QUALITY AND RELIABILITY ENGINEERING INTERNATIONAL
Qual. Reliab. Engng. Int. 2004; **20**:383–395 (DOI: 10.1002/qre.653)

Special Issue

Measurement System Analysis for Bounded Ordinal Data

Jeroen de Mast*,† and Wessel van Wieringen
Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA), Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands

The precision of a measurement system is the consistency across multiple measurements of the same object. This paper studies the evaluation of the precision of measurement systems that measure on a bounded ordinal scale. A bounded ordinal scale consists of a finite number of categories that have a specific order. Based on an inventory of methods for the evaluation of precision for other types of measurement scales, the article proposes two approaches. The first approach is based on a latent variable model and is a variant of the intraclass correlation method. The second approach is a non-parametric approach, the results of which are, however, rather difficult to interpret. The approaches are illustrated with an artificial data set and an industrial data set. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: intraclass correlation coefficient; Gauge R&R study; attribute data; kappa method

1. INTRODUCTION

Measurement system analysis (MSA) seeks to describe, categorize, and evaluate the quality of measurements, improve the usefulness, accuracy, precision, and meaningfulness of measurements, and propose methods for developing new and better measurement instruments[1]. In this article we study the evaluation of *precision* (or *consistency*) of measurement systems. By precision we mean the extent to which we find similar results if we measure (the properties of) the same object multiple times with the same or comparable measuring instruments.

How precision is addressed depends on the field where the measurement system is used. Industrial statistics concentrates on measurement spread[2,3], whereas in psychometrics the focus is on reliability[4]. In both fields the precision of the measurement is assessed by means of an experiment using the fundamental principles of experimental design. In this paper we consider a simple design where m repeated measurements are obtained from n objects with the same measurement system in randomized order. The observations are denoted X_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$.

A scale is the target range of a measurement system. An ordinal scale is a countable set with a defined order but without a distance metric. In a bounded ordinal scale the number of categories is finite. A discrete scale is an ordinal scale with a distance metric imposed. The concept of distance distinguishes an ordinal scale from a discrete scale. For both scales a statement of the form '$a < b$' makes sense (as opposed to nominal scale), but, unlike on discrete scales, '$a - b$' has no meaning on an ordinal scale. Examples of bounded ordinal measurements are quality judgments of the form 'good', 'mediocre', or 'bad', and ratings in classes I, II, III, and IV.

*Correspondence to: Jeroen de Mast, Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA), Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands.
†E-mail: jdemast@science.uva.nl

Received 1 December 2003
Revised 25 May 2004

This article concentrates on the evaluation of the precision of bounded ordinal measurement systems. First, an overview is given of existing methods used in the assessment of precision for other measurement scales. Based on these methods, two approaches are developed for bounded ordinal measurements. These approaches—a latent variable approach and a non-parametric approach—are illustrated with an artificial data set and an industrial example.

2. INVENTORY OF CURRENT METHODS

2.1. Intraclass correlation coefficient

The social sciences interpret precision as *reliability*, which is the degree of object variation relative to the total observed variation or, equivalently, the correlation among multiple measurements of the same object. Reliability is often expressed in the form of an intraclass correlation coefficient (ICC)[5,6].

The observations X_{ij} are generally assumed to follow the model

$$X_{ij} = Z_i + \varepsilon_{ij} \tag{1}$$

with $Z_i \sim \mathcal{N}(\mu_p, \sigma_p^2)$ the true value of object i and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ the stochastic measurement error. The model states that the distribution of the measurement error is symmetrical around and independent of the object's true value.

The ICC is the correlation between different measurements X_{ij} and X_{ik} of a single object i. Under model (1) we have

$$\text{ICC} = \frac{\text{Cov}(X_{ij}, X_{ik})}{\sqrt{\text{Var}(X_{ij}) \cdot \text{Var}(X_{ik})}} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2} \tag{2}$$

ICC expresses measurement reliability as can be seen from the right-hand side of (2): a ratio of the variance of interest over the total variance. Under model (1), ICC can only assume values in the interval [0, 1], 1 corresponding to perfect reliability and 0 to a measurement system which is no more consistent than chance.

One-way analysis of variance gives the estimates for the variance components in (2). Denoting by MS_w and MS_b the within and between groups mean squares, respectively, a biased but consistent estimator of ICC is[6]

$$\widehat{\text{ICC}} = \frac{\text{MS}_\text{b} - \text{MS}_\text{w}}{\text{MS}_\text{b} + (m-1)\text{MS}_\text{w}}$$

Note that this estimate is only acceptable if the objects $i = 1, \ldots, n$ are sampled randomly from the population. If this is not the case, σ_p^2 should be estimated from a historical sample (in practice, it will be easier to estimate $\sigma_{\text{total}}^2 = \sigma_p^2 + \sigma_e^2$, because in general it is not possible to obtain measurements without measurement spread).

2.2. Gauge R&R

Industrial statistics interprets precision as *measurement spread*[2,3,7]. The model underlying the Gauge R&R equals model (1) of the ICC method. The measurement spread is the standard deviation σ_e of repeated measurements of a single object. In standard Gauge R&R studies this standard deviation is split into a component due to the measurement system itself (repeatability) and a component due to additional sources of variation such as operators (reproducibility). The measurement spread is compared to the process spread (including measurement spread), as is done by the Gauge R&R statistic[8]:

$$\text{Gauge R\&R} = \frac{\sigma_e}{\sigma_{\text{total}}} \tag{3}$$

with $\sigma_{\text{total}} = \sqrt{\sigma_p^2 + \sigma_e^2}$.

