Measurement 95 (2017) 473–479

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

The statistical evaluation of binary tests without gold standard: Robustness of latent variable approaches

Thomas Akkerhuis [a,*], Jeroen de Mast [a], Tashi Erdmann [b]

[a] Department of Operations Management, Amsterdam Business School, University of Amsterdam, P.O. Box 15953, 1001 NL Amsterdam, The Netherlands
[b] Shell Technology Centre Amsterdam, P.O. Box 38000, 1030 BN Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 20 June 2016
Received in revised form 12 October 2016
Accepted 15 October 2016
Available online 17 October 2016

Keywords:
Binary measurement
Binary data
Precision
Reproducibility and repeatability
Pass/fail inspection

ABSTRACT

Binary tests classify items into two categories such as reject/accept or positive/negative. Such tests are usually evaluated in terms of their misclassification probabilities *FAP* (false acceptance probability) and *FRP* (false rejection probability). A common complication arises when there is no gold standard or reference standard. Various methods based on latent variable modelling have been proposed for this situation. We present the results of a simulation study in which these methods are tried in a range of scenarios, to study how robust they are to departures from the assumptions on which they are based.

The study convincingly shows that in general, the ambition of estimating *FAP* and *FRP* without gold standard is unattainable, since all methods easily derail when assumptions are not precisely met. The study also shows that the random components of the *FAP* and *FRP* can be reliably estimated by a straightforward modification of one of the tested methods.

1. Introduction

Binary tests are common in industrial processes, and classify items in two categories such as 'reject' ($Y = 0$) or 'accept' ($Y = 1$). Examples include visual quality inspections and automated tests where some parts fail and others pass. Diagnostic and screening tests in medicine, yielding the results 'positive' or 'negative', are closely related. We conceive of such tests as a form of measurement ("binary measurement", see Suzuki et al. [1]), and thus, these classifications aim to reflect an underlying true state X of the items called the measurand [2,3], which can be 'truly defective' ($X = 0$) or 'conforming to specifications' ($X = 1$).

A measurement system analysis (MSA) experiment is an experiment to evaluate how reliably the test results Y reflect the measurand X. Measurement error is generally defined as the discrepancy between measured and true value. Binary scales are equipped with only the simplest of algebraic structures and in particular, subtraction and addition are usually not meaningfully defined [4]. Consequently, it is problematic to define discrepancy in terms of a difference $Y - X$ (or derived statistics such as standard deviations). For binary measurements, therefore, the measurement error is usually expressed as a misclassification ($Y \neq X$), and a statistical evaluation is in terms of the misclassification probabilities:

$$FAP = P[Y = 1|X = 0] \quad \text{(False Acceptance Probability)},$$

$$FRP = P[Y = 0|X = 1] \quad \text{(False Rejection Probability)}.$$

In medicine, one sometimes works with the complements of these probabilities: the sensitivity and specificity. Traditional methods for estimating the *FAP* and *FRP* (as described in AIAG [5], Pepe [6], Danila et al. [7,8] and many other articles), require a so-called gold standard or reference standard: a higher-order, authoritative test that is accepted to constitute a faithful representation of the measurand [9]. A common problem occurs when such gold standard is not available, in which case the true state of the items is practically unobservable. Reasons for this include the absence of a sufficiently capable authoritative test, ambiguity of the specifications (such as when human perception is involved), prohibitive cost, or damage resulting from applying the gold standard to the tested items.

When a gold standard is unavailable, the traditional solution is to fit a latent class model, in which *FAP* and *FRP* are assumed constant in the subpopulations of defective and conforming items (see Hui and Walter [10], Boyles [11], Van Wieringen and De Mast [12], Danila et al. [13] and many others). However, this assumption has been discredited as generally not realistic [14–17]. For example, if there are various degrees of defectiveness such that some parts are harder to judge than others, this assumption is violated and estimators may have a substantial bias.

* Corresponding author.
E-mail address: t.s.akkerhuis@uva.nl (T. Akkerhuis).

http://dx.doi.org/10.1016/j.measurement.2016.10.043
0263-2241/

To allow for variability in the misclassification probabilities in each subpopulation, more complex latent variable models have been proposed. In industrial statistics, Danila et al. [18] treat *FAP* and *FRP* as random effects with beta distributions, and Erdmann et al. [19] explicitly attribute variability in the misclassification probability to an underlying continuous property of the items. Also in medical statistics various more complex approaches have been suggested [20,21]. However, Albert and Dodd [22] warn that these models are not robust against model misspecification, which is typically difficult to detect from the binary observations. Mathematical analysis by Akkerhuis [23] reveals that the parameters *FAP* and *FRP* are in general unidentifiable from the binary observations when a gold standard is unavailable. This suggests that the estimation of *FAP* and *FRP* is inherently problematic without a gold standard.

This paper compares the main approaches proposed in industrial statistics for the evaluation of binary tests without a gold standard, to learn how robust they are to model misspecification. The study is based on a crossed design where methods *M1*, *M2* and *M3* are applied in a range of scenarios *S1*, *S2*, …. The estimation bias of the methods in the various scenarios is analyzed with three goals:

- Establish which method is most reliable across a range of realistic scenarios.
- Empirically establish how problematic estimation of *FAP* and *FRP* is without gold standard.
- If the problems turn out to be severe: identify alternative ways to evaluate binary tests.

The next section explains the set-up of the study in detail. Section 3 presents the findings of the study, which results in specific conclusions discussed in the final section.

2. Theory and methods

The current literature describes three classes of approaches for estimating misclassification probabilities of binary tests under absence of a gold standard: traditional latent class methods (*M1*), latent class random effects approaches (*M2*), and approaches based on characteristic curves (*M3*). In the comparison study we try these methods *M1*, *M2* and *M3* in a number of scenarios (*S1*, *S2*, …) to learn how sensitive they are to violations of their assumptions.

