The Statistical Evaluation of a Binary Test Based on Combined Samples

TASHI P. ERDMANN, THOMAS S. AKKERHUIS, and JEROEN DE MAST

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

STEFAN H. STEINER

University of Waterloo Waterloo, Ontario N2L 3G1, Canada

The statistical evaluation of the reliability of binary tests and inspections is a challenging endeavor. In this paper, we propose an approach for the common situation where the true condition of the inspected items is unobservable ("gold-standard unavailable"), the probabilities of false acceptance and false rejection vary across items, and rejections are relatively rare. Our approach fits a latent-variable model, where the variability in misclassification probabilities is driven by a continuous property of a part. To deal with the low prevalence of rejections, we propose sampling items from multiple sources. The performance and properties of the estimators are assessed using simulation, asymptotic approximations, and a real-life case at a car-parts manufacturer.

Key Words: Accuracy of Diagnostic Tests; Latent-Variable Model; Measurement Reliability; Measurement System Analysis; Pass/Fail Inspection.

1. Introduction

Binary tests and inspection systems classify items in two categories, such as 'reject' ($Y = 0$) or 'accept' ($Y = 1$). Examples include leak tests, visual quality inspections, and inspections based on Go/No-Go gauges. Diagnostic and screening tests in medicine are closely related. We conceive of such tests as a form of measurement, and thus, these classifications aim to reflect an underlying empirical condition X of the items that is called the measurand (ISO (1995)). The measurand can be a dichotomous property, but in many cases, it is a continuous or more complex condition (De Mast et al. (2011)). Measurement-system analysis (MSA) studies how reliably test results Y reflect the measurand X and, for binary measurements, this is often expressed in terms of misclassification probabilities (De Mast et al. (2014)).

The motivating example for this paper (Section 4) concerns an optical inspection system in a plant for car parts. The measurand is the misalignment X of a clip and the pad to which it is to be fastened. A part is considered 'good' if misalignment is below the upper specification limit (USL), and 'defective' otherwise. The reliability of the inspections can be quantified by the false-acceptance and false-rejection probabilities,

$$\text{FAP} = P[Y = 1 \mid X > \text{USL}],$$
$$\text{FRP} = P[Y = 0 \mid X \leq \text{USL}].$$

Recent literature in industrial and medical statis-

Dr. Erdmann is a Statistical Consultant in the Statistics & Chemometrics team at Shell Global Solutions International. His email address is tashi.erdmann@shell.com.

Mr. Akkerhuis is Consultant and PhD Student at IBIS UvA. His email address is t.s.akkerhuis@uva.nl.

Dr. De Mast is Principal Consultant and Professor of Methods and Statistics for Operations Management at IBIS UvA. His email address is j.demast@uva.nl.

Dr. Steiner is Professor and Chair in the Department of Statistics and Actuarial Science, and Director of the Business and Industrial Statistics Research Group at the University of Waterloo. His email address is shsteiner@uwaterloo.ca.

tics has shown that the statistical evaluation of the reliability of binary tests is challenging (e.g., Kraemer (1987), Feinstein (2002), Irwig et al. (2002), Knottnerus et al.(2002), De Mast et al. (2011), Danila et al. (2012), De Mast et al. (2014)). In this paper, we deal with a typical combination of three challenges. The first challenge is that the true conditions of the items are practically unobservable. This rules out the use of traditional methods for estimating the FAP and FRP (as described in AIAG (2003), Pepe (2003), Danila et al. (2008)), which assume the availability of a so-called gold standard: a higher-order, authoritative measurement that is accepted to constitute a faithful representation of the measurand.

For situations where no gold standard is available, the literature proposes methods based on latent class models (Hui and Walter (1980), Boyles (2001), Van Wieringen and De Mast (2008), Danila et al. (2010), and many others). Such methods are essentially based on the premise that the FAP and FRP are constant across items in the populations of defective and good items, respectively. Technically, such methods assume that the inspections Y are independent and identically distributed conditional on whether the items are defective or good. However, this assumption is often too simplistic: De Mast et al. (2011) show that a measurand related to a continuous property, such as misalignment in our example, violates such conditional independence assumptions. Artificially treating such measurands as dichotomous leads to biased estimators (possibly substantially so). Thus, the second challenge is that we want to avoid the assumption that the FAP and FRP are constants in the populations of defective and good items.

The third challenging aspect of the situation that we consider is that the prevalence of defects in industrial processes is typically very low. Consequently, a sample from the population of all produced items would contain no or only very few defective items, resulting in impractically large standard errors in the estimated FAP.

This combination of three challenges is quite common, and this paper proposes an approach based on latent-trait models. A similar model was introduced for ordinal classifications in De Mast and Van Wieringen (2010) and is related to item-response theory models (Lord (1980)). In the latent-trait model, the probability of rejection is not assumed constant within the populations of defective and good items. Instead, it is a function of an unobserved continuous property X, and the dependence is modeled by means of a characteristic curve $q(x) = P[Y = 0 \mid X = x]$. This curve is estimated from data collected in an MSA experiment, as is, under some assumptions, the distribution function $F_X(x) = P[X \leq x]$ of the measurand in the population of items. FAP and FRP or comparable metrics can be determined from these two functions, as shown later. Due to the third challenge (defects are extremely rare in the population of items), a random sample of items will contain no or only very few defective items. Estimation of the relevant part of the characteristic curve q, however, requires a sample with fair numbers of good and defective items. Our solution involves combining samples from various origins (the total items population, the stream of rejected items, and historical data about the rejection rate) and incorporating the sampling origin in the estimation algorithm to correct for a potential bias.

The purpose of this paper is to elaborate this approach for binary inspections and to explore to what extent it provides an effective solution to the combination of the three mentioned challenges. We elaborate our approach in the next section. The third section presents an evaluation of the approach on the basis of simulation and asymptotics. Section 4 describes the motivating case about optical inspections of car parts, and we draw conclusions in the final section.

2. Methodology

2.1. Statistical Model

Here we present our experimental model. We consider MSA experiments in which a sample of items $i = 1, \ldots, I$ is (repeatedly) appraised by each of one or more appraisers $a = 1, \ldots, A$. Repeated appraisals of item i by appraiser a are indexed by $k = 1, \ldots, K_{ai}$. The subscript on K_{ai} indicates that the number of repetitions may differ over appraisers and items, and $K_{ai} = 0$ means that appraiser a did not measure item i. We denote the result of the kth appraisal by appraiser a for item i as Y_{aik}. Thus, the outcome of the MSA experiment is the vector of zeros and ones given by $\boldsymbol{Y} = \{Y_{aik}\}_{a=1,\ldots,A;\, i=1,\ldots,I;\, k=1,\ldots,K_{ai}}$.

The binary appraisals under study aim to reflect an unobservable, continuous property X. Without loss of generality, we assume that we only have an upper specification limit (USL). An item i is 'good' if $X_i <$ USL, and the intended inspection outcome in that case is $Y_{aik} = 1$ ('accept'). If $X_i \geq$ USL, the item

is 'defective' and the intended outcome is $Y_{aik} = 0$ ('reject'). The X_i are assumed to be independently distributed and, in the population of produced items, have a normal distribution. The location and scale of the latent X-continuum are arbitrary and, without loss of generality, we set $E(X_i) = 0$ and $\mathrm{Var}(X_i) = 1$. Without such restrictions, the latent-variable model is unidentifiable. Thus, the density of X_i in the population of items is the standard normal denoted ϕ.

We assume that, besides X_i, there are no other properties of the items and no environmental factors that induce dependencies among repeated appraisals. In particular, we assume that, conditional on X_i, the $Y_{ai1}, \ldots, Y_{aiK_{ai}}$ are independent and identically Bernoulli distributed. Careful experimental design may enable this assumption to be fulfilled, for example, by experimental randomization and, in the case of human appraisers, by presenting the items in a way they cannot be recognized from the previous inspection.