The intraclass correlation coefficient and the Gauge R&R are essentially the same:

$$\text{ICC} = 1 - (\text{Gauge R\&R})^2$$

The main difference is that ICC expresses the ratio of measurement spread and total spread in terms of variances and the Gauge R&R in terms of standard deviations. Proportions suggest that the numerator plus its complement add up to the denominator. This holds for variances ($\sigma_e^2 + \sigma_p^2 = \sigma_{total}^2$), but not for standard deviations ($\sigma_e + \sigma_p \neq \sigma_{total}$, in general), which makes ICC the more natural choice (Wheeler[9] makes a similar observation).

An alternative evaluation of the measurement system is to consider $5.15\,\sigma_e$. This value represents the width of a 99% confidence interval on an object's true value, given a single measurement. Often, this interval is compared to the distance between the tolerance limits on the characteristic (the so-called *P/T ratio*). A third alternative is to determine the discriminatory power of the measurement system. Suppose we have two objects and corresponding measurements X_1 and X_2. It can be decided that the two objects are not identical (with 99% confidence) if $|X_1 - X_2| > 2.575\sqrt{2}\,\sigma_e$. Objects whose true values are more than $2.575\sqrt{2}\,\sigma_e$ apart will be distinguished in this sense with at least 50% probability. Taking $5.15\,\sigma_{total}$ to represent the range of the measured products, the measurement system can distinguish between $\sqrt{2}\,\sigma_{total}/\sigma_e$ categories.

2.3. *Kappa*

A concept that is related to precision is *agreement*. Two measurements of an object agree if they are identical. Cohen[10] introduces a measure of agreement called the kappa, which is nowadays frequently used to evaluate measurement systems on nominal scales. The kappa, denoted κ, is a measure of agreement corrected for agreement by chance, which has the form

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{4}$$

Here P_o is the observed proportion of agreement and P_e the expected proportion of agreement. The kappa attains the value 1 when there is perfect agreement, 0 if all observed agreement is due to chance and negative values when the degree of agreement is less than is to be expected on the basis of chance.

For the simple case where $m = 2$ (two measurements per object) the terms in (4) are computed as[10]

$$P_o = \sum_k p_{1,2}(k,k) \quad \text{and} \quad P_e = \sum_k p_1(k)\,p_2(k)$$

where k ranges over all categories of the scale. Here P_o is the observed proportion of objects with agreeing measurements 1 and 2 and $p_{1,2}(k,k)$ denotes the proportion of objects that have been categorized as k by measurements 1 and 2. P_e is the expected proportion of agreement based on independence of measurements 1 and 2. $p_1(k)$ and $p_2(k)$ denote the marginal proportions of both measurements and categories k. For $m \geq 3$ generalizations are given in the literature (see Conger[11]).

2.4. *Non-parametric methods*

If one does not want to make distributional assumptions as in model (1), one may resort to non-parametric methods[12] such as Kendall's tau[13]. Precision is interpreted as *consistency* between different rankings of a series of objects. Let r_{i1} and r_{i2} be the rank numbers of object i in two rankings 1 and 2. Let P and Q be the numbers of agreeing and opposite rankings; that is,

$$\begin{aligned} P &= \#\{h, i : (r_{h1} < r_{i1}, r_{h2} < r_{i2}) \text{ or } (r_{h1} > r_{i1}, r_{h2} > r_{i2})\} \\ Q &= \#\{h, i : (r_{h1} < r_{i1}, r_{h2} > r_{i2}) \text{ or } (r_{h1} > r_{i1}, r_{h2} < r_{i2})\} \end{aligned} \tag{5}$$

Then τ is the difference between P and Q divided by the absolute value of their maximum difference (the total number of pairs one can form). This is

$$\tau = \frac{P - Q}{n(n-1)/2}$$

τ measures the rank correlation between two rankings. As a non-parametric analogue to the usual product moment correlation coefficient it represents the extent to which there exists a monotonous relationship between two variables[13]. One speaks of a perfect positive monotonous relationship when for every pair of objects i and j we have $(r_{i1} - r_{j1})(r_{i2} - r_{j2}) > 0$. Negative monotony is defined analogously. τ can only assume values in the interval $[-1, 1]$, where 1 corresponds to a perfect positive monotonous relationship, -1 to a negative relationship and 0 to no relationship at all (i.e. a random ranking process).

Another non-parametric measure of rank correlation is Spearman's ρ_s[13]. At the core is the sum of squares of the differences in rank number of two rankings for each individual object. This is scaled such that ρ_s equals 1 in the case of identical rankings and -1 if the rankings are each other's reverse:

$$\rho_s = 1 - \frac{6\sum_i (r_{i1} - r_{i2})^2}{n^3 - n}$$

ρ_s treats the ranks as if they were the true units of measurement, assuming a discrete scale instead of an ordinal one.

τ and ρ_s are concerned with correlation between two rankings. For the case involving $m > 2$ rankings, Kendall[13] defined his coefficient of concordance as

$$W = \frac{\sum_{i=1}^{n} (R_i - \frac{1}{2}m(n+1))^2}{\frac{1}{12}m^2(n^3 - n)}$$

Here r_{ij} is the ranking of object i by ranking j and $R_i = \sum_{j=1}^{m} r_{ij}$. The rationale underlying this definition is the analogy to the analysis of variance. This is also its criticism, as rankings are not independent of each other, but are assigned in conjunction with each other. Therefore, it has been proposed to use the average τ of all possible pairs of measurements instead.

