WORKING PAPER SERIES

WORKING PAPER NO 6, 2008

Swedish Business School at Örebro

Comparison of methods in the analysis of dependent ordered catagorical data

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

Comparison of methods in the analysis of dependent ordered catagorical data

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

Rating scales for outcome variables produce categorical data which are often ordered and measurements from rating scales are not standardized. The purpose of this study is to apply commonly used and novel methods for paired ordered categorical data to two data sets with different properties and to compare the results and the conditions for use of these models.
The two applications consist of a data set of inter-rater reliability and a data set from a follow-up evaluation of patients. Standard measures of agreement and measures of association are used. Various loglinear models for paired categorical data using properties of quasi-independence and quasi-symmetry as well as logit models with a marginal modelling approach are used. A nonparametric method for ranking and analyzing paired ordered categorical data is also used.
We show that a deeper insight when it comes to disagreement and change patterns may be reached using the nonparametric method and illustrate some problems with standard measures as well as parametric loglinear and logit models. In addition, the merits of the nonparametric method are illustrated.

JEL classification: C14
Keywords: Agreement, ordinal data, ranking, reliability, rating scales

Introduction

Outcome variables such as pain, mood, functioning, quality of life, quality of received care, ability, etc. are common in clinical research and in health evaluation studies. These outcomes are often assessed using a rating scale. The data set consists therefore of categories and often these categories are ordered. Furthermore, subjective assessments based on scales and judgments are qualitative and the measurements are not standardized. There are many types of rating scales such as a verbal descriptive scale, a Likert-scale, and a visual analogue scale among others.

The choice of a rating scale is a part of the operationalization process. A crucial point is then the quality of the rating scale chosen, in terms of validity and reliability related to the specific study. The criteria of inter- and intra-rater agreement in reliability studies are often used. Research in medical and health science often concerns assessment of change after some intervention by a rating scale that fulfils the required quality. In the analyses of reliability and change, the dependencies in data must be considered.

The purpose of this study was to apply commonly used methods and a nonparametric approach for the analysis of paired ordered categorical data proposed 1993 [1] to two different data sets and to compare and interpret the results and their conditions for use. The first data set concerned inter-rater reliability while the second data set concerned change in patients' social outcome between two occasions. Inter-rater reliability was assessed by means of agreement. Sometimes the same measures and models could be used in both types of study purposes, but specialized methods for analysis of agreement have been developed, and other tests were only meaningful for analysis of change.

Data

The data set of inter-rater reliability concerns agreement in judging biopsy slides for carcinoma of the uterine cervix. The purpose was to study variability in classification and the degree of agreement in ratings among pathologists [2]. The data were originally used by Holmqvist [2] and then used by Landis and Koch in 1977 [3] for presenting methods for modelling agreement among more than two observers. The data has later been used frequently as illustrations in methodological papers and textbooks [4-7]. Originally seven pathologists classified each of 118 biopsy slides based on the most involved lesion in terms of carcinoma in situ of the uterine cervix. The ordered categories were 1) negative, 2) atypical squamous hyperplasia, 3) carcinoma in situ, 4) squamous carcinoma with early stromal invasion, and 5) invasive carcinoma. In this example we used only two of the pathologists' ratings, see table 1.

Table 1. Cross-classification of two pathologists' ratings of 118 biopsy slides labeled X and Y[4].

Pathologist Y	*Pathologist X*					*Total frequency*	*Cumulative frequency*
	1	*2*	*3*	*4*	*5*		
1	22	2	2			26	26
2	5	7	14			26	52
3		2	36			38	90
4		1	14	7		22	112
5			3		3	6	118
Total frequency	27	12	69	7	3	118	
Cumulative frequency	27	39	108	115	118		

The other data set stems from a study of individual and group changes in patients' social outcome after aneurysmal subarachnoid haemorrhage between two occasions [8]. Sixty-three patients who were operated on, in acute stage, for ruptured cerebral arterial

aneurysm were recruited for a neurological, neuropsychiatric and neuropsychological follow-up evaluation. Global social outcome was one of the study variables, operationalised by the Swedish version of the eight point form of the Glasgow Outcome Scale (S-GOS). The eight ordered categories were 1) dead, 2) vegetative state, 3) severe disability: low, 4) severe disability: high, 5) moderate disability: low, 6) moderate disability: high, 7) good recovery: low, and 8) good recovery: high. The result of the change in social outcome is shown in table 2.