In this section, we present a novel statistical modelling framework that is general enough to describe all methods *M1*, *M2* and *M3* under study as special cases. Also the test scenarios will be defined in terms of this modelling framework. The descriptions of the methods are cursory, only highlighting the main idea, but full descriptions can be found in the references.

2.1. Statistical modelling framework

When a test is applied in regular production, it produces results $Y_i = 0$ (rejection) or $Y_i = 1$ (acceptance) for tested items $i = 1, 2, \ldots$. The unobservable true states of the items are $X_i = 0$ (truly defective) or $X_i = 1$ (conforming to specifications), and the (unknown) defect rate is $p = P[X_i = 0]$. The probability that an item i is rejected is $R_i = P[Y_i = 0]$, which depends on X_i and possibly on other properties affecting the measurement.

To determine the error probabilities of the test, one executes an MSA experiment, in which a sample of I items are tested J times, producing the results $Y_{ij} \in \{0, 1\}$. When a gold standard is available, the corresponding true values $X_1, \ldots, X_I$ are established, and a comparison of Y_{ij} to X_i allows the calculation of *FAP* and *FRP*. In the problem under consideration, however, a gold standard is unavailable and the true states X_i of the items in the experiment are unobservable. In fact, one can only observe the per-item rejection counts $U_i = \sum_{j=1}^{J}(1 - Y_{ij}) \in \{0, 1, \ldots, J\}$, from which the rejection probabilities R_i can be estimated.

We will specify statistical models in terms of the R_i, avoiding the unobservable X_i. The rejection probabilities of items vary from 0 to 1 and have a statistical distribution $F_R(r) = P[R_i \leqslant r], r \in [0, 1]$. We assume that given an item's rejection probability, repeated tests are independent: conditional on the event $\{R_i = r\}$, the $Y_{i1}, \ldots, Y_{ij}$ are i.i.d. Bernoulli (with parameter $1 - r$) and the rejection counts U_i have a binomial (J, r) distribution. A special case is the traditional latent class model, where the rejection probabilities assume only two values: $R_i = FRP$ for all conforming items and $R_i = 1 - FAP$ for all defective items (with *FRP* and *FAP* two constants).

The distribution of rejection probabilities can be interpreted as a mixture of two component distributions: $F_R(r) = pF_R^0(r) + (1 - p)F_R^1(r)$, with

$$F_R^0(r) = P[R_i \leqslant r | X_i = 0]$$
(distribution of rejection probabilities of defective items),

$$F_R^1(r) = P[R_i \leqslant r | X_i = 1]$$
(distribution of rejection probabilities of conforming items).

Fig. 1 gives an example. The solid curve is the density f_R of rejection probabilities in the population of items. The gray and white areas show the components pf_R^0 and $(1 - p)f_R^1$ associated with defective and conforming items. In this example (produced by scenario *S2a* explained later), the rejection probabilities of conforming and defective items are clearly separated and $f_R(r) \approx 0$ around $r = 0.6$.

The *FAP* and *FRP* are the mean probabilities of misclassification of conforming or defective items:

$$FAP = 1 - E[R_i | X_i = 0],$$

$$FRP = E[R_i | X_i = 1].$$

In MSA studies for quantitative measurements, the measurement error is often decomposed into systematic and random measurement error (or similar concepts such as trueness/precision). Recent literature has proposed a similar decomposition for binary MSA [19,23]. Let $\tilde{X}_i = 0$ if $R_i > 0.5$ and $\tilde{X}_i = 1$ if $R_i \leqslant 0.5$ be the modal (most likely) outcome for item i. Then, $\tilde{X}_i \neq X_i$ implies that test results for item i are systematically off, and $Y_{ij} \neq \tilde{X}_i$ is a random deviation of a test result from the modal outcome. Akkerhuis [23] shows how the misclassification probabilities can be decomposed; the terms corresponding to random measurement error are

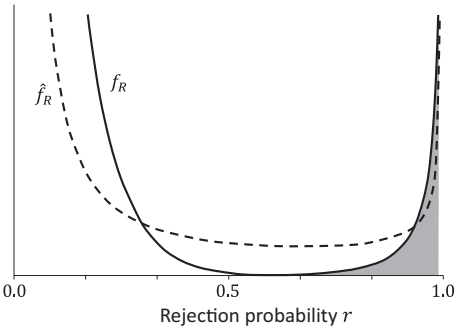

Fig. 1. Density f_R of rejection probabilities in scenario *S2a* (solid curve). Estimated pdf $\hat{f}_R$ fitted by method *M3* (dashed curve).

$$IAP = 1 - E[R_i|\tilde{X}_i = 0] = P[Y_{ij} = 1|\tilde{X}_i = 0]$$
(Inconsistent Acceptance Probability),

$$IRP = E[R_i|\tilde{X}_i = 1] = P[Y_{ij} = 0|\tilde{X}_i = 1]$$
(Inconsistent Rejection Probability).

They represent the probability that a test result randomly deviates from the most likely result for the item. These statistics have a longer history in binary MSA, as measures for the repeatability of binary tests; c.f. De Mast [17].

2.2. Methods included in the comparison

In the study we compare the sensitivity to model misspecification of three methods for estimating the *FAP* and *FRP*, and/or the *IAP* and *IRP*, proposed in the recent literature in industrial statistics.

