Research Article

Quality and
Reliability
Engineering
International

(wileyonlinelibrary.com) DOI: 10.1002/qre.1214

Published online 4 May 2011 in Wiley Online Library

Assessment of Binary Inspection with a Hybrid Measurand

Tashi P. Erdmann*[†][‡] and Jeroen de Mast[§]

This paper addresses issues that arise in measurement system analysis of a binary measurement system if the measurand is a hybrid between a dichotomy and a continuum. A case study is presented, which illustrates methods to assess the error rates of binary measurements with such a hybrid measurand. The case study concerns pass/fail inspection of laptop screens for scratches, where the measurand is the presence or absence of scratches. If a scratch is present, the measurand corresponds with a continuum of scratch sizes, but if no scratch is present, the measurand corresponds with a point. It is argued that if the measurand is a hybrid, a standard logistic regression model is not suitable to estimate the characteristic curve relating the reject probability with the measurand. Several alternative specifications for the characteristic curve are introduced and compared. We conclude that many of the methods currently used for assessment of a binary measurement system with a hybrid measurand are unsuited. This is a remarkable conclusion, given the frequent occurrence in industry of leak tests, inspections for defects, and other binary measurement systems with a hybrid measurand. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: binary measurement; inspection errors; measurement system analysis; misclassification rates; pass/fail inspection

1. Introduction

Binary measurement classifies items into either of two categories, $Y=R$ ('reject') or $Y=A$ ('accept'), reflecting a property X of the item that is not observed directly. A common industrial application is pass/fail inspection, where X is a quality characteristic such as the state of a light bulb (functional or defective), or the amount of discoloration of a food product. The property X underlying the measurement Y is often referred to as the 'true value,' but is called the 'measurand' by the *Guide to the Expression of Uncertainty in Measurement* (*GUM*)[1]. As De Mast *et al.*[2] note, X can be binary itself, as in the case of light bulbs being functional or defective. But in many cases, X is a continuous property.

Measurement system analysis (MSA) studies the quality of measurements. The quality of binary measurements can be expressed as error rates. The probability that a defective item passes is the *False Acceptance Probability* (*FAP*), and the probability that a good item fails is the *False Rejection Probability* (*FRP*). Recently, a flow of literature has appeared regarding MSA of binary measurements, describing methods to determine the *FAP* and *FRP*, such as Boyles[3], Pepe[4], Van Wieringen and Van den Heuvel[5], Danila *et al.*[6], Van Wieringen and De Mast[7], Danila *et al.*[8], and Beavers *et al.*[9]. Some of these methods treat the measurand as binary, whereas others treat it as continuous. De Mast *et al.*[2] analyze the complications brought about if a method treats X as binary when it is actually continuous, suggesting that the choice of method should depend on the measurand being binary or continuous. Moreover, when the measurand is continuous, it is often desirable to know the rejection probability for any value of the measurand $X=x$, as this allows one to make statements about measurement reliability without reference to a specific population of items.

In many cases, however, the measurand X is neither binary nor continuous, but a hybrid. A leak test is an example; the outcome $Y=A$ reflects the point $X=0$, whereas the outcome $Y=R$ corresponds to a continuum $X>0$, with X being the size of the leak. We call this a hybrid measurand.

The overview of De Mast *et al.*[2] shows that there are currently no methods intended for situations with a hybrid measurand. Treating it as binary creates the same complications as treating a continuous measurand as binary. In particular, it creates an intrinsic reason for violation of the assumption that measurements be conditionally i.i.d. (independent and identically distributed). This results in a biased estimate of *FAP* unless defective items are randomly sampled, but it is difficult to obtain a random sample

Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA), Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands

Correspondence to: Tashi P. Erdmann, Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA), Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands.

[†] *E-mail: t.p.erdmann@uva.nl*

[‡] *PhD Student in Industrial Statistics and Consultant at IBIS UvA.*

[§] *Associate Professor in Industrial Statistics and Principal Consultant at IBIS UvA.*

with sufficient defective items. Treating it as continuous, for example by logistic regression, is more informative, but requires modifications of the used models.

The purpose of this paper is, based on a somewhat simplified but yet realistic case study, to explore the problem of binary MSA with a hybrid measurand and give options toward a solution. The case study concerns visual inspection for scratches on a laptop screen. We propose and compare alternative models for the relationship between *Y* and *X*, and for the distribution of the measurand, allowing assessment of the quality of measurement. Throughout this paper we will assume that a gold-standard measurement is available, allowing us to obtain the measurand (or reference value) of all items in the sample. It is often expensive to obtain such a gold-standard measurement, and therefore it is advantageous to use only a small number of items in the MSA experiment.

The following section introduces the case study. In Section 3, the problems occurring with standard methods are discussed. In Sections 4 and 5 we propose alternative methods and apply them to the scratch example. In Section 6 conclusions are drawn.

2. Case study: visual inspection for scratches

In order to illustrate the difficulties that arise when the measurand is a hybrid, we present an example of an MSA experiment involving a pass/fail inspection. We have in mind a visual inspection of laptop displays for scratches. A display is good if there is no scratch ($X=0$), and defective if there are one or more scratches with positive scratch size ($X>0$). Since we are not in a position to do such an experiment involving real scratches, we designed a similar experiment that mimics such scratch tests. Instead of real scratches, the computer shows white bitmaps, some of which have a gray curve representing a scratch. All such curves have the same length and width, but varying graynesses and varying locations on the screen. We left the situation of multiple scratches out of consideration. The curves' grayness *X* mimics the varying depths of real scratches. It is measured in percents, varying from 10 to 46% in steps of 4%. The scratches are located on ten different locations on the screen.