Table 2. Cross-classification of levels in social outcome assessed by the Swedish version of Glasgow Outcome Scale (S-GOS) for 63 patients at discharge and after 3 months follow-up.

After 3	*At discharge*								
months	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Total*
1			3						3
2									0
3			4						4
4			2	1					3
5			4	2	13				19
6			3	2	9	1	1		16
7			2	3	6	3	2		16
8						1		1	2
Total	0	0	18	8	28	5	3	1	63

A common feature in the two studies was the paired observations. The same patient was evaluated independently twice. In the inter-rater reliability study, two raters independently judged each patient's biopsy slide and in the social outcome study, each patient was assessed regarding social outcome at discharge from hospital and three months after discharge. Although independently evaluated, the data for the patients are dependent, as each patient is assessed twice.

Method

In this study we used several summary measures, models, and tests for comparisons. Bivariate agreement and change may be explored by tabular and graphical means. A plot in which the cumulative relative frequencies were plotted against each other was used. This plot was called a Relative Operating Characteristic (ROC) curve of the systematic inter-rater differences[1, 9]. This application differs from applications in diagnostic test procedures where the more usual definition of ROC (Receiver Operating Characteristic) curve is used. The way the ROC curves are used here was also proposed by P. M. E. Altham in the discussion of a paper by McCullagh[10].

To better understand disagreement or change and their sources, one may choose to investigate the joint distribution. We have used the measure of percentage agreement. The percentage agreement may be extended to handle restrictions due to marginal heterogeneity. The ratio of the percentage agreement and the maximal attainable percentage agreement, given the distributions of the m categories, is called the agreement ratio. In the 1960s measures of agreement were developed further. One of the most commonly used is Cohen's kappa measure [11]. The kappa measure quantifies agreement beyond what is expected by chance under the hypothesis of independence. Kappa max was defined by Cohen [11] in order to determine the maximum value of kappa, given the marginal frequencies. The kappa measure has been extended to treat nominal variables with more than two categories and to treat ordinal variables. If observations are classified in more than two categories, the possibility for disagreement increases. A weighted kappa measure has been proposed [12] with different types of weights.

Other measures used in agreement studies are measures of concordance and association for ordered categorical variables, such as Goodman-Kruskal's gamma, Kendall's tau-b and Spearman's rank-order correlation[13, 14]. Measures of concordance may be used in assessment of order consistency. These summary measures were also used in the comparison. In general, measures of association are only adequate as measures of agreement if the marginal distributions are similar.

Although well established in applied statistics, loglinear and logit models for more than two categories are not commonly used in research in medicine and social science [15]. Loglinear models treat the variables symmetrically, i.e. they 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 the joint distribution of the variables. Loglinear models can be modified to reflect situations in which there are square tables where the categories in the rows exactly correspond to the categories in the columns, and within this framework we may also model dependent data. The models may be further extended to model categorical variables with an ordered structure [4, 16-20].

Some level of disagreement is common when using subjective scales. Models focus then on describing strength of agreement and detecting patterns of disagreement. Based on loglinear models, agreement was assessed by symmetry, quasi-symmetry and marginal homogeneity parameterizations beginning in the mid 1960s and continuing into the early 1990s. Darroch and McCloud [21] 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. Since logit and loglinear estimates may be greatly influenced by sparse tables with many zero cells, sometimes ML estimates do not even exist, and the asymptotic approximations of the chi-square statistics may be problematic, we have added a small constant (0.0005) to each zero cell not smoothing too much in the social outcome study[5, 22].

Several approaches to testing the equivalence of the marginal distributions are common. Some nonparametric tests apply to comparisons between independent groups, e.g. the Wilcoxon-Mann-Whitney test and Kruskal-Wallis test, and others apply to the evaluation of change in paired studies, such as the sign test and McNemar's test. In the case of quantitative data, Wilcoxon's signed-rank test would also be an option.

