Y \le i)}{1-P(Y \le i)}\right] = \log\left[\frac{P(Y \le i)}{P(Y > i)}\right] = \alpha_i + \boldsymbol{\beta}'\mathbf{x} \quad \text{, i=1,…, I-1}$$

A consequence of the model formulated as above, is constant effect for each category – the log cumulative odds ratio is proportional to the distance of two covariate settings for each category and thus the model is called the proportional odds model [7]. It implies stochastic ordering and linearity of the response. The model is invariant to reversed order of the response scale and allows collapsing the categories. The odds ratios refer to the entire scale in terms of cumulative frequencies. Among other advantages no scores on response are required for the fit of the model [6], and it requires only one parameter for each predictor which means

fewer degrees of freedom and a more parsimonious model. The assumptions of stochastic ordering and proportional odds are not valid when there is shift in variability. This is a location and scale problem and then one has to use models which incorporate varying dispersion, which leads to non-linear models. For cumulative probabilities other links than the logit link are available as well, e.g. probit, complementary log-log link [7].

The *adjacent categories logit model*, which may be written

$$\log\left[\frac{P(Y=i)}{P(Y=i+1)}\right] = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x} \quad , \text{i=1,…, I-1}$$

is related to the baseline category model. The adjacent categories model with a common effect can be expressed as baseline category model and as such reducing the number of parameters. The effects refer to individual and adjacent response categories and not to the entire scale for the response as in the cumulative logit model. Like the cumulative logit model it implies stochastic ordering. By simplifying the model letting the parameters $\boldsymbol{\beta}_i$ be fixed and assigning scores to the response then there is correspondence to the loglinear linear-by-linear association model [14, 15]. The results from the model are not invariant to the choice of and the number of response categories. Since ordinal data has an inherent ordered structure the crucial is to choose the first category, representing the best or the worst status.

The *continuation ratio logit model*

$$\log\left[\frac{P(Y=i)}{P(Y\geq i+1)}\right] = \log\left[\frac{P(Y=i+1)}{P(Y\leq i)}\right] = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x} \quad , \text{i=1,…, I-1}$$

is a useful model when there is some sequential mechanism that determines the outcome [4, 6, 16]. If modelling separate effects for each category on the response the multinomial likelihood factors into a product of binomial likelihoods for the separate likelihoods and separate fitting of models for different continuation-ratio logits is equivalent to simultaneous fitting [6]. If reversing the category order it becomes a different model and collapsing categories also results in a different model. In a discussion of the paper of Liu and Agresti [6],

Tutz put forward advantageous properties of the continuation-ratio model regarding using category specific effects compared to cumulative type of models, particularly in connection with more complex models such as marginal and conditional models.

Modelling clustered ordinal data

In modelling dependent, i.e. clustered, ordered categorical data, two major classes of asymmetric models are common. These are marginal models for which effects are averaged over all clusters at particular levels of predictors, and cluster-specific models for which effects apply at the cluster level. Cluster-specific models are also called conditional models due to the method to estimate the fixed effects by conditioning on the cluster-specific effects. For example, Agresti and Natarajan [5], surveys various ways of modelling these types of data including other methods as well, such as Bayesian methods, semi-parametric and nonparametric methods. Generalized estimating equation (GEE) estimation method of marginal models is marginal modelling of generalized linear models (GLM), and generalized linear mixed models (GLMM) of random effects is an extension of GLM.

Marginal modelling

Marginal models focus at the marginal distributions and not explicitly on the individual response. There are three different approaches of estimation.

Maximum likelihood

The maximum likelihood approach for fitting marginal logit models is problematic. The model refers to marginal distributions when likelihood refers to joint multinomial probabilities. So, if the number of categories $i = 1,\ldots,I$, the number of responses $t = 1,\ldots T_k$ within clusters $k = 1,\ldots n$, or the number of predictors is large the ML approach is not practical. The lack of a simple multivariate distribution for describing marginal moments and correlations for categorical responses in contrast to the multivariate normal distribution is also a problem.

The ML methods used are based on generalized loglinear formulation of marginal logit models which incorporates model constraints together with identifiability constraints. The ML fit is then the solution to Lagrangian likelihood equations using Newton-Raphson algorithms, see [17-20]. There are several alternative fitting approaches [19, 20]. Other ML methods define multivariate logistic models with one-to-one correspondence with the joint multinomial probabilities and marginal model parameters as well as higher order parameters of joint distributions; see [21-23]. With the ML approach one has a likelihood function and hence the possibility to use likelihood ratio tests and tests of goodness of fit.

GEE

One successful alternative to ML estimation in marginal modelling is a multivariate generalization of quasi-likelihood which links the mean vector to a linear predictor and specifies how the covariance matrix depends on the mean vector [24]. The method requires a working guess of the covariance structure but it is not necessary to assume a particular multivariate distribution. By further assuming that the multivariate distribution belongs to the exponential family with that mean vector and covariance matrix, the parameter estimates then are the solution of the likelihood equations. These equations are called the set of generalized estimating equations (GEE) [25]. The method applies to the marginal distribution for each Y_t. The parameter estimates are consistent even if the covariance structure is miss-specified. The GEE method is very attractive as it is computational more simple than ML methods are. As the generalized quasi-likelihood not requires the multivariate joint distribution, the full likelihood function is not specified. Thus, likelihood based methods for tests of fit, comparing models, and parameter tests and interval estimation are not available. Inference is based on Wald statistics. The most common link functions in applications of marginal models of ordinal data with GEE are cumulative logit and cumulative probit.

WLS

An alternative estimating method to ML is the weighted least squares (WLS) which is an extension to the ordinary least squares (OLS) to handle correlated response and non constant variances. The advantage in modelling clustered data is that it is a simple estimation process but it also has very severe limitations. For example, it is only applicable to categorical explanatory variables and the sample sizes must be large, the contingency tables must be small and not to sparse. It is not so often used nowadays when computer routines are available

for ML solutions. The WLS was made popular for categorical data by Grizzle, Starmer and Koch in 1969 [26] and further elaborated in Koch et.al. [27].

Conditional modelling

An alternative way of modelling dependent ordered categorical data is to explicitly model the joint distribution, i.e. modelling cluster-level terms in the model. It is not recommended to use ordinary ML to estimate the cluster-specific parameters as fixed effects. One can handle the large number of cluster-specific parameters by treating them as nuisance parameters and either eliminating them by condition on their sufficient statistics or treating them as random effects, varying randomly among clusters.

Conditional maximum likelihood

The conditional ML has several drawbacks compared to random effects but is in some instances advantageous to consider. First, in retrospective sampling the clusters are not randomly sampled and then one may introduce bias by modelling the cluster-specific effects with a random model. Secondly, if a random effect model is appropriate it may be difficult to check the assumed parametric distribution. Some drawbacks with conditional ML is restriction to inference about within-cluster effects, no information about $\{u_k\}$, less efficient than the random effects approach for estimating the fixed effects, restricted to canonical links [28, 29]. The restriction to canonical link functions for ordinal data excludes the possibility to use the cumulative logit models [5].

