UNIVERSITY OF GÖTEBORG
Department of Statistics

RESEARCH REPORT 1992:4
ISSN 0349-8034

SEPARATION OF SYSTEMATIC AND RANDOM ERRORS IN ORDINAL RATING SCALES

by

Elisabeht Svensson, Sture Holm

Statistiska institutionen
Göteborgs Universitet
Viktoriagatan 13
S 411 25 Göteborg
Sweden

SEPARATION OF SYSTEMATIC AND RANDOM ERRORS IN ORDINAL RATING SCALES

Elisabeth Svensson, Sture Holm

Department of Statistics
University of Göteborg
Viktoriagatan 13
S-411 25 GÖTEBORG
Sweden

SEPARATION OF SYSTEMATIC AND RANDOM ERRORS IN ORDINAL RATING SCALES

Abstract:

The aim of this paper is to introduce a new rank method which enables us to separate the inconsistency of repeated measurements into random and systematic differences and to quantify this lack of consistency in a few measures. The key of the separation approach is to make a particular type of ranking of the repeated judgements in the same experimental unit. It means that cases which have the same classification from one rater will be internally ranked according to the classifications from the other. This enables us to extract the random variation. The variance of the rank differences between the judgements is a suitable measure of the random interrater variability.

The systematic differences are described by empirical measures of relative position and of relative concentration. These measures are normed into the interval [-1,1].

Our method has been applied to several medical rating scales both for construction and analysis. We use one of the data sets as an illustration.

INTRODUCTION

Measuring instruments based on ordered categories i e observer-rating and self-rating scales, create ordinal data. Characteristic features such as that the labels are replaceable, that there are unequal unmeasurable distances between the categories and that there are no standardized rules for the lengths of the categories , imply that the observations are not easily attached to some model structure. This motivates a nonparametric approach to the analysis.

The purpose of an ordered categorical measure is to find a rank order of the objects and to discriminate into distinct levels of a scale. The rater is thus forced to judge objects into discrete categories. When two raters independently classify individuals from the same population into discrete categories, they may agree or disagree. A slight disagreement will be unmeasurable - covered within the discrete categories. A more obvious disagreement, however, will result in judgements into different categories and the disagreement is measurable. This disagreement can include both random and systematic differences. The reason for disagreement might be that the descriptions of the discrete ordered categories do not satisfactory fit all the individuals distinctly or the measuring situation may influence the judgements. Furthermore the raters may have different ideas about the bounds of the categories or they may differ in the interpretation of the descriptions.

In many models for continuous data there exist methods for separating systematic and random errors. The usual methods for analysing ordinal scale data do not include possibilities for separating the variability into such components.

The aim of this paper is to present a nonparametric analysis, where the different types of error can be separated and to introduce simple characterizations of the random and systematic errors.

THE RANKING APPROACH

Consider a situation where n individuals or objects from the same population are independently classified by two observers into one of m ordered categories in order to assess the inter-rater variability.

The probability of rating a randomly chosen individual to the i:th category by judgement 1 and to the j:th category by judgement 2 is denoted by p_{ij} . The numbers of judgements in the (i,j):th cell , x_{ij}, have a multinomial distribution with parameters n and p_{ij}, (i,j=1,....,m)

The number of observations judged into the i:th category of rater 1 equals

$$x_{i.} = \sum_{j=1}^{m} x_{ij}$$

and the number of observations judged into the j:th category of rater 2 is

$$x_{.j} = \sum_{i=1}^{m} x_{ij}$$

The basic notations are shown in figure 1.