We define $q_a(x) = P[Y_{iak} = 0 \mid X_i = x]$ as the characteristic curve for appraiser a. Our initial choice for $q_a(x)$ is defined by the logistic function

$$\log\left(\frac{q_a(x)}{1 - q_a(x)}\right) = \alpha_a(x - \delta_a), \quad \alpha_a > 0. \qquad (1)$$

The curve's inflection point δ_a, which is also the point where $q_a(x) = 0.5$, can be interpreted as the decision threshold that appraiser a appears to apply: items with $X_i > \delta_a$ are more likely to be rejected than accepted. The parameter $\alpha_a > 0$ is a discrimination parameter for appraiser a, determining the steepness of the curve. Larger values of α_a correspond to a steeper curve and better reliability of the inspection results. The characteristic curve in Equation (1) is symmetric about δ_a. Figure 1 illustrates the model.

The misclassification probabilities are not equal for all items, but depend on the measurand X_i. The average probabilities (weighted by the density of X_i in the population of items) for appraiser a are:

$$\mathrm{FAP}_a = \int_{\mathrm{USL}}^{\infty} (1 - q_a(x))\phi(x)dx \Big/ \int_{\mathrm{USL}}^{\infty} \phi(x)dx,$$
$$\mathrm{FRP}_a = \int_{-\infty}^{\mathrm{USL}} q_a(x)\phi(x)dx \Big/ \int_{-\infty}^{\mathrm{USL}} \phi(x)dx. \qquad (2)$$

In the situation we consider in this paper, the measurand is unobservable and, as a consequence, the USL is ill defined. Consequently, also FAP_a and FRP_a are ill defined. We propose to work instead with an alternative proposed by De Mast and Van Wieringen (2010). The inconsistent acceptance probability (IAP_a) and inconsistent rejection probability (IRP_a) are the probabilities that appraiser a's classification is inconsistent with his or her own decision threshold δ_a.

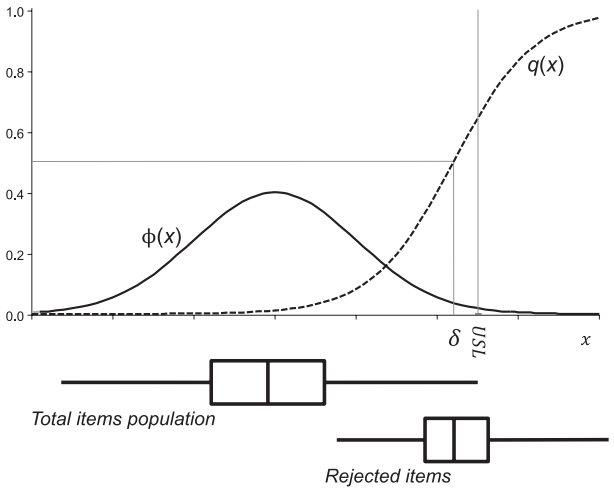

FIGURE 1. Probability Density $\phi(x)$ and Characteristic Curve $q_a(x)$. Box plots represent the distribution of X in a representative sample from the total items population and in a sample from the rejected items.

$$\mathrm{IAP}_a = P[Y = 1 \mid X > \delta_a]$$
$$= \int_{\delta_a}^{\infty} (1 - q_a(x))\phi(x)dx \Big/ \int_{\delta_a}^{\infty} \phi(x)dx,$$
$$\mathrm{IRP}_a = P[Y = 0 \mid X \le \delta_a]$$
$$= \int_{-\infty}^{\delta_a} q_a(x)\phi(x)dx \Big/ \int_{-\infty}^{\delta_a} \phi(x)dx. \qquad (3)$$

Whereas FAP_a and FRP_a express both the systematic component of classification error (that is, $|\delta_a - \mathrm{USL}|$) and the random component (the degree to which classifications randomly deviate from an appraiser's own δ_a), IAP_a and IRP_a express the random component only. This can be seen from the following decomposition of FRP_a, where we assume that $\delta_a \le \mathrm{USL}$ (and similar decompositions can be given for FAP_a and for $\delta_a > \mathrm{USL}$):

$$\mathrm{FRP}_a = P[Y = 0 \mid X \le \delta_a]P[X \le \delta_a \mid X \le \mathrm{USL}]$$
$$+ P[Y = 0 \mid \delta_a < X \le \mathrm{USL}]$$
$$\times P[\delta_a < X \le \mathrm{USL} \mid X \le \mathrm{USL}].$$

The last term is the contribution to FRP_a due to systematic classification error, which is determined by the distance between δ_a and USL. The first term is IRP_a (both terms are multiplied by probability

weights). Without gold-standard measurement, IAP_a and IRP_a, rather than FAP_a and FRP_a, are often all that one could hope to estimate.

The relevant aspects of measurement reliability can now be framed in technical terms. Random measurement error can be quantified by IAP_a and IRP_a. Systematic differences between appraisers can be quantified by differences in their decision thresholds δ_a. Further, the percentage of defective items reaching the customer is $P[X > \mathrm{USL} \mid Y = 1]$ and the percentage of falsely rejected ('good') items in the stream of rejects is $P[X \leq \mathrm{USL} \mid Y = 0]$. If USL is known (or assumed to equal the δ_a of an appraiser a who is taken as reference standard), these probabilities can be determined from

$$\begin{aligned} &P[X > \mathrm{USL} \mid Y = 1] \\ &\quad = \int_{\mathrm{USL}}^{\infty} (1 - q_a(x))\phi(x)dx \Big/ \int_{-\infty}^{\infty} (1 - q_a(x))\phi(x)dx, \\ &P[X \leq \mathrm{USL} \mid Y = 0] \\ &\quad = \int_{-\infty}^{\mathrm{USL}} q_a(x)\phi(x)dx \Big/ \int_{-\infty}^{\infty} q_a(x)\phi(x)dx. \end{aligned}$$

2.2. Sampling Strategy and Estimation

We propose maximum-likelihood estimators for the parameters α_a, δ_a; $a = 1, \ldots, A$ of the characteristic curves q_a (model (1)). First, we describe the sampling strategy that we propose. A representative sample from the total population of items is the obvious choice, but on second thought, this turns out to be problematic. As mentioned in the Introduction, rejection rates in typical manufacturing processes are low and the steep part of q_a (around its inflection point δ_a) will usually be in the remote tail of the density ϕ of X in the items population. Consequently, a representative sample from the total items population of reasonable size will contain no or only a few x values close to or to the right of δ_a. This results in very large standard errors in the estimated IAP_a. Figure 1 illustrates the point: the first box plot below the graph shows the quartiles of the probability density of X for items in such a sample (the box represents the 25%, 50%, and 75% quartiles and the end points of the whiskers delineate a 99% interval). It can be seen that only a small fraction or even none of the items have values in the steep part of $q_a(x)$.

To estimate the characteristic curves close to and to the right of δ_a with acceptably small standard errors, one needs a sample with more evenly spread x-values. We propose to take samples from various sampling sources and combine them to have a more balanced sample. Besides the total population of items (*Tot*-sample), we also propose sampling from the population of rejected items (*Rej*-sample) and to incorporate aggregate information from historical data (*His*-sample). It is essential that, during the period over which the items are sampled, the circumstances are constant. This may be realized by ensuring the three subsamples are taken over the same time period.

The idea to sample items from the stream of rejects was proposed by Danila et al. (2010). They called the resulting data a conditional sample, as the sampling distribution of items sampled from the population of rejected items is obtained by conditioning on the initial rejection decision. This approach typically leads to a sample that has a larger proportion of items in the steep part of the characteristic curve (see the box plot labeled "Rejected items" in Figure 1). For an item i sampled from the stream of rejected items, let $d_i \in \{1, \ldots, A\}$ be the appraiser who rejected it, and let $Y_{d_i i0} = 0$ denote the event of this rejection. Given that item i has been rejected by appraiser d_i, the sampling distribution of X_i is

$$\begin{aligned} F^{d_i}(x) &= P[X_i \leq x \mid Y_{d_i i0} = 0] \\ &= \int_{-\infty}^{x} \phi(t)q_{d_i}(t)dt \Big/ \int_{-\infty}^{\infty} \phi(t)q_{d_i}(t)dt, \end{aligned} \tag{4}$$

with probability density f^{d_i}.