Svensson [1] and Svensson and Holm [9] studied various aspects of analysis of dependent ordered categorical data and developed new measures of agreement. Svensson[23] showed that these methods also could be used for studies of change in outcomes between two occasions. By means of an augmented ranking approach, an observed disagreement in paired ordinal data was separated and measured in terms of systematic and occasional disagreement. This constituted a foundation for a nonparametric method. Two types of systematic disagreement were identified and measured. When there is a systematic shift to higher categories by the second (Y) rater (occasion) compared with the first (X), or the reverse, we have a case of stochastic ordering, and the parameter of systematic disagreement in position was defined by Svensson [1, 9] as

$$\gamma = P(X < Y) - P(Y < X),$$

The empirical measure of relative position (RP) estimates the probabilities by the corresponding relative frequencies [1, 9]. In the case of a systematic shift in concentration of the repeated classifications to central categories in one occasion (X) compared with another (Y), the parameter of systematic disagreement in concentration, defined as

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

for any independent random variables X_l, Y_l , X_k, Y_k , X_m, and Y_m will be estimated by the empirical measure of relative concentration, RC. In this case the ROC curve is S-shaped along the diagonal line [1, 9].

The level of occasional or individual disagreement is related to the pattern of total monotonic joint ranking given the observed marginal distributions. This pattern was called the rank transformable pattern of agreement (RTPA) and is the expected pattern of paired classification when there is a total agreement in the ordering of all pairs of assessments. A measure of dispersion of observations from the best possible agreement in ordering, given the marginal distribution, was called the relative rank variance (RV). A measure of the closeness of observations to the best possible agreement in ordering when the marginal heterogeneity is taken into account was the augmented rank-order agreement coefficient r_a [1, 9, 13].

Some of the results have been presented elsewhere, e.g. some results from the reliability application were presented in [7] and [4, 5] and some of the results from the change in

social outcome application were presented in [8]. The summary measures, tests, and models used in this paper have been calculated and estimated by SAS software, version 9 of the SAS system for Windows (SAS Institute Inc., Cary, NC., USA); Stata Statistical Software, Release 8.0 (College Station, TX: Stata Corporation); and SPSS software version 14 for Windows (SPSS Inc. Chicago, Ill, USA).

Results

The inter-rater reliability study

Disagreement may be caused by systematic and by random events. Systematic variations, which reflect bias, are revealed by the marginal frequencies. Looking at the marginal frequencies in table 1, pathologist Y used the entire range of the scale evenly, but pathologist X had strong preference to the mid value.

For the pathologists, the kappa measure of agreement was about 0.50 and the weighted kappa with quadratic weight was 0.78. The proportion of perfectly agreeing pairs, i.e. the percentage agreement, was 64 % (table 3), and the agreement ratio 90%. The maximum achievable kappa given the marginals was 0.63.

The Goodman-Kruskal's gamma, Kendall's tau-b and Spearman's rank-order correlation indicated relatively high relationship and gamma had a higher value as it measures the excess of concordant pairs over discordant pairs of all such pairs, not including tied pairs. Due to the many tied observations and a number of observations off the main diagonal, Kendall's tau-b was lower.

Table 3. Traditional agreement and association measures applied to the data set in Table 1.

Measure	*Value*
Percentage agreement (PA)	63.6%
Kappa	0.498
Weighted kappa	0.779
Kappa max	0.627
Agreement ratio (PA/PAmax)	(63.6/70.3) = 90.4%
Goodman-Kruskal gamma	0.923
Spearman rank-order correlation	0.781
Kendall tau-b	0.715

The model of marginal homogeneity showed signs of poor fit to the data when modelled by means of loglinear models (table 4). So did the model of symmetry. The model of quasi-symmetry showed good fit, but it is a model which treats the variable as nominal. Of the models for ordinal classification, it was the quasi-uniform association model and agreement plus uniform association model which had the best fit. The latter model is easier to interpret. The results of some models in table 4 have previously been reported by Agresti [4]

For the model of agreement plus uniform association [4], the estimated parameters together with their estimated asymptotic standard errors gave evidence of extra association beyond that due to exact agreement (β) and extra agreement beyond that due to the baseline association between ratings (δ). The degree of distinguishability of categories *i* and *i+1* was described by $\log \tau_{i,i+1} = \beta + 2\delta$, estimated to 1.15+2*1.067=3.284.