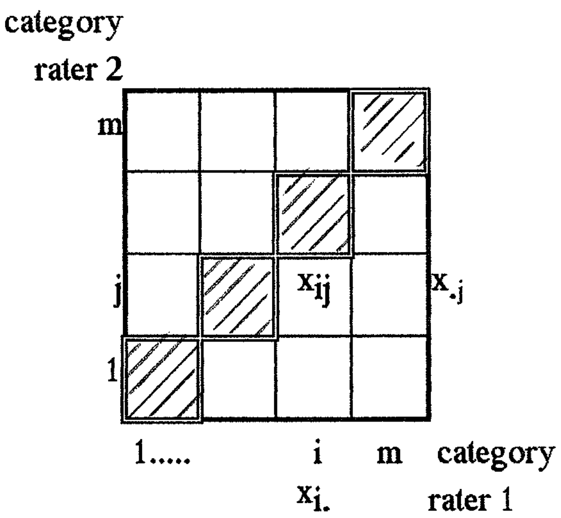

Figure 1: Basic notations in a contingency /an agreement table with x_{ij} observations in the (ij):th cell, $1 \leq i,j \leq m$. The agreement diagonal is marked.

We will now introduce a particular type of ranking, which enables us to separate the different sources of the inter-rater variability for two raters into random and systematic differences.

In ranking the observations of judgement 1 we use the convention of making the internal ranking of the i:th category according to the ranks of judgement 2. The observations in a cell (i,j) then get the following mean rank from judgement 1

$$\bar{R}^{(1)}_{ij} = \sum_{i_1=1}^{i-1} x_{i_1 .} + \sum_{j_1=1}^{j-1} x_{i\, j_1} + \tfrac{1}{2}(1+x_{ij})$$

In the same way we can define the mean rank of the judgement 2 for the (ij):th cell as

$$\bar{R}_{ij}^{(2)} = \sum_{j_1=1}^{j-1} x_{\cdot j_1} + \sum_{i_1=1}^{i-1} x_{i_1 j} + \frac{1}{2}(1+x_{ij})$$

$\bar{R}_{ij}^{(1)}$ and $\bar{R}_{ij}^{(2)}$ are defined only for all (ij) such that $x_{ij} \geq 1,\ 1 \leq i, j \leq m$.

Further details are given in an earlier research report[1].

Example 1:

This small example of paired classifications of 10 observations into one of three ordered categories, K,L or M illustrates the ranking approach. In figure 2 , there is a slight disagreement close to the main diagonal.

category rater 2				
M		1	2	3
L	2	2		4
K	3			3
	5	3	2	category
	K	L	M	rater1

Figure 2. A hypothetical example of inter-rater disagreement where the judgements have a concentrated band character.

The judgements have different marginal distributions, but the observations are concentrated to a band. We say that the judgements have a **concentrated band character.**

The two raters differ in two of the classifications resulting in observations into the cells (K,L) and (L,M). Five of the ten objects are classified to the category K of rater 1. The internal ranking of those objects is not observable in the marginal sum of rater 1, but according to the judgement of rater 2, our ranking approach will give two of the five objects a higher rank order - those who are judged to category L of rater 2. The ranks are displayed in table I.

$\bar{R}_{ij}^{(1)}$	rater 1	rater 2	$\bar{R}_{ij}^{(2)}$
2	K	K	2
2	K	K	2
2	K	K	2
4.5	K	L	4.5
4.5	K	L	4.5
6.5	L	L	6.5
6.5	L	L	6.5
8	L	M	8
9.5	M	M	9.5
9.5	M	M	9.5

Table I: A hypothetical example showing the ranking approach when there is a common order of the observations

Despite of the disagreement there exists a distinct common order of the ten observations appearing as equal mean ranks , $\bar{R}_{ij}^{(1)} = \bar{R}_{ij}^{(2)}$ for all (ij) with $x_{ij} \geq 1$. This means that the reason for the observed disagreement probably is that the raters have different ideas of the bounds of the categories K and L. Thus there is a pure systematic difference between the two judgements. Our particular type of ranking will thus reveal such a systematic difference.

Definition:

Two sets of judgements of the same n individuals are called **rank transformable** if $\bar{R}_{ij}^{(1)} = \bar{R}_{ij}^{(2)}$ for all (ij) such that $x_{ij} \geq 1$.