The experiment was set up as follows. Twenty appraisers, assumed to be randomly selected from the population of appraisers, each inspected 100 different laptop screens, in random order. They sat in front of the screen with their eyes approximately 80 cm from the screen, and judged within 10 s whether a screen had a scratch on it. The full factorial design in ten levels of grayness and ten locations (100 combinations in total) was divided into two blocks of 50 combinations each. Half of the appraisers inspected the combinations in one block, and the other half the combinations in the other block. In addition, the appraisers inspected 50 samples that had no scratch.

For each value of *X*, the number of rejected and accepted screens are displayed in Table I. A peculiarity of the results is that at $X=30$ fewer screens were rejected than might be expected. At that level of grayness, 14 out of the 100 scratches were missed, whereas at $X=34$ only two were missed. This is visible in Figure 1 as a peculiar kink in the curve that would result from connecting the points that give the empirical rejection fraction for each value of *X*. We have no explanation for this kink, but

Table I. Data from the scratch experiment

Appraiser:			A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
X	*Y*	*Freq.*																				
0	*R*	18	1	0	2	0	3	0	0	2	0	1	0	1	2	0	0	4	0	0	0	2
10	*R*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	*R*	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
18	*R*	19	0	1	1	3	1	1	1	3	1	1	0	0	0	2	1	1	0	0	1	1
22	*R*	63	2	2	2	5	1	2	4	3	3	4	1	3	2	5	5	3	5	2	4	5
26	*R*	84	3	5	3	5	4	5	5	5	5	5	3	4	2	5	3	4	4	4	5	5
30	*R*	86	4	3	1	5	5	4	5	4	4	5	3	5	4	5	5	5	4	5	5	5
34	*R*	98	5	5	5	5	5	4	5	5	5	5	4	5	5	5	5	5	5	5	5	5
38	*R*	98	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	5	4	5
42	*R*	99	5	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	5	5
46	*R*	100	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
0	*A*	982	49	50	48	50	47	50	50	48	50	49	50	49	48	50	50	46	50	50	50	48
10	*A*	100	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
14	*A*	99	5	5	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5
18	*A*	81	5	4	4	2	4	4	4	2	4	4	5	5	5	3	4	4	5	5	4	4
22	*A*	37	3	3	3	0	4	3	1	2	2	1	4	2	3	0	0	2	0	3	1	0
26	*A*	16	2	0	2	0	1	0	0	0	0	0	2	1	3	0	2	1	1	1	0	0
30	*A*	14	1	2	4	0	0	1	0	1	1	0	2	0	1	0	0	0	1	0	0	0
34	*A*	2	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
38	*A*	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
42	*A*	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
46	*A*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

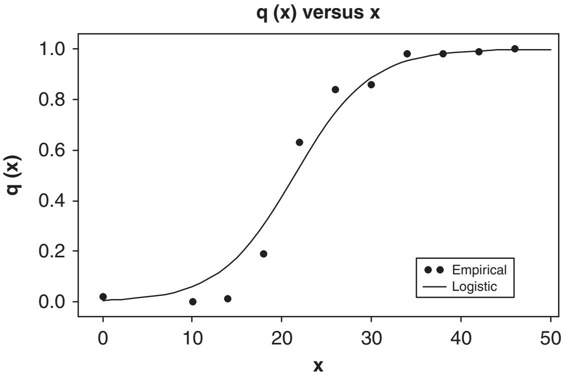

Figure 1. Plots of the characteristic curve

we know that it is not an artifact of the location of the scratches on the screen, because the location on the screen was varied according to the experimental design described above.

3. Complications with the current methods

Borrowing the setup of De Mast *et al.*[2], we note that the probability of rejecting a screen is a function, named *characteristic curve*, of the scratch size x:

$$q(x)=P(Y=R|X=x)$$

Typically, $q(x)$ is an S-curve. Let $F_X^d(x)=P(X\leqslant x|X>0)$ be the distribution of scratch sizes in the subpopulation of defective items. The measurement system could be evaluated in terms of its error rates $FRP=q(0)$ and

$$FAP=\int_{x>0} q(x)f_X^d(x)\,\mathrm{d}x \tag{1}$$

that is, FAP is the average $q(x)$ over the interval $x>0$, weighted by $f_X^d(x)$. Note that FAP depends on the population of items. In some applications, one needs to specify the measurement quality without reference to the population of items, as it is unknown. For example, a manufacturer of leak testers must specify the performance of its products without knowing the population distribution of leaks. In such cases, the measurement system could be better characterized by stating a limit value x° such that $q(x^{\circ})=0.90$ (or some other probability).

Two out of the methods described by De Mast *et al.*[2] are candidates for our situation. The first is nonparametric estimation of error rates based on sample proportions, in which the measurand is treated as binary. The measurement system is evaluated in terms of $q(x)$, with $x=0$ (no scratch) or $x=1$ (scratch), instead of a continuum of positive x-values representing scratch size. For each of the 100 screens in the MSA experiment, the measurand (scratch or no scratch) is given as well as the evaluations by 20 appraisers. The FRP and FAP are estimated from sample proportions. $\widehat{FRP}=0.018$ and $\widehat{FAP}=0.352$, because good screens were incorrectly rejected 18 times (out of 1000 appraisals) and scratched screens were accepted 352 times (out of 1000 appraisals). This analysis is misguided; since the defective items were not sampled randomly, but selectively at equidistant values for x at 4% increments, the estimate of FAP is inconsistent. In particular:

$$E(\widehat{FAP})=\int_{x>0} q(x)f_X^s(x)\,\mathrm{d}x$$

where $F_X^s(x)$ is the sampling distribution of X in the subsample of defective items, (i.e. the distribution determined by the sampling mechanism). Since the defective items were not sampled randomly, $E(\widehat{FAP})\neq E(FAP)$ as defined in Equation (1), unless $q(x)$ is constant for positive x.

