A Comparison of Methods for the Evaluation of Binary Measurement Systems
Wieringen, Wessel N. van; van den Heuvel, Edwin

Published in:
Default journal

IMPORTANT NOTE: You are advised to consult the publisher’s version (publisher’s PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2005

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 07-11-2017
A Comparison of Methods for the Evaluation of Binary Measurement Systems

Wessel N. van Wieringen† and Edwin R. van den Heuvel§
Institute for Business and Industrial Statistics of the University of Amsterdam, Amsterdam, The Netherlands

Many quality programs prescribe a measurement system analysis (MSA) to be performed on the key quality characteristics. This guarantees the reliability of the acquired data, which serve as the basis for drawing conclusions with respect to the behavior of the key quality characteristics. When dealing with continuous characteristics, the Gauge R&R is regarded as the statistical technique in MSA. For binary characteristics, no such universally accepted equivalent is available. We discuss methods that could serve as an MSA for binary data. We argue that a latent class model is the most promising candidate.

Keywords Kappa; Intraclass correlation coefficient; Latent class model; Log-linear model; Measurement system analysis.

INTRODUCTION

It may happen that the quality of a part does not meet a desired level. The improvement of the quality of such a part is achieved by following the method of empirical research. Keystone to this method is the collection of measurements on which conclusions are based. Thus, the measurements are guiding in how to improve the quality. The quality of the measurements determines the usefulness of the measurements to empirical investigation. The quality of measurements is directly related to the uncertainty of measurement. The uncertainty of measurement (and consequently the quality of measurement) is assessed through a measurement system analysis (MSA) study. An MSA study prescribes an experiment that is designed to allow for the assessment of the uncertainty of measurement.

Traditionally (Wheeler and Lyday, 1989), an experiment for the evaluation of a measurement system is designed such that n parts are being judged by m raters, preferably repetitively, say l times. When dealing with continuous measurements, it is assumed that the outcome of the experiment can be modeled by a two-way, random-effects model. Let \( X_{ijk} \) the \( k \)th judgment of operator \( j \) on part \( i \), the model then is

\[
X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},
\]

where \( \mu \) is the overall mean common to all observations, \( \alpha_i \sim N(0, \sigma^2_{\alpha}) \), \( \beta_j \sim N(0, \sigma^2_{\beta}) \), \( \gamma_{ij} \sim N(0, \sigma^2_{\gamma}) \) and \( \epsilon_{ijk} \sim N(0, \sigma^2_{\epsilon}) \), and the random variables are independent for \( i = 1, \ldots, n, j = 1, \ldots, m \) and \( k = 1, \ldots, l \).

Then, the variance component \( \sigma^2_{\alpha} + \sigma^2_{\beta} + \sigma^2_{\gamma} \) is the operationalization of the uncertainty of measurement. It provides the information needed to construct a confidence interval for the true value of a part’s quality.

However, in practice, one often encounters measurement systems that are noncontinuous; for instance, a quality characteristic that only assumes two values, a so-called binary measurement. In an industrial setting, such a situation may arise when quality of a part is evaluated through visual inspection, resulting in parts labeled either defective or nondefective. Assuming the aforementioned experimental design is also the most sensible for binary measurements, it is not clear how the outcome of the experiment should be modeled. Moreover, this model should give rise to an operational definition of the uncertainty of measurement in the binary context.

In this article, we model the outcome of an MSA experiment of binary measurement system by means of a latent class model (in line with Boyles, 2001). Within this latent class model, we give an operational...
comparison of uncertainty of measurement. Next, we compare the latent class model with alternative approaches\(^1\) that are currently used to evaluate the measurement system, namely the kappa statistic, the intraclass correlation coefficient, and log-linear models. This comparison shows the best way to analyze an MSA study for binary measurements. We conclude with a case study that illustrates all techniques discussed.

### LATENT CLASS MODEL

Consider an MSA experimental design as described in the introduction, where \(n\) parts are being judged by \(m\) raters only once (for the purpose of comparing several techniques, it is most illustrative to disregard repetitions). Ideally, the sample of parts involved is a good representation of the parts one expects to encounter in later investigations. So far, we do not know the quality of measurements. Therefore, theoretically, we have no information of the true state, henceforth called \(Y\), of any part. As parts are selected randomly, we regard \(Y\) as a random variable, taken to be Bernoulli distributed with unknown parameter \(\theta = P(Y = 1)\), the probability of the part being of good quality.