When two sets of judgements are rank transformable, there always exists a common ranking. The judgements will essentially describe the individual interpretation of the measuring scale. The observations will have a concentrated band character and the observed disagreement is due to a **systematic difference** between the raters.

If, on the other hand, there is no distinct common order among the individuals, there will appear different mean ranks $\bar{R}_{ij}^{(1)}$ and $\bar{R}_{ij}^{(2)}$ for some cells (ij). We consider this being a random difference between the judgements.

By means of our ranking approach it is possible to catch the minimal systematic difference between the paired judgements and thus separate the variability into its systematic component and a remaining part which we will consider to be the random component of variability. It is thus possible to get a detailed description of the variability. The disagreement patterns displayed in the next example will illustrate some situations of variability.

Example 2:

In figure 3 there are given three hypothetical disagreement patterns of inter- rater classifications of 100 objects into three categories (K,L,M). These examples have the same agreement of 80 percent and the same value of kappa[2] (=0.7) in spite of the difference in variability pattern.

rater 2 M		10	30	40
L	10	20		30
K	30			30
	40	30	30	100
	K	L	M	rater 1

Figure 3A. Hypothetical example of a disagreement pattern from the inter - rater judgements of 100 objects where the judgements have a concentrated band character

The classifications in the disagreement pattern of figure 3A have different marginal distributions, but the observations have a concentrated band character. The observed disagreement is caused by systematic differences only.

rater 2 M			30	30
L	10	20		30
K	30	10		40
	40	30	30	100
B	K	L	M	rater 1

rater 2 M	3	4	30	37
L	7	20	3	30
K	30	2	1	33
	40	26	34	100
C.	K	L	M	rater 1

Figures 3 B and C. Hypothetical example of disagreement patterns from the inter rater judgements of 100 objects where the judgements have equal marginal distributions but no concentrated band character (B) and where the judgements have different marginal distributions and no concentrated band character (C)

The observations in the figures 3B and 3C do not have a concentrated band character. The two judgements shown in figure 3B have the same marginal distribution but the objects classified into the cells (K,L) and (L,K) get different mean rank values from the two raters, i e $\bar{R}^{(1)}_{KL} \neq \bar{R}^{(2)}_{KL}$ and $\bar{R}^{(1)}_{LK} \neq \bar{R}^{(2)}_{LK}$. The observed disagreement is caused only by random differences between the raters.

The disagreement pattern of figure 3C have different marginal distributions indicating systematic differences between the raters. Some of the cells (ij) have different mean rank values, see figure 4 , revealing also random differences between the raters.

category rater 2				
M	39/ 65	64.5/ 68.5	85.5/ 85.5	
L	34/ 37	52.5/ 50.5	69/ 62	
K	15.5/ 15.5	41.5/ 31.5	67/ 33	
	K	L	M	category rater 1

Figure 4 The mean ranks of the disagreement pattern from figure 3C written as $\bar{R}_{ij}^{(1)} / \bar{R}_{ij}^{(2)}$

Conclusions from the ranking approach.

Our ranking approach applied to the examples indicates that there are different reasons for the inconsistency. The conclusions from our ranking approach are summarized below

$\bar{R}^{(1)}_{ij} = \bar{R}^{(2)}_{ij}$ differing marginal distributions $\sum_{i=1}^{\nu}\sum_{j=1}^{m} x_{ij} \neq \sum_{i=1}^{m}\sum_{j=1}^{\nu} x_{ij}$ $\nu = 1,2..(m-1),\ 1 \leq i,j \leq m$	* The paired judgements are **rank transformable** * there exists a common ordering among the objects * the observations have a **concentrated band character** The observed disagreement has * a **systematic difference** * no observed random difference between the raters
$\bar{R}^{(1)}_{ij} \neq \bar{R}^{(2)}_{ij}$ equal marginal distributions	* the observed disagreement is caused by **random differences** between the raters. * no observed systematic difference between the raters
$\bar{R}^{(1)}_{ij} \neq \bar{R}^{(2)}_{ij}$ differing marginal distributions	* The observed disagreement is caused by **random and systematic** differences between the raters. * The systematic differences are determined by the different marginal distributions

MEASURES OF RANDOM ERROR

A difference in the mean rank values of an observation in a cell means that two objects are ranked in reversed order in one judgement relative the other. One possible measure of the random differences is thus the empirical probability of such a reversed rank classification order.