2.5. *Other alternatives*

Alternative methods, which we do not discuss in this paper, can be found in Dunn[12], Feldstein and Davis[14], Agresti[15], Uebersax and Grove[16] and Vanleeuwen and Mandabach[17].

3. *MSA FOR BOUNDED ORDINAL DATA*

Modifying the methods discussed in the preceding section for application with bounded ordinal data, we develop two main approaches. The choice between them relates to the distinction between the situation where one deals with a scale that is intrinsically bounded and ordinal, and the situation where one is in fact dealing with a continuous variable that is mapped by the measurement system onto a bounded ordinal scale. In the first situation one cannot use methods based on standard deviations and correlations, because these methods assume a distance metric on the measurement scale. One has to resort to non-parametric methods. In the second situation, the ordinal scale can be equipped with a distance metric, which it inherits via the map (formed by the measurement system) from the underlying continuous scale. This enables the use of methods based on standard deviations and correlations (ICC and Gauge R&R). The underlying continuous scale need not be known and the object's true value is treated as a latent variable. The kappa method will be shown to reduce to a variant of the ICC method.

3.1. Modification of the ICC method

As shown in Section 2.2 the ICC method is essentially the same as the Gauge R&R method and for this reason we do not discuss the Gauge R&R method separately.

When trying to apply the ICC method to bounded ordinal data, we come across two problems, which relate to:

(1) a distance metric for the measurement scale;
(2) distributional properties of the measurement error.

Problem 1. Ordinal scales only have an order defined, not a distance metric. The ICC method, however, makes use of standard deviations and correlations, which are only defined for measurement scales for which there is a well-defined distance metric. Not until the ordinal scale is extended with a metric can we apply ICC type methods. In effect, this extension transforms an ordinal scale into a discrete scale.

Problem 2. The standard ICC method (as well as the Gauge R&R method) assumes that: (a) the measurement error is symmetrically distributed around an object's true value; and (b) that this distribution is the same, whatever the true value is (as reflected in model (1)). Both assumptions (a) and (b) are natural in the study of measurement error and we wish to introduce similar assumptions for the bounded ordinal case. Neither assumption can, however, be retained for bounded scales in a straightforward form: the measurement error of objects close to a bound will be skewed away from the bound.

In order to adapt the ICC method for use with bounded ordinal data, it is unavoidable to make bold assumptions on both issues. It appears possible to derive both a distance metric and a distribution for the measurement error if one is prepared to assume that underlying the measurements there is a continuous variable (the 'true' value of the object). Below, we study how to adapt the ICC method in this situation for use first with ordinal but unbounded data, and then for use with bounded ordinal data. If one is not willing to assume a continuous true value that underlies the measurements, one has to resort to non-parametric methods.

3.1.1. ICC for unbounded ordinal data

Let Z denote the true value of the measured property of an object. We assume that $Z \in \mathbb{R}$. Moreover, we assume that Z has a normal distribution:

$$Z \sim \mathcal{N}(\mu_{\mathrm{p}}, \sigma_{\mathrm{p}}^2) \tag{6}$$

Z is not observed; instead we measure an ordinal variable X, which assumes a value in $\mathbb{D}$. $\mathbb{D}$ is an infinite countable set, whose categories are labelled $\ldots, 1, 2, 3, \ldots$. By reporting X instead of Z, the measurement system maps $\mathbb{R}$ onto $\mathbb{D}$ and adds a stochastic component due to measurement error. The map $RD : \mathbb{R} \to \mathbb{D}$, $RD(Z) = \lceil Z \rceil$ represents a measurement system which is not subject to measurement error ($\lceil . \rceil$ is the *ceiling* function). Its reverse is $DR(k) = k - \frac{1}{2}$, for $k \in \mathbb{D}$. The measurement error in X can be modelled by specifying the distribution of X conditional on Z, which is of the form $P(X = k \mid Z) = p_k(Z)$, $k \in \mathbb{D}$, with p_k dependent on the true value Z of the measured object.

In order to apply a method analogous to the ICC method, we have to define a distance metric on $\mathbb{D}$. We propose to interpret the categories of $\mathbb{D}$ as equidistant by taking the distance between any two successive categories as 1. In this way, $\mathbb{D}$ inherits the distance metric of the domain of Z, in that $|k - \ell|_{\mathbb{D}} = |DR(k) - DR(\ell)|$, for any $k, \ell \in \mathbb{D}$.

Furthermore, we have to make assumptions about the distribution of X. We propose to assume that

$$p_k(Z) = \int_{t=DR(k-1/2)}^{DR(k+1/2)} f_{\mu=Z;\sigma_{\mathrm{e}}}(t)\, \mathrm{d}t \tag{7}$$

with $f_{\mu;\sigma}$ the density of the normal distribution with mean μ and standard deviation σ. Thus, the distribution of X given Z is a discretized form of a normal distribution. Combining (6) and (7) we find

$$\begin{aligned} P(X=k) &= \int_{t=-\infty}^{\infty} p_k(t) f_{\mu_\mathrm{p};\sigma_\mathrm{p}}(t)\,\mathrm{d}t \\ &= \int_{u=DR(k-1/2)}^{DR(k+1/2)} f_{\mu_\mathrm{p};\sqrt{\sigma_\mathrm{p}^2+\sigma_\mathrm{e}^2}}(u)\,\mathrm{d}u \end{aligned} \tag{8}$$