Let a sample of \(m\) raters be involved in the MSA experiment. Associated with each rater, we define a random variable \(X_j(Y)\), the judgment of rater \(j\), which depends on \(Y\), the true state of the judged part. \(X_m\) is also Bernoulli distributed, with parameter \(\pi_j(y) = P(X_j = 1|Y = y)\), where we have dropped (for notational convenience) the argument \(Y\) in \(X_j(Y)\). The unconditional probability that a randomly selected part is judged as \(x_j \in \{0, 1\}\) by rater \(j\) is then given by

\[
p_j(x_j) = P(X_j = x_j) = P(X_j = x_j|Y = 0)P(Y = 0) + P(X_j = x_j|Y = 1)P(Y = 1)
\]

(2)

Latent class analysis distinguishes between a manifest variable (the judgment of a rater) and an unobserved, latent variable (the true value of the part). The latter is used to explain the correlated structure in the (observed) former. Crucial to this approach is that it assumes conditional independence. That is, given the realization of the latent variable, the manifest variables are independent of one another. Conditional independence can be formulated as

\[
P(X_1, X_2, \ldots, X_m|Y) = \prod_{j=1}^{m} P(X_j|Y), \quad (3)
\]

(i.e., given the true state of the part, the raters judge independently).

Because both the observed and latent variable are Bernoulli, the unconditional probability that rater \(j\) judges a part as good can be written as in Eq. 2. Using this and Eq. 3, we can specify the model underlying the latent class analysis. Hereto, let \(X\) denote the \(n \times m\) matrix, containing the data from the MSA experiment, with \(X_{ij}\) the judgment of rater \(j\) on part \(i\), given by:

\[
X = \begin{pmatrix}
X_{11} & \ldots & X_{1m} \\
\vdots & \ddots & \vdots \\
X_{n1} & \ldots & X_{nm}
\end{pmatrix}
\]

We repeat that, for illustration purposes, we assume the operators to judge each part only once. However, the method described here is naturally extended to the situation where multiple judgments are made by each operator, as one expects in a regular MSA experiment for continuous measurements.

Then, the likelihood function of the joint response of the raters of the sample, \(X\), becomes

\[
L(X; \theta, \pi_1(1), \ldots, \pi_m(0)) = \prod_{i=1}^{n} \left( (1-\theta) \prod_{j=1}^{m} (\pi_j(0))^{X_{ij}} (1-\pi_j(0))^{1-X_{ij}} \right)^{\lambda_{ij}} + \theta \prod_{j=1}^{m} (\pi_j(1))^{X_{ij}} (1-\pi_j(1))^{1-X_{ij}}, \quad (4)
\]

where we substituted \(P(Y = 1) = \theta\) and \(P(X_{ij} = 1|Y = y) = \pi_j(y)\) for all \(j\) and \(y\). In addition, we impose restrictions on the model to ensure identification because this is not automatically guaranteed. To see this, choose any vector of parameters \(\Psi' = (\theta', \pi_1'(1), \ldots, \pi_m'(1), \pi_1'(0), \ldots, \pi_m'(0))^T\), and define

\[
\Psi^* = (1-\theta', \pi_1'(0), \ldots, \pi_m'(0), \pi_1'(1), \ldots, \pi_m'(1))^T. \quad (5)
\]

Then, for any response pattern, due to symmetry in the density function:

\[
P(X = x; \Psi') = P(X = x; \Psi^*). \]

\(^1\)We realize that the automotive industry (Automotive Industry Action Group, 2002) has prescribed a way to conduct an attribute gauge study. It assumes the qualitative evaluation of a part can be compared with an underlying continuous measurement. However, as knowledge of this continuous measurement is not always at hand, we disregard it here.
Therefore\(^2\) in the particular case where each operator makes only one judgment, we need to involve at least three operators and require that \(\theta \in (0, 1)\) and \(1 \geq \pi_j(1) > \pi_j(0) \geq 0\) for all \(j\). The restriction \(\pi_j(1) > \pi_j(0)\) follows naturally because it merely states that the probability of operator \(m\) judging a good part as such is higher than the probability that he judges a bad part as good. Now, Eq. (4) plus these restrictions enable one to use a maximum likelihood procedure to estimate all parameters, confer Bartholomew and Knott (1999) and Boyles (2001).

To find a maximum likelihood estimate for \(\Psi\), instead of applying the traditional Newton-Raphson algorithm, the so-called E-M algorithm is used (see the Appendix). It has been shown that the sequence of estimates produced by the E-M algorithm converges to a maximum of the likelihood function, confer McLachlan and Krishnan (1997).