2.2.1. M1 traditional latent class method

Approaches based on a latent class model have been well studied in literature. These methods fit latent class models in which *FAP* and *FRP* are constants in the subpopulations of conforming and defective items [11,10,12,13,26]. Method *M1* in our comparison fits a model with $R_i = FRP$ if $X_i = 1$ and $R_i = 1 - FAP$ if $X_i = 0$. The three model parameters *FRP*, *FAP* and $p = P[X_i = 0]$ are commonly estimated by a maximum likelihood or Bayesian algorithm (e.g. Gadrich and Bashkansky [24]). Under the simple setting of *M1* we have, under the mild assumption that $FAP < 0.5$ and $FRP < 0.5$, that $IAP = FAP$ and $IRP = FRP$. In view of this, the estimated *IAP* and *IRP* are identical to the estimated *FAP* and *FRP*.

2.2.2. M2 latent class random effects approach

An extension of the traditional latent class method allows *FAP* and *FRP* to be random effects, drawn from distributions F_R^0 and F_R^1. Danila et al. [18] propose a method based on a mixture of two beta distributions: F_R^0 is a beta distribution with mean $\mu_0 = 1 - FAP$ and dispersion ϕ_0, and F_R^1 is a beta distribution with mean $\mu_1 = FRP$ and dispersion ϕ_1. The five parameters μ_0, μ_1, ϕ_0, ϕ_1 and p are fitted by a maximum likelihood algorithm. Based on the beta distributions in each subpopulation, the density of the rejection probabilities in the entire items population is the mixture

$$f_R(r) = p\frac{r^{\mu_0\frac{1-\phi_0}{\phi_0}-1}(1-r)^{(1-\mu_0)\frac{1-\phi_0}{\phi_0}-1}}{\mathrm{Beta}\left(\mu_0\frac{1-\phi_0}{\phi_0}, (1-\mu_0)\frac{1-\phi_0}{\phi_0}\right)} + (1-p) \times \frac{r^{(1-\mu_1)\frac{1-\phi_1}{\phi_1}-1}(1-r)^{\mu_1\frac{1-\phi_1}{\phi_1}-1}}{\mathrm{Beta}\left((1-\mu_1)\frac{1-\phi_1}{\phi_1}, \mu_1\frac{1-\phi_1}{\phi_1}\right)}.$$

FAP and *FRP* are estimated from the fitted μ_0 and μ_1. As introduced originally, the method does not produce estimates for *IAP* and *IRP*. However, these can be obtained from the fitted density $\hat{f}_R$ by

$$\widehat{IRP} = \frac{\int_0^{0.5} r\hat{f}_R(r)dr}{\int_0^{0.5}\hat{f}_R(r)dr}, \text{ and}$$

$$\widehat{IAP} = \frac{\int_{0.5}^1 (1-r)\hat{f}_R(r)dr}{\int_{0.5}^1 \hat{f}_R(r)dr}.$$

2.2.3. M3 approaches based on characteristic curves

Another approach models the source of variability among items more explicitly [17,19]. It assumes that the dichotomous true state masks an underlying continuum where some items are more defective than others (and, possibly, some items are more conforming than others). Thus, the true state X_i is determined by a latent continuous quality characteristic $Z_i \in \mathbb{R}$ in the sense that $X_i = 1$ if $Z_i \leqslant L$ and $X_i = 0$ otherwise (where L is the limit defining for which values of Z_i an item is out of specification).

Variability in the rejection probabilities, then, is related to variability in Z_i by means of a characteristic curve (here a logistic curve with inflection point at $z = \delta$ and slope α)

$$q(z) = P[Y_{ij} = 0|Z_i = z] = (1 + \exp[-\alpha(z-\delta)])^{-1}.$$

Writing $F_Z(z) = P[Z_i \leqslant z]$ for the distribution of Z_i in the population of items, the distribution of rejection probabilities is

$$F_R(r) = P[q(Z_i) \leqslant r] = F_Z(q^{-1}(r)).$$

Erdmann et al. [19] take F_Z to be the standard normal distribution. The model's parameters α and δ are estimated by a maximum likelihood algorithm. From the fitted characteristic curve, *IAP* and *IRP* are determined from

$$\widehat{IRP} = \int_{-\infty}^{\hat{\delta}} \hat{q}(z)\phi(z)dz \Big/ \int_{-\infty}^{\hat{\delta}} \phi(z)dz, \text{ and}$$

$$\widehat{IAP} = \int_{\hat{\delta}}^{\infty} (1-\hat{q}(z))\phi(z)dz \Big/ \int_{\hat{\delta}}^{\infty} \phi(z)dz.$$

The approach does not offer a way to estimate *FAP* and *FRP*.

For all three methods, recent literature has investigated the suitability of various sampling strategies. For the MSA study, one needs a sample $i = 1, \ldots, I$ of items. The most straightforward sampling strategy is to take a representative sample from the total population of relevant items. However, given the usually very low defect rates of industrial processes, samples thus collected typically contain no or only very few items with $R_i > 0.5$. Recent studies have investigated the effect of various alternative sampling strategies on the precision of the estimates (Danila at al. for approaches of type *M1* [13] and *M2* [18]; Erdmann et al. for *M3* [19]). Consistently, these studies show that the best sampling strategy is to take a representative sample from the stream of rejected items, which typically produces a sample with more evenly spread R_i (note that under realistic *FAP*, *FRP* and p even the stream of rejected items contains more good items, rejected falsely, than defective items; see De Mast et al. [17]). However, such a sample is not representative for the total population of items, and to correct for the resulting bias in the estimators, the likelihood contributions in the maximum likelihood procedures should be calculated conditional on the event that they have been rejected initially (see Appendix A for the details of this procedure).

A further improvement can be obtained by augmenting the data from the MSA experiment with an estimate for the rejection rate $P[Y_i = 0]$ based on historical data. Even if such historical data are not available, they can be recorded as a by-product in the time period when the sample is collected from the stream of rejected items. Recent papers call such data *baseline data* and demonstrate how their incorporation in the estimation procedure makes estimates more precise [13,18,19]. It is this setup, where the MSA study involves a sample of I items taken from the rejection stream plus a historical estimate for the rejection rate, that we have implemented for all methods *M1*, *M2* and *M3*.