Instead of the setup of our experiment, the literature[2, 6] offers three standard setups, but these are either difficult to apply or require a large sample size. One setup is to take random subsamples from the subpopulations of good and defective items separately. The proper way to do this is to use the gold-standard measurement to measure all items that are produced until a sufficiently large amount of defective items is obtained. In this way, two subpopulations are created: one of good items and the other of defective items. Then, from each of these subpopulations a sample is taken. However, if the defect rate is low, as is usually the case, it takes an unrealistically large effort to create such subpopulations. For example, if the defect rate is 1% and one wants to obtain subsamples of 30 items, thousands of items will need to be measured by the gold-standard measurement procedure. Plan I by Danila *et al.*[6] solves this problem by sampling from the streams of rejected and accepted items instead. The measurand of the sampled items is determined by applying the gold-standard measurement. Then, one calculates the proportion of rejected items that were actually good, and the proportion of accepted items that were actually defective. If one also knows the historical reject rate, one can calculate the error rates of the measurement system by applying Bayes' Law. However, Plan I only uses the original evaluation as accepted or rejected; it does not allow for repeated measurements of the same item and therefore it requires large sample sizes of, say, 300 items per subsample. Another approach (Plan II) is to sample directly from the population of all items, but, in case of a low defect rate, this requires an unrealistically large sample size in order to include sufficient defective items. Again, thousands of items need to be measured by the gold-standard measurement. Of these three standard setups, only Plan I is applicable if the defect rate is low.

Besides the sampling problems discussed above, a second disadvantage of the nonparametric approach based on sample proportions is that it is less informative than logistic regression. It does not give the complete characteristic curve $q(x)$, but only the proportions *FRP* and *FAP*. The latter depends on the population of items per Equation (1). Therefore, in applications where the population distribution of items is not fixed, the nonparametric approach is not applicable. In such applications one is typically interested in the limit value x°.

In summary: nonparametric estimation of error rates, treating the measurand as binary, is unsatisfactory, as sampling schemes facilitating unbiased estimation are difficult to apply or require large sample sizes, and the reduction of the characteristic curve to an *FAP* and *FRP* is an unsatisfactory summary of the stochastic behavior.

The other method offered by De Mast *et al.*[2] is logistic regression, which treats the measurand as continuous. One specific variant of this approach is AIAG's analytic method[10]. It estimates a parametric model $q_\theta(x)$ for the characteristic function, with parameter vector θ, from a sample obtained by selecting a number of items such that their measurands are more or less equidistant. Each item j is classified m_j times (AIAG[10] recommends $m_j=20$ for all j). The number of rejects for covariate value x_j is r_j. The parameter vector θ is then estimated by maximizing the log likelihood function:

$$l(\theta)=\sum_j (r_j \log q_\theta(x_j)+(m-r_j)\log(1-q_\theta(x_j)))$$

In order to characterize the measurement system, *FRP* can be estimated as $q_{\hat{\theta}}(0)$, but *FAP* is more difficult to estimate. Departing from Equation (1), for q_θ one substitutes $q_{\hat{\theta}}$, and for F_X^d one fits a parametric distribution function $\hat{F}_X^d(x)$, whose parameters are estimated from a random sample of defective items. Natural choices for $\hat{F}_X^d$ are nonnegative distributions such as the exponential, lognormal, and Weibull distributions. Note that it is difficult to obtain a random sample of defective items. It cannot be taken randomly from the rejected items because if the defect rate is low, a substantial part of the rejected items will actually be good, and difficult to judge defective items will be underrepresented, as shown by De Mast *et al.*[2]. A possible solution is to select items randomly from the population of all items, apply a gold-standard measurement to determine whether each item is good or defective, and continue until a sufficient number of defective items are obtained. If the defect rate is low, an unrealistically large number of items need to be measured with the gold-standard measurement, for the same reasons as described above in the context of standard setups for nonparametric estimation of error rates based on sample proportions. One could avoid estimating $F_X^d(x)$ altogether if one is not interested in the *FAP* and instead evaluate the measurement system in terms of a limit value x°.

The standard logistic regression model for $q(x)$ is the distribution function of a standard normal or a logistic distribution with a linear function of x as its argument:

$$q^{\text{Prob}}(x)=\Phi(\alpha+\beta x) \tag{2}$$

$$q^{\text{Log}}(x)=\frac{1}{1+e^{-(\alpha+\beta x)}} \tag{3}$$

both with $\beta>0$.

This standard model is not appropriate if the measurand is a hybrid, as in the scratch example. The particular forms of q^{Prob} and q^{Log} imply point symmetry in their inflection point, and fix $q^{\text{Prob}}(0)$ and $q^{\text{Log}}(0)$ given the location and slope at the inflection point; these properties will be shown to be unsuited for situations with a hybrid measurand.

We illustrate these problems by applying standard logistic regression, based on model (3) and with the restriction $x\geqslant 0$, to the example. To simplify matters, we treat the appraisers as interchangeable, and treat the results of the 20 appraisers as replications. The maximum likelihood estimates are $\hat{\alpha}=5.15$ and $\hat{\beta}=0.24$. The estimation results are summarized in Table II, and graphically displayed in Figure 1. For each value of x, the empirical rejection fraction is also displayed.