**OPERATIONAL DEFINITION**

The latent class approach also allows for a natural operationalization of the uncertainty of measurement. The uncertainty of measurement of binary measurements is that of misclassification. That is, a part may have been measured as belonging to one category, whereas preferably the evaluation of the measurement system—an operational definition of the uncertainty of measurement to be related to the probability of misclassification. However, the probability of misclassification itself depends on the sample one studies, whereas preferably the evaluation of the measurement system is independent of the sample. Compare the continuous case where the variance component (which is the operationalization of uncertainty of measurement there) is independent of the mean and variation of the parts in the sample. Therefore, we adopt from Uebersax (1988) the terms sensitivity defined as the probability that a good part is rated as such, and specificity as the probability that a bad part is rated as bad. These definitions coincide with the indices \(\pi_j(1)\) and \(1 - \pi_j(0)\). This operational definition allow us, given the quality of the sample, to calculate the probability of misclassification. Thus, giving a clear insight into the consequences of this measurement system (i.e., the number of parts judged wrongly). Its importance is eminent in the context of acceptance sampling on attributes when meeting requirements as AQL, LQL, AOQL, etc. Also, in the context of quality improvement knowledge of the probability of misclassification is of importance when it comes to drawing conclusions with respect to key factors influencing the key quality characteristic.

Once the parameters \(\pi_1(0), \ldots, \pi_m(0)\) and \(\pi_1(1), \ldots, \pi_m(1)\) have been estimated, it is straightforward to calculate the probability of misclassification. Assuming, for simplicity (other distributions only ask for small changes), that during regular production all raters judge an equal share of the parts, then, for any quality \(\theta\) of the sample, the probability of misclassification is

\[
P(\text{incorrect decision}) = \frac{1}{m} \sum_{j=1}^{m} \left( \theta(1 - \pi_j(1)) + (1 - \theta)\pi_j(0) \right)
\]

In addition, on the individual part level, one can indicate which category the part most likely originates from. That is category \(y\) that maximizes: \(P(Y = y | X_1, X_2, \ldots, X_m)\).

Besides these favorable features of the latent class model, it has two main drawbacks. It requires many measurements to give an accurate estimation of the parameters. However, this is inherent to binary data. Second, for the estimation of the parameters an algorithm is needed (hardly a problem in the present computer age).

**ALTERNATIVE METHODS**

**Measure of Agreement**

An alternative to the latent class model is Cohen’s kappa (Cohen, 1960). This statistic has been proposed as a useful statistic for the evaluation of categorical measurement systems, confer Boyles (2001), Futrell (1995), and Dunn (1989). Recently, the statistical software package Minitab reports the kappa in its Attribute Gauge R&R study for the evaluation of the measurement system.

Many measures representing the quality of the measurement system have been proposed, confer Goodman and Kruskal (1954) and the review papers of Landis and Koch (1975a) and (1975b). Cohen (1960) introduced a measure of agreement called the kappa. There is agreement if two measurements (on the same part) are equal. The kappa represents the
degree of agreement between two raters, based on how they classify a sample of parts into a number of categories. Cohen’s observed agreement may be due to chance. Hence, Cohen defined the kappa, denoted \( \kappa \), corrected for agreement by chance and normalized, as

\[
\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{7}
\]

Here, \( P_o \) is the observed proportion of agreement and \( P_e \) the expected proportion of agreement due to chance. Kappa attains the value one when there is perfect agreement, zero if all observed agreement is merely due to chance, and negative values when the amount of agreement is less, then is to be expected on the basis of chance. Often, the observed proportion is used to evaluate the judgment process. However, \( P_o \geq \kappa \), and thus \( P_o \) always gives a better impression of the measurement system than when chance, agreement is taken into account.

As a comparison, consider an exam with only multiple-choice questions. Its grades are calculated in accordance with Eq. 7. That is, the proportion of questions the examinee answered correctly, \( \text{Pe}_i \) got right had he chosen his answers randomly, \( P_e \). This difference is scaled such that grades fall in the aimed range.

Cohen (1960) specified, for any pair of raters \( j_1 \) and \( j_2 \), the terms in Eq. 7 as

\[
P_o = \sum_{x=0}^{1} p_{i,j_1,j_2}(x, x) \quad \text{and} \quad P_e = \sum_{x=0}^{1} p_{i,j_1}(x) p_{i,j_2}(x).
\]

Here, \( P_o \) is the proportion of parts with matched judgments of raters \( j_1 \) and \( j_2 \) and \( p_{i,j_1,j_2}(x, x) \) denoted the proportion parts that have been judged as \( x \) by raters \( j_1 \) and \( j_2 \). \( P_e \) is the expected proportion of agreement based on the individual marginal distributions of each rater. The marginal proportion for rater \( j \) and category \( x \) is denoted by \( p_{j,x}(x) \).

Due to the way \( P_e \) is calculated, \( \kappa \) may give values that are counterintuitive. For instance, suppose that all raters measure most parts in the same category (small part variation). Then, \( \kappa \) is small, as \( P_e \) is large. Thus, \( \kappa \) confounds to some extent uncertainty of measurement of the measurement system with part variation. Similarly, let one rater measure almost all parts in one category, and the other rater almost all of them in a different category (systematic rater difference). Then, \( P_e \) approaches its minimum and causes a relatively high \( \kappa \). Thus, whereas \( \kappa \) is designed to measure systematic rater differences, it ignores them to some extent. These are called the paradoxes of the kappa, confer Cicchetti and Feinstein (1990) and Feinstein and Cicchetti (1990). In this context, it has been argued (Brennan and Prediger, 1981) to define agreement by chance as completely random (i.e., the raters assign the parts to any category with equal probability).