Random effect

Random effects may be modelled by the generalized linear mixed model (GLMM) which is an extension of GLM that permit random effects as well as fixed effects in the linear predictor. Some specific random effect mixed models has been in use since 1970s, for example the negative binomial model for count data and beta binomial model for binary data [30, 31]. Since the Rasch models and other item response models were introduced these has been further developed in the framework of random effects models [5, 32]. By the GLMMs these, and other mixed models (not all), are included in a general framework.

GLMM

Agresti et.al. [32] give an overview of modelling cluster-level random effects by generalized linear mixed models (GLMM). In a general form the linear predictor of the model may be written as

$$g(\mu_{kt}) = \mathbf{x}'_{kt}\boldsymbol{\beta} + \mathbf{z}'_{kt}\mathbf{u}_k$$

for a link function $g(\cdot)$ and where intercepts are treated as random. The vector $\mathbf{x}_{kt}$ is a column vector of explanatory variables for the *k:th* cluster and *t:th* observation in the cluster, and $\boldsymbol{\beta}$ is a column vector of fixed effect parameters. Often, the observation unit in a cluster is a subject, but it could also be a time point in a longitudinal study for subject then constituting a cluster. The column vector $\mathbf{z}_{kt}$ denotes the design vector for the random effects $\mathbf{u}_k$.

Typically, the random effects are assumed to be multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, i.e. $N(\mathbf{0}, \boldsymbol{\Sigma})$. Characteristic for a GLMM is that the model is built up in two stages. Conditionally on the random effects it is a GLM, and the random effects is specified by an assumption of their multivariate distribution. The most common link functions in applications of random effects models are cumulative logit, cumulative probit, continuation-ratio logits, adjacent-categories logits, and other cumulative links such as complementary log-log link.

By introducing cluster specific random effects the variability of $\mathbf{u}_k$ induces a correlation between responses within a cluster. This is so because heterogeneity of the observations [28, 32]. Thus, the model incorporates the dependencies between observations within a cluster. As the variances of $\mathbf{u}_k$ increases, the correlation increases. The estimated random effects may be of interest to assess the observation unit heterogeneity and also for prediction of probabilities and odds ratios. Whether the assumption of multivariate normality of the random effects is valid is of no concern regarding the estimates of the fixed effects, but if the aim is to estimate and make inference about the random effects the distributional assumption is important.

One of the drawbacks of GLMM is the difficulty by which the parameters are estimated within the framework of maximum likelihood. This is due to the complex form of the likelihood function, in which one has to integrate out the random effects. Two types of methods are distinguished, numerical or Monte Carlo methods to solve the integral in the likelihood function, or approximate ML methods. Numerical or Monte Carlo methods are

complex and difficult to program and implement in statistical software but they converge to the ML estimates. Approximate ML methods, such as penalized quasi-likelihood and marginal quasi-likelihood methods are much easier to implement, but can be very biased in some cases [29, 32, 33]. An approach that uses a combination of a fully multivariate Taylor expansion and a Laplace approximation has been proposed by Raudenbush et al. [34]. Bayesian approaches also exits.

Inference about $\boldsymbol{\beta}$ is based on the asymptotic normality of the ML estimate of $\boldsymbol{\beta}$. Hypothesis involving $\boldsymbol{\beta}$ may be tested using asymptotic likelihood ratio tests. Testing the variance components is more troublesome. The likelihood ratio statistic does not necessarily have an asymptotic chi-square distribution under the null hypothesis when it involves parameter on the boundary of the parameter space (σ^2=0) [32]. Sometimes, interest may be on estimates of the random effects. These estimates may be obtained by empirical Bayes methods [35].

Log linear models

To describe association between categorical variables we also have the association models within the class of loglinear models [36, 37]. Special cases are the models of linear-by-linear association, row effects, column effects, and row and column effects [36, 38, 39]. These are models which all treat the variables symmetrically, i.e. make no distinction between response and explanatory variables, in contrast to logit models in which one response variable depends on one or several explanatory variables. Loglinear models focus on associations and interaction in their joint distribution but connection exists between them [7]. A general structural form for a loglinear model of association between two variables X and Y may look like

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \mu_i \nu_j$$

for categories *i* and *j*. The form of the interaction term identifies some common models. When $\mu_i\upsilon_j = \beta u_i \upsilon_j$ with $\{u_i\}$ and $\{\upsilon_j\}$ being fixed monotone constants the model is the linear-by-linear association model [36, 38]. When using them with equally spaced scores they relate to adjacent category models [28]. The linear-by-linear association model is an extension of the

saturated log linear model for two variables. One advantage is that the model has one parameter instead of (I-1)(J-1).

One may generalize the linear-by-linear model to apply to scores as parameters rather than fixed ones. If, for example, the row scores $\{\mu_i\}$ are treated as parameters and the column scores $\{\upsilon_j\}$ are retained as fixed monotone scores we have a row-effect model. In a contingency table setting that is to say that X (row variable) is treated as nominal and Y (column variable) is treated as ordinal. Instead to having one parameter reflecting the ordered structure in the linear-by-linear model we have now (I-1) parameters. Letting the Y variable be the nominal and X the ordinal variable and letting the categories for the columns be represented by parameters we have the column-effect model. Both models may be formulated in adjacent-categories logit form [28].

Further generalizations of the linear-by-linear models are to replace both the row scores and the column scores by parameters. For identification purpose location and scale constraints are required. The model is no longer a log-linear model, but rather a log-multiplicative. The model may be used to estimate scores and these may in turn be used to describe distinguishability of categories [40].

Log linear models has been modified to reflect situations in which there are square tables and when the categories in the rows exactly corresponds to the categories in the columns and within this framework we may also model dependent data. The models are further extended to model categorical variables with ordered structure.

Measures of order consistency

When evaluating and comparing rating scales it may concern assessment of construct validity. Construct validity refers to how closely two measurement instruments with different operational definitions for the same theoretical concept to be measured are related, that is, the extent of interchangeability of the scales [12]. If the scales are interchangeable the different scaling methods should produce the same ordering of individuals, i.e. there should be a high level of order consistency between the scales [12]. Assessment of order consistency may be achieved by measures of concordance.

As Kruskal in [41], page 818, says: *"Some measures of association reflect aspects of concordance (greater values of X go with greater values of Y), while other measures reflect aspects of connection that do not take the sense or direction into account"* . Measures of association based on concordance or correlation expresses both direction and strength. It is a matter of sign and distances which is illustrated by Daniels [42] in his unification of measures in the correlation family. The correlation ρ is also a measure of concordance, in which the distances between pairs of observations are weights.