Besides the *Tot* and *Rej* samples, a third data source is a historical dataset of (single) inspection results ('reject' or 'accept') of a large number of items, typically summarized as a total count and a number of rejected items. Danila et al. (2010, 2012) called this baseline data and showed that it substantially increases the precision of the estimators in the latent class models they discussed. A historical rejection rate is typically easy to obtain. Even if it is not readily available, it can be obtained during the collection of the *Rej*-sample for the MSA study (which typically involves several thousand inspections before a sufficient number of rejected items are obtained).

Conditional on X_i, the likelihood of all outcomes for item i is

$$\begin{aligned} &P[Y_{1i1} = y_{ai1}, \ldots, Y_{AiK_{Ai}} = y_{AiK_{Ai}} \mid X_i = x_i] \\ &\quad = \prod_{a=1}^{A} \prod_{k=1}^{K_{ai}} q_a(x_i)^{1-y_{aik}}(1 - q_a(x_i))^{y_{aik}}. \end{aligned}$$

This expression uses the conditional independence of repeated appraisals. To obtain the unconditional

probability we integrate out the latent variable X_i weighted by its probability density. For items in the *Tot*-sample and *His*-sample, this is $f^0 = \phi$. For items in the *Rej*-sample, this is f^{d_i}, as in Equation (4). This gives

$$P[Y_{1i1} = y_{ai1}, \ldots, Y_{AiK_{Ai}} = y_{AiK_{Ai}}] = \int_{-\infty}^{\infty} f^{d_i}(x) \prod_{a=1}^{A} \prod_{k=1}^{K_{ai}} q_a(x_i)^{1-y_{aik}} (1 - q_a(x_i))^{y_{aik}} dx,$$

where $d_i = 0$ for items in the *Tot*- and *His*-samples and $K_{ai} = 1$ if i is in the *His*-sample. The resulting log likelihood, with parameter vector $\boldsymbol{\theta} = \{\alpha_a, \delta_a\}_{a=1,\ldots,A}$, is

$$L(\boldsymbol{\theta} \mid \boldsymbol{y}) = \sum_{i=1}^{I} \log \left[\int_{-\infty}^{\infty} f^{d_i}(x) \prod_{a=1}^{A} \prod_{k=1}^{K_{ai}} q_a(x_i)^{1-y_{aik}} \times (1 - q_a(x_i))^{y_{aik}} dx \right],$$

where $\boldsymbol{y} = \{y_{aik}\}_{a=1,\ldots,A;\, i=1,\ldots,I;\, k=1,\ldots,K_{ai}}$ represents the data. For concise representations of the experimental outcomes $\boldsymbol{Y}$ and efficient calculation of the log likelihood, we determine an equivalent expression in terms of response-pattern frequencies. A response pattern $\boldsymbol{R}_i = (\sum_{k=1}^{K_{ai}} Y_{aik}, \sum_{k=1}^{K_{ai}} (1 - Y_{aik}))_{a=1,\ldots,A}$ is an $A \times 2$ matrix, in which the elements $R_i[a,1]$ are the number of times item i is accepted by appraiser a and $R_i[a,2]$ is the number of times it is rejected by appraiser a. As before, let $d_i = 0$ if item i is in the *Tot*- or *His*-sample and $d_i \in \{1, \ldots, A\}$ be the appraiser who initially rejected the item if i is in the *Rej*-sample. The response-pattern frequencies are $e(\boldsymbol{r}, d) = \{\#i \mid \boldsymbol{R}_i = \boldsymbol{r}, d_i = d\}$ with $d \in \{0, \ldots, A\}$ and $\boldsymbol{r}$ an $A \times 2$ matrix with elements in $\mathbb{N}$. By tabulating $e(\boldsymbol{r}, d)$, the data can be displayed concisely, and the log likelihood can be rewritten to contain fewer integrals,

$$L(\boldsymbol{\theta} \mid \boldsymbol{y}) = \sum_{d=0}^{A} \sum_{\boldsymbol{r} \in \mathbb{N}^{A \times 2}} e(\boldsymbol{r}, d) \times \log \left[\int_{-\infty}^{\infty} f^{d}(x) \prod_{a=1}^{A} q_a(x)^{r[a,2]} \times (1 - q_a(x))^{r[a,1]} dx \right]. \quad (5)$$

We maximize the log likelihood with the interior point algorithm (Mehrotra (1992), as implemented in the function "FindMaximum" in software package Mathematica 8 (2010)), and we find starting values using the Nelder-Mead algorithm (Nelder and Mead (1965), as implemented in "NMaximize" in Mathematica 8). The integrals are approximated numerically using adaptive quadrature (Rice (1975), as implemented in "NIntegrate" in Mathematica 8). Once the parameters α_a and δ_a have been estimated, they can be plugged into $q_a(x)$ in Equation (3) to obtain the estimates for IAP_a and IRP_a.

Note that the convergence time rapidly increases with the number of appraisers A. The increasing number of parameters and exponentially increasing number of response patterns in A leads to a large number of integrals to evaluate. We find that, for $A > 1$, the optimization may take a half hour and even more for some choices of the functional form for $q_a(x)$.

Standard errors of the maximum likelihood estimators $\hat{\boldsymbol{\theta}}$ and of $\widehat{\mathrm{IAP}}_a$ and $\widehat{\mathrm{IRP}}_a$ can be obtained by bootstrapping, but in view of the above, this becomes practically undoable if $A > 1$. We recommend approximating the covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$ of $\hat{\boldsymbol{\theta}}$ on the basis of the observed Fisher information matrix,

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}} = \left(-\frac{\partial^2}{\partial \boldsymbol{\theta}^2} L(\boldsymbol{\theta} \mid \boldsymbol{y}) \Big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \right)^{-1}. \quad (6)$$

The covariance matrix of $\widehat{\mathrm{IAP}}_a$ and $\widehat{\mathrm{IRP}}_a$ is then approximated by

$$\hat{\boldsymbol{\Sigma}}_{(\widehat{\mathrm{IAP}}_a, \widehat{\mathrm{IRP}}_a)} = \left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2} (\mathrm{IAP}_a, \mathrm{IRP}_a) \Big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \right) \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}} \times \left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2} (\mathrm{IAP}_a, \mathrm{IRP}_a) \Big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \right)^{\mathrm{T}}, \quad (7)$$

which uses a linear approximation to the functions $\mathrm{IAP}_a(\boldsymbol{\theta})$, $\mathrm{IRP}_a(\boldsymbol{\theta})$ defined in Equation(3).

2.3. Model Diagnostics

For assessment of the fit of the model, we can use standard techniques from the latent-variable modeling literature. We briefly mention them here and demonstrate their use in the car-parts case described in Section 4. For residual analysis, one may compare the observed frequencies $e(\boldsymbol{r}, d)$ of the response patterns to the frequencies $\eta(\boldsymbol{r}, d)$ predicted by the fitted model, either as raw differences or as Freeman-Tukey variance-stabilized residuals (Formann (2003)). To test for lack of fit, one can apply a likelihood-ratio test (which is known as the G-test when response

pattern frequencies are concerned). This test is preferred over the Pearson χ^2 test when one or more of the expected cell counts are below 5 (Kallenberg et al. (1985)). The test statistic is based on the likelihood ratio

$$G = 2\sum_{d=0}^{A}\sum_{\boldsymbol{r}\in\mathbb{N}^{A\times 2}} e(\boldsymbol{r},d)\log\frac{e(\boldsymbol{r},d)}{\eta(\boldsymbol{r},d)} \sim \chi^2(df).$$

The number *df* of degrees of freedom of the chi-square distribution is the difference between the number of parameters needed for a fully saturated model and the number of parameters in the model ($2A$ for logistic curves). The calculation is straightforward but confusing to define for the general case; we will instead demonstrate the calculation for the case in Section 4. By simulation, we have evaluated the test's power in detecting one form of lack of fit in particular, namely, that q is not symmetric about δ, but has different slopes to the left and right of δ (details are in the supplementary material at http://www.asq.org/pub/jqt/). The test turns out to have useful power for detecting departures from symmetry. If asymmetry is detected, one could fit a loglogistic function instead,

$$q(x) = \frac{1}{(1+(\alpha(x-\mu))^{\beta}}\mathbf{1}_{x\geq\mu}, \tag{8}$$

where the indicator function $\mathbf{1}_{x\geq\mu}$ indicates that the curve is zero for $x<\mu$. This family of curves is more flexible and can adapt to varying degrees of asymmetry. We demonstrate this in the case in Section 4.

