Special Issue Article

Quality and Reliability Engineering International

(wileyonlinelibrary.com) DOI: 10.1002/qre.2023

Published online 13 June 2016 in Wiley Online Library

Quantifying the Random Component of Measurement Error of Nominal Measurements Without a Gold Standard

T. S. Akkerhuis*† and J. de Mast

It is well known that measurement error of numerical measurements can be divided into a systematic and a random component and that only the latter component is estimable if there is no gold standard or reference standard available. In this paper, we consider measurement error of nominal measurements. We motivate that, on a nominal measurement scale too, measurement error has a systematic and a random component and only the random component is estimable without gold standard.

Especially in literature about binary measurement error, it is common to quantify measurement error by 'false classification probabilities': the probabilities that measurement outcomes are unequal to the correct outcomes. These probabilities can be split up in a systematic and a random component. We quantify the random component by 'inconsistent classification probabilities' (*ICP*s): the probabilities that a measurement outcome is unequal to the modal (instead of correct) outcome. Systematic measurement error then is the event that this modal outcome is unequal to the correct outcome.

We introduce an estimator for the *ICP*s and evaluate its properties in a simulation study. We end with a case study that demonstrates not only the determination and use of the *ICP*s but also demonstrates how the proposed modeling can be used for formal hypothesis testing. Things to test include differences between appraisers and random classification by a single appraiser.

Keywords: nominal measurement; measurement error; dirichlet distribution

1. Introduction

Nominal measurement systems are devices, persons, or persons operating devices that classify objects on a nominal scale. They yield a measurement outcome Y, which takes on one of C possible values (categories or classes). An example from industry (the context of this paper) is an operator who assigns a failure mode to products that failed quality inspection. Because many decisions in life are a result of measurement, a measurement system should be accompanied by a quantification of its reliability in terms of measurement error. The experiment and subsequent analysis needed to arrive at such quantification is called 'measurement system analysis' (MSA). In this paper, we look at MSA for nominal measurement systems, or 'nominal MSA'.

In symbols, measurement error is often expressed as the conditional event $\{Y = a|X \neq a\}$, with Y a measurement outcome and X the true value. X is obtained by applying a 'reference standard'or 'gold standard': an authoritative procedure to arrive at the true value of the object. We consider the case that this standard is not available: the 'gold standard unavailable case'. This may be because a gold standard is too costly to apply, has a negative impact on the object under measurement, or simply does not exist, for example for lack of a clear definition.

Past literature about nominal MSA has focused on the concept of agreement, which originated in the social sciences (Cohen [1]). Agreement is the event that two measurements on the same object i yield equal outcomes, $\{Y_{i1} = Y_{i2}\}$. A popular measure of agreement is the kappa (κ) statistic (Fleiss [2]), which can be found in industry standards such as AIAG [3] and influential works in medicine like Pepe's [4] book. Log-linear modeling is a related approach (e.g. Tanner and Young [5] and Agresti [6]). There have been many criticisms regarding agreement studies (De Mast [7]; Erdmann *et al*. [8]), and moreover, agreement quantifies a different concept than measurement error: even if appraisers agree on some decision, it is not necessarily the correct decision X.

Proneness to measurement error can be quantified using probabilities of false classification, $FCP_a = P[Y = a|X \neq a]$. These are the probabilities that measurement of an object gives a while the true value of this object is something else. For the binary case, Akkerhuis *et al*. [9,10] have shown that false classification probabilities are in general not identifiable without gold standard, even though many attempts have been made. The goal of this paper is to quantify the random component of the *FCP*s in terms of probabilities of 'inconsistent classification' (*ICP*s) on scales with more than two classes.

Institute for Business and Industrial Statistics, University of Amsterdam
**Correspondence to: T. S. Akkerhuis, IBIS UvA, Department of Operations Management, Plantage Muidergracht 12, 1018TV Amsterdam, The Netherlands.*
† E-mail: t.s.akkerhuis@uva.nl

We define the probabilities of inconsistent classification as $ICP_a = P\left[Y = a|\tilde{X} \neq a\right]$. These probabilities condition on the most likely outcome $\tilde{X}_{ij}$ of a measurement of object i by appraiser j. That is, $\tilde{X}_{ij}$ is the measurement outcome that object i would receive most frequently from appraiser j during repeated measurement. This is comparable with common practice for numerical MSA, where random measurement error is conceived as deviation from the expected measurement outcome as opposed to deviation from the true value. In the nominal case, random variables do not have a statistical expectation, and we replace it by the mode. Because it does not condition on a true value, but on something that is observable, it is an identifiable property.

There are other ways to quantify random measurement error. In some contexts, a quantification like $P\left[Y \neq a|\tilde{X} = a\right]$ or even $P\left[\tilde{X} = a|Y \neq a\right]$ or $P\left[\tilde{X} = a|Y \neq a\right]$ may be useful as well. The modeling we propose will allow calculation of any of these probabilities, but for conciseness, we focus on $ICP_a = P\left[Y = a|\tilde{X} \neq a\right]$ because it is the most natural extension of the False/Inconsistent Acceptance/Rejection Probabilities in binary MSA (for example, Akkerhuis *et al*. [9,10]): probabilities of a specific outcome, conditional on the event that this outcome is False/Inconsistent.

Akkerhuis *et al*. [10] have developed a method to estimate *ICP*s in the binary case, and this paper presents an extension. Estimation of *ICP*s for ordinal MSA has been explored by De Mast and Van Wieringen, [11] but their approach relies on the assumption of a univariate continuum underlying the measurement scale, which is not reasonable for unordered, nominal scales.

In Section 2, we introduce notation, modeling, estimation, and diagnostics. Moreover, we illustrate how *ICP*s represent the random components of the *FCP*s. Section 3 explores the statistical properties of the proposed estimators, and Section 4 demonstrates the method using data from a real life case.

2. Statistical modeling

In the first subsection, notation and modeling of the measurement outcomes are introduced. The second subsection covers estimation of parameters and, consequently, of the *ICP*s, followed by diagnostic tests in the third subsection. The final subsection illustrates what it means that *ICP*s are random components of the *FCP*s.

2.1. Modeling of measurement outcomes

In a typical MSA experiment, I objects are measured K times by J appraisers on a scale with C classes, leading to $I \times J \times K$ measurement outcomes $\boldsymbol{Y} = \left\{Y_{ijk}\right\}_{i=1,\ldots,I;j=1,\ldots,J,k=1,\ldots,K}$.

Objects are modeled by probability vectors $\boldsymbol{P}_{ij}$ that contain classification probabilities. Conditional on a realization of this vector $\boldsymbol{p}_{ij}$, the repeated measurements of object i by appraiser j follow a multinomial distribution with parameters K and $\boldsymbol{p}_{ij}$. Note that $\boldsymbol{P}_{ij}$ is indexed by j to indicate that the distribution of appraisals of some object i may be different for each of the J appraisers.