The most commonly used measures of association for ordered categorical data are measures of differences between probabilities of concordant and discordant pairs. Examples of these are Kendall's tau-a, Kendall's tau-b, Stuart's tau-c, Goodman-Kruskal's gamma, Somers' delta [12]. The difference in these measures lies in the handling of ties. A standard non-parametric measure of association is also the Spearman rank-correlation ρ_s. In the definition of measures using concordance and discordance no scores, rank or other, need to be used. But Kendall's tau may be formulated as a Pearson product-moment correlation between signed indicators of X's and Y's, and Spearman's rank-correlation is the Pearson product-moment correlation with the ranks instead of the actual variates [41-43]. The different ways of adjusting for tied observations affect the possibility of attaining the limiting values of the measures, viz. -1 and +1. They are also affected by the frequency distributions [12]. In order to attain the limiting values Kendall's tau-b require all observations on the main diagonal in a square (m x m) table or untied observations while Stuart's tau-c requires that the number of observations n is a multiple of the number of cells in the longest diagonal and all observations uniformly distributed on the m cells. Somers' delta attain the limiting values provided a strict monotonicity in the dependent variable and Goodman-Kruskal's gamma requires monotonicity but not necessarily strict monotonicity, i.e. gamma equals one when there is total order consistency. If we have total order consistency only gamma attain one while the other measures requires further conditions. Both tau-b and delta are less in absolute value than gamma, but gamma seems more dependent than tau-b on the number of categories and the way that they are defined [12, 14].

Measures of agreement

As Agresti [28], pages 431-432, points out, *"With subjective scales, agreement is less than perfect. Analyses focus on describing strength of agreement and detecting patterns of disagreement. Agreement and association are distinct facets of the joint distribution. Strong agreement requires strong association, but strong association can exist without strong agreement."* A further requirement for strong agreement is similar marginal distributions. Landis and Koch [44], page 113, summarized the distinction between association and agreement: *"Whereas measures of association reflect the strength of the predictable relationship between ratings of the two observers, measures of agreement pertain to the extent to which they classify a given subject identically into the same category. As such, agreement is a special case of association"*. So in general, a measure of association is not adequate as a measure of agreement if not there is similar marginal distributions. But if a measure of association is formulated as the extent to which the data cluster on the main diagonal it may serve as a measure of agreement. Later, we will see that a novel ranking system and a way to construct the best common ordering of paired classification make it possible to construct measures of association that reflect agreement and order consistency.

An intuitive and simple approach to evaluate agreement is to compare the number of exact agreement to the total number of observations in a ratio, i.e. the sum of the observations on the main diagonal to the total number of observations. One disadvantage of this measure is that some of the agreement may have arisen due to chance only [45]. The development of other agreement measure, such as Cohen's kappa has this as its starting point. Perfect agreement can only occur when the marginals are homogeneous, so bias corrected measures has been devised [10, 12].

Cohen's kappa

Kappa is a measure of agreement adjusted for chance expected agreement. Kappa has the form

$$\kappa = \frac{\pi_o - \pi_e}{1 - \pi_e}$$

where π_o is the observed probability of agreement and π_e is the expected probability of agreement under some hypothesis, such as independence. By this, the kappa measure

quantifies agreement beyond what is expected by chance under the hypothesis of independence.

Originally, kappa was applied for a 2 x 2 table in which results for two equally skilled raters judging outcome of a variable for a group of individuals were recorded [46]. Equally skilled raters imply no bias in classification of the individuals (marginal homogeneity). The kappa statistic does not get hold of any bias, which occurs when two raters uses the scale differently. Even in situations when marginal homogeneity occurs, the kappa measure is dependent on the prevalence of the attribute being measured. Different tables may give rise to the same value [47]. The maximum value of kappa is + 1, which is attainable when perfect agreement occurs and the off diagonal cells are zero. That implies equal marginal frequencies. It may be of interest to determine the maximum value of kappa given the marginal frequencies, so kappa max was defined in [46] but is not commonly used.

In assessing inter rater agreement it is equally important to measure the level of agreement and bias. In a 2 x 2 table marginal homogeneity can be tested by the McNemar's test. Kappa is extended to treat nominal variables with more than two categories and to treat ordinal variables.

Weighted kappa

If observations are classified in more than two categories, the possibility for disagreement increases. Then a weighted kappa measure with different types of weights has been proposed [48]. The different weighting systems have the purpose of adjusting for the seriousness of different levels of disagreement. It amounts to assigning numerical values to the categories [49]. A linear weighting system may assign maximum weights (=1) to perfect agreement and minimum weights (=0) to the most extreme disagreement observations and equidistant weights in-between. The weighted kappa with quadratic weights has been shown to equal the intra class correlation coefficient, irrespective of the marginal distributions [50]. Marginal heterogeneity is not as easily tested as in the case of a 2 x 2 table. Models of symmetry may be used.

Models of symmetry, quasi-symmetry and quasi-independence for agreement

Based on log-linear association models agreement was assessed by symmetry, quasi-symmetry and marginal homogeneity parameterizations beginning in the mid 1960s and continuing into the early 1990s. Extension to exploit ordinality in the classifications was carried out in the 1980s, starting with McCullagh in 1978, and Goodman 1979 and 1985 [36, 37, 51, 52]. Darroch and McCloud [53] defined the degree of distinguishability as a measure of category distinguishability and argued that the kappa measure was unsatisfactory in the framework of a quasi-symmetry model. Becker [54], Agresti [3], and Schuster and von Eye [55] gave compact and concise overviews of different log-linear models to analyse agreement. Some fundamental log-linear and association models are extended via arguments from Darroch and McCloud about the property of quasi-symmetry. Adding parameters to account for association and agreement, either constant or uniform association and agreement, or different parameters for the categories make modelling flexible. Although research and development of log-linear models for agreement is ongoing, much more efforts seem to be on logit and related asymmetric models.

Svensson´s methods

Svensson has devised and developed methods for comprehensive evaluation of paired dependent ordered categorical data [9-11, 13]. The methods exploit the rank invariant properties of the data and assume nothing further. Thus, the methods are non-parametric. The systematic part (bias) of the difference between the paired observations is clear from the marginal distributions in a contingency table. The type of systematic difference may be visualized by a type of curve, called ROC (Relative Operating Characteristic) or Q-Q plot, see figure 1. This way of illustrate two responses has been used in psychology [56], and this application differs from applications in diagnostic test procedures where Receiver Operating Characteristic curves is used. If there is a constant shift to higher frequencies on the second variable (occasion) compared to the first variable or the reverse, we have a case of stochastic ordering, and Svensson [9, 13] define the parameter of the systematic disagreement in position empirically measured by the measure of relative position (RP). In this case the ROC curve is concave or convex and we have a case when links for cumulative odds may also be used for estimation and test, see figure 1a. If there is a systematic shift in concentration of the classification to central categories on one variable compared to the other, Svensson [9, 13]

defined the parameter of systematic disagreement in concentration to represent it. The parameter is measured by the empirical measure of relative concentration (RC). Then the ROC curve is S-shaped along the diagonal line, see figure 1b.