The empirical probability of this event is

$$T = \frac{1}{n(n-1)} \sum_{\substack{k=1 \\ k\neq l}}^{n} \sum_{l=1}^{n} I_{k \between l}$$

where $I_{k \between l}$ indicates the reversed rank classification order of two observations k and l. This is an estimate of the parameter

$$\Theta = E[T] = P[\, X_1 < X_2 \,\Lambda\, Y_1 > Y_2 \,] + P[\, X_1 > X_2 \,\Lambda\, Y_1 < Y_2 \,]$$

where X_i denotes the judgement from rater 1 and Y_i that from rater 2 of the i:th object (i = 1,2). We suppose here that the ratings of the objects are independent.

The variance of the empirical probability of the reversed rank classification order is

$$\mathrm{Var}\,[T] = \frac{1}{n(n-1)} \{2(\Theta - \Theta^2) + 4(n-2)(\Psi - \Theta^2)\}$$

where $\Psi = E[I_{1 \between 2} \cdot I_{1 \between 3}]$

The empirical disagreement measure T, however, expresses only the relative frequency of reversely ordered observations and does not take into account any distance between the disagreed categories.

Another measure of the random error, taking into account the magnitude of the mean rank differences, is the empirical variance of the rank difference of an observation. It can be shown that this variance will not exceed $n^2/3$. We norm the variance to the interval [0,2] and we denote this normed variance RV

$$RV = \frac{6}{n^3} \sum_{i=1}^{m} \sum_{j=1}^{m} (\bar{R}_{ij}^{(1)} - \bar{R}_{ij}^{(2)})^2 \ x_{ij}$$

This measure of the random error has the following interpretations:

RV=0 No measurable random differences between the judgements, the observed disagreement is caused by systematic differences only

RV=1 The agreement between the two judgements equals what may be randomly caused

RV=2 A total systematic disagreement with all observations in the disagreement diagonal; one rater uses the categories in the reverse order of the other

$1<RV\leq2$ A reverse transformation of one judgement relative the other will give a better agreement than the observed one.

Corresponding to the empirical measures discussed here, there are parameters determined analogously by the true probabilities.

Example 3:

The measures of random error are calculated for the three disagreement patterns in the example 2. The rank transformable case in figure 3A have no observable random error, thus RV = T = 0.

The disagreement pattern of figure 3B has equal marginal distributions, indicating no systematic error. The observed disagreement is caused by a small random error, RV=0.01 and T=0.02 ($\hat{\sigma}_T = 0.008$).

The more dispersed disagreement of figure 3C has a probability of reversed ordered classification, T=0.03 ($\hat{\sigma}_T = 0.01$) and a relative variance RV=0.02. Since there are unequal marginal distributions, there are also systematic differences between the judgements.

By means of the ranking approach the different sources of disagreement are separated and it is possible to quantify the error components. Our scheme for the empirical measures of systematic and random errors in ordinal rating scales is shown in figure 5.