In symbols, we have

$$\left\{Y_{ijk}|\boldsymbol{P}_{ij} = \boldsymbol{p}_{ij}\right\}_{k=1,\ldots,K} \sim MN\left(K, \boldsymbol{p}_{ij}\right),$$
$$\boldsymbol{p}_{ij} = \left\{p_{ij1}, \ldots, p_{ijC}\right\}, \sum_{c=1}^{C} p_{ijc} = 1.$$

For instance, $\boldsymbol{p}_{ij} = \{0.01, 0.98, 0.01\}$ means that appraiser j mostly classifies object i as 2, but sometimes inconsistently as 1 or 3. If $\boldsymbol{p}_{ij} = \{0, 1, 0\}$, the appraiser will classify i as 2 with certainty. If $\boldsymbol{p}_{ij} = \{0.01, 0.49, 0.50\}$, there is doubt between the categories 2 and 3, but classification 1 is almost never given. The modal outcome $\tilde{X}$ is defined as the position of the largest element, that is, $\tilde{X}_{ij} = \arg\max_c p_{ijc}$.

We model the $\boldsymbol{P}_{ij}$ as draws from a probability distribution $F_{j,\boldsymbol{P}}\left(p_{ij1}, \ldots, p_{ijC}\right) = P\left[P_{ij1} \leq p_{ij1}, \ldots, P_{ijC} \leq p_{ijC}\right]$. The Dirichlet distribution is a natural choice, as it is often used as a prior for multinomial data in Bayesian statistics and allows for explicit calculations. Its density function is given by the following:

$$f_{j,\boldsymbol{P}}\left(p_{ij1}, \ldots, p_{ijC}\right) = \frac{\Gamma\left(\sum_{c=1}^{C} \beta_j \alpha_{jc}\right)}{\prod_{c=1}^{C} \Gamma\left(\beta_j \alpha_{jc}\right)} \prod_{c=1}^{C} p_{ijc}^{\beta_j \alpha_{jc} - 1} \text{ for } \sum_{c=1}^{C} p_{ijc} = 1$$
$$\alpha_{jc} > 0, \sum_{c=1}^{C} \alpha_{jc} = 1, \beta_j > 0.$$

This distribution $F_{j,\boldsymbol{P}}$ can be conceived as a mixture distribution $F_{j,\boldsymbol{P}} = \sum_{c=1}^{C} P[X = c] \times F^c_{j,\boldsymbol{P}}$, where the components $F^c_{j,\boldsymbol{P}} = P\left[P_{ij1} \leq p_{ij1}, \ldots, P_{ijC} \leq p_{ijC}|X = c\right]$ are the distributions of $\boldsymbol{P}_{ij}$ among objects with true value c. However, as both X and its distribution are unknown, only the aggregate distribution is identifiable, and not its components.

Where the $\boldsymbol{P}_{ij}$ are the expected relative frequencies of classifications of an object i ($p_{ijc} = P\left[Y_{ijk} = c|\boldsymbol{P}_{ij} = \boldsymbol{p}_{ij}\right]$), the $\boldsymbol{\alpha}_j$ are the expected relative frequencies in the population of objects: $\alpha_{jc} = E\left[P_{ijc}\right] = E_{\boldsymbol{P}_{ij}}\left[P\left[Y_{ijk} = c|\boldsymbol{P}_{ij} = \boldsymbol{p}_{ij}\right]\right]$.

The scalar β_j is a concentration parameter. For $\beta_j \to 0$, probability mass in $f_{j,\boldsymbol{P}}$ will concentrate only around the C unit vectors $(1,0,0,\ldots)$, $(0,1,0,\ldots)$, etc. The realizations of $\boldsymbol{P}_{ij}$ will then almost surely all be unit vectors. Consequently, the distribution of $\{Y_{ij1},\ldots,Y_{ijK}|\boldsymbol{P}_{ij}=\boldsymbol{p}_{ij}\}$ becomes degenerate, leading to perfectly consistent classifications. The C unit vectors are drawn in proportions given by $\boldsymbol{\alpha}_j$.

If, however, $\beta_j \to \infty$, probability mass in $f_{j,\boldsymbol{P}}$ will concentrate only around $\boldsymbol{\alpha}_j$, leading to realizations of $\boldsymbol{P}_{ij}$ that are almost identical (and equal to $\boldsymbol{\alpha}_j$) for each object i. This means that measurement outcomes do not at all depend on the objects under study, and we call these measurements 'uninformative'. In conclusion, a low β_j implies better consistency of measurements, while a high β_j means poor consistency and uninformative measurements.

If $\beta_j \to \infty$, and measurements are uninformative, categorization becomes purely random guessing. Naturally, we would expect the distributions of $\boldsymbol{P}_{ij}$ then to approach an uniform distribution. That is, if $\beta_j \to \infty$, we would expect to find that approximately $\boldsymbol{\alpha}_j = \left(\frac{1}{C},\ldots,\frac{1}{C}\right)$ (this is the maximum entropy and thus maximally uninformative distribution for the Y_{ijk}; De Mast [8]). However, it may be that because of some artifact in the measurements or because of behavioral tendencies of the appraisers, purely random categorizations have a non-uniform distribution (that is, $\beta_j \to \infty$ while $\boldsymbol{\alpha}_j \neq \left(\frac{1}{C},\ldots,\frac{1}{C}\right)$).

2.2. *Estimation of inconsistent classification probabilities*

We propose estimation of the parameters in $f_{j,\boldsymbol{P}}$ by maximum likelihood. As shorthand notation, we use $\boldsymbol{\theta}_j = (\alpha_{j1},\ldots,\alpha_{jC},\beta_j)$. Conditional on a realization $\boldsymbol{p}_{ij}$, the outcome of repeated measurement of object i by appraiser j has likelihood

$$P\left[Y_{ij1}=y_{ij1},\ldots,Y_{ijK}=y_{ijK}\middle|\boldsymbol{P}_{ij}=\boldsymbol{p}_{ij}\right]=\prod_{c=1}^{C}p_{ijc}^{\#_k\{Y_{ijk}=c\}}.$$

Here, $\#_k\{Y_{ijk}=c\}$ is the number of occurrences of classification c in repeated measurements $\{Y_{ij1},\ldots,Y_{ijK}\}$, and we assume that Y_{ijk} are independent conditional on a realization of $\boldsymbol{P}_{ij}$. In symbols:

$$\{Y_{ijk}|\boldsymbol{P}_{ij}=\boldsymbol{p}_{ij}\}_{i=1,\ldots,I,j=1,\ldots,J,k=1,\ldots,K} \text{ are independent.}$$