3. Suitable Sample Sizes and Robustness Against Model Misspecification

We conducted a number of simulation studies to assess the properties of the estimators proposed in the previous section. The aims of these studies were to establish guidelines for choosing sample sizes and to assess the robustness of the estimators against model misspecification. The studies are limited to the case of a single appraiser ($A=1$). Finite-sample properties were established by Monte Carlo simulation. In addition, we derived the asymptotic distribution of the maximum-likelihood estimators, using that, for $I\to\infty$, $\hat{\boldsymbol{\theta}} \sim \mathrm{N}(\boldsymbol{\theta},\boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \left(-E_{\boldsymbol{y}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}^2}L(\boldsymbol{\theta}\mid\boldsymbol{y})\right]\right)^{-1}. \tag{9}$$

For the interested reader, a detailed report of the studies can be found in the supplementary material at http://www.asq.org/pub/jqt/. Here, we briefly summarize the results.

3.1. Optimal Proportions of Sample Sources

The total sample consists of $I = I^{Tot}+I^{Rej}+I^{His}$ items, with I^{Tot}, I^{Rej}, and I^{His} the sizes of the *Tot*, *Rej*, and *His* samples. First, we investigated which proportion of sample sources is optimal. As noted before, a large historical dataset of rejections is typically available or can be obtained while collecting the *Rej* sample. For this reason, we assumed a large *His*-sample size of $I^{His}=100{,}000$. To see how much the *His* sample contributes, however, we also investigated the case that $I^{His}=0$.

The study demonstrated that the most precise estimates of IAP and IRP are obtained when the sample consists solely of a *Rej* sample (that is, $I^{Tot}=0$) supplemented with a large *His* sample. Further details of the study are as follows. We considered sample sizes $I^{Tot}\in\{0,10,\ldots,200\}$ and $I^{Rej}=200-I^{Tot}$. In both the data-generating process and the estimated model, a standard normal distribution was used for X and a logistic curve with $\alpha=5$, $\delta=2$ for q. These choices imply true probabilities of IAP $=0.2154$ and IRP $=0.0125$. Note that the IAP is (much) larger than the IRP, which is typical when δ is in the far tail of the distribution of X. Conditional on $X>\delta$, $\phi(x)$ has the most probability mass close to δ (which corresponds to the range of hard-to-judge items), while conditioning on $X\leq\delta$, most probability mass is around 0 (with easy-to-judge items).

The precision of the estimators in each scenario was determined from their asymptotic covariance matrix (9) and also from Monte Carlo simulation (2500 runs per scenario). Precision of the estimators was quantified as the width of empirical 95% confidence intervals for IAP and IRP (based on the 2.5 and 97.5 percentiles of the 2500 realizations of $\widehat{\mathrm{IAP}}$ and $\widehat{\mathrm{IRP}}$.

Figure 2 gives plots of the 95% confidence interval width for IAP and IRP as a function of I^{Rej} with and without a *His* sample. The lines in the figure are based on asymptotic standard errors and the diamonds on Monte Carlo simulation. The conclusions were corroborated for other values of α and δ (based on a study with a more limited range of values for I^{Tot}, I^{Rej} and I^{His}). Further details are in the supplementary material.

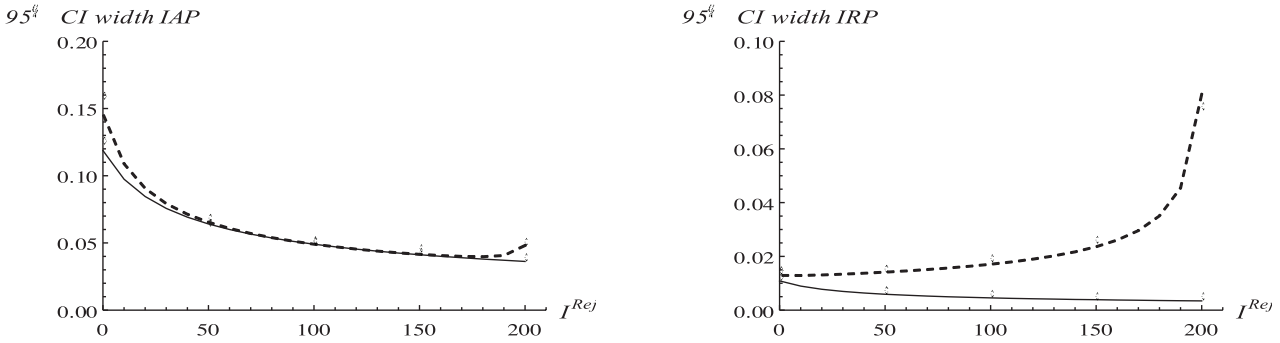

FIGURE 2. 95% Confidence Interval Widths for an Inspection System with $\alpha = 5$, $\delta = 2$, IAP = 0.2154, IRP = 0.0125 for Various I^{Rej}, I^{His}, ($I^{Tot} = 200 - I^{Rej}$). Lines: results obtained through the asymptotic distribution of the MLE (dashed: $I^{His} = 0$; solid: $I^{His} = 10^5$). Diamonds: results obtained by simulation.

3.2. Precision as a Function of I^{Rej} and K

Following the recommendation above, the MSA experiment should involve a sample of I^{Rej} items from the stream of rejects, which are appraised K times, and a historical rejection rate estimated from I^{His} appraisals. To choose an appropriate sample size I^{Rej} and number of repeated appraisals K, we determine their effect on the precision of the estimators.

As before, we set $I^{His} = 100{,}000$. By Monte Carlo simulation, we investigated 7×7 combinations in the ranges $I^{Rej} \in \{50, \ldots, 200\}$ and $K \in \{3, \ldots, 15\}$. In addition, we derived asymptotic results for 16×13 combinations in the same ranges. We obtained results for all four combinations of $\alpha = 5$, 12 and $\delta = 2$, 3, which we think represent typical binary inspection systems in industry in terms of reliability and rejection rate. Figure 3 gives level plots of the empirical 95% confidence interval width as a function of I^{Rej} and K for $\alpha = 5$ and $\delta = 2$. Level plots of the precision for other (α, δ) combinations can be found in the supplementary material.

The results obtained by simulation (dashed lines) are close to the asymptotic results (solid lines). We conclude that the asymptotic properties provide a useful approximation to the finite-sample properties of the estimators. This is an important result for the calculation of standard errors for the estimators.

The marginal effects of I^{Rej} and K are diminishing. In particular, for these values of α, δ, it seems inefficient to do more than $K = 7$ repeated appraisals. The required sample size depends on the desired precision, and plots such as in Figure 3 can be used as a reference. In general, we recommend $I^{Rej} \geq 150$ because, in the considered ranges, this ensures a confidence-interval width of at most 50% of the IRP or IAP (assuming $K = 7$), which seems reasonable to us.