Thus, in line with a contingency table setting, Cohen assumes an independence model based on the individual marginal proportions. He then constructs a sample statistic, \( \kappa \), of which he claims measures the degree of agreement between the raters. As suggested by Futrell (1995) and Dunn (1989), this statistic can be used to evaluate the quality of measurements. However, the relation between degree of agreement and uncertainty of measurement has never been specified. This, together with \( \kappa \)'s paradoxes, makes it unsuitable for the evaluation of the quality of measurements.

If one uses the \( \kappa \) method, the following criteria can be applied that apparently guarantee a reliable measurement system. Landis and Koch (1977) give a table (see Table 1) that expresses the relationship between the value of \( \kappa \) and the corresponding evaluation of the measurement system.

Although as they suggest themselves, their classification is rather arbitrary. Another approach is to test \( H_0 : \kappa = 0 \) against \( H_A : \kappa \neq 0 \), thus testing whether agreement is substantial, or merely due to chance. For more on test procedures and moments of the \( \kappa \), check Everitt (1968) and Hubert (1977).

**KAPPA FOR MULTIPLE RATERS**

It is not straightforward to extend the kappa statistic to more than two raters case. We point out shortly how this is done. Because at least two people are necessary for agreement, Fleiss (1971) suggested the degree of agreement may be expressed in terms of the proportion of agreeing pairs. If there are \( m \) raters, then the maximum possible number of agreeing pairs per part equals \( \frac{1}{2} m(m - 1) \). To estimate the proportion of agreeing pairs per part, Fleiss (1971) proposed the sum of the number of agreeing pairs per category. Then, to get an overall estimate for the observed proportion of agreement, we take the sum over the parts of all these proportions, scaled with the total number of possible agreeing pairs. In formula

\[
P_o = \frac{1}{nm(m - 1)} \left( \sum_{i=1}^{n} \sum_{x=0}^{1} n_i(x)(n_i(x) - 1) \right),
\]

with \( n_i(x) \) the number of times part \( i \) has been classified as \( x \). The expected proportion of agreement is
Comparison of Methods for Evaluation of Binary Measurement Systems

given by

\[ P_e = \frac{2}{m(m-1)} \sum_{j_1,j_2=1}^{m} \sum_{x=0}^{1} \frac{p_{j_1}(x)p_{j_2}(x)}{C_0} \]

It should be clear that each pair of raters enters the sum only once. \( P_e \) estimates, under the assumption of independence, the probability that two randomly selected raters classify a part into the same category, based on the individual marginal proportions of the raters. Note that we have adopted Conger (1980) here instead of Fleiss (1971). The main difference with Fleiss (1971) is that Conger (1980) allows the raters to have different marginal distributions and calculates \( P_e \) without rater replacement. This has the advantage that it is conceptually in line with Cohen (1960). This is illustrated by the fact that it equals the average of all Cohen’s (1960) pairwise kappa’s, if either there is independence between all raters or their marginal probabilities are equal. One may generalize this by looking at other tuples of agreeing raters.

KAPPA STATISTIC VERSUS LATENT CLASS MODEL

The latent class model is a model for the outcome of an MSA experiment, whereas \( \kappa \) is merely a sample statistic. Using the latent class model, we can rewrite \( \kappa \) in terms of the parameters of the latent class model. We limit ourselves to two raters, mainly to avoid cumbersome notational issues. Then, the observed agreement is the probability that both raters make the same judgment:

\[ P_o = P(X_1 = X_2) = \sum_{x,y=0}^{1} |1 - y - \theta|(x - \pi_1(y))(x - \pi_2(y)) \]

and the expected proportion of agreement (i.e., the probability that by chance the raters judge similarly), which is given by

\[ P_e = \sum_{x=0}^{1} P(X_1 = x)P(X_2 = x) = \sum_{x=0}^{1} p_1(x)p_2(x), \]

with \( p_j(x) \) given by Eq. (2). It should now be clear how to reformulate Eq. (7) in terms of the latent class parameters. This yields a \( \kappa \) that depends on the \( \pi_j(y) \) and on the process parameter \( \theta \). We have displayed this graphically, for an arbitrary choice of the \( \pi_j(y) \), by plotting \( \kappa \) against \( \theta \), see Figure 1.

Thus, for one and the same measurement system \( \kappa \) can differ substantially from one MSA experiment to another depending on the quality of the sample involved. This may result in an evaluation of the measurement system that is at odds with evaluations based on MSA experiments involving other samples.