In order to understand intuitively the assumption (7), one could think of a measurement error $\varepsilon \in \mathbb{R}$, which has a $\mathcal{N}(0, \sigma_\mathrm{e}^2)$ distribution. ε is added to the true value Z and then discretized: $X = RD(Z+\varepsilon)$, which results in the distribution given in (8). Thus, (8) is the discrete analogue of (1), and we have retained symmetry of measurement error and independence of the distribution of the measurement error of the true value Z. Analogous to the standard ICC method, we define measurement reliability as

$$\mathrm{ICC} = \frac{\sigma_\mathrm{p}^2}{\sigma_\mathrm{p}^2 + \sigma_\mathrm{e}^2} \tag{9}$$

In order to estimate ICC we have repeated measurements $X_{i1}, X_{i2}, \ldots, X_{im}$ of objects $i = 1, 2, \ldots, n$. Following standard ICC methodology, one would estimate ICC from a ratio of mean squares. For discretized data, mean squares have, however, a bias. Correcting for this bias (see the derivation in Appendix A), the estimate becomes

$$\widehat{\mathrm{ICC}} = \frac{\mathrm{MS_b} - \mathrm{MS_w} + (m-1)/12\,m}{\mathrm{MS_b} + (m-1)\mathrm{MS_w} - (m^2-m+1)/12\,m} \tag{10}$$

3.1.2. ICC for bounded ordinal data

Next, we study how to modify the ICC in the case of bounded ordinal data. We assume that $\mathbb{D}$ is a finite set, whose categories are labelled $1, 2, \ldots, a$. We could assume a bounded domain for the true value Z as well. This would, however, make it impossible to retain the assumption of the distribution of the measurement spread being independent of and symmetrical around the true value: for values close to the bounds, the measurement spread would be skewed away from the bound, thus violating both assumptions. Instead, we retain $\mathbb{R}$ as the domain of the true value Z and define the map $LRD : \mathbb{R} \to \mathbb{D}$,

$$LRD(Z) = \left\lceil \frac{a\exp(Z)}{1+\exp(Z)} \right\rceil \tag{11}$$

Its reverse (for $k \in \mathbb{D}$) is

$$LRD(k) = \log\left(\frac{k-1/2}{a-k+1/2}\right) \tag{12}$$

LRD is similar to the logistic transformation that is used in logistic regression. For Z we retain model (6). For the measurement error we have

$$P(X=k \mid Z) = p_k(Z) = \int_{t=LRD(k-1/2)}^{LRD(k+1/2)} f_{\mu=Z;\sigma_\mathrm{e}}(t)\,\mathrm{d}t \tag{13}$$

An equation similar to (8) could be derived. In the domain of Z, the distribution of the measurement spread is independent of and symmetrical around an object's true value. In the centre of the domain, the map *LRD* approximates *RD*. Towards the bounds, more and more of the $\mathbb{R}$ domain is condensed in classes of $\mathbb{D}$ and the extreme classes of $\mathbb{D}$ cover all values of Z smaller or larger than a certain value. In our opinion, this behaviour reflects how bounded ordinal measurement scales in reality are often implicitly defined: they are distinctive in a relevant subdomain of true values, whereas values more to the extremes are combined in the two extreme

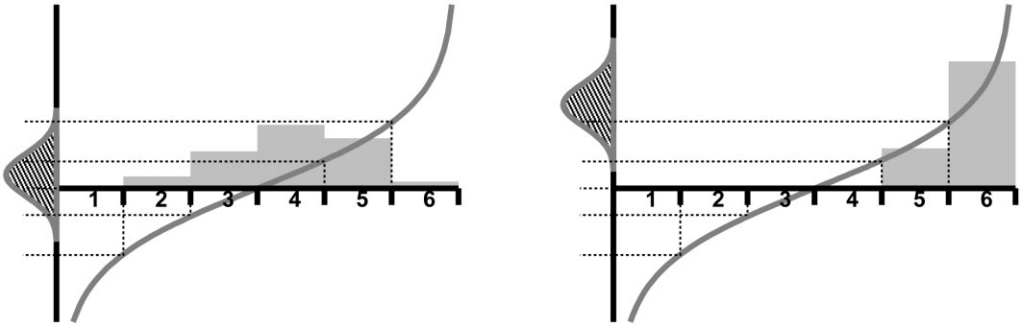

Figure 1. Relation between $\mathbb{R}$ and $\mathbb{D}$

categories, which cover all values beyond a certain lower and upper point. The distribution of the measurement error is illustrated in Figure 1. The graph has $\mathbb{D} = \{1, \ldots, 6\}$ on its x-axis and $\mathbb{R}$ on its y-axis. The curve shows how values in $\mathbb{R}$ and $\mathbb{D}$ are related. The histogram shows the distribution $p_k(Z)$, $k = 1, \ldots, a$, of a measurement X for a single object. This distribution can be derived by imagining a true value Z on the y-axis to which a normally distributed and zero-mean error is added (this hypothetical distribution is indicated by the Gaussian curve on the y-axis). The graph on the right shows the distribution of X given an object that has a large true value Z. Note that this model implies that the measurement system is more consistent in the extreme classes, meaning that the really good and really bad objects can be judged with high precision.