2.3. Scenarios

To analyze the sensitivity of methods *M1*, *M2* and *M3* to model misspecification, their estimation bias was compared in 10 scenarios. Scenario *S1* is a situation where measurements result from the simple latent class model (as in method *M1*). In scenarios *S2a*, *S2b* and *S2c*, data are generated from random effects models (as in *M2*),

in which the R_i have beta or logit-logistic distributions with little (*S2a*) or substantial (*S2b,c*) dispersion. Scenarios *S3a* through *S3d* (related to *M3*) are based on a standard normal continuous characteristic Z. This characteristic determines the rejection probability through a steep logistic characteristic curve in *S3a*, and a flat logistic curve in *S3b*. In *S3c* the characteristic curve is an asymmetric, cumulative Weibull distribution, and in *S3d* it is a cumulative Weibull distribution with an infimum set at $\inf_z q(z) = 0.01$ instead of 0.00.

Scenarios *S4a* and *S4b* incorporate the effects of nuisance variables that may affect the outcomes of the tests. In real life, these could be the effects of conditions under which the tests are performed (such as light conditions for visual inspections) or characteristics of the appraisers (such as fatigue and motivation). Such nuisances may leverage measurement error, and scenarios *S4a* and *S4b* allow us to study whether methods are robust against their effects.

We model nuisance variables as a standard normal random variable V_{ij}, which may assume a new value for each item i and each repeat j. In scenario *S4a*, it affects the characteristic curve's slope in its inflection point:

$$P[Y_{ij} = 0|Z_i = z, V_{ij} = v] = q(z, v) = (1 + \exp[-10e^{v}(z - 2)])^{-1}$$

and in *S4b* it affects the location of the curve's inflection point:

$$P[Y_{ij} = 0|Z_i = z, V_{ij} = v] = q(z, v) = (1 + \exp[-10(z + v - 2)])^{-1}$$

Further details of the investigated scenarios are presented in Appendix A.

2.4. Implementation details

In our study, the parameters of the models prescribed by methods *M1* (p, *FAP* and *FRP*), *M2* (p, μ_0, μ_1, ϕ_0 and ϕ_1) and *M3* (α and δ) are estimated by maximum likelihood, as is often used for fitting these models. We implemented the recommended MSA setup where I items are sampled from the stream of rejected items and tested J times, and in addition, historical test results are available for I^{His} baseline data.

As the study focusses on robustness against model misspecification rather than the determination of appropriate sample sizes, the comparison study treats sample sizes as neutrally as possible. Recommendations in the literature for the number of repeats are in the range of $J = 7$ (Erdmann et al. [19] for method *M3*) to $J = 10$ (Danila et al. [18] for *M2*), with $J = 3$ as a minimum to ensure identifiability (Van Wieringen and De Mast [12] for *M1*). In our study, we have worked with $J = 7$. The size of the sample of items is left out of consideration by studying the limit behavior of estimators when I approaches infinity. For this reason, estimators converge to the true values when the model assumed in a scenario is the same as the model assumed in the method (such as method *M1* applied in scenario *S1*). The number I^{His} of historical data is taken in a fixed proportion to the sample size I, namely, $I^{His} = I/P[Y_{ij} = 0]$. This is the ratio that one finds if no historical data are available, and instead, the baseline data are collected as a by-product of obtaining the sample of I rejected items. Namely, in order to obtain I rejected items, one has to produce an expected number of $I/P[Y_{ij} = 0]$ items. When historical data are available, the number I^{His} of baseline data will almost always be larger.

Likelihood functions and other details of the implementation are given in Appendix A. For our calculations, we used Mathematica 8 [25]. Loglikelihood functions are optimized with the interior point algorithm (as implemented in the function "FindMaximum") or the Nelder-Mead algorithm (as implemented in "NMaximize"). The integrals are approximated numerically using adaptive quadrature (as implemented in "NIntegrate").

3. Results

Applying the methods *M1*, *M2* and *M3* across the selected scenarios reveals to what extent they are robust against violations of the assumptions on which they are based. Table 1 presents the results. We discuss our findings and motivate conclusions for the robustness of methods *M1*, *M2* and *M3*.

3.1. Traditional latent class method (M1)

Our findings for the performance of the traditional latent class method *M1* confirm earlier findings in the medical statistics literature [15,22]. Namely, such methods only work when the rejection probabilities are constant in the subpopulations of conforming and defective items (scenario *S1*), but seriously derail in other scenarios, where *M1* gives badly biased estimates for *FAP*, *FRP*, *IAP* and *IRP*. Fig. 2 shows a typical example. Here, the solid line is the density f_R of rejection probabilities produced by scenario *S2b*, where the rejection probabilities of conforming and defective items follow beta distributions. The implied true misclassification probabilities are $FAP = 1 - \int r f_R^0(r)dr = 0.05$ and $FRP = \int r f_R^1(r)dr = 0.05$. Further, $IAP = 1 - \int_{0.5}^{1} r f_R(r)dr = 0.1293$ and $IRP = \int_0^{0.5} r f_R(r)dr = 0.0255$. Method *M1* yields $\hat{p} = 0.0882$, $\widehat{FAP} = \widehat{IAP} = 0.1220$ and $\widehat{FRP} = \widehat{IRP} = 0.0357$ (which gives the point masses $P[R_i = 0.0357] = 0.9118$ and $P[R_i = 0.878] = 0.0882$). Note that $\widehat{FAP}$ and $\widehat{FRP}$ are both strongly biased. The estimates for *IAP* and *IRP* are more accurate, but in other scenarios (*S3a*, *S3b*, *S3c*) the biases are enormous. In this and similar situations, *M1* approximates a continuous density f_R of rejection probabilities by two point masses, and the resulting estimates are unreliable.