Given the fact that kappa depends on the process parameter \( \theta \), one may argue that the criteria on the kappa, as proposed in Landis and Koch (1977), should be adjusted accordingly, confer Elffers (2001). It remains unclear how this should be done. Moreover, because the criteria themselves are arbitrary, so will their adjustments be.

The latent class model is a model for the outcome of the MSA experiment, whereas kappa is merely a summary statistic, trying to summarize all aspects of a measurement system into one number, an almost impossible task. Moreover, when it indicates that the measurement system is not up to standard, it provides no clues as to how this has arisen. The latent class model yields information about the individual rater performances, thus given insight how discrepancies between judgments come about.

INTRACLASS CORRELATION COEFFICIENT

For continuous measurement systems, the intraclass correlation coefficient is often used to assess the uncertainty of measurement (Shrout and Fleiss, 1979). The intraclass correlation coefficient measures the correlation among multiple measurements of the same part. The intraclass correlation coefficient can also be used to evaluate the uncertainty of measurement of binary measurement systems.

For binary measurements, the intraclass correlation coefficient is called the \( \phi \) coefficient and (for two raters) defined as

\[ \phi = \frac{\text{Cov}(X_{11}, X_{12})}{\sqrt{\text{Var}(X_{11}) \cdot \text{Var}(X_{12})}} = \frac{P(X_{11} = 1, X_{12} = 1) - p_1 p_2}{\sqrt{p_1 (1 - p_1) p_2 (1 - p_2)}}. \]

As other product moment correlation coefficients \( \phi \) only assumes values in the interval \([-1,1]\).

The \( \phi \) coefficient is estimated by replacing all the terms in the righthand side of Eq. (9) by their corresponding estimates: \( p_1 = \frac{1}{n} \sum_{i=1}^{n} X_{1i}, p_2 = \frac{1}{n} \sum_{i=1}^{n} X_{2i} \), and \( P(X_{11} = 1, X_{12} = 1) \) by \( \frac{1}{n} \sum_{i=1}^{n} X_{1i} X_{2i} \).

For the situation involving \( m > 2 \) raters, Fleiss (1965) and Bartko and Carpenter (1976) propose to
evaluate the reliability by means of the average of the 
\( \phi \) coefficients of all possible rater pairs, where they 
assume that \( p_j = p \) for \( j = 1, \ldots, m \). The \( \phi \) coefficient 
for multiple raters is then estimated by \( \phi = (P - p^2)/(p - p^2) \), where

\[
P = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}
\]

and

\[
P = \frac{2}{nm(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} X_{ij} X_{ij'}
\]

When using \( \phi \) as the statistic representing the quality of measurements, from Wheeler and Lyday (1989) one can distract the following criteria for \( \phi \) (Table 2).

The criteria in Table 2 apply to intraclass correlation coefficients for continuous measurements. We assume they can be used for \( \phi \) coefficient.

**INTRACLASS CORRELATION COEFFICIENT VERSUS LATENT CLASS MODEL**

As with the kappa statistic, we use the latent class model to study the intraclass correlation coefficient and restrict ourselves to the two rater case. To this extent, the numerator of \( \phi \) becomes

\[
\text{Cov}(X_{i1}, X_{i2})
\]

\[
= \sum_{x_1, x_2} \left( (x_1 - p_1(1))(x_2 - p_2(1)) \times P(X_1 = x_1, X_2 = x_2)
\]

\[
= \theta(1 - \theta)(p_1(1) - p_1(0))(p_2(1) - p_2(0)),
\]

and its denominator is \( \sqrt{\text{Var}(X_{i1}) \cdot \text{Var}(X_{i2})} \), where
The overall formula for $\phi$ is not a transparent expression. We therefore resort to visual means to portray the relation between $\phi$ and the parameters of latent class model. The surface in Figure 2 represents $\phi$ against $\pi_1(1)$ and $\pi_2(1)$ (to obtain a three-dimensional graph we have fixed $\theta$ and taken $\pi_j(0) = 1 - \pi_j(1)$ for all $j$). This corresponds with the intuitive idea of $\phi$: $\phi = 1$ if the raters measure similarly (i.e., $\pi_j(1)$ equals 1 for all $m$) and $\phi = 0$ if the raters both rate randomly (i.e., $\pi_j(1) = 1/2 = \pi_2(1)$).

Figure 3 shows that $\phi$ (like $\kappa$) depends on $\theta$. Although $\phi$ measures, given a sample, the correlation between two variables, this dependence may lead to different evaluations of the measurement process.

Furthermore, like $\kappa$, $\phi$ is a summary statistic, providing only aggregated information, which is of limited use when the measurement system needs improvement.