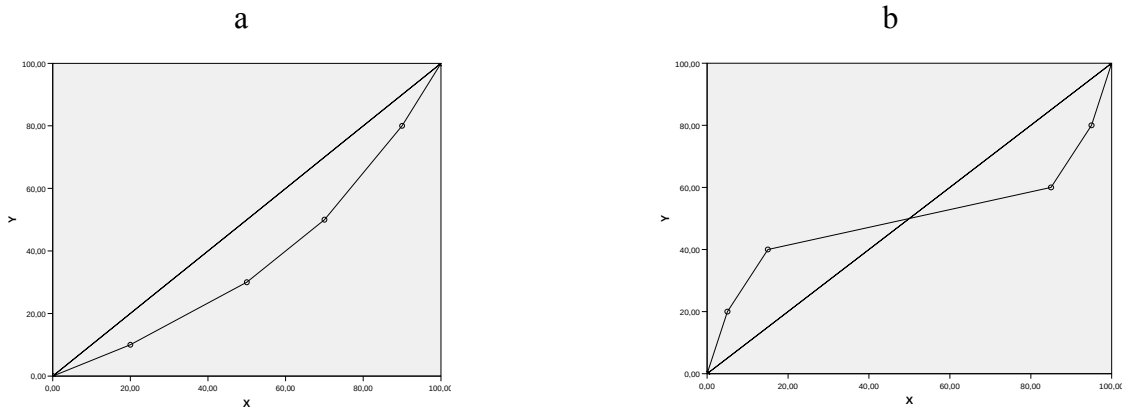

Figure 1. Relative Operating Characteristic curves illustrating, a) systematic disagreement in position measured by the measure of relative position (RP), and, b) systematic disagreement in concentration measured by the measure of relative concentration (RC).

More specifically, for the pairs of ordered classifications (X_k, Y_k) and (X_l, Y_l) the parameter of systematic disagreement in position is defined [9, 13]

$$\gamma = P(X < Y) - P(Y < X) = \sum_{\nu=1}^{m} p_{\nu}^{(Y)} P_{\nu-1}^{(X)} - \sum_{\nu=1}^{m} p_{\nu}^{(X)} P_{\nu-1}^{(Y)}$$

where $p_{\nu}^{(X)}$ and $p_{\nu}^{(Y)}$ are the νth category probabilities for classification X and Y, respectively, and $P_{\nu}^{(X)}$ and $P_{\nu}^{(Y)}$ are the cumulative νth category probabilities for the ν=1,…, m categories. The parameter satisfies $-1 \le \gamma \le 1$. The empirical measure of relative position (RP) replaces the probabilities by the corresponding relative frequencies.

The parameter of systematic difference in concentration has its departure from the expression

$$P(X_l < Y_k < X_m) - P(Y_l < X_k < Y_m),$$

for any independent random variables X_l, Y_l , X_k, Y_k , X_m, and Y_m with distributions $P(X = \upsilon) = p_{\upsilon}^{(X)}$ and $P(Y = \upsilon) = p_{\upsilon}^{(Y)}$ [9, 13]. This expression is normalized by its bounds and the parameter is

$$\delta = \frac{1}{\min(p_0 - p_0^2, p_1 - p_1^2)}\left[\sum_{\upsilon=1}^{m} p_\upsilon^{(Y)} P_{\upsilon-1}^{(X)}(1 - P_\upsilon^{(X)}) - \sum_{\upsilon}^{m} p_\upsilon^{(X)} P_{\upsilon-1}^{(Y)}(1 - P_\upsilon^{(Y)})\right],$$

where $p_0 = P(X_k < Y_l) = \sum_{\upsilon=1}^{m} p_\upsilon^{(Y)} P_{\upsilon-1}^{(X)}$ and $p_1 = P(Y_k < X_l) = \sum_{\upsilon=1}^{m} p_\upsilon^{(X)} P_{\upsilon-1}^{(Y)}$.

It is shown in [9] that $-1 \le \delta \le 1$. Substituting the probabilities by their corresponding relative frequencies yields the empirical measure of relative concentration (RC). Situations when there is a systematic shift in concentration occur now and then. Models that imply stochastic ordering is then not applicable.

If the difference (disagreement or change) is purely systematic the marginal frequencies completely determines the pattern of systematic change in pairs of observations. Thus, given a contingency table with observed data, it is possible to construct a table with purely systematic difference by pairing off the two set of marginal frequencies. Such a construction Svensson [9, 13] called the rank transformable pattern of agreement (or change), RTPA (or RTPC). We then have a table with cell frequencies that are consistent with pure systematic difference. The pairing off procedure may change an individual position in the table, i.e. an individual may have been placed in other categories, but none of the individual observation has changed its ordering relative to the other observations. By comparing the actual cell frequencies with the cell frequencies in the RTPA (RTPC) we get an idea of the extra individual difference in addition to the systematic difference. A rank transformable pattern of agreement exists for any pair of two marginal distributions, including homogeneous distributions, so there are examples of tables with no systematic part of difference and a significant individual part. Complete agreement, or no change, is a table with all observations on the main diagonal in a square table. If observations in a table differ from the RTPA, then it is a sign of individual difference.

An extension of the common way to rank observations is to rank them tied to the paired observation [9, 13]. Each observation is ranked on one of the observation in the pair while considering the position of the other observation in the pair. The result is two rank values for each of the observation, and in the contingency table we construct two rank values in each cell, one rank value for the row variable conditional on the position of the column variable and the other rank value for the column variable conditional on the position of the row

variable. This ranking is called the augmented ranking approach. When the two rank values differ there is a sign of individual dispersion of pattern of change from the systematic difference. The table with RTPA has equal augmented rank values in each cell. A measure of individual dispersion from RTPA is then the relative rank variance RV. To be specific, the parameter of the Variance of the Relative Rank Difference [9] for a probability distribution p_{ij}, i=1,…, m and j=1,…,m with $\sum_{i=1}^{m}\sum_{j=1}^{m} p_{ij} = 1$, is defined as

$$\tau^2 = \sum_{i=1}^{m}\sum_{j=1}^{m} p_{ij}(q_{ij}^{ul} - q_{ij}^{lr})^2$$

where $q_{ij}^{ul} = \sum_{i_1<i}\sum_{j_1>j} p_{i_1 j_1}$ and $q_{ij}^{lr} = \sum_{i_1>i}\sum_{j_1<j} p_{i_1 j_1}$, and where the upper case acronyms *ul* and *lr* stands for upper left and lower right, respectively. To make the parameter to have the attractive interval [0, 1] a normalizing factor of 6 is multiplied

$$\psi = 6\tau^2$$

The empirical measure of random differences between two ordered categorical judgments on the same individual, called the relative rank variance is obtained by substituting the probabilities by their observed relative frequencies [9, 13].