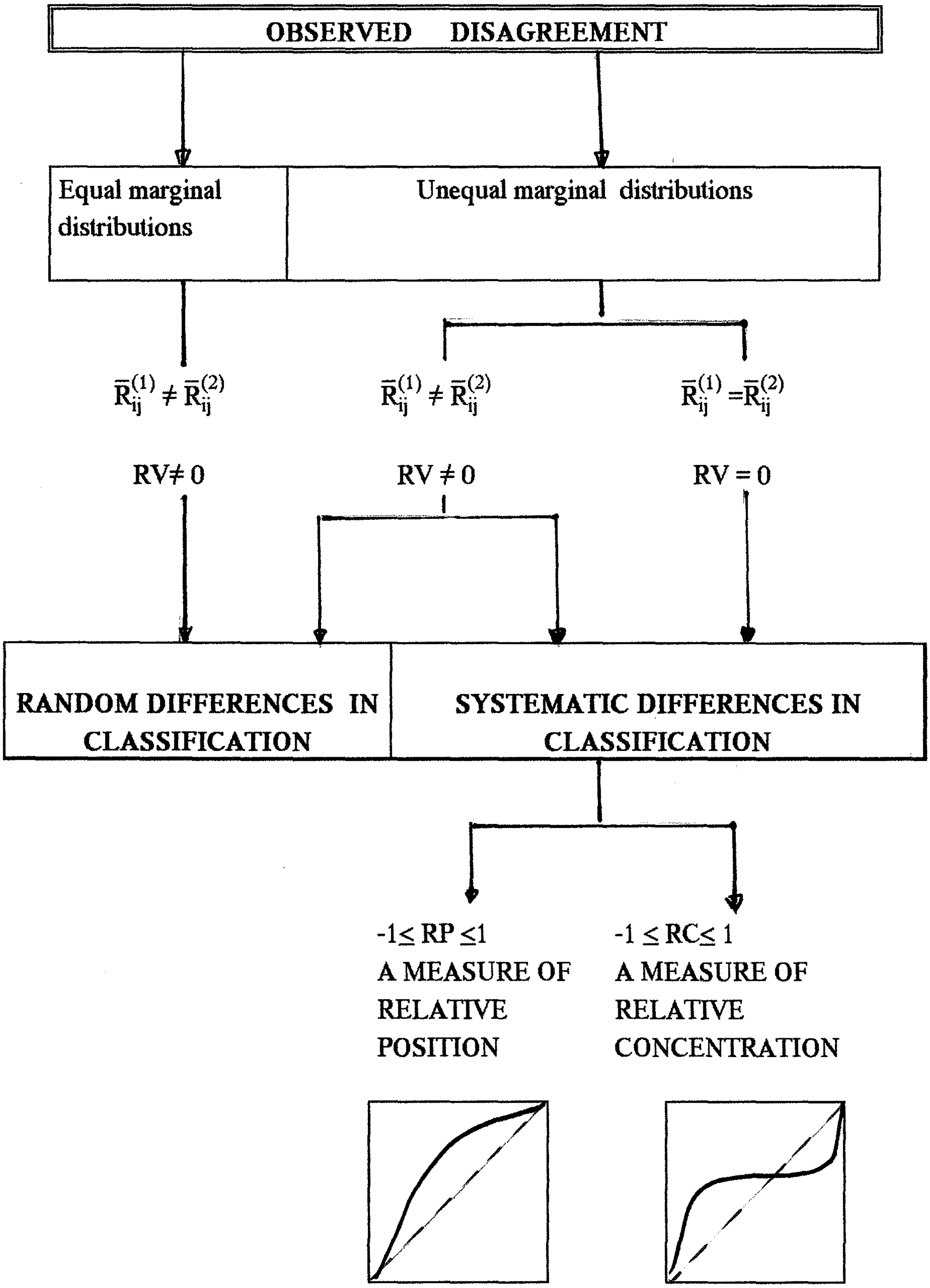
OBSERVED DISAGREEMENT
Equal marginal distributions
Unequal marginal distributions
$\bar{R}_{ij}^{(1)} \neq \bar{R}_{ij}^{(2)}$
$\bar{R}_{ij}^{(1)} \neq \bar{R}_{ij}^{(2)}$
$\bar{R}_{ij}^{(1)} = \bar{R}_{ij}^{(2)}$
RV≠ 0
RV ≠ 0
RV = 0
RANDOM DIFFERENCES IN CLASSIFICATION
SYSTEMATIC DIFFERENCES IN CLASSIFICATION
$-1 \leq RP \leq 1$
A MEASURE OF RELATIVE POSITION
$-1 \leq RC \leq 1$
A MEASURE OF RELATIVE CONCENTRATION