3.2. Approaches based on characteristic curves (M3)

The method *M3* based on a characteristic curve model produces reliable estimates for *IAP* and *IRP* under scenarios *S3a–b*, which are instances of its native model. However, the method turns out to lack robustness in many other scenarios and produces a poor fit. The performance is particularly poor in scenarios where there is little variation in the rejection rates R_i of the items (*S1*, *S2a*, *S2c*, *S3c* and *S3d*). We illustrate the problem with the typical example of scenario *S2a* (Fig. 1 in Section 2). The solid curve is f_R with $FAP = FRP = 0.0500$, $IAP = 0.0540$ and $IRP = 0.0497$. There is little variation in rejection probabilities, which can be seen from f_R being almost zero for rejection probabilities between 0.4 and 0.8. The characteristic curve approach *M3* turns out to be unable to fit this low degree of variation (*M3* produces the dashed curve with $\hat{\alpha} = 4.56$ and $\hat{\delta} = 1.43$). *M3* does not offer a way of estimating *FAP* and *FRP*, but instead, the method yields the estimates $\widehat{IAP} = 0.200$ and $\widehat{IRP} = 0.033$, which are rather far off.

3.3. Latent class random effects method (M2)

The latent class random effects approach *M2* gives a good fit of f_R across scenarios, which is a promising finding. The *M2* method fits a model with 5 parameters, 3 more than *M3*, which may explain the relatively good performance. For many scenarios, however, the estimated *FAP* and *FRP* are substantially off. To illustrate, the true f_R in scenario *S2c* is a mixture of logit-logistic distributions, but it is approximated rather well by the mixture of beta distributions fitted by method *M2*. However, the estimated *FAP* and *FRP* (0.1891 and 0.0373) are far off (true values 0.0500 and

Table 1
Overview of the results when methods *M1–3* are applied in scenarios *S1–4*.

Scenario	Method	FAP	FRP	IAP	IRP
S1	*TRUE*	*0.0500*	*0.0500*	*0.0500*	*0.0500*
Constant *FAP*	M1[a]	0.0500	0.0500	0.0500	0.0500
and *FRP*	M2	0.0500	0.0500	0.0500	0.0500
	M3	NA	NA	0.2153	0.0358
S2a	*TRUE*	*0.0500*	*0.0500*	*0.0540*	*0.0497*
Beta distributions	M1	0.0535	0.0692	0.0535	0.0692
(little dispersion)	M2[a]	0.0500	0.0500	0.0540	0.0497
	M3	NA	NA	0.2001	0.0330
S2b	*TRUE*	*0.0500*	*0.0500*	*0.1293*	*0.0255*
Beta distributions	M1	0.1220	0.0357	0.1220	0.0357
(substantial dispersion)	M2[a]	0.0500	0.0500	0.1293	0.0255
	M3	NA	NA	0.1359	0.0200
S2c	*TRUE*	*0.0500*	*0.0500*	*0.0995*	*0.0405*
Logit-logistic	M2	0.1891	0.0373	0.0970	0.0399
Distributions	M3	NA	NA	0.1765	0.0280
S3a	*TRUE*	*0.5360*	*0.0000*	*0.0733*	*0.0022*
Std normal F_Z	M1	0.0206	0.4179	0.0206	0.4179
Steep logistic	M2	0.1598	0.0004	0.0799	0.0014
Characteristic curve	M3[a]	NA	NA	0.0733	0.0022
S3b	*TRUE*	*0.0252*	*0.0183*	*0.2863*	*0.0278*
Std normal F_Z	M1	0.2673	0.1584	0.2673	0.1584
Flat logistic	M2	0.5614	0.0205	0.2846	0.0275
Characteristic curve	M3[a]	NA	NA	0.2863	0.0278
S3c	*TRUE*	*0.5587*	*0.0000*	*0.0538*	*0.0005*
Std normal F_Z	M1	0.0135	0.5895	0.0135	0.5895
Cum. Weibull	M2	0.1030	0.0000	0.0540	0.0005
Characteristic curve	M3	NA	NA	0.0476	0.0012
S3d	*TRUE*	*0.5531*	*0.0100*	*0.0538*	*0.0105*
Std normal F_Z	M1	0.0370	0.0284	0.0370	0.0284
cum. Weibull	M2	0.1040	0.0100	0.0540	0.0105
characteristic curve	M3	NA	NA	0.1439	0.0071
S4a	*TRUE*	*NA*	*NA*	*0.1522*	*0.0137*
Outcomes affected by	M1	0.2217	0.0120	0.2217	0.0120
Nuisance variable *V*	M1[b]	0.1628	0.0157	0.1628	0.0157
	M2	0.1843	0.0161	0.1534	0.0137
	M2[b]	0.2231	0.0120	0.1689	0.0132
	M3	NA	NA	0.1894	0.0111
	M3[b]	NA	NA	0.2142	0.0133
S4b	*TRUE*	*NA*	*NA*	*0.3659*	*0.0675*
Outcomes affected by	M1	0.0731	0.0125	0.0731	0.0125
Nuisance variable *V*	M1[b]	0.5468	0.0440	0.5468	0.0440
	M2	0.0744	0.0153	0.0809	0.0088
	M2[b]	0.7173	0.0520	0.3650	0.0672
	M3	NA	NA	0.0812	0.0088
	M3[b]	NA	NA	0.3547	0.0678

NA: The *M3* methods do not offer a way to estimate *FAP* and *FRP*.
[a] Methods marked with a star fit the exact same model as in the scenario, and therefore converge to the true parameter values (for $I \to \infty$).
[b] Method applied with experimental randomization.