This assumption has the interpretation that, besides the $\boldsymbol{P}_{ij}$, there are no factors that induce dependencies among the $Y_{ijk}, k=1,\ldots,K$. As $\boldsymbol{P}_{ij}$ is not observed directly, we integrate it out over its probability distribution and obtain the unconditional likelihood:

$$P\left[Y_{ij1}=y_{ij1},\ldots,Y_{ijK}=y_{ijK}\right]=\int_0^1\cdots\int_0^1 f_{j,\boldsymbol{P}}\left(p_{ij1},\ldots,p_{ijC}\right)\times$$

$$\prod_{c=1}^{C}p_{ijc}^{\#_k\{Y_{ijk}=c\}}\mathrm{d}p_{ij1}\ldots\mathrm{d}p_{ijC}=\frac{\Gamma\left(\sum_{c=1}^{C}\beta_j\alpha_{jc}\right)}{\prod_{c=1}^{C}\Gamma\left(\beta_j\alpha_{jc}\right)}\times\frac{\prod_{c=1}^{C}\Gamma\left(\beta_j\alpha_{jc}+\#_k\{Y_{ijk}=c\}\right)}{\Gamma\left(\sum_{c=1}^{C}\beta_j\alpha_{jc}+\#_k\{Y_{ijk}=c\}\right)}.$$

The loglikelihood is obtained by aggregating log-probabilities over objects:

$$L_j\left(\{\alpha_{jc}\}_{c=1,\ldots,C},\beta_j\right)=\sum_{i=1}^{I}\log\frac{\Gamma\left(\sum_{c=1}^{C}\beta_j\alpha_{jc}\right)}{\prod_{c=1}^{C}\Gamma\left(\beta_j\alpha_{jc}\right)}\times\frac{\prod_{c=1}^{C}\Gamma\left(\beta_j\alpha_{jc}+\#_k\{Y_{ijk}=c\}\right)}{\Gamma\left(\sum_{c=1}^{C}\beta_j\alpha_{jc}+\#_k\{Y_{ijk}=c\}\right)}. \quad (1)$$

The number of parameters is $C+1$ (one of which is trivial as the C elements in $\boldsymbol{\alpha}_j$ sum to 1) for each appraiser. We found in our analyses that the Nelder–Mead algorithm [12] gives stable estimates. Instead of maximizing L_j with respect to $\alpha_{j1},\ldots,\alpha_{jC},\beta_j$ under the restriction that $\sum_c \alpha_{jc}=1$, we recommend to maximize L_j with respect to $\alpha_{j1}\beta_j,\ldots,\alpha_{jC}\beta_j$.

ICP_{ja} can be calculated using the indicator function (denoted 1):

$$ICP_{ja}=P\left[Y_{ijk}=a|\tilde{X}_{ij}\neq a\right]=\frac{P\left[Y_{ijk}=a\right]-P\left[Y_{ijk}=a,\arg\max_c p_{ijc}=a\right]}{1-P\left[\arg\max_c p_{ijc}=a\right]}=$$

$$\frac{\int_S f_{j,\boldsymbol{P}}\left(p_{ij1},\ldots,p_{ijC}\right)1_{\{\arg\max_c p_{ijc}=a\}}\mathrm{d}\boldsymbol{p}_{ij}-\int_S f_{j,\boldsymbol{P}}\left(p_{ij1},\ldots,p_{ijC}\right)p_{ija}1_{\{\arg\max_c p_{ijc}=a\}}\mathrm{d}\boldsymbol{p}_{ij}}{1-\int_S f_{j,\boldsymbol{P}}\left(p_{ij1},\ldots,p_{ijC}\right)1_{\{\arg\max_c p_{ijc}=a\}}\mathrm{d}\boldsymbol{p}_{ij}}.$$

These C integrals are not easily approximated by numerical quadrature because of the complex intersection of the support S of $f_{j,\boldsymbol{P}}$ and the range of $\boldsymbol{P}_{ij}$ that satisfies the indicator condition. We therefore recommend the following transformation and corresponding Jacobian determinant:

$$p_{ijk} = \begin{cases} \frac{p^*_{ijk}}{1+\sum_{c=1}^{C} p^*_{ijc}-p^*_{ija}} & k \neq a \\ \frac{1}{1+\sum_{c=1}^{C} p^*_{ijc}-p^*_{ija}} & k = a \end{cases}, \det\left[\frac{\partial p_{ijk}}{\partial p^*_{ijk}}\right] = \frac{1}{\left(1+\sum_{c=1}^{C} p^*_{ijc}-p^*_{ija}\right)^2}.$$

By integrating $\left\{p^*_{ijc}\right\}_{c=1}^{C} \setminus \left\{p^*_{ija}\right\}$ over the unit hypercube, the conditions $\arg\max_c p_{ijc} \neq a$ and $\sum_{c=1}^{C} p_{ijc} = 1$ are automatically satisfied.

For $C > 5$, standard adaptive numerical quadrature turned out computationally prohibitive, and Monte Carlo integration provided a better alternative.

2.3. *Diagnostics and tests*

To verify the fit of the model, we employ response patterns $\boldsymbol{E}_{ij} = \left(\#_k \left\{Y_{ijk} = c\right\}\right)_{c=1,\ldots,C}$. An example of such a pattern is $\boldsymbol{e}_{ij} = (1,0,2)$: object i is classified once as 1 and twice as 3 in three repeated measurements by appraiser j.

The data from the MSA study are summarized in response pattern frequencies F:

$$F_{j,\boldsymbol{E}}(\boldsymbol{e}) = \#_i \left\{\boldsymbol{E}_{ij} = \boldsymbol{e}\right\}.$$

For example, $F_{j,\boldsymbol{E}}((0,2,0)) = 3$ means that 3 objects are classified as 2 consistently by appraiser j. Goodness of fit is verified by comparing $F_{j,\boldsymbol{E}}$ with the expected frequencies (in which $e[c]$ is the cth element of $\boldsymbol{e}$):

$$\begin{aligned} E\left[F_{j,\boldsymbol{E}}(\boldsymbol{e})\right] &= I \times \int_0^1 \cdots \int_0^1 f_{j,\boldsymbol{P}}(p_{ij1},\ldots,p_{ijC}) \frac{K!}{\prod_{c=1}^{C} e[c]!} \times \prod_{c=1}^{C} p_{ijc}^{e[c]} \mathrm{d}p_{ij1} \ldots \mathrm{d}p_{ijC} \\ &= I \times \frac{K!}{\prod_{c=1}^{C} e[c]!} \frac{\Gamma\left(\sum_{c=1}^{C} \beta_j \alpha_{jc}\right)}{\prod_{c=1}^{C} \Gamma\left(\beta_j \alpha_{jc}\right)} \frac{\prod_{c=1}^{C} \Gamma\left(\beta_j \alpha_{jc} + e[c]\right)}{\Gamma\left(\sum_{c=1}^{C} \beta_j \alpha_{jc} + e[c]\right)}. \end{aligned}$$

Note that in a high variety of possible response patterns, expected frequencies are often low, which makes the standard chi-squared test unreliable as shown by Kallenberg *et al.* [13]. We therefore use the likelihood ratio test, which is also known as the G-test when it comes to fitting response pattern frequencies:

$$G_j = 2 \sum_{\text{all } \boldsymbol{e}} F_{j,\boldsymbol{E}}(\boldsymbol{e}) \times \log \frac{F_{j,\boldsymbol{E}}(\boldsymbol{e})}{E\left[F_{j,\boldsymbol{E}}(\boldsymbol{e})\right]} \sim \chi^2 (\#\text{ unique } \boldsymbol{E} - 1 - C).$$

unique $\boldsymbol{E} - 1$ is the number of nontrivial parameters needed for a saturated model, and C is the number of parameters in the proposed model (although there are $C+1$ parameters, one is trivial because of the restriction $\sum \alpha_{jc} = 1$). The estimates for $\boldsymbol{\alpha}_j$ and β_j can be used to test against uninformative categorization, or to compare appraisers.