The measurement system's reliability is defined as in (9). Due to the nonlinearity of *LDR*, mean squares give heavily biased estimators for the variances in (9). To derive suitable estimators, we consider the statistics $N_{ik} = (\#X_{ij},\ j = 1, \ldots, m : X_{ij} = k)$, for $i = 1, \ldots, n$ and $k \in \mathbb{D}$. Regarding the true values Z_i as fixed for the moment, and given that for a single product i the tuple $(N_{i1}, \ldots, N_{ia})$ has a multinomial distribution, we can compute the log-likelihood L.

$$\begin{aligned} L &= \sum_{i=1}^{n} \log P(N_{i1} = n_{i1}, \ldots, N_{ia} = n_{ia}) \\ &= \sum_{i=1}^{n} \log\left(\frac{m!}{n_{i1}! \cdots n_{ia}!}\right) + \sum_{i=1}^{n} \sum_{k=1}^{a} n_{ik} \log(\Phi(A(+)) - \Phi(A(-))) \end{aligned}$$

with Φ the cumulative standard normal distribution function,

$$A(+) = \frac{LDR(k + 1/2) - Z_i}{\sigma_e} \quad \text{and} \quad A(-) = \frac{LDR(k - 1/2) - Z_i}{\sigma_e}$$

We find estimates for $Z_1, \ldots, Z_n$ and σ_e^2 from

$$\hat{Z}_1, \ldots, \hat{Z}_n, \hat{\sigma}^2_{e,ml} = \arg\max \sum_{i=1}^{n} \sum_{k=1}^{a} n_{ik} \log(\Phi(A(+)) - \Phi(A(-)))$$

In order to correct for the bias that maximum likelihood estimators are subject to in general, we work with

$$\hat{\sigma}_e^2 = \frac{m}{m-1} \hat{\sigma}^2_{e,ml} \tag{14}$$

Next, we estimate σ_p^2 by

$$\hat{\sigma}_p^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{Z}_i - \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_i\right)^2 - \frac{\hat{\sigma}_e^2}{m} \tag{15}$$

The sample ICC is given by

$$\widehat{\mathrm{ICC}} = \frac{\hat{\sigma}_\mathrm{p}^2}{\hat{\sigma}_\mathrm{p}^2 + \hat{\sigma}_\mathrm{e}^2}$$

An alternative representation of the results is to report for objects with true values $Z_k = LDR(k)$, $k = 1, \ldots, a$, the distribution of the measurements $p_\ell(Z_k)$, which is computed by substituting $\hat{\sigma}_e$ for σ_e in (13). These $p_\ell(Z_k)$ give the probability that an object that should be rated k is in fact rated $\ell = 1, 2, \ldots, a$.

3.2. *Modification of the kappa method*

The main problem of the kappa method when dealing with bounded ordinal data is that it only uses the value information. Values are interpreted as mere labels, ignoring the order information. In effect, ordinal data are downgraded to nominal data, and therefore the kappa method does not take along in its evaluation of an ordinal measurement system, one of the system's most important aspects.

It has been suggested[18] that instead one should use the *weighted* κ (of which κ is a special case) when dealing with ordinal data. This statistic takes into account that some types of disagreement may be considered more important than others, and that this should be reflected by assigning weights. For the weighted kappa the expected and observed proportion of agreement are defined as:

$$P_\mathrm{o} = \sum_{k,\ell=1}^{a} w(k,\ell)\, p_{1,2}(k,\ell) \quad \text{and} \quad P_\mathrm{e} = \sum_{k,\ell=1}^{a} w(k,\ell)\, p_1(k)\, p_2(\ell)$$

with $0 \le w(k,\ell) \le 1$, and $w(k,k) = 1$. Krippendorff[19] proposed quadratic weights:

$$w(k,\ell) = 1 - \frac{(k-\ell)^2}{(a-1)^2} \quad \text{for } k \text{ and } \ell \text{ in } \mathbb{D}$$

In effect these weights define a distance metric on $\mathbb{D}$. Based on quadratic weights, and assuming model (1) for the (ordinal) data, *weighted* κ is a biased estimate of ICC as defined in (2). Thus, the method reduces to a variant of the ICC method.

3.3. *Modification of non-parametric methods*

By studying Kendall's τ in the situation of a bounded ordinal scale, we apply the theory of *rankings* to *ratings*. The main difference is that for rankings it is not possible for two objects to fall into the same category (so called 'ties'). Ratings can be regarded as rankings, with the complication of ties (which are usually unavoidable for ratings). When dealing with ratings, they should be converted to rank numbers. To obtain rank numbers from the ratings X_{ij}, order for each j the X_{ij}, $i = 1, \ldots, n$, from small to large, where X_{ij} that are equal are left in an arbitrary order. Next, let r_{ij}, $i = 1, \ldots, n$ be the rank numbers of the ordered X_{ij}, where rank numbers for ties are averaged. This is expressed as

$$r_{ij} = \sum_{k=1}^{X_{ij}-1} M_{jk} + (1 + M_{j,X_{ij}})/2$$

with $M_{jk} = (\#X_{ij},\ i = 1, \ldots, n : X_{ij} = k)$, for $j = 1, \ldots, m$ and $k \in \mathbb{D}$. When ties are present, τ should be modified as follows:

$$\tau = \frac{P - Q}{\sqrt{n(n-1)/2 - T_1}\sqrt{n(n-1)/2 - T_2}}$$

Table I. Artificial dataset

	1	2	3	4	5	6		1	2	3	4	5	6
1	3	2	3	2	2	3	**16**	4	3	3	4	4	3
2	3	3	4	4	4	4	**17**	3	3	3	3	3	3
3	4	3	3	4	3	3	**18**	4	4	4	4	4	4
4	2	2	3	3	2	3	**19**	2	3	2	2	3	3
5	3	3	2	3	2	2	**20**	3	3	4	3	3	4
6	4	4	4	4	4	4	**21**	2	3	2	2	2	3
7	3	3	3	4	3	4	**22**	2	1	2	2	2	2
8	2	2	2	2	3	2	**23**	3	3	2	3	3	2
9	3	2	2	3	2	2	**24**	4	4	3	4	4	4
10	2	2	2	2	2	2	**25**	2	2	2	3	2	1
11	2	2	2	3	2	2	**26**	4	4	3	4	4	3
12	3	3	3	4	3	3	**27**	2	2	2	2	2	2
13	4	4	4	5	4	5	**28**	3	3	2	2	3	2
14	3	3	3	3	3	3	**29**	2	2	2	2	1	2
15	4	4	3	3	4	3	**30**	4	4	4	5	4	3

where

$$T_j = \frac{1}{2}\sum_{k=1}^{a} M_{jk}(M_{jk}-1), \quad \text{for } j = 1, 2$$

P and Q are defined as in (5), which implies that ties are not counted.

Likewise, W modified for the presence of ties is equal to

$$W = \frac{\sum_{i=1}^{n}(R_i - (1/2)m(n+1))^2}{(1/12)m^2(n^3-n) - (m/12)\sum_{j=1}^{m}\sum_{k=1}^{a}(M_{jk}^3 - M_{jk})} \tag{16}$$

with, as before, $R_i = \sum_{j=1}^{m} r_{ij}$.

4. EXAMPLES

4.1. Artificial dataset

We created data X_{ij}, $i = 1, \ldots, 30$, $j = 1, \ldots, 6$, on a bounded ordinal scale $\mathbb{D} = \{1, \ldots, 5\}$. The data are realizations of the model $X_{ij} = LRD(Z_i + \varepsilon_{ij})$, with $Z_i \sim \mathcal{N}(0, 0.49)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, 0.09)$ (see Table I). The true ICC equals $0.49/(0.49 + 0.09) = 0.845$.

Using (14) and (15) we find $\hat{\sigma}_e^2 = 0.082$ and $\hat{\sigma}_p^2 = 0.43$. Consequently, ICC is estimated as 0.839. Another way to present the results is by reporting a table such as Table II. This table displays the distribution of the measurements X given the true value Z for $Z = LDR(k)$, $k = 1, \ldots, 5$. For example, an object that should be rated in class 2 has a 3% probability to be rated 1, 91% to be rated 2 and 6% to be rated 3.

To demonstrate the effect of the proposed map (12), we analyse the data using another map, namely

$$PDR(k) = \Phi^{-1}\left(\frac{k - 1/2}{a}\right) \tag{17}$$

We find $\hat{\sigma}_e^2 = 0.030$, $\hat{\sigma}_p^2 = 0.16$ and $\widehat{\text{ICC}} = 0.842$ (note that the estimated variances cannot be compared to 0.49 and 0.09 in the model, since choosing a different map in the analysis implies a different scale for the underlying domain of Z). The distribution of the measurements X is estimated as specified in Table III.

Kendall's τ can only be computed for pairs of columns, but it has been suggested to use the average of the pairwise τs as a measure of precision. Computing τ for all pairs of columns we find: (1, 2) 0.79; (1, 3) 0.66;

Table II. Distribution of X given Z

True value		Measurement X				
Z	Class	1	2	3	4	5
−2.20	1	1.00	0.00	0.00	0.00	0.00
−0.85	2	0.03	0.91	0.06	0.00	0.00
0.00	3	0.00	0.08	0.84	0.08	0.00
0.85	4	0.00	0.00	0.06	0.91	0.03
2.20	5	0.00	0.00	0.00	0.00	1.00

Table III. Distribution of X given Z, based on analysis using *PDR*

True value		Measurement X				
Z	Class	1	2	3	4	5
−1.28	1	0.99	0.01	0.00	0.00	0.00
−0.52	2	0.03	0.91	0.06	0.00	0.00
0.00	3	0.00	0.07	0.86	0.07	0.00
0.52	4	0.00	0.00	0.06	0.91	0.03
1.28	5	0.00	0.00	0.00	0.01	0.99

Table IV. Printer assembly data

	1	**2**	**3**	**4**	**5**	**6**		**1**	**2**	**3**	**4**	**5**	**6**
1	4	4	4	4	2	4	**14**	2	1	1	2	1	1
2	1	4	2	3	3	4	**15**	4	4	3	4	4	1
3	1	1	1	2	2	4	**16**	1	4	3	3	4	4
4	2	2	2	4	2	4	**17**	1	1	1	2	3	4
5	1	1	3	4	2	4	**18**	1	2	1	1	4	2
6	1	2	2	2	4	1	**19**	1	1	1	1	4	4
7	4	4	3	4	1	4	**20**	4	3	3	4	4	4
8	2	2	2	3	3	4	**21**	3	4	2	3	4	4
9	2	4	4	4	3	4	**22**	2	1	1	3	4	4
10	1	1	2	2	2	4	**23**	2	2	1	3	4	4
11	4	4	3	4	2	4	**24**	4	2	2	3	4	2
12	2	4	1	1	4	4	**25**	1	4	1	1	3	4
13	1	1	2	2	4	3	**26**	1	1	2	2	4	4

(1, 4) 0.72; (1, 5) 0.77; (1, 6) 0.54; (2, 3) 0.60; (2, 4) 0.63; (2, 5) 0.81; (2, 6) 0.63; (3, 4) 0.70; (3, 5) 0.66; (3, 6) 0.83; (4, 5) 0.65; (4, 6) 0.56; (5, 6) 0.62. The average of these values is 0.68. W, computed from (16), is 0.78.