The rank transformable pattern of agreement is a pattern of total agreement of ordering all pairs of observations. It is an expected pattern of paired classifications when there is a total agreement in the ordering of all pairs [12, 57]. RV is a measure of dispersion of observations from the best possible agreement in ordering (RTPA), given the marginal distribution. In the repeated measures situation it is a measure of dispersion of observations from the ordered-preserved group change in categories between occasions [58]. A measure of the closeness of observations to the best possible agreement in ordering when the marginal heterogeneity is taken into account is the augmented rank-order agreement coefficient r_a [10]. It is a measure of the consistency of the observations to the rank transformable pattern of agreement. The augmented rank-order agreement coefficient (r_a) is also the Pearson product-moment correlation coefficient based on pairs of augmented mean ranks.

In the presentation of the concepts of concordance and association we saw that the classical measures differ in how they corrected for ties. If we take into account the information on the mutual relationship between the paired judgements on the same variable for the same individual we have in the augmented ranking approach some new measures of concordance for agreement evaluation may be defined. The parameter of the measure of disorder (Θ_D) and a coefficient of monotonic agreement (Θ) are defined as [12, 57]:

$$\Theta_D = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}(q_{ij}^{ul} - q_{ij}^{lr})}{1 - \sum_{j=1}^{I}\sum_{j=1}^{J} p_{ij}^2}$$

where the scales has I categories for the row variables X_k and J categories for the Y_k variables, k=1,…,n and I and J not necessary the same. Due to the fact that ties only occurs in identical pairs, the other observations are either disordered or ordered, so a measure of excess ordered pairs over disordered is:

$$\Theta = 1 - 2\Theta_D$$

The range of possible values of the coefficient of monotonic agreement is $-1 \le \Theta \le 1$. The pattern of total order consistency, as defined by RTPA, implies no disordered observations, and thus $\Theta = 1$. The pattern of total inconsistency in order ($\Theta_D = 1$) result in $\Theta = -1$.

Summary of important aspects and conclusions

Models for dependent ordered categorical data are often considered as superior to tests and summary measures. Agresti [3] pointed out regarding modelling agreement that model-based approaches yield additional and more precise information than provided by summary measures. However, in reducing the degrees of freedom by modelling the risk for misspecification increases due to the restrictions put on the data. Models are also used for tests, and within the framework of the likelihood principle several different tests are available. Models may also be used for estimation of different aspects of the data and to predict, for example, probabilities. Residuals may be used for comparing different models. The models are most often fitted by the maximum likelihood principle even if it may be computationally

difficult in some cases. During the latest decade Bayesian approaches [5, 6, 59] to modelling has become popular. The risk of modelling is violations of fundamental assumptions.

Marginal models focus on marginal distributions and not explicitly on the individual responses. Conditional models explicitly model the joint distribution. The choice of approach is a question of the application at hand. Conditional models are most useful when the objective is to make inference specifically about subjects and when interests are in heterogeneity of the subjects. The mixed model approach takes care of within-cluster dependencies and between-subject heterogeneity. The models produce consistent parameter estimates with improved efficiency compared to fixed effects models and other models not taken care of the within-cluster dependencies. Conditional models produce estimates of subject-specific effects. When population-averaged effects are of primary interest it is more relevant to parameterize in such a way that the regression parameters has a direct marginal interpretation. By the GEE methodology estimation is easier, and consistent estimates of population-averaged effects are obtained even if the working guess of the correlation matrix is misspecified [29, 32]. In the discussion of which approach to use, Lee and Nelder [60] concluded that the conditional model approach is the basic model and that a marginal model may be deducted from any conditional model. The seemingly notable differences in regression parameters are caused by unidentifiable constraints on the random effects. There are at least two potential drawbacks with GLMM, the first has to do with the assumption of a particular distribution for the random effect. The assumption may be important if the purpose is to estimate and make inference about the random effects[32]. The second is the estimation of the parameters in the model. In applications where the dimension of the integral in the likelihood function is small, numerical integration such as Gauss-Hermite quadrature methods works well [29, 32]. But with higher dimension other methods are preferable, some which are not available in standard computer software. Both marginal models and conditional models uses custom link functions, with specifically defined ways of dichotomization the multinomial response.

Models for ordinal categorical data should exploit the rank invariant properties of the data. Cumulative models for ordinal data are not permutation invariant, which is as it should be, but models with e.g. logit and probit links are palindromic invariant. The models with complementary log – log link and log – log link are not palindromic invariant. This can be seen from the form of the distribution function (link function). Symmetrical distribution

function as links allows palindromic invariance [7]. Instead of using asymmetric GLM models for ordered categorical data describing stochastic ordering one may use scores in log linear models for that purpose. The result from the log-linear model is then not invariant to redefinition of categories or grouping categories together.

When applying maximum likelihood (ML) methods to ordered categorical data it is common to use some scoring system for the ordinal variables. Then the ML estimates are dependent of the choice of scores. Often more than one scoring system is plausible. Rank scores is always a possibility, but for example, when a variable is a classified quantitative variable other possibilities such as class midpoints may be used. One characteristic of ordered categorical data is its invariance under monotone-ordered transformations. Thus, the statistical methods should be unaffected by any kind of ordered re-labelling of scale categories [1, 11]. The interpretation of cumulative odds models is not dependent of assigning scores to the response categories, but these models may be motivated by the assumption that a continuous variable is underlying the response categories [7]. The models imply stochastic ordering and constant odds ratios which may be relevant in many applications, but not in cases where, for example, there is a difference in concentration in response categories. The cumulative odds models are invariant to how to choose response categories, grouping, merging, and splitting categories, and the same parameters with reference to the latent continuous response variable apply [7, 16]. Adjacent-categories models, and related log linear models, are not invariant to the number and choice of categories. In general, the continuation-ratio model are not invariant to merging categories [16]. Tutz in the discussion of the paper by Liu and Agresti [6] refers to relative merits of sequential models (e.g. continuation ratio logit model) over cumulative models regarding the possibility of allowing for category-specific effects $\boldsymbol{\beta}_j$. The cumulative model has restrictions that imply severe restriction on the parameters and for which values of x the model can hold. There are also numerical problems in fitting cumulative type models with category-specific effects. The drawbacks of the cumulative models in marginal or mixed models become more severe. But as Liu and Agresti [6] says in the reply, a reason that the sequential type of models are not so common in use is that results are not invariant to the choice of ascending or descending ordering of the response.