Table 4. Summary of likelihood-ratio chi-squared statistic (goodness-of-fit tests) for different models applied to the data set in Table 1.

Model		G^2*-statistic*	*Degrees of freedom*
N*	Independence	131.2	16
N	Quasi-independence	13.6	11
N	Symmetry	39.2	10
N	Quasi-symmetry	1.0	6
O*	Ordinal quasi-symmetry	37.8	9
N	Marginal homogeneity	38.2	4
O	Conditional symmetry	38.0	5
O	Uniform association	16.2	15
O	Quasi-uniform association	1.3	10
O	Agreement plus uniform association	8.4	14

* N=Model for nominal data. O=Model for ordinal data

Results from some additional tests of marginal homogeneity and symmetry applied to the data set of agreement shown in table 1 are shown in table 5. The tests of marginal homogeneity showed diverging results. The test of marginal homogeneity (1) did not reject the null hypothesis of marginal homogeneity but Stuart's test (3) [24] rejected the null hypothesis. Test 1 in table 5 is a cumulative logit marginal model, and imposes stochastic ordering of the marginal frequencies, which was clearly not the case with these data [5]. Stuart's test ignores the ordered structure of categories, which was also the case for the test of marginal homogeneity in table 4. Test 4 [25] resulted in not rejecting the null hypothesis of marginal homogeneity with addition of a linear trend in the scores. Bowker's test for symmetry [26] rejected the null hypothesis of symmetry. This test also ignores the ordering of categories.

Table 5. Summary of various models and tests of marginal homogeneity (MH) and symmetry applied to the data set in Table 1.

Model/Test	*Chi-square*	*Degrees of freedom*	*p-value*
1. Test of MH with marginal model	0.15	1	0.697
2. Bowkers test for symmetry	30.3	6	<0.001
3. Stuarts test for marginal homogeneity	29.1	4	<0.001
4. Linear trend in the log (RR)	1.3	1	0.249

In figure 1 the cumulative relative frequencies are plotted against each other. Connecting the points reveal an S-shaped curve around the diagonal line, indicating a difference in how the raters concentrate the ratings on the scale [7].

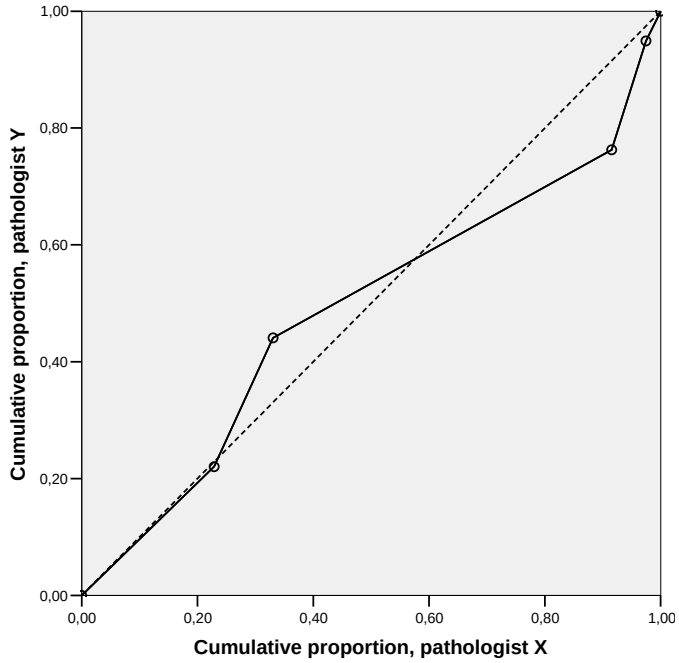

Figure 1. ROC curve for the systematic disagreement between pathologists in classification of biopsy slides from Table 1.

The systematic disagreement was a sign of disagreement in concentration (RC), see table 6. The S-shape of the ROC curve in figure 1 illustrates the descriptive conclusion. Significant systematic disagreement in concentration is neither consistent with marginal homogeneity nor with stochastic ordering. The raters evidently had different ideas of the

cut points between the categories in the middle of the scale. The occasional disagreement (RV) was small, and the joint classification was highly correlated with the rank transformable pattern of agreement, RTPA, $r_a = 0.984$.