To assess the goodness of fit we compute the chi-square statistics based on Pearson and deviance residuals[11]. The Pearson residual u_j of the j'th covariate value x_j equals

$$u_j=\frac{r_j-m_j\hat{q}(x_j)}{\sqrt{m_j\hat{q}(x_j)(1-\hat{q}(x_j))}}$$

where m_j is the total number of items with covariate value x_j of which r_j is the number of items rejected, and $\hat{q}(x_j)$ is the estimated reject probability. The squared deviance residual d_j^2 equals

$$d_j^2=2\left(r_j\ln\left(\frac{r_j}{m_j\hat{q}(x_j)}\right)+(m_j-r_j)\ln\left(\frac{m_j-r_j}{m_j(1-\hat{q}(x_j))}\right)\right)$$

Table II. Results of logistic regression

Coefficient	Coefficient	SE coef	*P*-value
Parametric fit of characteristic curve using logistic regression			
α	−5.153	0.267	0.0000
β	0.2403	0.012	0.0000
Log-likelihood	−355.5		
Goodness-of-fit tests			
Method	Chi-square	DF	*P*-value
Pearson	64.09	9	0.0000
Deviance	71.13	9	0.0000

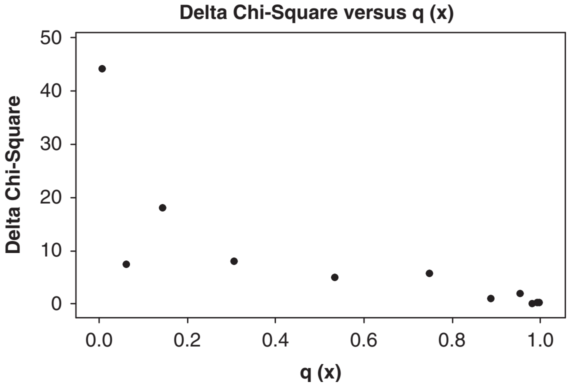

Figure 2. Plot of delta chi-square against the reject probability

The chi-square statistics are obtained by summing the corresponding version of squared residuals over all covariate values. They asymptotically follow a chi-square distribution with degrees of freedom equal to $J-k$, where J is the number of different covariate patterns and k is the number of parameters in the model. Both goodness-of-fit tests indicate that the true reject probabilities are significantly different from the estimated ones (*P*-value$=$0.000 for both tests). The model specification is clearly incorrect.

Further evidence of the inadequacy of the model is obtained by plotting *delta Pearson chi-square* against the reject probability $q(x)$ for each scratch size. Delta Pearson chi-square is a measure for the change in Pearson chi-square that results if a certain covariate pattern (scratch size) is left out of the data set. A plot of delta chi-square against the reject probability is shown in Figure 2.

A high delta chi-square indicates that for a certain scratch size the percentage of screens rejected is far from the estimated reject probability. Hosmer and Lemeshow[11] state that if the model fits well, the upper 95 percentile of the delta chi-square is approximately 4. In our case most of the scratch sizes have delta chi-squares higher than 4, indicating a very poor fit. Especially for the covariate pattern $x=0$ (screens with no scratch) the delta chi-square is extremely high (44.3). The empirical rejection fraction of good screens was 0.018 in the experiment, whereas the predicted false rejection probability is only 0.006. The False Negative Fraction $q(0)$ estimated by this model is too close to zero, because it is not allowed to vary independently from the location and slope of the curve. The model assumes symmetry and does not allow the slope to get close to zero for values of $q(0)$ that differ too much from 0, and therefore $q(0)$ is underestimated.

Clearly, both nonparametric estimation of error rates and the standard logistic regression model are unsuited for the analysis of the probability of detecting a scratch. In the following section more adequate methods are proposed to model the characteristic function $q(x)$.

4. Parametric solutions

We aim to modify the logistic regression model to overcome the problems described in the previous section, by exploring alternative functions for the characteristic curve $q(x)$. In order to identify and evaluate options for $q(x)$, we reflect on the physical mechanisms that determine the stochastics of the inspections. First, there is the possibility that an appraiser imagines seeing a scratch that is not actually there. Second, given that there is a scratch, the randomness in the inspection results is induced by such mechanisms as:

- Differences in eyesight between appraisers, and fluctuations in a single appraiser's eyesight over time, including fluctuations in concentration and motivation.
- Differences in circumstances, such as ambient light, angle of vision, distance to the screen.

- The randomness of the search process, where the appraiser's eyes are scanning the screen, trying to locate a possible scratch in a limited amount of time. As a matter of fact, even a relatively deep scratch may be missed because the appraiser, in the restricted amount of time given, happens not to scan the area where it is located.

We believe that beyond the characteristic curve's inflection point, scratch size is progressively less influential in the randomness of the inspection results—the probability of detecting a scratch is almost completely determined by the random outcome of the search for its location, and increases in scratch size have virtually no further effect. This is one of the motivations to consider mathematical functions for $q(x)$ that are flexible in the degree of asymmetry that they can represent.

In view of these physical mechanisms, ideally, we have a mathematical function $q(x)$ with sufficient degrees of freedom to represent these four aspects:

(1) The probability $q(0)$ that a good screen is rejected.
(2) The location of the inflection point $x^* = \{x : q''(x) = 0\}$.
(3) The slope at the inflection point $x^* = q'(x)$.
(4) The height of the inflection point $q(x^*)$, determining the amount of symmetry of the curve.

A natural general form for $q(x)$ is

$$q(x) = q(0) + (1 - q(0))G(x)$$

with G a nonnegative distribution function. We will refer to this form as a *zero-inflated* nonnegative distribution. That is, a nonnegative distribution rescaled such that $x = 0$ with positive probability. For the logistic distribution in Equation (3), we have $x^* = -\frac{\alpha}{\beta}$; $G(x^*) = \frac{1}{2}$; $g(x^*) = \frac{\beta}{4}$, where g is the derivative of G. The height of the inflection point is fixed at $G(x^*) = \frac{1}{2}$ and the curve is symmetrical, which makes this an unsuitable option for our purpose, as noted earlier. We discuss alternative options.