- To test for uninformative categorization, one can test $H_0 : f_{j,\boldsymbol{P}} = f^{\infty}_{j,\boldsymbol{P}}$ with $f^{\infty}_{j,\boldsymbol{P}} = \lim_{\beta_j \to \infty} f_{j,\boldsymbol{P}}$ and $\boldsymbol{\alpha}_j = \left\{\frac{1}{C}\right\}_{c=1}^{C}$.
- A comparison of appraisers j and k with respect to consistency can be performed by testing $H_0 : \beta_j = \beta_k$ for $j \neq k$.
- Systematic differences between appraisers can be detected by testing $H_0 : \boldsymbol{\alpha}_j = \boldsymbol{\alpha}_k$ for $j \neq k$.
- Interchangeability of appraisers can be tested by $H_0 : \boldsymbol{\alpha}_j = \boldsymbol{\alpha}_k, \beta_j = \beta_k$ for $j \neq k$.

The likelihood ratio test can be used for these purposes, where the likelihood (1) is optimized both with and without these restrictions. However, no appraiser is the same, and thus, for large enough samples, we will always find significant differences, however small. Because statistical significance does not imply practical relevance, we recommend to combine hypothesis testing with a judicious look at the actual sizes of the differences.

For evaluating the actual sizes of these differences, it can help to construct and plot confidence intervals. By calculating minus the inverse of the Hessian matrix of the loglikelihood and then taking the square roots of the diagonal elements, we find standard errors. Ninety-five percent confidence intervals are roughly given by the estimated value ± 1.96 times the standard error.

2.4. *Probabilities of mis- and inconsistent classification*

As claimed in the introduction, probabilities of inconsistent classification ICP_{ja} are the random components of the probabilities of false classification FCP_{ja}. We show the mathematical interpretation by following the reasoning in Akkerhuis *et al.* [9]

The modal outcome $\tilde{X}_{ij}$ allows decomposing the probability of misclassification FCP_{ja} as follows:

$$FCP_{ja} = P\left[Y_{ijk} = a, \tilde{X}_{ij} = a | X_i \neq a\right] + P\left[Y_{ijk} = a, \tilde{X}_{ij} \neq a | X_i \neq a\right].$$

The first term, $P\left[Y_{ijk} = a, \tilde{X}_{ij} = a | X_i \neq a\right]$, is the systematic component, where the modal outcome does not equal the true value but the actual outcome does equal the modal outcome. In $P\left[Y_{ijk} = a, \tilde{X}_{ij} \neq a | X_i \neq a\right]$ (the second term), we have that the measurement outcome randomly deviates from the modal outcome. The second term can be further decomposed as follows:

$$P\left[Y_{ijk} = a, \tilde{X}_{ij} \neq a | X_i \neq a\right] = P\left[Y_{ijk} = a | \tilde{X}_{ij} \neq a\right] \times c_1 - P\left[Y_{ijk} = a, \tilde{X}_{ij} \neq a | X_i = a\right] \times c_2,$$

with $c_1 = P[\tilde{X} \neq a]/P[X \neq a]$ and $c_2 = P[X = a]/P[X \neq a]$. This decomposition reveals an interesting fact that does not hold for numerical measurements. The first part is merely random error. However, a probability is subtracted for the event that a measurement outcome Y_{ijk} equals the true value X_i, but $\tilde{X}_{ij} \neq X_i$. That is, a random error cancels out a systematic error, an event that has zero probability for numerical measurements.

Consequently, we use the probability of inconsistent classification ICP_{ja} to quantify random measurement error: $ICP_{ja} = P\left[Y_{ijk} = a | \tilde{X}_{ij} \neq a\right]$. Note that in the absence of systematic measurement error ($\tilde{X}_{ij} = X_i$), we have that $ICP_{ja} = FCP_{ja}$.

3. Evaluation by means of simulation

In this section, we evaluate first-order and second-order properties of the proposed estimators of ICP_a by means of simulation. We vary the concentration parameter β and the number of classes C. We also investigate the effect of non-uniformity in $\boldsymbol{\alpha}$, to represent cases in which some classes are rare. Finally, we evaluate the relation between sample size I and number of repeated measurements K on the one hand, and standard errors of ICP_a on the other.

For ease of notation, in what follows, we omit subscripts j that index appraisers.

3.1. *Effect of different concentration β*

We study the situation of classifications on a three-point scale. For the case $\boldsymbol{\alpha} = \left\{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right\}$, we vary $\beta = 0.50, 0.55, \ldots, 2.00$ (high concentration to low concentration). In particular, for each value of β, we perform 1000 simulation runs in which we draw $I = 300$ realizations $\boldsymbol{p}_i$ from the Dirichlet distribution, perform $K = 10$ repeated measurements for each object i, optimize the likelihood to obtain estimates for $\widehat{\boldsymbol{\alpha}}, \widehat{\beta}$, and use these to calculate $\widehat{ICP}_a$. This gives us the empirical distribution of the estimators for each value of β.

Figure 1 contains plots of the average estimates and the associated standard errors, where interpolation between the points has been used. In the left plot, we see that, naturally, the higher β becomes (the lower the concentration), the higher the probability of inconsistent classification. For low values of β, draws for $\boldsymbol{P}_i$ are close to unit vectors, leading to highly consistent classifications. For high values of β, $\boldsymbol{p}_i$ is closer to $\boldsymbol{\alpha}_j$.

The right plot in Figure 1 shows that the relative standard error of the estimates is decreasing. Higher values of β can be estimated relatively precisely. This means that the mean increases faster in β than does the standard error.

3.2. *Effects of non-uniformities in $\boldsymbol{\alpha}$*

We consider a variety of situations in which $\boldsymbol{\alpha}$ is not uniform. In particular, we consider $\boldsymbol{\alpha} = \left\{q, \frac{1-q}{2}, \frac{1-q}{2}\right\}$ for $q = 0.05, 0.10, \ldots, 0.90$. For low values of q, there is one underrepresented class, and for high values of q, there are two underrepresented classes.