3.3. Robustness to Misspecification

Finally, we report on the estimation procedure's robustness against model misspecification. As mentioned before, the G-test is powerful in detecting asymmetry in $q(x)$, and Section 4 will demonstrate how to handle such situations with log-logistic characteristic curves. Therefore, we focus on misspecification of the distribution of X here. We simulated 20 scenarios where we drew realizations of X ($I^{Rej} = 200$, $I^{His} = 100{,}000$) from skewed and leptokurtic distributions instead of the normal (and applied a Bernoulli trial to simulate drawing from the stream of rejects for the *Rej* realizations). For each realization of X, we drew $K = 9$ realizations of Y by applying a logistic characteristic curve (with parameters α, δ). We fitted our model (with logistic curve for q and a standard normal distribution for X) to these data, and compared the estimated $\widehat{\text{IAP}}$ and $\widehat{\text{IRP}}$ with the true values.

Each scenario in Table 1 is determined by a distribution for X (standard normal, standard log normal, $\chi^2(1)$, $t(3)$, or $t(7)$) and values for α and δ. These choices imply a corresponding IAP, IRP, and reject probability $p_0 = P(Y = 0)$. We could not choose the scenarios such that they have the same IAP, IRP, and reject probability across distributions. Namely, given a distribution, there are only two free parameters (α and δ) and, in addition, some ratios of IAP and IRP are not possible for some distributions. Instead, rows in Table 1 were chosen to roughly represent the situations of poor and good measurement

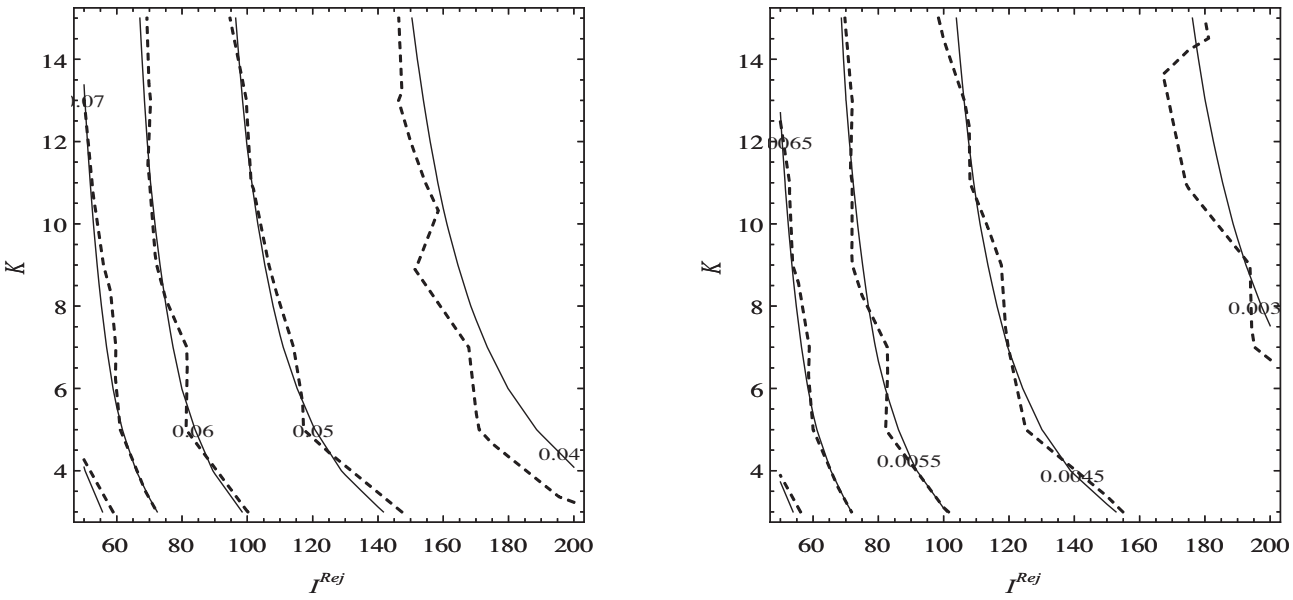

FIGURE 3. Level Plots of 95% Empirical Confidence Interval Widths for IAP and IRP Estimates for an Inspection System with $\alpha = 5$, $\delta = 2$, IAP = 0.2154, IRP = 0.0125. Solid lines: results obtained through the asymptotic distribution of the MLE. Dashed lines: results obtained by simulation.

reliability and medium and low reject rates. Poor and good measurement reliability are defined here in terms of the 99% part of the characteristic curve (the 'grey area' where inspection results are uncertain) compared with the 99% part of the distribution of X,

$$\frac{q^{-1}(0.995) - q^{-1}(0.005)}{F^{-1}(0.995) - F^{-1}(0.005)} = \rho.$$

We took $\rho = 0.30$ for poor reliability and $\rho = 0.15$ for good reliability and $p_0 = 0.005$ for a low reject rate and $p_0 = 0.010$ for a medium reject rate. Table 1 shows the α and δ values that give these results for each of the five distributions given in the columns, and also the implied IAP and IRP. We simulated each scenario 1000 times. The table gives the absolute deviation of the average $\widehat{\text{IAP}}$ and $\widehat{\text{IRP}}$ from their true values, and an asterisk indicates that a bias is significantly different from 0.0000 (based on a t-test).

We have also investigated robustness against multidimensional (multivariate normal and mixtures) distributions by Monte Carlo simulation and robustness against skew-normal and t-distributions based on the asymptotic distribution of the covariance estimators. We refer to the supplementary material for the details of these studies.

For all forms of misspecification that we investigated, biases in the estimates remain below 0.0100 for IAP and 0.0007 for IRP (except for one case of extreme negative skewness; in this case, the absolute bias of $\widehat{\text{IAP}}$ is larger but, relative to the true IAP, the bias is still modest because the true IAP is large for such cases). In all cases considered, biases remain small compared with the true IAP and IRP. In conclusion, the estimation procedure seems reasonably robust to misspecification of the distribution of X.

4. MSA Study of Car-Parts Manufacturing

4.1. Background of the MSA Study

We apply the proposed approach in an MSA experiment conducted to evaluate an optical inspection system at a car-parts manufacturer. The parts in question contain an integrated circuit, about 1 cm in diameter, on which several electronic components are connected and soldered. After assembly, the parts are inspected for soldering errors. This check is performed by a team of operators, but a visual inspection machine called Automated Optical Inspection (AOI) has been purchased to replace the operators' inspections in the future. Currently, the AOI is operational, but the operators still perform the inspections. This situation will continue until the AOI is deemed to function satisfactorily.

The AOI inspects a variety of properties of a sol-

TABLE 1. Bias (Absolute Value) of the Estimators of IAP and IRP (Based on $I^{Rej} = 200$; $K = 9$; $I^{His} = 100.000$) in 20 Scenarios

	$X \sim \mathrm{N}(0,1)$		$X \sim t(3)$		$X \sim t(7)$		$X \sim \log \mathrm{N}(0,1)$		$X \sim \chi^2(1)$	
	IAP	IRP	IAP	IRP	IAP	IRP	IAP	IRP	IAP	IRP
$\rho = 0.15$,	$\alpha = 13.7$,		$\alpha = 6.04$,		$\alpha = 10.1$,		$\alpha = 5.40$,		$\alpha = 8.96$,	
$p_0 = 0.005$	$\delta = 2.60$		$\delta = 5.87$		$\delta = 3.52$		$\delta = 13.2$		$\delta = 7.89$	
True value	0.1194	0.0009	0.0489	0.0003	0.0789	0.0005	0.0266	0.0002	0.0398	0.0002
Bias	0.0003	0.0000	0.0013*	0.0000*	0.0014*	0.0000*	0.0080*	0.0001*	.0010*	0.0000
$\rho = 0.15$,	$\alpha = 13.7$,		$\alpha = 6.04$,		$\alpha = 10.1$,		$\alpha = 5.40$,		$\alpha = 8.96$,	
$p_0 = 0.010$	$\delta = 2.35$		$\delta = 5.48$		$\delta = 3.02$		$\delta = 10.3$		$\delta = 6.65$	
True value	0.1120	0.0016	0.0589	0.0008	0.0824	0.0012	0.0312	0.0004	0.0404	0.0005
Bias	0.0006	0.0000	0.0010*	0.0001*	0.0009*	0.0001*	0.0011*	0.0000	0.0004	0.0000
$\rho = 0.30$,	$\alpha = 6.85$,		$\alpha = 3.02$,		$\alpha = 5.04$,		$\alpha = 2.70$,		$\alpha = 4.48$,	
$p_0 = 0.005$	$\delta = 2.67$		$\delta = 5.96$		$\delta = 3.60$		$\delta = 13.2$		$\delta = 7.93$	
True value	0.2012	0.0019	0.0875	0.0007	0.1366	0.0012	0.0504	0.0003	0.0744	0.0005
Bias	0.0001	0.0000	0.0041*	0.0001*	0.0062*	0.0001*	0.0006*	0.0000	0.0006*	0.0000*
$\rho = 0.30$,	$\alpha = 6.85$,		$\alpha = 3.02$,		$\alpha = 5.04$,		$\alpha = 2.70$,		$\alpha = 4.48$,	
$p_0 = 0.005$	$\delta = 2.67$		$\delta = 5.96$		$\delta = 3.60$		$\delta = 13.2$		$\delta = 7.93$	
True value	0.1912	0.0035	0.1030	0.0018	0.1422	0.0026	0.0583	0.0008	0.0753	0.0010
Bias	0.0001	0.0000	0.0070*	0.0002*	0.0065*	0.0002*	0.0007*	0.0001	0.0014*	0.0000*