The Weibull distribution is a flexible nonnegative distribution. Its distribution function is given by

$$G(x) = 1 - \mathrm{e}^{-(x/\beta)^{\alpha}}$$

with α, $\beta > 0$. The parameter α is called the shape parameter and β the scale parameter. The distribution is positively skewed for $0 < \alpha < 3.6$. The location, height, and slope at the inflection point are:

$$x^* = \beta\left(\frac{\alpha - 1}{\alpha}\right)^{\frac{1}{\alpha}}, \quad G(x^*) = 1 - \mathrm{e}^{\frac{1-\alpha}{\alpha}}, \quad g(x^*) = \frac{\alpha - 1}{\beta}\left(\frac{\alpha}{\alpha - 1}\right)^{\frac{1}{\alpha}} \mathrm{e}^{\frac{1-\alpha}{\alpha}}$$

A disadvantage of the Weibull distribution for our purposes is that the coordinates of the inflection point fix the slope $g(x^*)$, through the relation:

$$g(x^*) = \frac{(G(x^*) - 1)\ln((1 - G(x^*))}{x^*(1 + \ln(1 - G(x^*)))}$$

For positively skewed Weibull distributions ($\alpha < 3.6$), $g(x^*)$ can be at most $1.26/x^*$. The empirical rejection fractions of the scratch inspection, however, suggest a positively skewed distribution with x^* around 20 and $g(x^*)$ around 0.11, properties unobtainable by the Weibull.

A distribution which is capable of fitting these results is the log-logistic distribution, with the following distribution function:

$$G(x) = \frac{1}{1 + (x/\beta)^{-\alpha}}$$

with α, $\beta > 0$. The log-logistic distribution is skewed to the right for all parameter values, although technically its third moment only exists for $\alpha > 3$. The location, height, and slope at the inflection point are:

$$x^* = \beta\left(\frac{\alpha - 1}{\alpha + 1}\right)^{\frac{1}{\alpha}}, \quad G(x^*) = \frac{\alpha - 1}{2\alpha}, \quad g(x^*) = \frac{\alpha^2 - 1}{4\alpha\beta}\left(\frac{\alpha + 1}{\alpha - 1}\right)^{\frac{1}{\alpha}}$$

Again, the slope is fixed by the coordinates of the inflection point:

$$g(x^*) = \frac{G(x^*)(1 - G(x^*))}{x^*(1 - 2G(x^*))}$$

The logistic distribution ascends more steeply at the inflection point than the Weibull for any given coordinate pair of the inflection point obtainable by both distributions (i.e. $0 < G(x^*) < 0.5$).

A distribution which has an additional parameter is the generalized logistic distribution[12]. This distribution is less commonly used than the distributions discussed so far. It has the distribution function

$$G(x) = \frac{1}{(1 + \mathrm{e}^{-(\alpha + \beta x)})^{\gamma}}$$

with β, $\gamma > 0$. For $\gamma = 1$ this distribution is equal to the logistic distribution as in Equation (2). For $\gamma > 1$, the distribution is positively skewed. The location, height, and slope at the inflection point are: $x^* = (\ln\gamma - \alpha)/\beta$; $G(x^*) = 1/((1 + \gamma^{-1})^{\gamma})$; $g(x^*) = \beta/((1 + \gamma^{-1})^{\gamma+1}$.

Table III. Results of maximum likelihood estimation for six different models

Model	ZI Logistic	ZI Weibull	ZI Log Log	ZI Gen Log	ZI GEV	ZI Tr Weib
Coefficient						
Parametric fit of characteristic curve						
$q(0)$	0.01465	0.01534	0.01531	0.01548	0.01579	0.01583
α	−7.278	3.997	7.744	11.129	−5.854	1.011
β	0.3285	24.661	21.600	0.2475	0.2973	5.4145
γ	—	—	—	9500000	0.1637	16.9086
Log-likelihood	−338.7	−348.4	−329.8	−327.4	−325.4	−324.7
Goodness-of-fit test based on Pearson residuals						
Chi-square	36.60	86.98	17.71	13.82	9.45	7.81
DF	8	8	8	7	7	7
P-value	0.0000	0.0000	0.0235	0.0544	0.2219	0.3498
Goodness-of-fit test based on deviances						
Chi-square	37.49	56.93	19.78	14.76	11.00	9.52
DF	8	8	8	7	7	7
P-value	0.0000	0.0000	0.0112	0.0391	0.1388	0.2175
Aspects of inflection point						
x^*	22.15	22.95	20.89	19.95	19.19	16.97
$q(x^*)$	0.5073	0.5348	0.4441	0.3777	0.3232	0.0263
$q'(x^*)$	0.08092	0.06076	0.08974	0.08963	0.10901	0.17319

The three parameters allow the distribution to represent all three aspects separately. A disadvantage of the generalized logistic distribution is that the height of its inflection point $G(x^*)$ cannot be less than $e^{-1} \approx 0.37$. Furthermore, if γ is large such that $G(x^*) \approx e^{-1}$ and $g(x^*) \approx \beta e^{-1}$, the distribution has a nearly identical shape for any combination of α and γ that keeps the value of x^* constant. This creates problems regarding identification of γ and α if the height of the inflection point is close to e^{-1} or less.

A three-parameter distribution that is less restrictive for the height of the inflection point is the generalized extreme value distribution introduced by Jenkinson[13]. Its distribution function is:

$$G(x) = e^{-(1+\gamma(\alpha+\beta x))^{-1/\gamma}}$$

for $1+\gamma(\alpha+\beta x) > 0$ with $\beta > 0$. Special cases of this distribution are the Gumbel (limit for $\gamma \to 0$), Frechet ($\gamma > 0$) and reversed Weibull ($\gamma < 0$) distributions. The location, height, and slope at the inflection point are:

$$x^* = \frac{(\gamma+1)^{-\gamma} - \alpha\gamma - 1}{\beta\gamma}, \quad G(x^*) = e^{-\gamma-1}, \quad g(x^*) = \beta(\gamma+1)^{\gamma+1} e^{-\gamma-1}$$

By changing γ, the inflection point can obtain any height between 0 and 1. For $\gamma > 0$ it is lower than e^{-1}, the minimum height of the inflection point for the generalized logistic distribution.