For each q, we perform 1000 simulation runs to generate the empirical distribution of $ICP_a; a = 1, 2, 3$. We set $I = 300$, $K = 10$ and $\beta = 1$. Figure 2 shows, as a function of q, the means of the estimators of ICP_1 and ICP_2 (and ICP_3, which equals ICP_2 up to some simulation error), as well as their relative standard errors.

We see, first, that the average misclassification probability in favor of category 1 increases when q increases (top graphs in Figure 2). Put simpler: as α_1 increases, the probability of misclassifying something as 2 or 3 decreases (bottom graphs in Figure 2). This makes sense: if an appraiser thinks there are many 1s in the population, he or she will be more inclined to misclassify 2s or 3s as 1, and vice versa.

Second, from the standard errors, we see that the estimate for ICP_1 is most precise if $q \approx \frac{2}{3}$. Especially for low q, when an object from category 1 is very rare, ICP_1 is difficult to estimate precisely. For high q, this is the case as well, because then categories 2 and 3 are rare, and so a misclassification as 2 or 3 is rare as well.

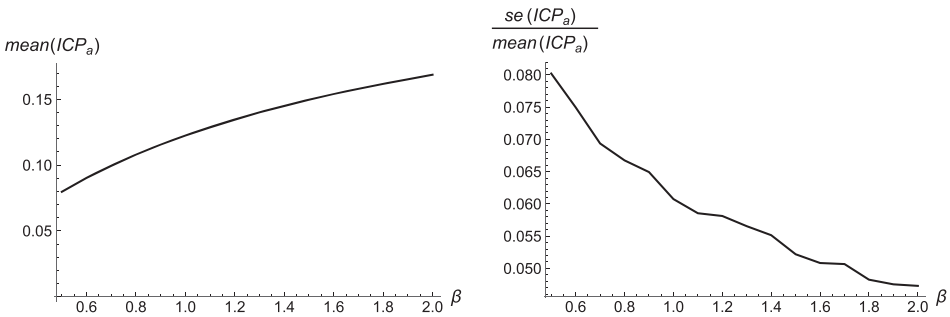

Figure 1. Average inconsistent classification probabilities ICP_a (left), and relative standard error of ICP_a (right) for different values of β

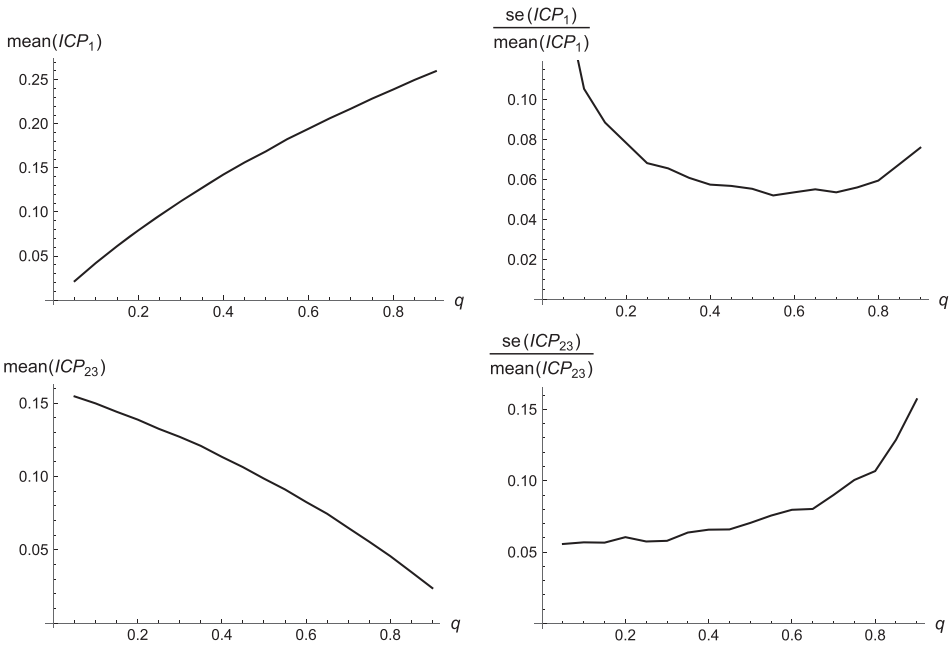

Figure 2. Average inconsistent classification probabilities ICP_a (left), and relative standard error of ICP_a (right) for different values of q. Top row: ICP_1; bottom row: ICP_2 and ICP_3

Table I. Average ICP_a (left), and relative standard error of ICP_a (right) for different values of C.

$\boldsymbol{\alpha}$	$(\frac{1}{2}, \frac{1}{2})$	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$
mean ICP_a	0.1823	0.1225	0.0924
rel. s.e. ICP_a	0.0107	0.0076	0.0064

3.3. *Effect of number of classes C*

We studied $\boldsymbol{\alpha} = \{\frac{1}{C}\}_{c=1}^{C}$ for $C = 2, 3, 4$. We set $I = 300$, $K = 10$, and $\beta = 1$. The estimates are given in Table I.

We see that the mean ICP_a are decreasing in the number of categories. Note that, in the case with three categories, the total probability of misclassification is $(3-1) \times 0.1225 = 0.2450$, but in the case with four categories, it is $(4-1) \times 0.0924 = 0.2772$, which is higher. The total probability of misclassification is thus increasing in C.

This means that, in order to maintain an equal probability of misclassification, the amount of concentration should increase (or: β should decrease) when there are more classes. Intuitively, this makes sense.

3.4. *Effect of sample size I and number of repeated measurements K*

We took the case $\boldsymbol{\alpha} = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$ and $\beta = 1$. We performed 1000 sets of simulations for the 9×7 grid defined by $K = 2$ to 18 in steps of 2 and $I = 120$ to 480 in steps of 60. The standard errors of ICP_a are given in Figure 3. Contour lines are derived by interpolation between the 63 grid points.

The solid contour lines in Figure 3 delineate points with constant standard errors of ICP_a. The dashed contour lines represent regions with constant $I \times K$, representing constant number of measurements. For a given amount of measurements $I \times K$, the standard errors can be optimized by moving along the dashed line into the region with lowest standard error. We see that, in the range explored in Figure 3, it is optimal to lower K to 3 or 4 and let the required precision determine I.

Note that optimizing the likelihood function can be seen as fitting a distribution of response patterns to the empirical distribution of response patterns. For that reason, we recommend choosing K such that the number of possible response patterns is larger than the number of parameters in F_P. In symbols, $\binom{C+K-1}{K} > C+1$. This means that, in the binary case, three repeated measurements are necessary (also see Van Wieringen and Van den Heuvel [14]), and with more classes, $K = 2$ is sufficient.