* Estimated bias significantly different from zero.

dered part and produces a binary decision in terms of 'accept' or 'reject' and, in the latter case, a specification of one or more failure modes. Our evaluation focused on the reliability of decisions about one failure mode in particular, namely, misalignment due to soldering faults. Here, the measurand is the misalignment X of a clip and the pad to which it is to be fastened. Due to the very small scale and uneven three-dimensional shape of the involved components, misalignment is very hard to measure directly. In addition, there is considerable ambiguity as to its precise definition, and there is no clearly defined upper specification limit (USL) that demarcates the acceptable range. The AOI's evaluation is based on a digital photo of the part, which is then analyzed by a proprietary algorithm.

The judgments by the team of operators were accepted by the company as the de facto standard and are used to determine whether the AOI's decisions are correct. However, because we found that there was occasional disagreement among the operators themselves, we could not treat them as a gold standard. Rather than fitting a characteristic curve for each operator individually, we treated the team of operators as a single appraiser and appraisals by individual team members as repetitions. Thus, the study involved $A = 2$ appraisers, the AOI ($a = 1$) and the team of operators ($a = 2$). The purpose of the reliability study was to evaluate the following:

- Variability of the AOI's decisions, represented by $\mathrm{IAP}_{\mathrm{AOI}}$ and $\mathrm{IRP}_{\mathrm{AOI}}$;
- Variability of the operators' decisions, represented by $\mathrm{IAP}_{\mathrm{opr}}$ and $\mathrm{IRP}_{\mathrm{opr}}$; and
- The systematic difference, if any, between the decisions of the AOI and the operators, represented by $\delta_{\mathrm{AOI}} - \delta_{\mathrm{opr}}$.

Due to a combination of changing objectives, advancing insight, practical limitations, and communication problems, the sample and dataset that we obtained are not optimal, but they do allow a useful analysis. Based on the results of the simulation studies reported in the previous section, a sample of $I^{Rej} = 150$ parts that had been rejected by the AOI

TABLE 2. Point Estimates and Standard Errors Based on the Original Analysis (150-Logistic), Log-Logistic Curves (150-loglgc), and After Removing One Anomalous Part (149-loglgc). Standard errors calculated by Equations (6) and (7)

	Estimate (150-logistic)	s.e.	Estimate (150-loglgc)	s.e.	Estimate (149-loglgc)	s.e.
δ_{AOI}	2.5800	0.0098	2.5500	0.0102	2.5600	0.0102
IAP_{AOI}	0.0673	0.0095	0.0728	0.0101	0.0695	0.0100
IRP_{AOI}	0.0004	0.0001	0.0001	0.0000	0.0001	0.0000
δ_{opr}	3.3700	0.0845	3.2100	0.0617	3.2200	0.0628
IAP_{opr}	0.2501	0.0254	0.0951	0.0379	0.0994	0.0386
IRP_{opr}	0.0004	0.0001	0.0000	0.0000	0.0001	0.0000

(the '*Rej*-sample') was collected. The parts in this sample were inspected $K_{a=1} = 7$ times by the AOI and once by each of $K_{a=2} = 3$ operators. The AOI can measure multiple parts simultaneously in different slots, and the parts were randomized over these slots during repeated measurements. A historical rejection rate of 0.050% for the AOI was also available (based on 1271 rejections out of $I^{His} = 254{,}200$ inspected parts). Moreover, $I^{Tot} = 100$ parts from the total parts population (the '*Tot* sample') were included, which were measured seven times by the AOI but not by the operators. The inclusion of these 100 parts cannot be motivated from the results reported in the previous section. At the time when the MSA experiment was designed, we believed this additional subsample would provide a crude check whether the AOI's behavior and its rejection rate, in particular, was in line with its normal performance (as reflected in the historical reject rate). The assumption that the inspections in the *Tot*, *Rej*, and *His* samples are comparable is an important one. However, in hindsight, we do not believe that the results from these 100 parts add much value in the evaluation of the AOI.

4.2. Statistical Analysis

The parameters (α_a, δ_a) for the AOI and the operators are estimated by maximizing the likelihood of Equation (5) using the interior point algorithm. The fitted parameters are $\hat{\alpha}_{\text{AOI}} = 26.69$ and $\hat{\delta}_{\text{AOI}} = 2.582$ for the AOI and $\hat{\alpha}_{\text{opr}} = 5.741$ and $\hat{\delta}_{\text{opr}} = 3.369$ for the operators. Table 2 (leftmost columns) gives the corresponding values of IAP, IRP, δ, and their standard errors.

To test goodness of fit, we determine the number df of degrees of freedom of the G-statistic. For the *Tot* sample (evaluated seven times by the AOI), the possible response patterns for a part are (0; 7), (1; 6), ..., (7; 0), and a saturated model has $8 - 1 = 7$ df. For the *His* sample, we have (0; 1) and (1; 0), so 1 df. For the *Rej* sample (seven repeats for the AOI, three for the operators), we have $8 \times 4 - 1 = 31$ df. In total, a saturated model has 39 df, while the fitted model has $2A = 4$ df. Thus, the G-statistic has 35 df.

The G-test rejects the fit ($G = 127$, df $= 35$, $p < 10^{-11}$). As often with real data, there are a few issues, which are best explained by discussing the results of the AOI first. Table 3 (columns headed *Observed*) presents response-pattern frequencies. The left column indicates how many of the 150 parts from the *Rej* sample were rejected 0, 1, ..., 7 times by the AOI. The next column shows the same for the 100 parts from the *Tot* sample.

Although the results from the *Tot* sample fit quite well, the fit is not that good for the *Rej* sample. In particular, the model overestimates the number of parts with a small number of rejections and underestimates the number of parts with six rejections. This can be seen from the raw differences between observed and predicted frequencies, or better, by calculating the Freeman-Tukey variance-stabilized residuals: −2.55, −2.62, −1.40, 0.97, 0.49, −0.32, 2.08, −0.15 (for zero to seven rejects). Because the collation of observed to predicted frequencies suggests asymmetry in the characteristic curve, we fit log-logistic curves instead (as in Equation (8)). The G-test confirms that this model fits better ($G = 42.4$, df $= 33$, $p = 0.13$). The estimated characteristics are in the middle columns of Table 2 (labeled *150-loglgc)*. The decision thresholds are obtained by solving $\hat{q}_a(\hat{\delta}_a) = 0.5$.