The last option that we consider is the translated Weibull distribution with a third parameter γ that determines the horizontal translation. Both the translated Weibull and generalized extreme value distributions have a discontinuity in one of their (higher order) derivatives at the point where they start increasing, $x = \gamma$ for translated Weibull and $x = -(\alpha\gamma+1)/\beta\gamma$ for the generalized extreme value distribution, although it may occur for negative x and thus fall outside the domain of $q(x)$. In the scratch example there is no physical explanation for such a discontinuity.

Table III gives the maximum likelihood estimation results for the six different specifications of $q(x)$ proposed: the zero-inflated logistic, Weibull, log-logistic, generalized logistic, generalized extreme value, and translated Weibull distribution. The resulting characteristic curves are shown in Figure 3. The figures also display the empirical rejection fractions for comparison.

The curves resulting from the zero-inflated logistic, Weibull, and log-logistic models do not seem to adequately model the asymmetry in $q(x)$ and the steep slope at its inflection point. This is confirmed by the goodness-of-fit tests. Both the chi-square based on Pearson residuals and the chi-square based on deviances indicate significant lack of fit for all three distributions. The zero-inflated Weibull distribution results in even a worse fit than the zero-inflated logistic distribution.

The curve resulting from the zero-inflated generalized logistic model is much closer to the empirical rejection fractions. However, because the height of the inflection point is close to $q(0)+(1-q(0))e^{-1}$, the parameters of this model are hard to identify by maximum likelihood due to the near-unidentifiability problem described in the previous section. After trying several different starting values for the parameter estimates, the highest log likelihood that we were able to obtain results in the estimates in the table. Despite the estimation problem, the zero-inflated generalized logistic distribution model gives a good fit. It has a high log likelihood and low goodness-of-fit chi-square statistics. The goodness-of-fit chi-square statistic based on Pearson residuals does not indicate significant lack of fit at the 5% significance level, but the one based on deviances does.

The generalized extreme value distribution and the translated Weibull distribution give the best fit. They have the highest log likelihoods and the lowest goodness-of-fit chi-square statistics. Both tests do not reject a good fit. The translated Weibull distribution fits slightly better according to these measures, but has some questionable properties. The characteristic curve fitted

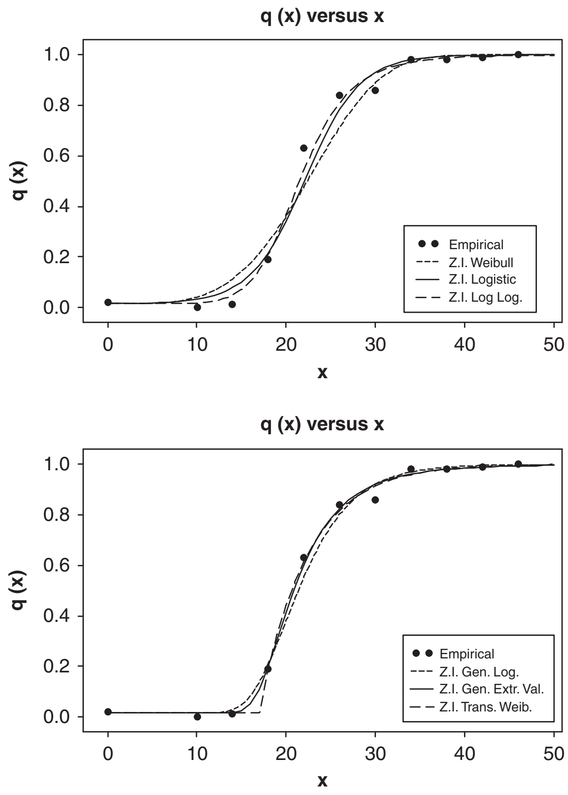

Figure 3. Plots of the characteristic curve for six different models

with the translated Weibull model has a kink at $x=16.9$, where the second-order derivative is discontinuous and jumps from zero to infinity. We cannot think of a physical explanation for such a discontinuity. Furthermore, the coordinates of the inflection point of the Weibull differ largely from those of all other models: they are located near this discontinuity at $\hat{x}^*=17.0$ and $\hat{q}(\hat{x}^*)=0.026$. The fitted generalized extreme value distribution does not have a discontinuity for positive x, and therefore seems the better choice. It leads to a characteristic function $\hat{q}(x)$ with $\hat{q}(0)=0.0158$, $\hat{x}^*=19.2$, $\hat{q}(\hat{x}^*)=0.323$, and $\hat{q}'(\hat{x}^*)=0.109$. *FRP* is estimated as $\hat{q}(0)=0.0158$ and if the distribution of scratch sizes in the population of defective items $F_X^d(x)$ were known, *FAP* could be estimated by substituting the fitted $q(x)$ in Equation (1). Note that if the defect rate is low, it is practically very difficult—if not impossible—to obtain $\hat{F}_X^d(x)$, because of the sampling problems explained in the previous section. Note that in our (artificial) experiment *FAP* cannot be evaluated, because there is no population of items. The limit value x° such that $q(x^\circ)=0.90$ is estimated as $\hat{x}^\circ=28.8$, so scratches with scratch size larger than 28.8 are detected with more than 90% probability.