Figure 5: The disagreement measures associated with the ranking approach

MEASURES OF SYSTEMATIC ERROR

Systematic differences between paired judgements are indicated by different marginal distributions. Analysis of marginal distributions is common in many other statistical approaches as well[3].

Different marginal distributions, meaning different cumulative category probabilities of the raters, determine a relative length of the categories. The more objects being judged to a category, the greater is the relative category frequency and the larger is this relative category length. Thus the empirical systematic error can suitably be illustrated by means of the ROC curve (relative operating characteristic).

Two hypothetical examples illustrating two typical forms of ROC curves from clinical practice are shown i figure 6.

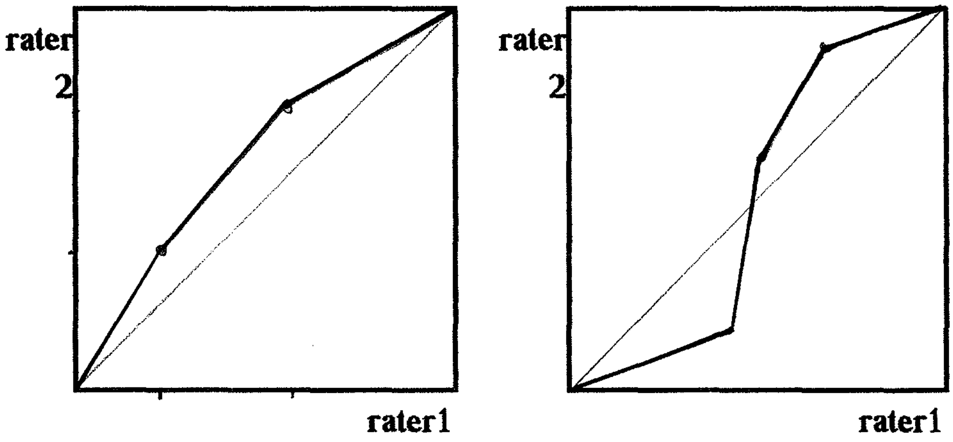

Figure 6: Two examples of systematic interjudgement differences in ROC curves

In figure 6A rater 2 judges systematically more objects to lower categories than does rater 1- the lower categories of rater 2 will thus have a larger relative length than the corresponding categories of rater 1. The judgements of rater 2 in figure 6B are more concentrated to the central categories compared to the judgements of rater 1.

The marginal distributions of two observers can differ in various ways depending on the type of systematic disagreement. It may be reasonable to describe the basic properties of the systematic errors by two measures- a **measure of relative position and a measure of relative concentration.**

Definition:

The relative position, RP, between two categorical classifications is defined

$$RP = \hat{P}(X<Y) - \hat{P}(Y<X) =$$

$$\sum_{\nu} [\ \hat{P}(X < \nu)\ \hat{P}(Y = \nu) - \hat{P}(Y < \nu)\ \hat{P}(X = \nu)] =$$

$$\sum_{\nu} [\ \hat{q}^{(1)}_{\nu-1} \cdot \hat{p}^{(2)}_{\nu} - \hat{q}^{(2)}_{\nu-1} \cdot \hat{p}^{(1)}_{\nu}\ ]$$

where

X denotes the judgement by rater 1 and Y denotes the judgement by rater 2 and the categories are labelled $\nu = 1 \ldots m$

$\hat{p}_{\nu}^{(\lambda)}$ denotes the category relative frequency of rater λ ,(λ =1,2)

$\hat{q}_{\nu}^{(\lambda)}$ denotes the cumulative category relative frequency of the rater λ,

RP>0 indicates that the classifications made by the observer 1 are systematically shifted to lower categories relative the classifications made by observer 2.