**LOG-LINEAR MODEL**

Instead of defining a measure for agreement, Tanner and Young (1985) model agreement. They use the rater measurements to construct a contingency table. The cells of this table are modeled by a log-linear model with two components: one representing the effect of chance, and the other representing the effect of rater agreement.

Let $X_i = (X_{i1}, X_{i2}, \ldots, X_{im})$ be the measurements of the $m$ raters on part $i$. For each $m$-tuple, $x = (x_1, x_2, \ldots, x_m)$ with $x_j \in \{0, 1\}$, define $n(x) = \sum_{j=1}^{m} \#(X_j = x)$. $n(x)$ is the number of times $m$-tuple $x$ appears in the measurement system analysis experiment. Tanner and Young assume $n(x)$ is strictly positive. This is a remarkable assumption. When dealing with a measurement system of high quality, one expects to find (mainly) the patterns $x = (0, 0, \ldots, 0)$ and $x = (1, 1, \ldots, 1)$. Therefore, one would expect patterns for which $n(x)$ equals 0.

Tanner and Young consider the $n(x)$ as the cells of a contingency table. Table 3 visualizes this for two raters. The main diagonal cells of the contingency table represent the agreement between the raters. Tanner and Young study agreement by comparing the frequencies in these diagonal cells to the expected cell count under an independence model (i.e., all raters measure independently and their marginal distributions yield the expected number of times $x$ will occur).

Conventionally, contingency tables are modeled by log-linear models. Therefore, the independence model is given by

$$\ln(\text{En}(x)) = u + \sum_{j=1}^{m} u_j(x_j). \quad (9)$$

Tanner and Young call $u$ the overall effect and $u_j(x_j)$ the effect of category $x_j$ of the $j$th rater. Eq. (9) is an alternative way of stating that the cell counts are explained by the marginal proportions of the raters. From this perspective, $u_j(x_j)$ can be viewed as the difference between the proportion of rater $j$ measuring a part as $x_j$ and the overall proportion. Added to Eq. (9) should be the restriction

$$\sum_{x_j=0}^{1} u_j(x_j) = 0 \quad \text{for all } j. \quad (10)$$

This makes Eq. (9) identifiable and ensures the marginal proportions sum to 1 for each rater.

A second term is added to Eq. (9), which accounts for the discrepancies between the observed and expected cell counts of the diagonal cells:

$$\ln(\text{En}(x)) = u + \sum_{j=1}^{m} u_j(x_j) + \delta(x), \quad (11)$$

with

$$\delta(x) = \begin{cases} c & \text{if } x \text{ is a diagonal cell} \\ 0 & \text{otherwise} \end{cases}$$

where $c$ is a constant that reduces the discrepancy between the observed and the expected cell count of
the diagonal cells. Tanner and Young interpret $c$ as the effect due to agreement among the raters.

Estimates of the parameters are obtained by a maximum likelihood procedure, where it is assumed that the contingency table can be described by a multinomial distribution. A significant discrepancy between the observed diagonal cells and their expected cell count under the independence model corresponds to the significance of the agreement. The significance of the discrepancy (and thus of the agreement) is assessed by testing whether Eq. (11) fits the data significantly better than Eq. (9).

**LOG-LINEAR MODEL VERSUS LATENT CLASS MODEL**

For the comparison between the log-linear model approach and the latent class model, we again limit ourselves to the two rater case. The model then becomes, in line with Eq. 11:

$$\ln\{En(x)\} = u + (-1)^{(x)}u_1 + (-1)^{(x)}u_2 + \delta(x),$$

where

$$\delta(x) = \begin{cases} c & \text{if } x \text{ is a diagonal cell} \\ 0 & \text{otherwise,} \end{cases}$$

To see whether the model actually describes what it claims, we have rewritten the agreement contribution $c$ in terms of the latent class parameters:

$$c = \frac{1}{4} \log \left( \frac{E(n(0,0)) \cdot E(n(1,1))}{E(n(0,1)) \cdot E(n(1,0))} \right)$$

with

$$E(n(x)) = n \left\{ \frac{1}{2} \sum_{j=1}^{2} (X_j - \pi_j(1))^2 + (1 - \frac{1}{2}) \sum_{j=1}^{2} (X_j - \pi_j(0))^2 \right\}.$$
As before this is a term that represents agreement and should preferably be independent of the process parameter \( \theta \). Plotting \( c \) against \( \theta \) (see Figure 4) reveals they are not unrelated. Here, the \( \pi_j(y) \) are fixed as for the kappa in Figure 1. As with the kappa, we consider this (read: \( c \) and \( \theta \) related) a compelling argument against the use of the previous approach for the evaluation of measurement systems on the basis of this agreement model.