5. Nonparametric solutions

If no parametric model for $q(x)$ fits well, the curve could be estimated by nonparametric logistic regression[14, 15]. The logistic regression model is generalized to

$$q(x)=\frac{1}{1+e^{-s(x)}} \tag{4}$$

where s is a smooth function, such as a cubic spline. A cubic spline is a twice differentiable continuous function consisting of piecewise cubic polynomials, whose pieces are separated by a sequence of breakpoints called *knots*. A useful function in the software program *S* to perform nonparametric logistic regression is the function *gam*[15].

Hastie and Tibshirani[15] fit s using a procedure that they call the *local scoring* algorithm, because it is based on the Fisher scoring procedure that is used to find the maximum likelihood estimates in linear logistic regression. They start with initial values of $s(x_i)$ for all observations i, and then compute adjusted values z_i, as follows:

$$z_i=s_{\text{old}}(x_i)+\frac{y_i-q_{\text{old}}(x_i)}{q_{\text{old}}(x_i)(1-q_{\text{old}}(x_i))}$$

The new function s_{new} is obtained by fitting z to x using a scatterplot smoother and $q_{\text{new}}(x)=1/(1+e^{-s_{\text{new}}(x)})$. Then in the next iteration s_{new} and q_{new} become s_{old} and q_{old}. This procedure is repeated until the change in deviance is less than a specified amount.

The scatterplot smoother that we will use is a cubic smoothing spline. The cubic smoothing spline that fits z to x is defined as the function s that minimizes the penalized residual sum of squares $RSS(s,\lambda)=\sum_i(z_i-s(x_i))^2-\frac{1}{2}\lambda\int(s''(t))^2\,dt$, where λ is the smoothing parameter determining the degree of smoothing. It can be shown that the solution to this minimization problem is a natural cubic spline with knots at all distinct values of x_i. That is, it is a cubic polynomial in each interval (x_i, x_{i+1}), its first two derivatives are continuous, and it is linear below x_1 and above x_n. Numerical algorithms to find fitted values of s are given by De Boor[16].

Hastie and Tibshirani[15] use the notion of effective degrees of freedom of a smoother in order to be able to compare them with parametric models. The fitted values of a smoothing spline at the observed points x_i can be written as linear combinations of the values of the dependent variable z_i

$$\hat{\mathbf{s}}=\mathbf{S}\mathbf{z}$$

where the matrix **S** depends only on the x_i and on λ, and is called the *smoother matrix*. The number of effective degrees of freedom is calculated as the trace of **S**. This is because of the analogy of **S** to the projection matrix in linear regression analysis that transforms the observed values of the dependent variable into the fitted values, often called the *hat matrix*. The effective degrees of freedom are determined by the smoothing parameter λ. If $\lambda=0$ the smoothing spline simply interpolates the data and the number of degrees of freedom of the curve equals the number of distinct values of x. If λ approaches infinity, the smoothing spline approaches the ordinary least squares line, and the number of degrees of freedom used approaches two. The smoothing parameter λ can be chosen to fix the degrees of freedom to a certain number. A common criterion is to choose it such that the cross-validation sum of squares is minimized[15].

An advantage of nonparametric logistic regression is that it can be generally applied to any data set, in contrast to the parametric models discussed in the previous section, which were chosen specifically for the data set in this example. A disadvantage is that the resulting curve is defined implicitly as the solution of a minimization problem and no explicit function of x is obtained[17].

Nonparametric logistic regression using a smoothing spline leads to the fitted values of $q(x)$ given in Table IV. Two splines are fitted with four and five effective degrees of freedom, respectively. The characteristic curves are plotted in Figure 4. The spline with four degrees of freedom does not fit well, as indicated by both chi-square statistics. The spline with five degrees of freedom gives a good fit. Its log likelihood and chi-square statistics are comparable to the generalized extreme value distribution, although it uses one (effective) degree of freedom more. Both chi-square statistics do not reject a good fit at the 5% level. The approximate characteristics of the inflection point are $\hat{x}^*=20.5$, $\hat{q}(\hat{x}^*)=0.425$, and $\hat{q}'(\hat{x}^*)=0.107$, indicating that the inflection point is located slightly higher and more to the right than for the parametric distributions with the best fit. The estimate of *FRP* is $\hat{q}(0)=0.0166$. Again, *FAP* cannot be evaluated for our experiment because we do not have a population of defective items.

Table IV. Results of nonparametric logistic regression

Nonparametric fit of characteristic curve using smoothing spline

Effective d.o.f.	5	4
0	0.0166	0.0149
10	0.0147	0.0282
14	0.0407	0.0668
18	0.1929	0.2145
22	0.5832	0.5428
26	0.8196	0.8015
30	0.9082	0.9150
34	0.9602	0.9647
38	0.9835	0.9855
42	0.9934	0.9941
46	0.9975	0.9976
Log-likelihood	−326.0	−331.8
Goodness-of-fit test based on Pearson residuals		
Chi-square	9.53	18.37
DF	6	7
P-value	0.1461	0.0104
Goodness-of-Fit test based on deviances		
Chi-square	12.12	23.70
DF	6	7
P-value	0.0594	0.0013
Aspects of inflection point (approximations)		
x^*	20.45	20.96
$q(x^*)$	0.425	0.451
$q'(x^*)$	0.107	0.090

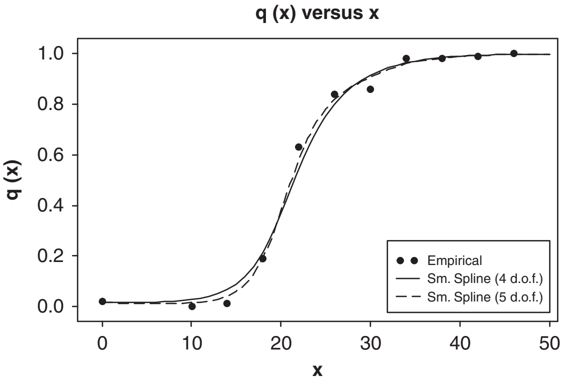

Figure 4. Plot of the characteristic curve for nonparametric logistic regression

The fitted $q(x^\circ)=0.90$ for $\hat{x}^\circ=29.5$, reasonably close to the value of x° based on the zero-inflated generalized extreme value distribution.