A second issue in the data is revealed by compar-

TABLE 3. Observed Rejection Frequencies (AOI Only) and Rejection Frequencies Predicted from a Fitted Logistic and Log-Logistic Curve

Number of rejections	Observed		Predicted (150-logistic)		Predicted (150-loglgc)	
	Rej-sample	Tot-sample	Rej-sample	Tot-sample	Rej-sample	Tot-sample
0	0	99	2.90	99.40	0.12	99.45
1	0	0	3.03	0.08	0.33	0.00
2	1	0	3.39	0.04	0.72	0.01
3	6	0	4.01	0.03	1.54	0.01
4	6	0	5.05	0.03	3.53	0.02
5	6	0	7.08	0.03	9.53	0.04
6	21	0	12.70	0.05	31.8	0.12
7	110	1	112.00	0.38	102.00	0.36

ing the results for the AOI with those of the operators (Table 4). One part (marked with an asterisk) was rejected three (out of three) times by the operators and only four (out of seven) times by the AOI. We consistently find that the AOI's decision threshold is stricter than that of the operators (Table 2) and, consequently, there is no realization for misalignment X for which this is a plausible response pattern (given the fitted model),

$$\max_x P_{\hat{\theta}_{\text{AOI}},\hat{\theta}_{\text{opr}}}\left(\mathbf{R}_i = \begin{pmatrix} 3 & 4 \\ 0 & 3 \end{pmatrix} \middle| X = x\right) \approx 4.7\times10^{-5}.$$

We were unfortunately not in a position to inspect the part in question and can only speculate that it may have had an abnormality or flaw that led the operators to reject it unanimously but that was, to some extent, different from the usual misalignment problems and, therefore, not convincingly picked up by the AOI's algorithms. For its being an anomaly amid the results of the other 149 parts, we removed the results for this part from the analysis and re-estimated the model.

The resulting parameter values of the log-logistic characteristic curves are $\hat{\alpha}_{\text{AOI}} = 60.2$, $\hat{\beta}_{\text{AOI}} = 1.26$, and $\hat{\mu}_{\text{AOI}} = 2.54$ for the AOI and $\hat{\alpha}_{\text{opr}} = 7.32$,

TABLE 4. Response Pattern Frequencies (Observed and Predicted from All Data and After Removing the Results of One Part)

Rejections by AOI	Observed					Predicted (150-loglgc)					Predicted (149-loglgc)				
	Rejections by operators					Rejections by operators					Rejections by operators				
	0	1	2	3	*Tot*	0	1	2	3	*Tot*	0	1	2	3	*Tot*
0	0	0	0	0	99	0.12	0.00	0.00	0.00	99.5	0.19	0.00	0.00	0.00	99.5
1	0	0	0	0	0	0.33	0.00	0.00	0.00	0.00	0.46	0.00	0.00	0.00	0.01
2	1	0	0	0	0	0.72	0.00	0.00	0.00	0.01	0.91	0.00	0.00	0.00	0.01
3	6	0	0	0	0	1.54	0.00	0.00	0.00	0.01	1.74	0.00	0.00	0.00	0.01
4	5	0	0	1*	0	3.52	0.00	0.00	0.00*	0.02	3.58	0.00	0.00	0.00	0.02
5	5	1	0	0	0	9.47	0.01	0.01	0.03	0.04	8.68	0.00	0.00	0.01	0.04
6	18	0	1	2	0	30.2	0.27	0.34	1.05	0.12	27.0	0.19	0.22	0.60	0.10
7	93	2	3	12	1	81.6	2.59	3.56	14.6	0.36	85.0	2.84	3.65	14.0	0.37
Historical rejection rate (AOI): 1,271 out of $I^{His} = 254{,}200$						Predicted rejection rate (AOI): 1,272 out of $I^{His} = 254{,}200$					Predicted rejection rate (AOI): 1,271 out of $I^{His} = 254{,}200$				

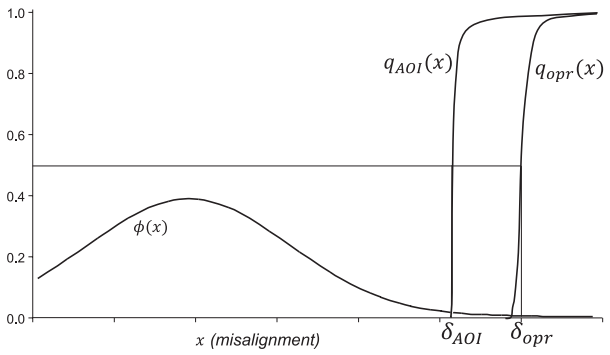

FIGURE 4. Fitted Log-Logistic Characteristic Curves of the AOI and Operators.

$\hat{\beta}_{\text{opr}} = 3.75$, and $\hat{\mu}_{\text{opr}} = 3.09$ for the operators. The fit has improved substantially, as is confirmed by the G-test ($G = 28.3$, df $= 33$, $p = 0.70$). Figure 4 shows the fitted characteristic curves of the AOI and the operators.

Table 2 (rightmost columns, labeled *149-loglgc*) presents the final estimated characteristics of the inspections. The key results of the analysis are as follows:

1. The AOI's repeatability can be characterized as $\widehat{\text{IAP}}_{\text{AOI}} = 0.0695$ and $\widehat{\text{IRP}}_{\text{AOI}} = 0.0001$.
2. The reproducibility of the operators is $\widehat{\text{IAP}}_{\text{opr}} = 0.0994$ and $\widehat{\text{IRP}}_{\text{opr}} = 0.0001$.
3. The systematic difference between the AOI and the operators is fairly substantial: $\hat{\delta}_{\text{AOI}} - \hat{\delta}_{\text{opr}} = -0.66$ (the AOI being stricter than the operators).

Further practical ramifications for the AOI are as follows. We could interpret the company's stance that the operators are the de facto standard by setting USL $= \delta_{\text{opr}}$. This allows the calculation of the AOI's misclassification probabilities using Equation (2): $\widehat{\text{FAP}}_{\text{AOI}} = \hat{P}[Y = 1 \mid X > \delta_{\text{opr}}] = 0.0064$, and $\widehat{\text{FRP}}_{\text{AOI}} = \hat{P}[Y = 1 \mid X \leq \delta_{\text{opr}}] = 0.0044$. The estimated fraction of nonconforming parts in the stream of parts accepted by the AOI is $\hat{P}[X > \delta_{\text{opr}} \mid Y = 1] = 3.79 \times 10^{-6}$, and the estimated fraction of conforming parts in the stream of rejects is $\hat{P}[X \leq \delta_{\text{opr}} \mid Y = 0] = 0.8822$ (these fractions are 45.9×10^{-6} and 0.0999 for the operators' inspections).

4.3. Discussion and Appraisal of the Car-Parts MSA Study

For the results reported above, we must make an important reservation, that they only hold if the complication that manifested itself in the one removed part is under control. It is an unsatisfactory situation that we cannot inspect the part in question. Instead, we cannot do more than recommend keeping the AOI under surveillance for a couple of weeks more. Any part found during this period that is accepted by the AOI, but rejected by the operators, should be closely inspected, as this inspection result is singular and indicative of a specific failure mode that the AOI's algorithms do not detect reliably. Also, those parts in the MSA experiment for which the AOI's seven appraisals were not in agreement could be inspected for clues to improve the AOI's inspection algorithms.

Once this complication is judged under control, we can conclude that the AOI has no worse reproducibility than the team of operators, but it turns out to act on a stricter decision threshold. Following the company's stance that the operators' decisions constitute the de facto standard, this implies that the decision threshold of the AOI should be adjusted. This could be accomplished by selecting from the *Rej* sample the five parts that have been rejected one or two times (out of three) by the operators and seven times by the AOI. Using these parts as a training set, the AOI should be adjusted until it rejects these parts about half the time. After effective adjustment, the AOI is expected to perform as well as the operators.