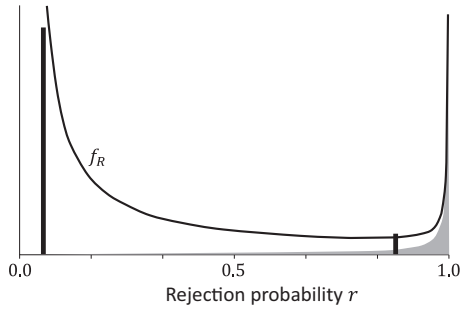

Fig. 2. True pdf f_R of rejection probabilities in scenario *S2b* (solid curve), decomposed into the densities for conforming (white) and defective (gray) items. Vertical bars represent the distribution of rejection probabilities fitted by method *M1*.

0.0500). The estimated *FAP* and *FRP* are further off in scenarios where the component densities f_R^0 and f_R^1 have more dispersion, and therefore, are less clearly separated.

Estimation of *FAP* and *FRP* becomes hopeless for the *S3* scenarios, where the true state X_i is a dichotomization of a continuous characteristic Z_i and the two component densities f_R^0 and f_R^1 are less clearly separated. For example, in Fig. 3 (scenario *S3b*), the density f_R of all items is fitted rather well by *M2*. The component densities f_R^1 for conforming items (that is, $Z_i \leqslant L$) and f_R^0 for defective items ($Z_i > L$) are separated by a vertical bound at $r = q(L) = 0.256$, but this is impossible to determine from the shape of f_R alone. Consequently, the estimated *FAP* and *FRP* are far off (true values: 0.02521 and 0.01825; estimates: 0.5614 and 0.0205).

Still, because of the good fit of f_R, method *M2* produces reliable estimates for *IAP* and *IRP* across the scenarios *S1–S3*. For example,

in scenario *S3b* (Fig. 3) $\widehat{IAP} = 0.2846$ and $\widehat{IRP} = 0.0275$ (true values are 0.2863 and 0.0278). In scenario *S2c*, the procedure gives 0.0970 and 0.0399 (true values 0.0995 and 0.0405).

3.4. Estimation in the presence of nuisance variables

The scenarios *S4a–b* explicitly incorporate the effects of a nuisance variable V_{ij} on the test results. The ideal way to deal with nuisance variables is of course to eliminate their effect in the test protocol, for example, by introducing countermeasures that keep nuisances constant. If this is not possible, we propose to handle nuisance variables by experimental randomization [4]. This means that in the MSA experiment each of the J repeats per item is done under a new realization of the nuisance variables, and such that these realizations are representative for normal circumstances. Experimental randomization was implemented in our study by drawing a new realization for V_{ij} from its assumed population distribution F_V for each repeat j and each item i. By the law of large numbers, we then have that the per-item rejection fractions U_i/J converge in probability to $\int P[Y_{ij} = 0|V_{ij} = v]f_V(v)dv = P[Y_{ij} = 0] = R_i$ (for $j \to \infty$), so the effects of the nuisance variable averages out.

When applying experimental randomization, the reliable performance of method *M2* in estimating *IAP* and *IRP* is maintained even under the presence of nuisance variables. In *S4b*, for instance, it gives $\widehat{IAP} = 0.3650$ and $\widehat{IRP} = 0.0672$ (true values 0.3659 and 0.0675). The importance of experimental randomization is illustrated by applying *M2* under a more careless setup, where only a single realization of V_{ij} is drawn per item i, say V_{i0}. The rejection fractions U_i/J now converge to $P[Y_{ij} = 0|V_{i0}]$ (for $j \to \infty$), which is generally not equal to R_i. In the simulations, *M2* gave $\widehat{IAP} = 0.0809$ and $\widehat{IRP} = 0.0088$ in scenario *S4b* (which are rather far off).

4. Discussion and conclusions

Given the ubiquity of binary tests and measurements in industry, medicine and beyond, and the often severe ramifications of false results, reliable methods for assessing their misclassification probabilities are important. The unavailability of a gold standard is a common complication, and despite a large volume of papers on the subject, there still is no approach that is generally accepted.

Latent variable methods are an obvious option to deal with an unobservable measurand. Literature has explored various directions to do so, but there is no consensus about the suitability of these methods. And in fact, a mathematical analysis by Akkerhuis [23] shows that generally, *FAP* and *FRP* are unidentifiable parameters when a gold standard is unavailable. In this study we have investigated empirically whether currently available methods are applicable, and if not, how severe the problems are.

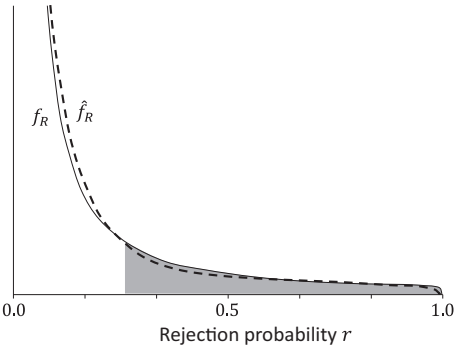

Fig. 3. True pdf f_R of rejection probabilities in scenario *S3b* (solid curve), with the densities of conforming and defective items marked by white and gray. Estimated pdf $\hat{f}_R$ fitted by method *M2* (dashed curve).

The first set of conclusions mirrors the results of a comparable study for methods proposed in medical statistics [22]: estimation methods for *FAP* and *FRP* easily become severely biased if their model assumptions are no precisely met. Such departures from model assumptions are difficult to detect from the binary test results, so this makes application of these estimation methods unreliable. Especially under the (realistic) scenarios *S3*, estimation of *FAP* and *FRP* becomes hopeless, with biases that are enormous.