4.2. *Printer assembly data*

The second example is a real dataset from a printer assembly line. After a printer has been assembled, its quality is tested by printing a grey area. This sample is visually inspected on uniformity by the operators. The samples are judged as *good*, *acceptable*, *questionable* or *rejected*. We code these categories as 1, 2, 3 and 4 respectively. In order to evaluate this inspection procedure, 26 samples (grey areas) were collected, which were judged six times. The data are given in Table IV.

We can imagine that underlying the operator judgments there is some continuous property *uniformity*, for which there is no known measurement method. We assume that this unobserved property has an unbounded domain, or at least that the bounds are removed far enough from the range of interest to make them irrelevant.

Table V. Printer assembly data: distribution of X given Z

True value		Measurement X			
Z	Class	1	2	3	4
−1.95	1	0.66	0.17	0.10	0.07
−0.51	2	0.39	0.21	0.19	0.21
0.51	3	0.21	0.19	0.21	0.39
1.95	4	0.07	0.10	0.17	0.66

Analysing the data, we find $\hat{\sigma}_e^2 = 4.06$, $\hat{\sigma}_p^2 = 0.867$ and $\widehat{\text{ICC}} = 0.18$ (using (17) instead of (12) we find the same value). Both from these results and from their implication as presented in Table V (probabilities being spread over several categories), we conclude that the inspection method is completely inadequate.

If one does not want to assume a continuous underlying variable, one could calculate W. Formula (16) yields 0.34. For each pair of columns one could compute τ, which yields: (1, 2) 0.45; (1, 3) 0.37; (1, 4) 0.59; (1, 5) −0.07; (1, 6) −0.04; (2, 3) 0.44; (2, 4) 0.36; (2, 5) 0.02; (2, 6) 0.12; (3, 4) 0.72; (3, 5) −0.17; (3, 6) 0.08; (4, 5) −0.26; (4, 6) 0.17; (5, 6) −0.22. The average value is 0.17.

5. DISCUSSION AND CONCLUSIONS

As is illustrated in the examples, τ and W are hard to interpret because it is difficult to assess the real-life implications of specific values. In part this is due to the fact that the statistics τ and W are not defined as estimators: they are given as sample statistics without a specified link to a parameter of the population distribution. These interpretation problems seem inherent to non-parametric methods. The modified ICC method, on the other hand, provides an easily interpretable evaluation. In particular, Tables II, III and V clearly demonstrate how a measurement system behaves in practice.

In the analysis of the printer assembly data (Table IV) it can be noted that the ratings in columns 1, 2, 3 and 4 have a moderate consistency, and that the ratings in column 5 and 6 are inconsistent mutually and with all other ratings (as can be concluded from the τ values which have been computed for all pairs of columns). For someone who is willing to improve the measurement system, this is an important indication. The ratings in columns 1 and 2 were made by a single rater, as were the ratings in columns 3 and 4, and 5 and 6. The ICC method facilitates only an overall evaluation. The possibility of a separate inter- and intra-rater evaluation would be a valuable extension of the method.

The existing methods for measurement system analysis cannot cope with measurement systems that measure on a bounded ordinal scale. The article proposes two approaches for this situation. The first approach requires bold assumptions. It defines a distance metric for the ordinal scale and a class of distribution functions in which the distribution of the measurement error is assumed. Both assumptions are derived from a latent variable model. Estimating the parameters of the distribution of the measurement error, precision can be evaluated as an intraclass correlation coefficient or from the estimated distribution of the measurement error. Given that the assumptions are approximately justified, the method is easily interpretable. If the assumptions cannot be justified, one has to resort to non-parametric methods, although the results of these are hard to translate into tangible implications.

REFERENCES

1. Allen MJ, Yen WM. *Introduction to Measurement Theory*. Brooks/Cole: Monterey, CA, 1979.
2. Montgomery DC, Runger GC. Gauge capability and designed experiments. Part I: Basic methods. *Quality Engineering* 1993; **6**:115–135.

3. Montgomery DC, Runger GC. Gauge capability and designed experiments. Part II: Experimental design methods and variance component estimation. *Quality Engineering* 1993; **6**:289–305.
4. Kerlinger FN, Lee HB. *Foundations of Behavioral Research* (4th edn). Harcourt: New York, 2000.
5. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores*. Addison-Wesley: London, 1968.
6. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing measurement reliability. *Psychological Bulletin* 1979; **86**:420–428.
7. Vardeman SB, VanValkenburg ES. Two-way random effects analyses and Gauge R&R studies. *Technometrics* 1999; **41**:202–211.
8. AIAG. *Measurement System Analysis; Reference Manual* (3rd edn). Automotive Industry Action Group: Detroit, MI, 2002.
9. Wheeler DJ. Problems with Gauge R&R studies. *ASQC Quality Congress Transactions* 1992; **46**:179–185.
10. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; **20**:37–46.
11. Conger AJ. Integration and generalization of kappas for multiple raters. *Psychological Bulletin* 1980; **88**:322–328.
12. Dunn G. *Design and Analysis of Reliability Studies*. Edward Arnold: London, 1989.
13. Kendall M, Gibbons JD. *Rank Correlation Methods* (5th edn). Edward Arnold: London, 1990.
14. Feldstein FL, Davis HT. Poisson models for assessing rater agreement in discrete response studies. *British Journal of Mathematical and Statistical Psychology* 1984; **37**:49–61.
15. Agresti AA. A model for agreement between ratings on an ordinal scale. *Biometrics* 1988; **44**:539–548.
16. Uebersax JS, Grove WM. A latent trait finite mixture model for the analysis of rating agreement. *Biometrics* 1993; **49**:823–835.
17. Vanleeuwen DM, Mandabach KH. A note on the reliability of ranked items. *Sociological Methods and Research* 2002; **31**:87–105.
18. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 1968; **70**:213–220.
19. Krippendorff K. Bivariate agreement coefficients for reliability of data. *Sociological Methodology*, Borgatta EF (ed.). Jossey-Bass: San Francisco, CA, 1970; 139–150.
20. Kendall M, Stuart A. *The Advanced Theory of Statistics* (4th edn), vol. 1. Charles Griffin: London, 1977; 77–82.