Table 6. Measures of systematic and random disagreement for the data set in table 1.

Measure		
Systematic disagreement		
in position	RP	0.028 (SE=0.037)
in concentration	RC	-0.127 (SE=0.054)
Random disagreement		
relative rank variance	RV	0.015 (SE=0.011)
augmented rank-order agreement	ra	0.984

The social outcome study

The individual and group changes in the social outcome for the 63 patients between discharge and the three months follow-up is shown in table 2. Eight ordered categories cross classified for 63 patients necessarily resulted in some zero cells. Unchanged outcome was seen in 22 of the 63 patients. The most obvious reason to difference in the marginal distributions was the higher frequencies in the higher range of the scale after 3 months; see the ROC-curve, figure 2. At group level (marginal level) there was a sign of higher social outcome three months after discharge.

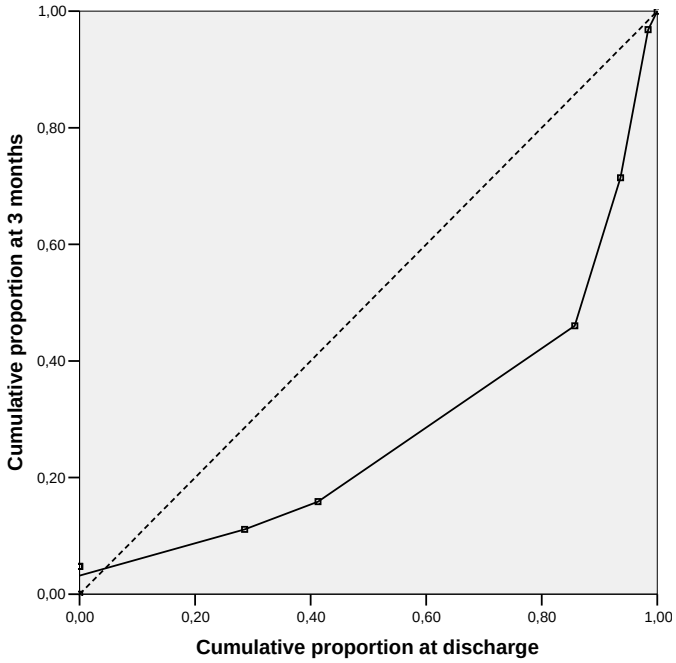

Figure 2. ROC curve for the systematic change in social outcome between discharge and at the three months follow-up from table 2.

The loglinear models for symmetry and marginal homogeneity as well as other tests of these hypotheses were highly significant; see table 7 and table 8. Notably the test of marginal homogeneity with the cumulative logit marginal model also rejected the null hypothesis. The assumption of stochastic ordering was consistent with the data. The models of quasi-independence and quasi-symmetry fitted well. These models assume nominal classification. Of the models for ordinal classification, the quasi-uniform association model and conditional symmetry model also fitted well. A simple applicable nonparametric test of equivalence of the marginal frequencies was the sign test, which was also highly significant (Table 8).

Table 7. Summary of likelihood-ratio chi-square statistic (goodness-of-fit tests) for different models applied to the data set in table 2. A small constant (0.0005) is added to each zero cell to get more stable results.

Model	*G^2 statistic*	*Degrees of freedom*
N* Independence	49.2	49
N Quasi-independence	23.8	41
N Symmetry	52.3	28
N Quasi-symmetry	0.05	21
O* Ordinal quasi-symmetry	24.4	27
N Marginal homogeneity	52.2	7
O Conditional symmetry	21.7	27
O Uniform association	21.4	48

* N=Model for nominal data. O=Model for ordinal data

Table 8. Summary of various models and tests of marginal homogeneity (MH) and symmetry for the data set in table 2.

Model/Test	*Chi-square*	*Degrees of freedom*	*p-value*
Sign test	25.0	1	<0.001
Test of MH with marginal model	21.2	1	<0.001
Bowkers test for symmetry	38.0	12	<0.001
Stuarts test for marginal homogeneity	33.4	6	<0.001
Linear trend in the log (RR)	22.3	1	<0.001

According to the nonparametric method by Svensson [1, 9] the systematic part of change, i.e. the change on group level, was dominating, see table 9. This change consisted of a statistical significant change to the better on the S-GOS scale for the group as a whole. Table 2 shows marked individual changes to the better. The RV was equal to 0.16, which means that besides the significant improvement for the group, the patients were heterogeneous in recovery according to the significantly non-zero RV value. The correlation of data to the pattern of total monotonic bivariate ranks, given the marginal heterogeneity, was 0.84. This information regarding recovery on both group and individual levels were not easily seen utilizing other methods.