Log linear models sometimes have connections to some asymmetric GLMs and share the drawbacks of those. Some log linear models may be formulated as logit models, for example,

association plus linear-by-linear model and other models with quasi-symmetric structure useful for agreement studies. When log linear models are formulated as logit models the design matrix is specified by scores or the rows or columns are parameterized by restrictions for the ordering of the categories (row effect model or column effect model). In logit models it is not crucial to score the categories of the response, but by the log linear formulation we have the drawback to find and determine scores for the categories. It is often difficult to justify a particular choice of scores if the response is an ordinal categorical variable [55]. Whether the parameterization is to be regarded as a possibility or a drawback may vary from one situation to another.

Models that originate from a fully parameterized quasi-symmetry model [54] are good starting points for analyzing square contingency tables with categories in the rows exactly corresponding to the categories in the columns. Models for agreement are often based on log linear models having a quasi-symmetric structure [55]. It is difficult to find one single model which gets hold of both systematic and individual patterns as well as association. But by quasi-symmetric models it is possible to test for marginal homogeneity [3, 54], and by comparison of models which are hierarchical ordered it is also possible to test for marginal homogeneity. Parameter estimates of differences of marginal frequencies are not obtained by log linear models, but by using conditional modelling (not CML) estimates are obtainable. If the sole purpose is to estimate parameters for marginal frequencies marginal models may be used. It is easily seen that it is problematic to specify and find the best model as well as to estimate and interpret all possible models. This is most often avoided by using Svensson's approach.

In the application of log linear models for agreement in reliability studies it is of great relevance to estimate the degree of distinguishability of categories. Recently, an extension to the log linear models for agreement, i.e. association plus agreement models which satisfy the quasi-symmetry property has been proposed [61]. These parameterize different variations of distinguishabilities between adjacent categories. The models become more and more complicated to parameterize and interpret, e.g. odds ratios and degree of distinguishability, and the models and parameters become numerous which may make the user confused. To compare models they have to be hierarchical for the LR-test and one has to be observant of the multiple test situations for p-values. To further expand the models it is possible to estimate the degree of distinguishability by the log-multiplicative RC models which are not easily

estimated. Log-multiplicative models may also be used for the simultaneous modelling of pairwise agreement on individuals and the overall agreement in using the scale [62].

Trying to capture the many aspects of change, association or agreement by one single measure is problematic. Measures which are not model-based are attractive as they often are easy to understand and applicable irrespective of model. The most used measure of agreement - the kappa measure - has many limitations, such as its dependency of the marginal frequencies, it does not exploit the order structure, its sensitivity of the prevalence of the phenomenon under study, it elucidate only perfect agreement and it indicates nothing about bias. The extension of the ordinary kappa measure to be applicable to ordered category by weights induces the problem of choosing the weights. The many different weighting possibilities may be looked upon as a drawback. Different weights result in different values of the kappa measure. Many articles have pointed on the shortcomings of the kappa measures [3, 13, 47, 49, 53, 62-64].

Measures of concordance and association have limitation as measures of agreement. Association is not always the same as agreement and the different ways of adjusting for tied observations affect the possibility of attaining the limiting values of the measures. The measures are also affected of the frequency distributions [12].

The measures from Svensson´s approach are not based on specific assumptions. They are specifically designed for paired data and exploit the ordered structure in ordinal categorical data and thus rank-invariant. The measures may be used in different designs. Due to the augmented bivariate ranking approach it is possible to evaluate both systematic and individual change or disagreement. The measures are easy to use and interpret. In contrast to the traditional measures of concordance, in the novel measure based on Svensson´s approach the limits are attainable irrespective of the number of possible response categories and of the type of scaling and the category distributions [12, 57].

There are however some limitations. The Svensson´s approach applies so far only to paired designs, or in multiple designs, to pairwise comparisons. Methods for simultaneous comparisons of more than two repeated observations are to be developed. All the necessary statistical properties for inference are not yet completely known. Using Jackknife estimates of

standard errors and rely on asymptotic normality makes Wald type of inference as a pragmatic option as well as bootstrap techniques for confidence interval estimation.

Acknowledgment

The work was supported by grants from Centre for Research and Development Uppsala University and County Council of Gävleborg.

Table 1. Summary of fundamental assumptions and implications of methods reviewed in this paper.

METHOD APPROACH	ASSUMPTIONS AND ISSUES	IMPLICATIONS	ESTIMATION
Baseline category model	Nominal. Canonical link. Stochastic ordering[5]. Individual response categories[5].	Location and scale parameters[5]. Permutaion invariant.	ML. Available software.
Cumulative model	No scores necessary [6, 7]. Not canonical link. Stochastic ordering, proportional odds [5, 7]. Underlying continuous response categories [5]. Monotone cut-points for positive probabilities [7].	Location and scale parameters by reformulation, shift in varaiability not taken care of as is[5, 7]. Palindromic invariant for common link functions [7]. Dependent OR [9]. Possible to redefine response categories, merging or splitting[6, 7].	ML. No computational difficulties[7]. Patterns of zeros in sparse tables causes problems [7]. LR, Score and Wald tests. Good knowledge and available software.
Adjacent categories model	No scores necessary. Canonical link. Stochastic ordering[5]. Individual response categories[5].	Location and scale parameters[5]. Connections to certain ordinal log linear models via scores [6]. Not invariant to choice of and number of response categories [7]. Not possible to merging or splitting of response categories.	ML. Available software.
Continuation ratio model	No scores necessary. Canonical link. Stochastic ordering in many cases[7]. Sequential mechanism [16]. Proportional odds. Extend possibilities of modelling vs cumulative models[6].	Location and scale parameters[5]. Groupings of categories [5]. Not palindromic invariant [9, 65]. Independent OR [9]. Not possible to merging or splitting of response categories. Useful for data which cannot sensible be grouped[7].	ML. Available software.
Marginal models, GEE	See different GLM models [66]. Exponential family. Working guess of covariance structure[25]. Missing observations are missing completely at random [25].	Consistent estimators of β and Var(β)[25]. Population specific.	ML, conditional, WLS and quasi likelihood. Computational difficulties with marginal models but not with GEE. Model and identifiability constraints for marginal models with ML. Full likelihood function not specified. LR and goodness of fit tests for marginal models with ML, but Wald tests only for GEE. Available software.
Conditional cluster specific models	See different GLM models[66]. Linear logit[66]. No distributional assumptions about the random subject effects[5, 66].	Consistent estimates [66]. Limited to models with canonical links and to within cluster effects [5]. Subject specific.	Conditional ML. Available software.