Taken together with the mathematical analysis of Akkerhuis [23], these results constitute a strong case that the estimation of the *FAP* and *FRP* of a binary test without gold standard is in general an unattainable ambition. Especially if one is not satisfied with evaluating a binary test in terms of *IAP* and *IRP*, this is a strong motivation to go quite some way in finding a gold standard.

By analogy with MSA studies for numerical measurements, Akkerhuis [23] shows how the total misclassification errors *FAP* and *FRP* can be decomposed into systematic errors and random errors. The latter can be quantified as the *IAP* and *IRP*. Our current study shows that *IAP* and *IRP* can reliably be estimated by a modification of method *M2*, and that even the presence of nuisance effects can be handled by carefully randomizing repeated tests. If *IAP* and *IRP* are poor, this means that there is too much random variability in the test system's results for them to be informative, and even a good calibration will not save such system. If *IAP* and *IRP* are good, this means that the test system's discriminative capability is sufficient for discerning worse from better items: the test results are reproducible. For many applications, this will in itself be a useful result. If good reproducibility is not enough, and an application demands that also the systematic error is controlled, the system needs a calibration. This is problematic without a gold standard. For test systems with good *IAP* and *IRP*, a rough calibration will typically do, and it may be possible to do so on the basis of a training set of items that are rated as defective or conforming by consensus among experts.

The model fitted in *M2* is a mixture of two beta distributions, reflecting its original purpose of estimating *FAP* and *FRP*. For the purpose of evaluating a test system in terms of its *IAP* and *IRP*, the class of distributions that is fitted to the data does not need to be a mixture of two distributions corresponding to conforming and defective items. This leaves a wide class of candidate distribution functions open for exploration, with the aim of finding a class of functions that has a limited number of parameters, but that nevertheless can reliably fit a wide range of scenarios.

All in all, the study offers bad and good news. Without gold standard, the estimation of the error probabilities *FAP* and *FRP* is usually not possible, but it is possible to determine the random components *IAP* and *IRP* of these errors. Although this conclusion is in line with theory and practice in MSA studies for numerical measurements, it represents quite a drastic turn in the development of techniques for binary MSA.

Appendix A

Each of the methods *M1*, *M2* and *M3* fits a model by maximum likelihood estimation on the basis of a sample of initially rejected items augmented with baseline data. We discuss here the details of this estimation procedure.

Data are generated by the models defined in each of the scenarios *S1–4*, and probabilities calculated under these data generating models are denoted P_{Sc}. The models fitted to the data depend on the methods *M1–3*, and probabilities calculated under these fitted

models are denoted P_θ, with θ the parameter vector of the fitted model. The results of the MSA study are the rejection counts $U_i = \sum_{j=1}^{J}(1 - Y_{ij})$, which can be aggregated in the response pattern frequencies $E_j = \{\#i : U_i = j\}$ for $j = 0, \ldots, J$ (with realizations e_j). For the I^{His} baseline data we have $E_0^{His} = \{\#i \in \{1, 2, \ldots, I^{His}\} : Y_i = 0\}$ is the number of rejected items, and E_1^{His} is the number of accepted items (and E_0^{His}/I^{His} is the historical rejection rate). The loglikelihood of the data is

$$L(\boldsymbol{\theta}|\boldsymbol{E} = \boldsymbol{e}) = \sum_{j=0}^{J} e_j \log[P_{\theta,\mathrm{Rej}}[U_i = j]] + \sum_{j=0}^{1} e_j^{\mathrm{His}} \log[P_\theta[Y_i = j]].$$

Here,

$$P_\theta[Y_i = 1] = \int_0^1 r f_R(r|\boldsymbol{\theta})\mathrm{d}r, \text{ and } P_\theta[Y_i = 0] = \int_0^1 (1 - r) f_R(r|\boldsymbol{\theta})\mathrm{d}r,$$

with f_R as implied by the model underlying *M1*, *M2* respectively *M3*. Further, since the I items in the sample are taken from the stream of rejected items, the likelihood is calculated conditional on this initial rejection:

$$P_{\theta,\mathrm{Rej}}[U_i = j] = P_\theta[U_i = j|\text{item initially rejected}] = \frac{\int_0^1 f_R(r|\boldsymbol{\theta}) \binom{J}{j} r^{j+1}(1 - r)^{J-j}\mathrm{d}r}{\int_0^1 f_R(r|\boldsymbol{\theta}) r \mathrm{d}r}.$$

The aim of the comparison study is to learn what can be estimated with a sufficiently large sample size I. We take sample size (that is, sampling error) out of consideration by studying what can be estimated if the size of the sample of items approaches infinity ($I \to \infty$). Also, we fix the ratio of I and the number I^{His} of baseline data at $I/I^{His} = P_{\mathrm{Sc}}[Y_i = 0]$. We believe that often, many more baseline data are available, depending on how long the test system has been in use and whether results are logged. But a lower bound is obtained when the baseline data are collected as a by-product of obtaining the sample for the MSA study. Namely, in order to obtain I rejected items, the expected total number of items to be inspected is $I/P_{\mathrm{Sc}}[Y_i = 0]$, and these can be used as baseline data. Dividing the loglikelihood by I and substituting this ratio we obtain

$$\frac{L(\boldsymbol{\theta}|\boldsymbol{E} = \boldsymbol{e})}{I} = \sum_{j=0}^{J} \frac{e_j}{I} \log[P_{\theta,\mathrm{Rej}}[U_i = j]] + \frac{1}{P_{\mathrm{Sc}}[Y_i = 0]} \sum_{j=0}^{1} \frac{e_j^{\mathrm{His}}}{I^{\mathrm{His}}} \log[P_\theta[Y_i = j]].$$