Table 9. Measures of systematic group changes and individual changes for the data set in table 2.

Measure		
Systematic change for the group		
in position	RP (SE)	0.44 (0.06)
in concentration	RC (SE)	-0.20 (0.14)
Individual changes		
relative rank variance	RV (SE)	0.16 (0.06)
Homogeneity to the group change	ra	0.84

Summary and conclusion

Statistical models are often considered to be more useful than single summary measures in many circumstances. In regard to modeling agreement, Agresti [22] pointed out that model-based approaches yield additional and more precise information than that provided by summary measures. However, the risk for misspecification increases due to the restrictions put on the data. Models may also be used for tests, and within the framework of the likelihood principle, several different tests are available. Furthermore, models may also be used for estimation of different aspects of the data and for example to predict probabilities. The models presented here were all parametric, which means that they were dependent on distributional assumptions for inferences. Intrinsic patterns in a cross classification may be modeled with several different models for different aspects of them or with one, more complex, model. But then the great advantage of parsimonious models is lost. The required knowledge of the interpretations of the different models and how to parameterize the different aspects of agreement/disagreement may hamper the use of the models.

In table 4 and table 7 we report comparisons of some loglinear model by means of the likelihood-ratio chi-squared statistic G^2. The G^2 statistic may be poor in approximating the chi-squared distribution for testing goodness-of-fit for a specific model when tables are sparse or contain sample zeros, but it may still be adequate in comparing nested, unsaturated models [4, 5]. The conclusions of loglinear modeling in the inter-rater reliability application were that there was an excess agreement beyond baseline association and extra association beyond that due to perfect agreement. The hypothesis of marginal homogeneity was rejected. Thus, essentially the same conclusions were reached by several models and tests as when calculating the measures proposed by Svensson. In addition, the Svensson method indicated a systematic disagreement in concentration as the cause of marginal heterogeneity in the first application. In the second application, several loglinear models were plausible and the hypothesis of marginal homogeneity was rejected. The Svensson measures showed both systematic and occasional changes.

Logit link models for ordered categorical data imply stochastic ordering. These models are not applicable when the marginal heterogeneity is due to disagreement or change in concentration. Furthermore, disagreement or individual changes may be substantial even if there is unbiased rating or no systematic change.

In the inter-rater reliability application, where there was a difference in concentration of the marginal frequencies, different tests of marginal homogeneity gave different conclusions. The reason was assumption of stochastic ordering, and also, misspecification. The misspecifications were, for example linear scores or not modeling the ordinal structure of data. Thus, different, relatively common tests of the same

hypothesis may result in contradictory results and results dependent upon assumptions and specification.

Originally, kappa was applied for a 2 x 2 table in which the results for two equally skilled raters judging outcome of a variable for a group of individuals were recorded [11]. The kappa statistic therefore does not catch any bias, which occurs when two raters use 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.

The different measures of association which were used are not in general adequate as measures of agreement. High level of association can exist without strong agreement [5, 27]. The measures differ in how they handle ties and the possibility of attaining the limiting values [-1, 1]. Kendall's tau-b requires all observations on the main diagonal in a square table to obtain the limit but Goodman-Kruskal's gamma requires monotonicity.

Svensson's measures and joint frequency tables revealed a deeper insight when it comes to disagreement and change patterns than most other methods did. And even if elaborative modeling by loglinear models may reveal certain patterns, they are sensible of distributional assumptions, which is not the case for Svensson's measures. The loglinear models accounting for the ordinality uses scores. The estimates are thus dependent of the choice of scores. Many of the models state the scoring system, i.e. integer spaced scores. The parameters and derived measures of association are then interpreted in terms of differences in the scores. In contrast, Svensson's method is based on ranks and is distribution free. Based on the augmented ranking approach, the

possibility of separating and quantifying a systematic and an individual part of the disagreement or change expand the interpretation of the results. In clinical research of interventions or treatment effects it is important to identify the types of change, either as a group level effect and/or as an individual effect, for planning the implementation of intervention or treatment.