METHOD APPROACH	ASSUMPTIONS AND ISSUES	IMPLICATIONS	ESTIMATION
(Random effect models) GLMM	See different GLM models [66]. Given the random effects, GLMM is a GLM. Some distribution, often normal, for the random effect[5].	Subject specific. Relatively flexible description of the nature of association[32].	Numerical integration and ML or approximate ML. Computational difficulties. Estimation complexities. LR. Problems with tests of $\sigma^2 = 0$. Available software.
Log linear association models	Choice of scores[40]. Poisson, multinomial. Scores, but not for RC models. Quasi-symmetry for agreement [53].	Focus on associations and interactions in joint distribution. Exact fit on the main diagonal undesirable [54]. Simple interpretations through OR. Not invariant to choice and number of categories [67]. Both permutation invariant (nominal) and palindromic invariant (ordinal)[7, 52]. Not permutation invariant in some applications[7]. Not possible to merging or splitting response categories [7]. Flexible description of the nature of association[3, 40]. Parameter interpretation plus residual description [40]. Investigate the structure of agreement in the data[68]. Different models for agreement and change[54].	Iterative ML. Goodness of fit tests of nested sequence of models [49]. LR, score tests [40]. In sparse tables, LR not χ^2 [40, 54]. Good knowledge and available software.
Kappa	Unbiased raters, dependences on frequencies, number of categories, marginal distributions. Nominal. No scores. For perfect agreement: marginal homogeneity[46].	Not detecting bias. Ignores the degree of disagreement[49]. Not possible to merging or splitting response categories. OR parameters not easy to get [3]. If strong assumptions it describes both the pattern and strength of agreement[3]. Mixes together bias and rank-order differences [49]. Not flexible; single number [49].	Good knowledge and available software.
Weighted kappa	Ordinal. Scores [49]. Setting of the weights[48]. Assign numerical values[49].	Different weighting system for flexible description of the nature of association [48].	
Measures of concordance association: Kendall's Tau-b, Tau-c, Gamma, Delta	Ordinal. No scores. No specific assumption [41].	Tau-b ordinal invariant[41]. Ways of adjusting for ties affects attaining limits [12]. Affected by merging or splitting response categories. In special cases OR parameters[14].	Good knowledge and available software.
Spearmans roh	Ordinal. No scores. No specific assumptions [41].	Ordinal invariant[41]. Limitation in agreement studies; $r_s < 1$ if heterogeneity [10].	Good knowledge and available software.

METHOD APPROACH	ASSUMPTIONS AND ISSUES	IMPLICATIONS	ESTIMATION
Svensson´s approach: RP, RC, RV, ra, MA	Ordinal. No scores. No specific assumptions [9].	Rank invariant [9]. Affected by merging or splitting response categories. If $r_a = 0$ one classification is reversed relative the other [10]. RV is a measure of disagreement when marginal heterogeneity is taken into account. r_a is an agreement measure and also a correlation [10]. MA agreement measure and also a correlation and concordance[10]. Flexible description of the nature of association[10]. Easy to interpret.	No computational difficulties. Allows for small data sets and zero cells[6]. Not so well known. Limited software.

References

1. Stevens, S.S., *On the theory of scales of measurements.* Science, 1946. **103**: p. 677-680.
2. Agresti, A., *A survey of models for repeated ordered categorical response data.* Statistics in Medicine, 1989. **8**(10): p. 1209-1224.
3. Agresti, A., *Modelling patterns of agreement and disagreement.* Statistical Methods in Medical Research, 1992. **1**(2): p. 201-218.
4. Agresti, A., *Modelling ordered categorical data: recent advances and future challenges.* Statistics in Medicine, 1999. **18**: p. 2191-2207.
5. Agresti, A. and R. Natarajan, *Modeling clustered ordered categorical data: A survey.* International Statistical Review, 2001. **69**(3): p. 345-371.
6. Liu, I. and A. Agresti, *The Analysis of Ordered Categorical Data: An Overview and a Survey of Recent Developments.* Sociedad de Estadistica e Investigacion Operativa. Test, 2005. **14**(1): p. 1-73.
7. McCullagh, P., *Regression Models for Ordinal Data.* Journal of the Royal Statistical Society, Series B, 1980. **42**(2): p. 109-142.
8. Sutradhar, B.C., *An overwiev on regression models for discrete longitudinal responses.* Statistical Science, 2003. **18**(3): p. 377-393.
9. Svensson, E., *Analysis of systematic and random differences between paired ordinal categorical data*. 1993, Stockholm: Almqvist & Wiksell International.
10. Svensson, E., *A Coefficient of Agreement Adjusted for Bias in Paired Ordered Categorical Data.* Biometrical Journal, 1997. **39**(6): p. 643-657.
11. Svensson, E., *Application of a rank-invariant method to evaluate reliability of ordered categorical assessments.* Journal of Epidemiology and Biostatistics, 1998. **3**(4): p. 403-409.
12. Svensson, E., *Concordance between ratings using different scales for the same variable.* Statistics in Medicine, 2000. **19**(24): p. 3483-3496.
13. Svensson, E. and S. Holm, *Separation of systematic and random differences in ordinal rating scales.* Statistics in Medicine, 1994. **13**(23-24): p. 2437-2453.
14. Agresti, A., *Analysis of ordinal categorical data.* 1984, New York: John Wiley & Sons.
15. Manor, O., S. Matthews, and C. Power, *Dichotomous or categorical response? Analysing self-rated health and lifetime social class.* International Journal of Epidemiology, 2000. **29**: p. 149-157.
16. Tutz, G., *Sequential models in categorical regression.* Computational statistics and data analysis, 1991. **11**: p. 275-295.
17. Aitchison, J. and S.D. Silvey, *Maximum-Likelihood estimation of parameters subject to restraints.* The Annals of Mathematical Statistics, 1958. **29**(3): p. 813-828.
18. Haber, M., *Maximum likelihood methods for linear and log-linear models in categorical data.* Computational statistics and data analysis, 1985. **3**: p. 1-10.
19. Lang, J.B., *Maximum likelihood methods for a generalized class of log-linear models.* The Annals of Statistics, 1996. **24**(2): p. 726-752.
20. Lang, J.B. and A. Agresti, *Simultaneously modeling joint and marginal distributions of multivariate categorical responses.* Journal of the american statistical association, 1994. **89**(426): p. 625-632.