**EXAMPLE**

At an engine manufacturer components are examined on dirt, for too much dirt might cause an engine to breakdown. For the purpose of examination, a tape is affixed to the component. The tape is detached and put under a microscope, magnified 30 times, and photographed. The photograph is compared with a number of references, ranging from clean to contaminated. These references are divided into two categories, one representing the approved (clean) surfaces, and the other the rejected (contaminated) surfaces. A rater decides which reference the photograph resembles best.

<table>
<thead>
<tr>
<th>Table 3 Results of the Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
</tr>
<tr>
<td>Rater B</td>
</tr>
<tr>
<td>Rater A</td>
</tr>
</tbody>
</table>

![Figure 3. \( \phi \) vs. \( \theta \).](image-url)
indirectly judging whether the component is suitable for production or needs to be cleaned first.

To evaluate the uncertainty of measurement of this measurement system we have set up an experiment where 34 raters judged 20 parts, according to the previously described procedure, in random order. Note that per component we gathered only one tape, which then was judged by all raters, thus excluding sample issues. The data have been reproduced in Table 4. We now apply the discussed techniques.

Using the E-M algorithm as in McLachlan and Krishnan (1997), in Eq. which maximizes the likelihood function in Eq.(4), we end up with $\theta = 0.13$, $\hat{\pi}_A(1) = 0.99$, $\hat{\pi}_B(1) = 0.99$, $\hat{\pi}_C(1) = 0.89$, $\hat{\pi}_A(0) = 0.42$, $\hat{\pi}_B(0) = 0.20$, and $\hat{\pi}_C(0) = 0.50$. These estimated parameters show the individual rater performances. All operators are good at judging a good part as such. Operators $A$ and $C$ have a tendency to mistake bad parts for good parts, giving too optimistic an impression of the sample.

We realize that all previous comparisons of techniques are illustrated by the two rater case, and that the example involves three raters. This inconsistency is due to the following reasons: (1) only for three (or more) raters is the latent class model identifiable, (2) all methods are also applicable to the multiple rater case, and (3) the differences between the latent class model and the alternative techniques also hold for the multiple rater case.

---

**Figure 4.** Agreement contribution $c$ against $\theta$. 

---
are given. They hardly deviate from the observed frequencies.

Also in Table 5 are the expected frequencies according to the log-linear agreement model (including $c$, with $\hat{c} = 0.63$). When we incorporate this term in the model, the Pearson $\chi^2$ goodness-of-fit statistic changes from 0.58 to 0.79. Neither exceeds the corresponding $x$-level of 0.05. It is unclear what this means for the quality of the measurement system.

The intraclass correlation coefficient and kappa for the three raters combined and each possible pair are given in Table 6. All these indices are way off their ideal value. This provides no information on the individual rater level to see who needs more attention to improve the measurement process now that it has been evaluated negatively.

All the previous methods confirm what the “eyeball test” already suggests, namely a rather poor measurement system for the engine components data. It is only the latent class method that gives a clear picture of the consequences of applying this measurement system in practice and provides clues for improving it.

CONCLUSION

We briefly touched on the subject of measurement system analysis when dealing with continuous measurements. Then, we noted that MSA for binary measurements is hardly dealt with in literature, although often encountered in practice. For the MSA experiment with binary measurements, we adopted the design of the continuous setting: each rater involved in the experiment judges all selected parts, preferably repetitively. We introduced the latent class model to model the outcome of such an experiment. This model involves parameters that have a clear interpretation. Furthermore, in the paradigm of this model, we gave an operational definition of uncertainty of measurement for binary measurements, relating directly with the parameters of the model. Once all parameters have been estimated, we have a clear insight into the consequences of applying this measurement system. This we propose should serve as the basis for the evaluation of the measurement system.

We compared the latent class model with alternative approaches. This comparison showed that $\kappa$, $\phi$, and $\delta$ (of the log-linear model approach) are sample statistics that—compared with the latent class model—give little insight in the quality of measurement (e.g., a low kappa gives no clues how to improve the measurement system, or how many parts are judged wrongly). Moreover, $\kappa$, $\phi$, and $\delta$ depend on the quality of the sample involved in the MSA experiment (whereas the latent class method does not). This may result in different evaluations of the quality of measurement of the same measurement system when different samples are used. Clearly, we prefer the latent class model for evaluation of binary measurement systems.
Disregarded here is the fact that there are two sides to a reliable measurement system. The Gauge R&R for continuous data distinguishes between repeatability and reproducibility. For binary data, we have within- and between-rater consistency. Although this article only deals with the latter, the former is as important. To evaluate the repeatability, the raters need to judge the parts repetitively. If repeated measurements are available, it is hard to see how \( \kappa, \phi \), and the log-linear model approach would deal with these additional data. Only the latent class model has a natural extension to experiments with repeated judgments.