APPENDIX A. BIAS OF MEAN SQUARE ESTIMATORS WITH DISCRETE DATA

We study a sequence of random variables Z_i, $i = 1, 2, \ldots, n$ which have a normal distribution with mean μ and variance σ^2. The measurement system maps Z_i onto a discrete scale with class width 1. Values of Z_i in the interval $[k - \frac{1}{2}, k + \frac{1}{2})$ are mapped onto k, $k = \ldots, -1, 0, 1, 2, \ldots$. The discretized version X_i of Z_i has a discrete distribution given by

$$P(X_i = k) = \int_{k-1/2}^{k+1/2} f_{\mu;\sigma}(t)\,\mathrm{d}t, \quad k = \ldots, -1, 0, 1, 2, \ldots$$

Estimating μ by $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$, we study the bias of $S^2 = (1/(n-1))\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ as an estimator of σ^2. It can be shown that this bias is given by

$$\mathrm{E}S^2 - \sigma^2 = \sum_{k=-\infty}^{\infty} (k-\mu)^2 \left(\Phi\left(\frac{k+1/2-\mu}{\sigma}\right) - \Phi\left(\frac{k-1/2-\mu}{\sigma}\right)\right) - \sigma^2 \tag{A1}$$

The bias as given by formula (A1) above depends on σ and trunc(μ). For various values the bias is given in Table AI. From the perspective of an experimenter, trunc(μ) is uniformly distributed in [0, 1). For small σ^2 (coarse resolution) the expected bias is 0.083. For larger σ^2 (fine resolution) the bias approximates 0.083 regardless of trunc(μ). The value 0.083 is $\frac{1}{12}$ of Sheppard's correction[20]. We see that, irrespective of σ,

$$\mathrm{E}S^2 \approx \sigma^2 + \tfrac{1}{12}$$

It follows that $\mathrm{MS_w} - \frac{1}{12}$ is an unbiased estimator of σ_e^2.

Table AI. Bias of S^2 for various values of σ and trunc(μ)

	σ							
trunc(μ)	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	0.006	0.052	0.075	0.082	0.083	0.083	0.083	0.083
0.1	0.021	0.058	0.077	0.082	0.083	0.083	0.083	0.083
0.2	0.062	0.074	0.081	0.083	0.083	0.083	0.083	0.083
0.3	0.110	0.093	0.086	0.084	0.083	0.083	0.083	0.083
0.4	0.148	0.109	0.090	0.084	0.083	0.083	0.083	0.083
0.5	0.162	0.115	0.091	0.085	0.083	0.083	0.083	0.083
0.6	0.148	0.109	0.090	0.084	0.083	0.083	0.083	0.083
0.7	0.110	0.093	0.086	0.084	0.083	0.083	0.083	0.083
0.8	0.062	0.074	0.081	0.083	0.083	0.083	0.083	0.083
0.9	0.021	0.058	0.077	0.082	0.083	0.083	0.083	0.083
1.0	0.006	0.052	0.075	0.082	0.083	0.083	0.083	0.083

Since $m\bar{X}_{i\cdot}$ is normally distributed with mean $m\mu$ and variance $m^2\sigma_{\mathrm{p}}^2 + m\sigma_{\mathrm{e}}^2$ and since $m\bar{X}_i$ has the same resolution as the X_{ij}, we find

$$\mathrm{E}(m\mathrm{MS_b}) = m^2\sigma_{\mathrm{p}}^2 + m\sigma_{\mathrm{e}}^2 + \tfrac{1}{12}$$

Taking suitable linear combinations of $\mathrm{MS_b}$ and $\mathrm{MS_w}$ we obtain the estimator in (10).

Authors' biographies

Jeroen de Mast works as senior consultant at the Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA). He teaches courses in Six Sigma and supervises improvement projects in Dutch industry. This consultancy work is combined with scientific research in the field of industrial statistics. In 2002 he defended his PhD thesis *Quality Improvement from the Viewpoint of Statistical Method*. He is a member of the editorial board of *Quality Engineering* and since 2000 has been a member of the executive committee of the European Network for Business and Industrial Statistics (ENBIS).

Wessel N. van Wieringen obtained his MSc degree in Mathematics (Functional Analysis) at the University of Leiden in 1998. In 1999 he finished an MSc in the Mathematics of Nonlinear Models at the University of Edinburgh. In 2000 he joined the Institute for Business and Industrial Statistics of the University of Amsterdam as a PhD student and consultant. Ever since he has been occupied with research, resulting in the PhD thesis entitled *Statistical Models for the Precision of Categorical Measurement Systems*, as well as consultancy in industrial statistics. In 2004 he accepted a job as biostatistician at the Daniel den Hoed Cancer Clinic of the Erasmus Medical Centre.