4. Real-life case

4.1. *Background*

The case study comes from a manufacturer of decorative products. Because of a recent increase in competition, senior management has focused its strategy on product quality. This was not straightforward as no customer specifications were defined, and thus, the first

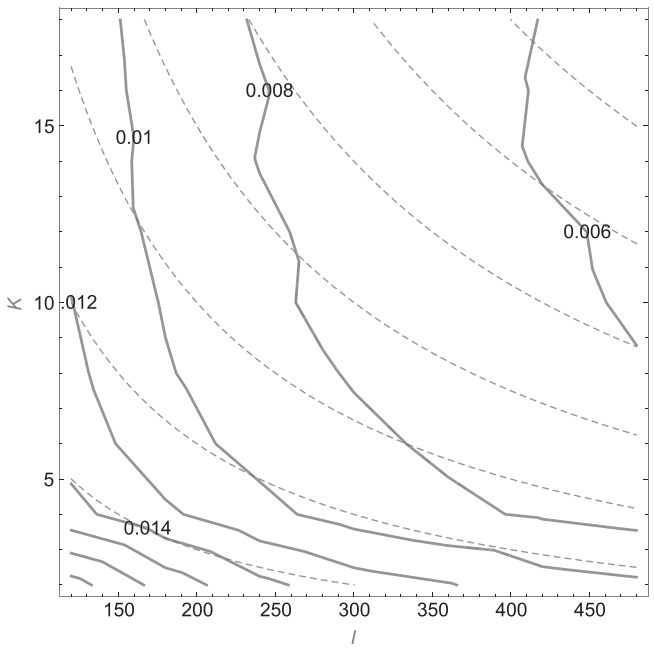

Figure 3. Relation between standard error, number of repeated measurements K, and sample size I. Dashed lines are regions with constant $I \times K$

step was to define quality. The decision was made to distill a definition from the judgments of line operators. Specifically, the goals were to

- find out whether there are systematic differences in appraisals between operators. These differences may serve as a basis for discussion when developing a definition for quality.
- identify the operator with highest repeatability, in order to work with him or her to create a quality inspection procedure that gives consistent measurement results.

The MSA study was performed on the production line for the casings for one of the products. Management had already determined that there are two types of poor quality: a shortcoming in the casing that may cause the product to malfunction or a visual shortcoming, which is almost equally relevant in a market for decorative products. We use a dataset containing measurements by three operators. Possible classifications are 'OK '(1), 'MALFUNCTION'(2), or 'VISUAL' (3). After the first round of measurements, appraisers were asked to classify the products again, 2 or 3 days later, in a different, randomized order. This was to minimize the impact of the memory effect.

The sample has an important, but also realistic, complexity, which is the result of (production) time and (personnel) cost limitations. The distribution of 1s, 2s, and 3s in the population of casings is, as is often the case in industry, very unbalanced, because the majority of products are conforming (1). Therefore, an extremely large sample would be required to make sure that parts with $\tilde{X}_{ij} = 2$ or 3 are sufficiently represented for estimation of all ICP_{ja}. In particular, this means that the recommended sample size will be beyond, and specifically to the right of the range as depicted in Figure 3. The manufacturer chose to take a sample of 60 casings that were (subjectively) considered hard to judge, because it was expected that these would provide the most information when comparing operators' judgments. The sample size of 60 was chosen because this corresponded to the maximum amount of man-hours that management wanted to devote to this experiment.

We are dealing with a non-random sample of hard-to-judge parts. These are parts with $\left|\max_c p_{ijc} - \min_c p_{ijc}\right|$ relatively small. That is, especially, the boundaries of the support of the Dirichlet distribution, where the $\boldsymbol{p}_{ij}$ are close to unit vectors, are underrepresented. Parts around these boundaries have a relatively low $P\left[Y_{ijk} = a | \tilde{X}_{ij} \neq a\right]$, and as the ICP_{ja} is an average of these probabilities over all products i, the estimate will turn out higher than if the sample were random. However, if appraiser $j = 1$ has $\boldsymbol{\alpha}$, β that are very much different from those of appraiser $j = 2$, this will be visible in any sample, random or non-random, in both the parameters and the ICP_{ja}. In light of the aims depicted earlier, this sample will thus still allow for comparing appraisers, even though the absolute values of the estimates are not representative for the entire population of products.

A more tractable approach for overcoming the challenge of an unbalanced population (with respect to X and/or $\tilde{X}$) would have been to perform conditional sampling (for example, Danila *et al.* [15] and Erdmann *et al.* [8]). Conditional sampling means letting appraiser j select a mix of parts that he or she has already ($k = 0$) classified as $Y_{ij0} = 1$, $Y_{ij0} = 2$ and $Y_{ij0} = 3$ and thus obtain a sample that is more balanced. To prevent biases, the conditional probability distributions $f_{j,\boldsymbol{p}}^{Y_0 = t}$ then need to be used in the likelihood (1).

The outcomes of the experiment are given in Table II. Some of the products seemed to be straightforward to classify (product 10 was clearly malfunctioning, and product 12 was clearly OK), while others were less easy to classify (for example product 21). In the next sections, we will fit the aforementioned model to the data and give interpretations.

Table II. Outcome of the experiment

Case	A	B	C	Case	A	B	C	Case	A	B	C
1	1,1	1,1	1,1	21	3,3	2,2	1,3	41	1,1	1,1	1,1
2	1,1	1,1	1,1	22	2,2	2,2	2,2	42	1,3	1,1	1,1
3	2,2	2,3	1,3	23	1,1	1,1	1,1	43	1,1	1,1	1,1
4	1,3	1,1	1,1	24	2,2	2,2	2,3	44	3,3	3,3	3,3
5	2,2	2,2	1,1	25	1,2	1,1	2,2	45	2,2	1,3	1,1
6	2,2	2,2	2,2	26	2,2	2,2	1,1	46	1,1	1,1	1,3
7	2,2	2,2	1,1	27	3,3	3,3	3,3	47	3,3	3,3	1,1
8	1,1	1,1	1,1	28	1,1	1,1	1,1	48	1,1	1,1	1,1
9	3,3	1,1	1,1	29	2,2	2,2	2,2	49	1,1	1,1	1,1
10	2,2	2,2	2,2	30	2,2	2,2	2,2	50	1,1	1,1	1,1
11	1,1	1,1	1,1	31	1,1	1,1	1,1	51	2,2	2,2	1,1
12	1,1	1,1	1,1	32	1,1	1,1	1,1	52	1,1	1,1	1,1
13	1,1	1,1	1,1	33	1,1	1,1	1,1	53	1,1	1,1	1,1
14	3,3	3,3	3,3	34	1,1	1,1	1,1	54	1,1	1,1	1,1
15	1,1	1,1	1,1	35	1,1	1,1	1,1	55	1,1	1,1	1,1
16	1,2	1,1	1,1	36	1,1	1,1	2,3	56	1,1	1,1	1,1
17	2,2	2,2	1,2	37	1,1	1,1	1,1	57	1,1	1,1	1,1
18	1,1	1,1	1,1	38	1,3	1,1	1,1	58	3,3	3,3	3,3
19	2,3	1,1	1,2	39	2,2	2,2	2,3	59	1,1	1,1	1,1
20	1,1	1,1	1,1	40	1,1	1,1	1,2	60	1,1	1,1	1,1

Categories: 1 = product is OK, 2 = product will lead to malfunction, and 3 = visual shortcoming.