If $I \to \infty$, the response pattern frequencies converge (with probability 1) to their expected values:

$$E_j/I \xrightarrow{a.s.} P_{\mathrm{Sc}}[U_i = j],$$

and also $I^{His} \to \infty$ and $E_j^{His}/I^{His} \xrightarrow{a.s.} P_{\mathrm{Sc}}[Y_i = j]$. The maximum of $L(\boldsymbol{\theta}|\boldsymbol{E} = \boldsymbol{e})$ is in the same location as that of $L(\boldsymbol{\theta}|\boldsymbol{E} = \boldsymbol{e})/I$. Therefore, for large sample sizes ($I \to \infty$), the parameter values found from maximizing $L(\boldsymbol{\theta}|\boldsymbol{E} = \boldsymbol{e})$ converge (with probability 1) to

$$\arg\max_{\boldsymbol{\theta}} \sum_{j=0}^{J} P_{\mathrm{Sc}}[U_i = j] \times \log[P_{\theta,\mathrm{Rej}}[U_i = j]] + \frac{1}{P_{\mathrm{Sc}}[Y_i = 0]} \sum_{j=0}^{1} P_{\mathrm{Sc}}[Y_i = j] \times \log[P_\theta[Y_i = j]], \quad (A.1)$$

which is the function that we maximize to find parameter estimates. To summarize, we leave the question of appropriate sample sizes out of consideration by studying the goodness of fit of methods *M1*, *M2* and *M3* when the sample size I becomes large, in which case the maximum likelihood estimators converge to the solutions from maximizing (A.1). These estimated model parameters $\hat{\boldsymbol{\theta}}$ yield a fitted density $\hat{f}_R$, from which estimates for *IAP*, *IRP*, and possibly *FAP* and *FRP* are derived.

References

[1] T. Suzuki, Y. Tsutsumi, H. Kawamura, Viewpoints to characterize precision evaluation methods in binary measurements, Measurement 46 (2013) 3710–3714.
[2] ISO, Guide to the Expression of Uncertainty in Measurement, first ed., International Organization for Standardization, Geneva, Switzerland, 1995.
[3] JCGM, International Vocabulary for Metrology – Basic and General Concepts and Associated Terms ISO, Geneva, Switzerland, 2008.
[4] J. De Mast, T. Akkerhuis, T.P. Erdmann, The statistical evaluation of categorical measurements: simple scales, but treacherous complexity underneath, Qual. Eng. 26 (2014) 16–32.
[5] AIAG, Measurement System Analysis: Reference Manual, third ed., Automotive Industry Action Group, Detroit, 2003.
[6] M.S. Pepe, The Statistical Evaluation of Medical Tests for Classification and Prediction, Oxford University Press, Oxford, UK, 2003.
[7] O. Danila, S.H. Steiner, R.J. MacKay, Assessing a binary measurement system, J. Qual. Technol. 40 (2008) 310–318.
[8] O. Danila, S.H. Steiner, R.J. MacKay, Assessing a binary measurement system with varying misclassification rates when a gold standard is available, Technometrics 55 (2008) 335–345.
[9] E. Bashkansky, T. Gadrich, Some statistical aspects of binary measuring systems, Measurement 46 (2013) 1922–1927.
[10] S.L. Hui, S.D. Walter, Estimating the error rates of diagnostic tests, Biometrics 36 (2008) 167–171.
[11] R.A. Boyles, Gauge capability for pass-fail inspection, Technometrics 43 (2001) 223–229.
[12] W. Van Wieringen, J. De Mast, Measurement system analysis for binary data, Technometrics 50 (2008) 468–478.
[13] O. Danila, S.H. Steiner, R.J. MacKay, Assessment of a binary measurement system in current use, J. Qual. Technol. 42 (2010) 152–164.
[14] P.M. Vacek, The effect of conditional dependence on the evaluation of diagnostic tests, Biometrics 41 (1985) 959–968.
[15] V.L. Torrance-Rynard, S.D. Walter, Effects of dependent errors in the assessment of diagnostic test performance, Stat. Med. 16 (1987) 2157–2175.
[16] L.M. Irwig, P.M.M. Bossuyt, P.P. Glasziou, C. Gatsonis, J.G. Lijmer, Designing studies to ensure that estimates of test accuracy are transferable, Br. Med. J. 324 (2002) 669–671.
[17] J. De Mast, T.P. Erdmann, W.N. Van Wieringen, Measurement system analysis for binary inspection: continuous versus dichotomous measurands, J. Qual. Technol. 43 (2011) 99–112.
[18] O. Danila, S.H. Steiner, R.J. MacKay, Assessing a binary measurement system with varying misclassification rates using a latent class random effects model, J. Qual. Technol. 44 (2012) 179–191.
[19] T.P. Erdmann, T.S. Akkerhuis, J. De Mast, The statistical evaluation of a binary test based on combined samples, J. Qual. Technol. 48 (2016) 54–67.
[20] Y. Qu, M. Tan, M.H. Kutner, Random effects models in latent class analysis for evaluating accuracy of diagnostic tests, Biometrics 52 (1996) 797–810.
[21] P.S. Albert, L.M. McShane, J.H. Shih, Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors, Biometrics 57 (2001) 610–619.
[22] P.S. Albert, L.E. Dodd, A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard, Biometrics 60 (2004) 427–435.
[23] T.S. Akkerhuis, Measurement System Analysis for Binary Tests PhD thesis, University of Amsterdam, 2016.
[24] T. Gadrich, E. Bashkansky, A Bayesian approach to evaluating uncertainty of inaccurate categorical measurements, Measurement 91 (2016) 186–193.
[25] Wolfram Research Inc., Mathematica, available at <wolfram.com>.
[26] D.P. Beavers, J.D. Stamey, B. Nebiyou Bekele, A Bayesian model to assess a binary measurement system when no gold standard system is available, J. Qual. Technol. 43 (2011) 16–27.