A drawback of Svensson's approach is that the statistical properties of the measures are not fully known. Jackknife estimates of standard errors and simulation results indicating asymptotic normality of the measures makes inferences of Wald type possible. Ongoing and future studies on asymptotic properties as well as small sample properties will increase the utility of the measures [28]. The measures are to be generalized to more than matched pairs. Implementation of algorithms in computer packages and introduction of the methods in the research society will further increase the utility.

Acknowledgment

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

References

1. Svensson, E., *Analysis of systematic and random differences between paired ordinal categorical data*. 1993, Stockholm: Almqvist & Wiksell International.
2. Holmquist, N.S., C.A. McMahon, and O.D. Williams, *Variability in Classification of Carcinoma in situ of the Uterine Cervix.* Archives of Pathology, 1967. **84**: p. 334-345.
3. Landis, R.J. and G.G. Koch, *An application of hierarchial kappa-type statistic in the assessment of majority agreement among multiple observers.* Biometrics, 1977. **33**(2): p. 363-374.
4. Agresti, A., *A model for agreement between ratings on an ordinal scale.* Biometrics, 1988. **44**(2): p. 539-548.
5. Agresti, A., *Categorical Data Analysis*. 2 ed. 2002, New York: Wiley and Sons.

6. Agresti, A. and R. Natarajan, *Modeling clustered ordered categorical data: A survey.* International Statistical Review, 2001. **69**(3): p. 345-371.
7. 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.
8. Svensson, E. and J.-E. Starmark, *Evaluation of Individual and group changes in social outcome after aneurysmal subarachnoid haemorrage: A long-term follow-up study.* Journal of Rehabilitation Medicine, 2002. **34**: p. 251-259.
9. 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.
10. McCullagh, P., *Regression Models for Ordinal Data.* Journal of the Royal Statistical Society, Series B, 1980. **42**(2): p. 109-142.
11. Cohen, J., *A Coefficient of Agreement for Nominal Scales.* Educational and Psychological Measurement, 1960. **20**(1): p. 37-46.
12. Cohen, J., *Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit.* Psychological Bulletin, 1968. **70**(4): p. 213-220.
13. Svensson, E., *A Coefficient of Agreement Adjusted for Bias in Paired Ordered Categorical Data.* Biometrical Journal, 1997. **39**(6): p. 643-657.
14. Svensson, E., *Concordance between ratings using different scales for the same variable.* Statistics in Medicine, 2000. **19**(24): p. 3483-3496.
15. 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.
16. 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.
17. Goodman, L.A., *Multiplicative models for square contingency tables with ordered categories.* Biometrika, 1979. **66**(3): p. 413-418.
18. 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.
19. 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.
20. Haberman, S.J., *Log-Linear Models for Frequency Tables with Ordered Classifications.* Biometrics, 1974. **30**(4): p. 589-600.
21. Darroch, J.N. and P.I. McCloud, *Category Distinguishability and Observer Agreement.* Australian Journal of Statistics, 1986. **28**(3): p. 371-388.
22. Agresti, A., *Modelling patterns of agreement and disagreement.* Statistical Methods in Medical Research, 1992. **1**(2): p. 201-218.
23. Svensson, E., *Ordinal invariant measures for individual and group changes in ordered categorical data.* Statistics in Medicine, 1998. **17**(24): p. 2923-2936.
24. Stuart, A., *A Test for Homogeneity of the Marginal Distributions in a Two-Way Classification.* Biometrika, 1955. **42**(3/4): p. 412-416.
25. Breslow, N.E. and N.E. Day, *Statistical Methods in Cancer Research.* Vol. 1. 1980, Lyon: International Agency for Research on Cancer.
26. Bowker, A.H., *A Test for Symmetry in Contingency Tables.* Journal of the american statistical association, 1948. **43**(244): p. 572-574.
27. 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.
28. Wahlström, H., *Nonparametric Tests for Comparing Two Treatments by Using Ordinal Data*. Örebro Studies in Statistics. Vol. 2. 2004, Örebro: Örebro University, University Library.