We recommend the latent class model as a useful tool in assessing the quality of a measurement system with a binary response. Hence, empirical investigations involving MSA will benefit greatly from the binary analog of the Gauge R&R.

REFERENCES


APPENDIX

Estimation by Means of the E-M Algorithm

The latent class model in this article involves a latent variable. A variable that is (by assumption) unobservable in nature. Theoretically, however, one could (by means of some hypothetical experiment) have information available about the true state of this variable. This lack of information can be viewed as a case of “missing data.” Here, only the observations of the raters are available, whereas the true state of the parts is not revealed.

To deal with the phenomena of missing data, we introduce a new variable:

\[
Z_{i,y} = \begin{cases} 
1 & \text{if part } i \text{ is truly a } y \\
0 & \text{if part } i \text{ is truly not a } y 
\end{cases}
\]

The vector \( Z \) is an indicator column, indicating from which category a part stems. Thus, where \( X \) is missing information, \( (X, Z) \) is the complete data from the experiment.
The corresponding likelihood function for the complete data situation is

\[ L_c = f_c((\mathbf{X}, \mathbf{Z}); \Psi) \]

\[ = \prod_{i=1}^{n} \left((1 - \theta) \prod_{j=1}^{m} (\pi_j(0))^{X_i} (1 - \pi_j(0))^{1-X_i} \right)^{Z_{i0}} \times \left(\theta \prod_{j=1}^{m} (\pi_j(1))^{X_i} (1 - \pi_j(1))^{1-X_i} \right)^{Z_{i1}}. \quad (1) \]

Because only either \(Z_{i0}\) or \(Z_{i1}\) can be equal to one, only one of them enters the likelihood function. This results in a complete-data log-likelihood function that has a rather convenient form.

To find the maximum likelihood estimates of \(\Psi\), instead of applying the traditional Newton-Raphson algorithm, the so-called E-M algorithm is used. The E-M algorithm approaches the problem of maximizing the incomplete likelihood function indirectly by exploiting the more convenient form of the complete data likelihood function.

The E-M algorithm can be described as follows:

**STEP 1**

Choose initial values for the parameter \(\Psi\) and specify a stopping criterion.

**STEP 2 (REFERRED TO AS THE E-STEP)**

Using the current fit of \(\Psi\) to replace \(\mathbf{Z}\) by its conditional expectation given \(\mathbf{x}\):

\[ \hat{Z}_{i0} = E_{\Psi}(Z_{i0}|X) \quad \text{and} \quad \hat{Z}_{i1} = E_{\Psi}(Z_{i1}|X) \]

In the present situation, these are probabilities that are complementary, (i.e., \(\hat{Z}_{i1} = 1 - \hat{Z}_{i0}\)). In fact, we substitute real numbers from the interval \([0,1]\) for \(\mathbf{Z}\), whereas their proper value is either 0 or 1.

Furthermore, using Bayes’ theorem, rewrite

\[ \hat{Z}_{i1} = E_{\Psi}(Z_{i1}|x) \]

\[ = \frac{P_{\Psi}(X|Y_i = 1) P(Y_i = 1)}{P_{\Psi}(X)} \] \quad (2)

Thus, given the present estimate of \(\Psi^{(k)}\), we calculate this probability by

\[ \hat{Z}_{i1}^{(k+1)} = \theta^{(k)} \prod_{j=1}^{m} \left(\pi_j^{(k)}(1)\right)^{X_i} \times \left(1 - \pi_j^{(k)}(1)\right)^{1-X_i} / \hat{Z}_{i0}^{(k+1)} \]

\[ \times \left(1 - \pi_j^{(k)}(1)\right)^{1-X_i} / \hat{Z}_{i1}^{(k+1)} \]

**STEP 3 (REFERRED TO AS THE M-STEP)**

The M-step consists of maximizing the log-likelihood function, which is now a linear combination of functions, each depending on only one of the parameters. Taking the first-order partial derivatives, equaling them to zero and solving these equations, we arrive at the following estimators:

\[ \theta^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} Z_{i1}^{(k+1)} \]

\[ \pi_j^{(k+1)}(0) = \frac{\sum_{i=1}^{n} X_{ij} Z_{i0}^{(k+1)}}{\sum_{i=1}^{n} Z_{i0}^{(k+1)}} \]

\[ \pi_j^{(k+1)}(1) = \frac{\sum_{i=1}^{n} X_{ij} Z_{i1}^{(k+1)}}{\sum_{i=1}^{n} Z_{i1}^{(k+1)}} \]

Thus, we obtain an MLE of \(\Psi\) using the complete data likelihood.

**STEP 4**

Go back step 2 until the stopping criterion of the algorithm is satisfied.

It has been shown that the sequence of estimates produced by the E-M algorithm converges to a maximum of the original (“incomplete data”) likelihood function, confer McLachlan and Krishnan (1997).