Even though the combination of samples that we obtained was not optimal, and despite the issues encountered during the analysis, we believe that the analysis is reasonable. The standard errors in Table 2 reveal that the estimates are reasonably precise (except for the estimated IAP_{opr}, which we discuss below). The results of various alternative analyses (*150-logistic, 150-loglgc, 149-loglgc*) are very close (again with the exception of IAP_{opr}). The estimated IAP_{opr} should be taken with fairly large confidence margins (s.e. $= 0.0386$, which is large in view of the estimated value of $\widehat{\text{IAP}}_{\text{opr}} = 0.0994$). This standard error would have been smaller if the *Rej* sample had been sampled from the population of items rejected *by the operators* instead of the AOI. This is because the AOI turns out to be substantially more strict than the operators (Figure 4) and, consequently, most of the items rejected by the AOI have X-values close to δ_{AOI} but to the left of δ_{opr}. This in turn implies that there is limited information in the sample that we obtained for fitting the middle and right part of q_{opr}. In hindsight, sampling 100 parts from the stream of items rejected by the operators would have been better than the 100 parts sampled from the total population (the *Tot* sample).

Besides the three analyses summarized in Table 2, we also fitted a model with a log-logistic (asymmetric) curve for the AOI, and a logistic (symmetric) curve for the operators. This alternative analysis and our final fit (the *149-loglgc* analysis) are equivalent in terms of goodness of fit ($G = 28.6$ instead of $G = 28.3$) and also the estimates are very close. The largest difference is for $\mathrm{IAP}_{\mathrm{opr}}$, which is estimated as 0.0774 instead of 0.0994. Note that this difference is small compared with the standard error of this estimate.

5. Discussion

In the recent academic literature, it is generally assumed that the traditional notion of constant misclassification probabilities across items and situations is too simplistic for most applications (see the references in the Introduction). Current literature explores various ways to accommodate this variability in the statistical models underlying the evaluation approaches. In our modeling, we have attributed the variability of the misclassification probabilities to the degree to which the condition is present that the inspections aim to detect. Thus, variability in the misclassification probabilities is linked by a characteristic curve to the stochastic properties of a random variable X, misalignment in the case study. This type of modeling is explored further in De Mast et al. (2014).

Two other approaches in the literature are closely related. The traditional approach for MSA under absence of a gold standard is to fit a latent class model (references given in the Introduction), which assumes that the rejection probabilities are constant in the subpopulations of good and defective items. However, this is now generally considered an implausible assumption in almost every situation (e.g., De Mast et al. (2011)).

Danila et al. (2012) propose an alternative approach based on a random-effects model. That model, like the model presented in this paper, also allows nonconstant error rates. However, variability in the error rates is not attributed to specific factors by means of link functions, but instead, an item's misclassification probabilities are assumed to be realizations of two beta distributions, one for good and one for defective items. An advantage of that model is that it allows the measurement outcomes to be affected by more than one property of the items. However, it does not distinguish between variability induced by properties of the items (the measurand) and variability induced by properties of the inspection system (the characteristic curves). We think that an advantage of the approach based on latent trait models is that it gives more informative results and can also be naturally extended to multiple appraisers or multiple inspection systems. Note how the fitted characteristic curves of the AOI and the operators improve the understanding of the reliability of the inspections and inspire ways to improve the AOI's performance. An objective of further research is to produce more detailed recommendations where one or the other approach is more promising.

A difficult problem in designing MSA studies, noted in the literature on the evaluation of industrial tests and medical tests as well, is the problem that the part of the characteristic curve q that is close to or to the right of δ is typically in the remote right tail of the distribution ϕ of X. As a consequence, simply sampling from the total items population provides an ineffective basis for estimating ϕ and q simultaneously. We believe that our approach, which combines samples from multiple sources, and takes the origin of each sample into account in the estimation procedure, offers an effective solution to this problem, as we have demonstrated for the specific case that we have described. Simulations show that the estimators $\widehat{\mathrm{IAP}}$ and $\widehat{\mathrm{IRP}}$ have the highest precision if only rejected items are included in the MSA experiment and the data are supplemented with a historical dataset. Furthermore, simulations show that this procedure is robust to certain forms of misspecification: even if the distribution of the measurand has heavy tails or is asymmetric or if the measurand is multidimensional, the estimators perform reasonably well in terms of bias and precision.

Acknowledgments

The authors thank Rob Cuperus and Peter Chen for their valuable contribution and pleasant collaboration in the case reported in this paper.

References

AIAG (2003). *Measurement System Analysis: Reference Manual*, 3rd ed. Detroit, MI: Automotive Industry Action Group.

Azzalini, A. (1985). "A Class of Distributions Which Includes the Normal Ones". *Scandinavian Journal of Statistics* 12, pp. 171–178.

Boyles, R. A. (2001). "Gauge Capability for Pass-Fail Inspection". *Technometrics* 43, pp. 223–229.

Danila, O.; Steiner, S. H.; and MacKay, R. J. (2008). "Assessing a Binary Measurement System". *Journal of Quality Technology* 40, pp. 310–318.

Danila, O.; Steiner, S. H.; and MacKay, R. J. (2010). "As-

sessment of a Binary Measurement System in Current Use". *Journal of Quality Technology* 42, pp. 152–164.

Danila, O.; Steiner, S. H.; and MacKay, R. J. (2012). "Assessing a Binary Measurement System with Varying Misclassification Rates Using a Latent Class Random Effects Model". *Journal of Quality Technology* 44, pp. 179–191.

De Mast, J.; Akkerhuis, T.; and Erdmann, T. P. (2014). "The Statistical Evaluation of Categorical Measurements". *Quality Engineering* 26(1), pp. 16–32.

De Mast, J.; Erdmann, T. P.; and Van Wieringen, W. N. (2011). "Measurement System Analysis for Binary Inspection: Continuous Versus Dichotomous Measurands". *Journal of Quality Technology* 43, pp. 99–112.

De Mast, J. and Van Wieringen, W. N. (2010). "Modeling and Evaluating Repeatability and Reproducibility of Ordinal Classifications". *Technometrics* 52, pp. 94–106.

Feinstein, A. (2002). "Misguided Efforts and Future Challenges for Research on 'Diagnostic Tests"'. *Journal of Epidemiology Community Health* 56, pp. 330–332.

Formann, A. K. (2003). "Latent Class Model Diagnostics—A Review and Some Proposals". *Computational Statistics & Data Analysis* 41, pp. 549–559.

Hui, S. L. and Walter, S. D. (1980). "Estimating the Error Rates of Diagnostic Tests". *Biometrics* 36, pp. 167–171.

Irwig, L. M.; Bossuyt, P. M. M.; Glasziou, P. P.; Gatsonis, C.; and Lijmer, J. G. (2002). "Designing Studies to Ensure that Estimates of Test Accuracy Are Transferable". *British Medical Journal* 324, pp. 669–671.

ISO (1995). *Guide to the Expression of Uncertainty in Measurement*, 1st ed. Geneva, Switzerland: International Organization for Standardization.

Kallenberg, W. C. M.; Oosterhoff, J.; and Schriever, B. F. (1985). "The Number of Classes in Chi-Squared Goodness-of-Fit Tests". *Journal of the American Statistical Association* 80, pp. 959–968.

Knottnerus, J. A.; Van Weel, C.; and Muris, J. W. M. (2002). "Evaluation of Diagnostic Procedures". *British Medical Journal* 324, pp. 477–480.

Kraemer, H. C. (1987). "The Methodological and Statistical Evaluation of Medical Tests: The Dexamethasone Suppression Test in Psychiatry". *Psychoneuroendocrinology* 12, pp. 411–427.

Lord, F. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Mathematica 8 (2010). Champaign, IL: Wolfram Research Inc. Available at: www.wolfram.com.

Mehrotra, S. (1992). "On the Implementation of a Primal-Dual Interior Point Method". *SIAM Journal on Optimization* 2, pp. 575–601.

Nelder, J. A. and Mead, R. (1965). "A Simplex Method for Function Minimization". *The Computer Journal* 7, pp. 308–313.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, UK: Oxford University Press.

Rice, J. R. (1975). "A Metalgorithm for Adaptive Quadrature". *Journal of the ACM* 22, pp. 61–82.

Van Wieringen, W. N. and De Mast, J. (2008). "Measurement System Analysis for Binary Data". *Technometrics* 50, pp. 468–478.