The estimated relative position of the judgements shown in the figure 6A is RP= - 0.24. Corresponding measure for the judgements in the figure 6B is RP = -0.03

Definition:

The relative concentration, RC, between two categorical classifications is defined

$$RC = \frac{1}{M} [\hat{P}(X_1 < Y_3 < X_2) - \hat{P}(Y_1 < X_3 < Y_2)] =$$

$$\frac{1}{M} \sum_{\nu} [\hat{P}(Y = \nu)\hat{P}(X < \nu)\hat{P}(X > \nu) - \hat{P}(X = \nu)\hat{P}(Y < \nu)\,\hat{P}(Y > \nu)] =$$

$$\frac{1}{M} \sum_{\nu} [\hat{p}_{\nu}^{(2)} \cdot \hat{q}_{\nu-1}^{(1)} (1 - \hat{q}_{\nu}^{(1)}) - \hat{p}_{\nu}^{(1)} \cdot \hat{q}_{\nu-1}^{(2)} (1 - \hat{q}_{\nu}^{(2)})\,)$$

where $M = \max_{i=0,1} \hat{p}_i(1-\hat{p}_i)$

$$\hat{p}_0 = \hat{P}(X < Y) = \sum_{\nu} \hat{P}(X < \nu)\, \hat{P}(Y = \nu)$$
$$\hat{p}_1 = \hat{P}(Y < X) = \sum_{\nu} \hat{P}(Y < \nu)\, \hat{P}(X=\nu)$$

$(X_i\, Y_i)$, i=1,2,3 denote independent pairs of judgements. Norming with M means that $-1 \leq RC \leq 1$. The extreme values correspond to the distribution of one of the classifications entirely concentrated in relation to the other.

The estimated measures of the relative concentration between the two judgements is RC=0.03 of figure 6A and is RC= 0.53 of figure 6B.

Example 4:

The disagreement pattern of the example 2, figure 3C has different marginal distributions, implying that there are systematic differences between the judgements. The estimated relative position is RP=0.06 and the estimated relative concentration is RC=0.04.

The empirical measures discussed here are to be considered as estimates of the distribution parameters defined by the same expressions with true probabilities inserted

APPLICATION EXAMPLE:

In an interobserver study about Fisher grade[4] two radiologists independently classified 59 cranial computerized tomography (CT) scans into one of four ordered categories in order to assess the presence and amount of subarachnoid blood. The observed frequencies of the judgements made by the two radiologists are shown in figure 7.

radiologist 2

F4	1		2	24	27
F3		3	9	9	21
F2		4	1	1	6
F1	3	2			5
	4	9	12	34	59

radiologist 1

Figure 7: Result of inter radiologist judgements of 59 CT scans according to Fisher grade

The judgements have different marginal distributions, but no concentrated band character. The observed disagreement is due to both random and systematic errors. The marginal distributions determine empirically the category bounds displayed in the ROC curve, see figure 8.

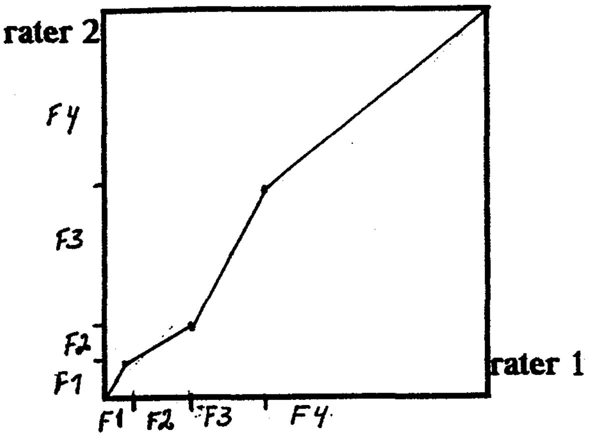

Figure 8: A ROC curve showing the empirically determined systematic inter-rater differences in the use of Fisher grade

Radiologist 1 classifies more scans to the highest Fisher level than does the other radiologist. By means of the empirical relative lengths of the categories , illustrated in the ROC curves it is easy to identify tho most important reasons to the variability in order to improve the validity of the scale.