21. Fitzmaurice, G.M. and N.M. Laird, *A likelihood-based method for analysing longitudinal binary responses.* Biometrika 1993. **80**(1): p. 141-151.
22. Glonek, G.F.V., *A class of regression models for multivariate categorical responses.* Biometrika, 1996. **83**(1): p. 15-28.
23. Glonek, G.F.V. and P. McCullagh, *Multivariate logistic models.* Journal of the royal statistical society. Series B (Methodological), 1995. **57**(3): p. 533-546.
24. Wedderburn, R.W.M., *Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method.* Biometrika, 1974. **61**(3): p. 439-447.
25. Liang, K.-Y. and S.L. Zeger, *Longitudinal Data Analysis Using Generalized Linear Models.* Biometrika, 1986. **73**(1): p. 13-22.
26. Grizzle, J.E., C.F. Starmer, and G.G. Koch, *Analysis of Categorical Data by Linear Models.* Biometrics, 1969. **25**(3): p. 489-504.
27. Koch, G.G., et al., *A General Methodology for the Analysis of Experiments with Repeated Measurement of Categorical Data.* Biometrics, 1977. **33**(1): p. 133-158.
28. Agresti, A., *Categorical Data Analysis*. 2 ed. 2002, New York: Wiley and Sons.
29. Tuerlinckx, F., et al., *Statistical inference in generalized linear mixed models: A review.* British Journal of Mathematical and Statistical Psychology, 2006. **59**: p. 225-255.
30. Crowder, M.J., *Beta-Binomial Anova for Proportions.* Applied Statistics, 1978. **27**(1): p. 34-37.
31. Lawless, J.F., *Negative Binomial and Mixed Poisson Regression.* The Canadian Journal of Statistics, 1987. **15**(3): p. 209-225.
32. Agresti, A., et al., *Random-Effects Modeling of Categorical Response Data.* Sociological Methodology, 2000. **30**: p. 27-80.
33. McCulloch, C.E., *Maximum Likelihood Algorithms for Generalized Linear Mixed Models.* Journal of the american statistical association, 1997. **92**(437): p. 162-170.
34. Raudenbush, S.W., M.-L. Yang, and M. Yosef, *Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation.* Journal of Computational and Graphical Statistics, 2000. **9**(1): p. 141-157.
35. Ten Have, T.R. and A.R. Localio, *Empirical Bayes Estimation of Random Effects Parameters in Mixed Effects Logistic Regression Models.* Biometrics, 1999. **55**: p. 1022-1029.
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**(367): p. 537-552.
37. Goodman, L.A., *The analysis of cross-classified data having ordered and/or unordered catecories:association models, correlation models, and asymmetry models for contingency tables with or without missing entries.* The Annals of Statistics, 1985. **13**(1): p. 10-69.
38. Haberman, S.J., *Log-Linear Models for Frequency Tables with Ordered Classifications.* Biometrics, 1974. **30**(4): p. 589-600.
39. Haberman, S.J., *Test for Independence in Two-way Contingency Tables Based on Canonical Correlation and on Linear-By-Linear Interaction.* The Annals of Statistics, 1981. **9**(6): p. 1178-1186.
40. Agresti, A., *A model for agreement between ratings on an ordinal scale.* Biometrics, 1988. **44**(2): p. 539-548.
41. Kruskal, W., H., *Ordinal Measures of Association.* Journal of the American Statistical Association, 1958. **53**(284): p. 814-861.

42. Daniels, H.E., *The relation between measures of correlation in the univers of sample permutations.* Biometrika, 1944. **33**(2): p. 129-135.
43. Hoeffding, W., *A class of statistics with asymptotically normal distribution.* The Annals of Mathematical Statistics, 1948. **19**(3): p. 293-325.
44. Landis, J.R. and G.G. Koch, *A review of statistical methods in the analysis of data arising from observer reliability studies (Part I).* Statistica Neerlandica, 1975. **29**: p. 101-123.
45. Altman, D.G., *Practical Statistics for Medical Research*. 1991, London: Chapman & Hall.
46. Cohen, J., *A Coefficient of Agreement for Nominal Scales.* Educational and Psychological Measurement, 1960. **20**(1): p. 37-46.
47. Nelson, J.C. and M.S. Pepe, *Statistical description of interrater variability in ordinal ratings.* Statistical Methods in Medical Research, 2000. **9**: p. 475-496.
48. Cohen, J., *Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit.* Psychological Bulletin, 1968. **70**(4): p. 213-220.
49. Banerjee, M., et al., *Beyond Kappa: A Review of Interrater Agreement Measures.* The Canadian Journal of Statistics, 1999. **27**(1): p. 3-23.
50. 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**: p. 613-619.
51. Goodman, L.A., *Multiplicative models for square contingency tables with ordered categories.* Biometrika, 1979. **66**(3): p. 413-418.
52. McCullagh, P., *A Class of Parametric Models for the Analysis of Square Contingency Tables with Ordered Categories.* Biometrika, 1978. **65**(2): p. 413-418.
53. Darroch, J.N. and P.I. McCloud, *Category Distinguishability and Observer Agreement.* Australian Journal of Statistics, 1986. **28**(3): p. 371-388.
54. Becker, M.P., *Quasisymmetric models for the analysis of square contingency tables.* Journal of the Royal Statistical Society, Series B (Methodological), 1990. **52**(2): p. 369-378.
55. Schuster, C. and A. von Eye, *Model for Ordinal Agreement Data.* Biometrical Journal, 2001. **43**(7): p. 795-808.
56. Altham, P.M.E., *A non-parametric measure of discriminability.* British Journal of Mathematical and Statistical Psychology, 1973. **26**: p. 1-12.
57. Svensson, E., *Comparison of the Quality of Assessments Using Continuous and Discrete Ordinal Rating Scales.* Biometrical Journal, 2000. **42**(4): p. 417-434.
58. Svensson, E., *Ordinal invariant measures for individual and group changes in ordered categorical data.* Statistics in Medicine, 1998. **17**(24): p. 2923-2936.
59. Agresti, A. and D.B. Hitchcock, *Bayesian inference for categorical data analysis.* Statistical Methods and Application, 2005. **14**: p. 297-330.
60. Lee, Y. and J.A. Nelder, *Conditional and Marginal Models: Another View.* Statistical Science, 2004. **19**(2): p. 219-238.
61. Valet, F., C. Guinot, and J.Y. Mary, *Log-Linear Non-Uniform Association Models for Agreement between two Ratings on an Ordinal Scale.* Statistics in Medicine, 2007. **26**(3): p. 647-662.
62. Perkins, S.M. and M.P. Becker, *Assessing Rater Agreement using Marginal Association Models.* Statistics in Medicine, 2002. **21**: p. 1743-1760.
63. Agresti, A., A. Ghosh, and M. Bini, *Raking Kappa: Describing Potential Impact of Marginal Distributions on Measures of Agreement.* Biometrical Journal, 1995. **37**(7): p. 811-820.

64. Brennan, P. and A. Silman, *Statistical methods for assessing observer variability in clinical measures.* British Medical Journal, BMJ, 1992. **304**: p. 1491-1494.
65. Kampen, J. and M. Swyngedouw, *The Ordinal Controversy Revisited.* Quality and Quantity, 2000. **34**: p. 87-102.
66. Conaway, M.R., *Analysis of Repeated Categorical Measurements with Conditional Likelihood Methods.* Journal of the American Statistical Association, 1989. **84**(405): p. 53-62.
67. Agresti, A., *A Survey of Strategies for Modeling Cross-Classifications Having Ordinal Variables.* Journal of the american statistical association, 1983. **78**(381): p. 184-198.
68. Tanner, M.A. and M.A. Young, *Modeling Agreement Among Raters.* Journal of the American Statistical Association, 1985. **80**(389): p. 175-180.