Table III. Estimated parameters of the full model

	α_{j1}	α_{j2}	α_{j3}	β_j
A	0.5832	0.2610	0.1558	0.2597
	(0.0600)	(0.0531)	(0.0434)	(0.1159)
B	0.6687	0.2216	0.1098	0.0741
	(0.0596)	(0.0525)	(0.0393)	(0.0553)
C	0.7211	0.1530	0.1259	0.5591
	(0.0525)	(0.0413)	(0.0378)	(0.2635)

Standard errors between brackets. Categories: 1 = product is OK, 2 = product will lead to malfunction, and 3 = visual shortcoming.

4.2. *Results*

The estimated parameters are in Table III, and a graphical representation of the α_{jk} is given in Figure 4. These show that all appraisers seem to agree that this sample consists mostly of acceptable products, although in a truly random sample, this is expected to be even more clearly so. *A* is more strict than *C* is. It seems moreover that, for the nonconforming products, most problems arise because of a malfunction, not a visual shortcoming, but this result may not hold for the total population of casings.

The estimates for β_j show that *B* has a particularly high repeatability, while *C* has a low repeatability. The point estimate for *C* lies outside the confidence intervals of β of *A* and *B*.

4.3. *Probabilities of inconsistent classification*

The estimates and standard errors of the *ICP*s are given in Table IV. As reflected in the high β_j, *C* uniformly has the highest probability of inconsistent classification. Moreover, for all appraisers, the most inconsistent classifications are among products that were classified as 3. This may be explained by esthetic value being very subjective when not objectively described. A classification 'OK' is uniformly most likely to be a classification that is not inconsistent with $\tilde{X}$.

As mentioned before, these figures are not based on a random sample, so the relative sizes of the *ICP* are more reliable than their absolute values.

4.4. *Hypothesis testing*

We perform various likelihood ratio tests on the parameters of the model. The null hypotheses with the corresponding loglikelihood values and *p*-values are given in Table V.

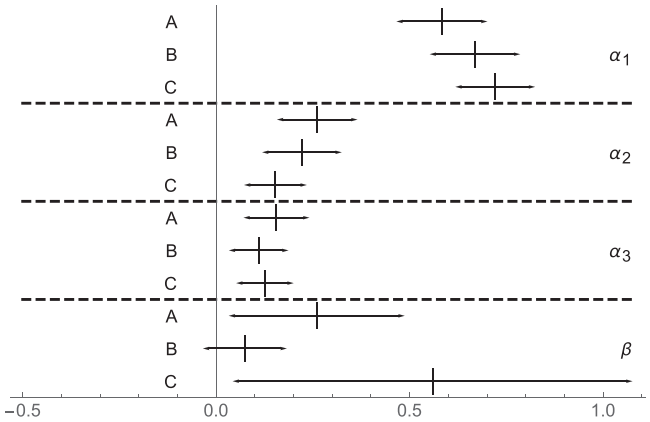

Figure 4. Estimated parameters full model. Vertical lines are point estimates, and arrows mark confidence interval. Categories: 1 = product is OK, 2 = product will lead to malfunction, and 3 = visual shortcoming

Table IV. Probabilities of inconsistent classification

j	ICP_{j1}	ICP_{j2}	ICP_{j3}
A	0.0636	0.1053	0.1183
B	0.0167	0.0364	0.0416
C	0.0809	0.1980	0.2034

Categories: 1 = product is OK, 2 = product will lead to malfunction, and 3 = visual shortcoming.

Table V. Tested hypotheses involving multiple appraisers

Type	H_0	logL	p-value
Unrestricted model		−218.91	—
Systematic	$\boldsymbol{\alpha}_A = \boldsymbol{\alpha}_B = \boldsymbol{\alpha}_C$	−220.75	0.4537
	$\boldsymbol{\alpha}_A = \boldsymbol{\alpha}_B$	−219.47	0.5739
	$\boldsymbol{\alpha}_A = \boldsymbol{\alpha}_C$	−220.54	0.1969
	$\boldsymbol{\alpha}_B = \boldsymbol{\alpha}_C$	−219.46	0.5806
Repeatability	$\beta_A = \beta_B = \beta_C$	−222.17	0.0389
	$\beta_A = \beta_B$	−220.11	0.1229
	$\beta_A = \beta_C$	−219.63	0.2327
	$\beta_B = \beta_C$	−222.16	0.0108
Exchangeability	$\alpha_A = \alpha_B = \alpha_C, \beta_A = \beta_B = \beta_C$	−223.81	0.1346
	$\boldsymbol{\alpha}_A = \boldsymbol{\alpha}_B, \beta_A = \beta_B$	−220.99	0.2469
	$\boldsymbol{\alpha}_A = \boldsymbol{\alpha}_C, \beta_A = \beta_C$	−220.85	0.2769
	$\boldsymbol{\alpha}_B = \boldsymbol{\alpha}_C, \beta_B = \beta_C$	−222.56	0.0634

Likelihood ratios calculated with respect to unrestricted model.

There is evidence that there are differences in repeatability ($p = 0.0389$), which seems mostly due to differences in repeatability between appraiser *B* and *C* ($p = 0.0108$). There is no indication of any systematic difference between appraisers.

Other test results are given in Table VI and are about informative versus uninformative measurements. The first three represent an inability to concentrate. If β_j is infinite, then each object will yield the same response pattern distribution. This can be conceived as some kind of guessing, which is ruled out by the low p-values. The latter three represent a special case of guessing, in which the guesses are uniformly distributed. These hypotheses are also ruled out.