Our measures of random and systematic errors are displayed in table III. The theoretical expression for the variance of the estimated mean of the squared rank differences (RV) is very complicated, but is conveniently estimated by the jackknife[5] technique. The jackknife-estimations of RC and RP are also given. The coefficient kappa, the maximum value of kappa permitted by the marginal distributions [2] and the percentage agreement, PA are also given.

The measures and the ROC curve show that the main reason for the disagreement is the systematic differences between the judgements.

Radiologist 2 judges systematically more scans to lower Fisher levels and has more observations concentrated to level 3 than has radiologist 1.

Table II. Assessment of the interobserver reliability on Fisher grade.

RANDOM ERROR		
The probability of the reversed rank classification order	T = 0.037	$\hat{\sigma}_T = 0.020$
Relative variance	RV = 0.04	$(\hat{\sigma}_{RV})_{jack} = 0.03$
SYSTEMATIC ERROR		
Relative position	RP = -0.084	$(\hat{\sigma}_{RP})_{jack} = 0.06$
Relative concentration	RC= 0.113	$(\hat{\sigma}_{RC})_{jack} = 0.06$
Alternative measures		
the coefficient kappa	$\kappa = 0.5$ $\kappa_m = 0.74$ $\kappa / \kappa_m = 0.68$	$\kappa_{max} < 1$ when there is a difference in marginal distributions κ / κ_m the marginally permitted agreement
Percentage agreement	PA=68%	

DISCUSSION

Our ranking approach enables us to separate the inconsistency into random and systematic errors and to quantify this lack of consistency in a few measures. These give more detailed descriptions of the variability than does the coefficient kappa and other common measures for ordered categorical data.

Our ambition is to develop non-parametric methods which are generally applicable to intra- and inter- rater problems as well as interscale problems and whose measures are easy to interpret and to use. The type of application depends on the design of the study.

By means of repeated measurements in the same experimental unit you may use the method in different validity and reliability assessments such as

- the **criterion validity** referring to the agreement to a gold standard or a criterion measure
- the **construct validity** referring to an inter-scale relation with the same theoretical and operational definitions
- the **predictive validity** referring to the agreement to outcomes in the future
- the **intra-rater reliability** concerning the short term consistency of the instrument
- the **inter-rater reliability** concerning the consistency of the raters.

The ranking approach allows for comparing different measurement instruments with unequal number of categories and it is also possible to compare a categorical and a continuous scale.

ACKNOWLEDGEMENT

We wish to thank Sven Ekholm, Department of Neuroradiology and Claes von Essen and Jan-Erik Starmark, Department of Neurosurgery, Sahlgren´s Hospital, Göteborg for access to data and for valuable discussions.

REFERENCES

1. Holm S, Svensson E. Statistical rank methods for ordinal categorical data. University of Göteborg, Dept of Statistics; Research Report 1991:3

2. Cohen J. A coefficient of agreement for nominal scales. Educational and psychological measurement 1960;XX:1

3. Becker MP, Agresti A. Log-linear modelling of pairwise interobserver agreement on a categorical scale. Statistics in Medicine 1992;11:101-114

4. Svensson E, von Essen C, Ekholm S, Johansson A, Holm S, Starmark JE. Interobserver variability in assessment of Fisher grade and preoperative hydrocephalus after aneurysmal subarachnoid hemorrhage. Application of a new statistical method for separation of systematic and random errors of variability. (To be published)

5. Efron B. The jackknife, the bootstrap and other resampling plans.Philadelphia: Society For Industrial And Applied Mathematics, 1982.