6. Summary and conclusions

The case study discussed in this paper shows that assessing a binary measurement system with a hybrid measurand is a difficult task. Even common binary inspections such as a leak test or an inspection for scratches are hard to assess. If *FRP* and *FAP* are determined using sample proportions of incorrectly rejected and incorrectly accepted items, the estimate of *FAP* critically depends on the randomness of the sample, but a sufficiently large random sample of defective items is practically difficult to obtain if the defect rate is low. This paper attempts to approach the binary MSA with an estimation of the characteristic curve $q(x)$ giving the reject probability as a function of the measurand x. Because standard logistic regression models are not appropriate for a nonnegative measurand such as scratch size, an appropriate asymmetric specification for $q(x)$ needs to be fitted to the data. In the case study discussed in this paper three specifications for $q(x)$ are found that reasonably model the specific shape of $q(x)$. They all have four parameters, creating the need for a large sample size. Another possible approach, which is more generally applicable, is nonparametric logistic regression, but the disadvantage is that it does not give an explicit function for $q(x)$. Once the function $q(x)$ has been estimated, *FRP* and a value x° such that $q(x^\circ)=0.90$ could be used to evaluate the measurement system. If one is interested in *FAP*, the distribution $F_X^d(x)$ of the measurand in the population of defective items needs to be estimated as well, again requiring a random sample of defective items, which is difficult to obtain.

Estimating $q(x)$ and $F_X^d(x)$ becomes even more difficult if the measurand is not a one-dimensional property such as grayness, but a multidimensional set of properties, for example: grayness, length of the scratch, number of scratches, etc. Reliably estimating *FAP*, *FRP*, or x° in such cases is a formidable challenge.

We conclude that the methods currently used for assessment of a binary measurement system with a hybrid measurand are often unsuited. This is a remarkable conclusion, given the frequent occurrence in the industry of leak tests, inspections for defects, and other binary measurement systems with a hybrid measurand. Methods aimed at expressing *FAP* and *FRP* will necessarily be bound to a certain study population of items, and often require unrealistic sampling plans. In order to correctly assess the quality of measurements, one either needs to (1) apply nonparametric estimation using sample proportions based on subpopulations of accepted/rejected rather than good/defective items, as is done in Plan I by Danila *et al.*[6], or (2) estimate the curve $q(x)$ rather than *FAP*. The curve $q(x)$ could be estimated based on repeated measurements at a number of values of x, using an appropriate zero-inflated distribution function as a model, or using nonparametric logistic regression.

References

1. ISO, *Guide to the Expression of Uncertainty in Measurement* (1st edn). International Organization for Standardization: Geneva, Switzerland, 1995.
2. De Mast J, Erdmann TP, Van Wieringen W. Pass/fail inspection: Continuous versus binary measurands. *Journal of Quality Technology* 2011; **43**(2).
3. Boyles RA. Gauge capability for pass-fail inspection. *Technometrics* 2001; **43**:223–229.
4. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, U.K., 2003.
5. Van Wieringen WN, Van den Heuvel ER. A comparison of methods for the evaluation of binary measurement systems. *Quality Engineering* 2005; **17**:495–507.
6. Danila O, Steiner SH, MacKay RJ. Assessing a binary measurement system. *Journal of Quality Technology* 2008; **40**:310–318.
7. Van Wieringen WN, De Mast J. Measurement system analysis for binary data. *Technometrics* 2008; **50**:468–478.
8. Danila O, Steiner SH, MacKay RJ. Assessment of a binary measurement system in current use. *Journal of Quality Technology* 2010; **42**:152–164.
9. Beavers DP, Stamey JD, Bekele BN. A Bayesian model to assess a binary measurement system when no gold standard system is available. *Journal of Quality Technology* 2011; **43**:16–27.

10. AIAG. *Measurement Systems Analysis*: *Reference Manual* (3rd edn). Automotive Industry Action Group: Detroit, MI, 2003.
11. Hosmer DW, Lemeshow S. *Applied Logistic Regression* (2nd edn). Wiley: New York, NY, 1989.
12. Zelterman D. Parameter estimation in the generalized logistic distribution. *Computational Statistics and Data Analysis* 1987; **5**:177–184.
13. Jenkinson AF. The frequency distribution of the annual maximum (or minimum) of meteorological elements. *Quarterly Journal of the Royal Meteorological Society* 1955; **81**:158–171.
14. Hastie T, Tibshirani R. Generalized Additive Models. *Statistical Science* 1986; **3**:297–310.
15. Hastie T, Tibshirani R. *Generalized Additive Models*. Chapman & Hall: London, U.K., 1990.
16. De Boor C. *A Practical Guide to Splines*. Springer: New York, NY, 1978.
17. Silverman BW. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society, Series B* (*Methodological*) 1985; **47**:1–52.

Authors' biographies

Tashi P. Erdmann is a consultant and researcher at the Institute for Business and Industrial Statistics of the University of Amsterdam. He is working on a PhD project about Measurement System Analysis. He obtained his MSc degree (cum laude) in econometrics at the University of Amsterdam.

Jeroen de Mast is associate professor at the University of Amsterdam, the Netherlands, and senior consultant at the Institute for Business and Industrial Statistics of the University of Amsterdam. He obtained a doctorate in statistics at the University of Amsterdam. He has coauthored several books about Six Sigma, and is a recipient of the ASQ 2005 Brumbaugh Award, as well as the 2005 ENBIS Young Statistician Award. He is a member of ASQ.