Note that $\beta_j \to \infty$ cannot be plugged into the likelihood directly. We used that

$$\lim_{\beta_j \to \infty} L_j\left(\{\alpha_{jc}\}_{c=1,\ldots,C}, \beta_j\right) = \log \prod_{i=1}^{I} \prod_{c=1}^{C} \alpha_{jc}^{\#\{Y_{ijk}=c\}} = \sum_{i=1}^{I} \sum_{c=1}^{C} \#\{Y_{ijk} = c\} \log \alpha_{jc}.$$

Table VI. Tested hypotheses and results

Null hypothesis	Likelihood ratio	*p*-value
$H_0 : \beta_A \to \infty$	62.95	0.000
$H_0 : \beta_B \to \infty$	78.74	0.000
$H_0 : \beta_C \to \infty$	31.35	0.000
$H_0 : \beta_A \to \infty, \boldsymbol{\alpha}_A = \{C^{-1}\}_{c=1}^{C}$	98.26	0.000
$H_0 : \beta_B \to \infty, \boldsymbol{\alpha}_B = \{C^{-1}\}_{c=1}^{C}$	142.9	0.000
$H_0 : \beta_C \to \infty, \boldsymbol{\alpha}_C = \{C^{-1}\}_{c=1}^{C}$	112.0	0.000

Table VII. Results of goodness of fit tests

Appraiser	G_j	p
A	0.504	0.777
B	4.545	0.103
C	4.772	0.092

Table VIII. Contributions to the goodness-of-fit G-statistic

$\boldsymbol{e}$	$F_{A,\boldsymbol{E}}(\boldsymbol{e})$	$E[F_{A,\boldsymbol{E}}(\boldsymbol{e})]$	G	$F_{B,\boldsymbol{E}}(\boldsymbol{e})$	$E[F_{B,\boldsymbol{E}}(\boldsymbol{e})]$	G	$F_{C,\boldsymbol{E}}(\boldsymbol{e})$	$E[F_{C,\boldsymbol{E}}(\boldsymbol{e})]$	G
(2, 0, 0)	32	32	0.0	40	39	1.6	41	39	4.2
(1, 1, 0)	3	3.8	−1.4	0	1.2	0.0	3	4.7	−2.8
(1, 0, 1)	3	2.2	1.7	1	0.6	1.0	3	3.9	−1.6
(0, 2, 0)	14	13	1.5	13	13	0.9	6	6.4	−0.8
(0, 1, 1)	1	1.0	−0.0	1	0.2	3.2	3	0.8	7.7
(0, 0, 2)	7	7.7	−1.4	5	6.2	−2.1	4	5.2	−2.1

4.5. *Diagnostics*

Because there are six possible response patterns (two repeated measurements over three classes), five parameters are necessary to fit all the associated frequencies (saturated model). In the proposed model, three parameters are needed. We can perform a goodness-of-fit test with $df = 5 - 3 = 2$, as described in Section 2.3. The test results are given in Table VII, and the actual and fitted response patterns are given in Table VIII.

It seems that all appraisers pass the goodness-of-fit test, although depending on the significance level, the results might be seen as weak evidence for a bad fit for *C*.

4.6. *Conclusions from the case study*

Although there are no significant systematic differences between appraisers, there were differences in repeatability. In particular, there was a significant difference between operators *B* and *C*, the former being more consistent.

As a result of the investigations, a discussion between the two operators was organized, with the task of formulating standard operating procedures in order to obtain consistent classifications for all operators. Moreover, parts 19, 21, 25, and 45 were classified least consistent and were brought to a discussion with other operators and management as well. Where no agreement could be obtained, higher management made a choice.

5. Conclusions

This paper proposes a method to quantify measurement error of nominal measurement systems in situations when no gold standard is available. Without a gold standard, the event of a measurement error is undetectable, and estimation of false classification probabilities $FCP_a = P[Y = a|X \neq a]$ is thus problematic (Akkerhuis *et al.* [9,10]). Therefore, this paper proposes a method to estimate the random component of measurement error only: the probabilities of inconsistent classification $ICP_a = P\left[Y = a|\tilde{X} \neq a\right]$, with $\tilde{X}$ the modal or most likely measurement outcome. We conceive the situation $X = \tilde{X}$ as an absence of systematic measurement error, which results in $ICP = FCP$. Although it may seem like an important limitation of *ICP* that it does not express systematic error, it is not unsurprising when compared with standard practices for numerical MSA without gold standard. In numerical measurement, it is a common practice to divide measurement error into a systematic and a random component. Systematic error is operationalized by 'bias': the difference between the expected measurement outcome and the true value $E[Y] - X$. In nominal measurement, 'difference' is not defined because of the measurement scale, but probabilities $P\left[\tilde{X} \neq X\right]$ offer another way of quantification. For both expressions, it is clear that its estimation is unattainable without a gold standard.

Random measurement error for numerical measurement is operationalized by the difference between measured and expected outcomes $Y - E[Y]$, where $E[Y]$ is comparable with the $\tilde{X}_i$ in the nominal case. It can be quantified by a gage R&R study without a gold standard being available. Because differences are undefined on a nominal scale, this paper proposes to quantify random measurement error of nominal measurement by probabilities of inequality $P\left[Y = a|\tilde{X} \neq a\right]$. In that sense, the approach proposed in this paper moves closer toward what is already common practice for numerical MSA, by acknowledging that, without a gold standard, only the random component of measurement error is estimable.

Acknowledgements

The authors thank Mark Vertogen for his valuable contribution and pleasant collaboration in the case reported in this paper.

References

1. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; **20**:37–46.
2. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971; **76**:378–382.
3. AIAG. Measurement System Analysis: Reference Manual 3rd ed. Automotive Industry Action Group: Detroit, MI, 2003.
4. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press: Oxford, UK, 2003.
5. Tanner MA, Young MA. Modeling agreement among raters. *Journal of the American Statistical Association* 1985; **80**:959–968.
6. Agresti A. A model for agreement between ratings on an ordinal scale. *Biometrics* 1988; **44**:539–548.
7. De Mast J. Agreement and kappa-type indices. *The American Statistician* 2007; **61**(2):148–153.
8. Erdmann TP, De Mast J, Warrens MJ. Some common errors of experimental design, interpretation and inference in agreement studies. *Statistical Methods in Medical Research* 2012; **24**:920–935.
9. Akkerhuis TS, De Mast J, Erdmann TP. The Statistical Evaluation of Binary Tests without Gold Standard: Robustness of Latent Variable Approaches, 2015. Submitted for publication.
10. Akkerhuis TS, De Mast J, Erdmann TP. Estimation of the Random Error of Binary Tests using Adaptive Polynomials, 2015. Submitted for publication.
11. De Mast J, Van Wieringen WN. Modeling and evaluating repeatability and reproducibility of ordinal classifications. *Technometrics* 2010; **52**:94–106.
12. Nelder JA, Mead R. A simplex method for function minimization. *The Computer Journal* 1965; **7**:308–313.
13. Kallenberg WCM, Oosterhoff J, Schriever BF. The number of classes in chi-squared goodness-of-fit tests. *Journal of the American Statistical Association* 1985; **80**:959–968.
14. Van Wieringen WN, Van den Heuvel ER. A comparison of methods for the evaluation of binary measurement systems. *Quality Engineering* 2005; **17**:495–507.
15. Danila O, Steiner SH, MacKay RJ. Assessment of a binary measurement system in current use. *Journal of Quality Technology* 2010; **42**:152–164.

Authors' biographies

T. S. Akkerhuis is Consultant and PhD Student at the Institute for Business and Industrial Statistics at the University of Amsterdam.

J. de Mast is Principal Consultant and Professor of Methods and Statistics for Operations Management at the Institute for Business and Industrial Statistics of the University of Amsterdam.