Quality Engineering

ISSN: 0898-2112 (Print) 1532-4222 (Online) Journal homepage: http://www.tandfonline.com/loi/lqen20

Assessing binary measurement systems and inspection protocols utilizing follow-up data

Stefan H. Steiner, Yi Lu & R. Jock Mackay

To cite this article: Stefan H. Steiner, Yi Lu & R. Jock Mackay (2016) Assessing binary measurement systems and inspection protocols utilizing follow-up data, Quality Engineering, 28:3, 329-336, DOI: 10.1080/08982112.2015.1086002

To link to this article: http://dx.doi.org/10.1080/08982112.2015.1086002

QUALITY ENGINEERING
2016, VOL. 28, NO. 3, 329–336
http://dx.doi.org/10.1080/08982112.2015.1086002

Taylor & Francis
Taylor & Francis Group

Assessing binary measurement systems and inspection protocols utilizing follow-up data

Stefan H. Steiner, Yi Lu, and R. Jock Mackay

Department of Statistics and Actuarial Science, Business and Industrial Statistics Research Group, University of Waterloo, Waterloo, Ontario, Canada

ABSTRACT
Industry uses inspection protocols to protect customers from receiving non-conforming product. The two error rates of these systems are the chance of shipping non-conforming product (customer's risk) and the chance of rejecting good product (producer's risk). We investigate the properties of two inspection protocols. In these protocols, the customer uses a gold standard measurement system that determines if received components are conforming. We show that with the first inspection protocol, we can estimate its error rates using only production data. With the second protocol, we propose adding a small measurement assessment study to allow estimation of the error rates.

KEYWORDS
measurement system analysis; binary measurement systems; inspection protocols; error rates; available data

Introduction

Binary measurement systems (BMS) are commonly used as diagnostic tools in medicine and as part of inspection systems in industry. Sometimes diagnostic tests are combined into protocols. For example, an invasive test such as a biopsy is carried out only on those patients with a positive result on a non-invasive screening test. There is an enormous effort (see Pepe 2003 for an overview) devoted to the study design and the subsequent estimation and comparison of the sensitivity and specificity of diagnostic protocols. Sensitivity is the chance that the test is positive when the subject has the disease. Specificity is the chance that the test is negative given the subject does not have the disease. The error rates of interest here are the complements of the sensitivity and specificity. One distinguishing feature of the assessment of a measurement system in an industrial context is that it is possible to measure the same unit many times. In a medical context, we may not be able to test a subject repeatedly because of ethical or compliance considerations.

In industry, many quality systems require periodic calibration and assessment of measurement systems for continuous characteristics that are important to the customer. The precision of the system is often estimated using a Gauge R&R study that involves repeated measurements of a sample of units. See the AIAG manual (2003) or Burdick, Borror, and Montgomery (2005). For the estimation of the properties of a BMS, many authors have considered various assessment plans, statistical models and estimation procedures that involve measuring some units $r \geq 1$ times. See Danila, Steiner, and MacKay (2008, 2010, 2012, 2013); de Mast, Erdmann, and Van Wieringen (2011), Burke et al. (1995), and Farnum (1994).

This article is motivated by the following example from the electronics industry. At an intermediate process step, a binary measurement system is used to screen a particular component. A component that passes inspection at the intermediate step moves to the next stage of the process. Components that fail inspection are re-measured with the BMS and, if they pass on the second attempt, they also move forward. Any component that fails both inspections is either scrapped or reworked. We call this Protocol A or Double Fail. The reason for the protocol is that the subsequent assembly process steps are expensive and, once completed, cannot be undone without destroying the entire device. After the subsequent process steps, each shipped component is inspected by an error-free gold standard measurement system (GSS) that determines among other tests, if the component conforms or not. As we will

CONTACT R. Jock Mackay rjmackay@uwaterloo.ca Department of Statistics and Actuarial Science, Business and Industrial Statistics Research Group, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

show, we can estimate the properties of the BMS and the inspection protocol without requiring additional measurements. Note, unlike the protocol described in Danila et al. (2010), the GSS is applied only to shipped components.

In a production environment, it is easy to understand why the operators repeat the inspection of a failed component. They selectively trust the BMS. As well, they do not have to deal with passed components. On the other hand, they would not consider re-inspecting a passed component. We have seen Protocol A (and its extension to even further inspections after multiple failures) applied in both the electronics and automotive sectors. In what follows, we also consider the simpler Protocol B (Single Fail) where a component that fails the first inspection is scrapped or reworked and only components that pass initial inspection are shipped to the customer. We provide a flow diagram of both protocols in Figure 1. For Protocol B we also show, inside a dashed box, the additional measurements we propose to allow assessment of the BMS and the inspection protocol. The reason for the additional measurements on the failed components in Protocol B is explained later.

To specify the assessment problem, we introduce some notation. We denote each component as conforming or not by the random variable X where

$$X = \begin{cases} 1 & \text{if the component is conforming} \\ 0 & \text{if the component is non - conforming} \end{cases}.$$

We can determine the value of X using the gold standard system once the component is part of a completed assembly. When the component is measured once by the BMS, we use the random variable Y to indicate the result, where

$$Y = \begin{cases} 1 \text{ if the BMS passes the component} \\ 0 \text{ if the BMS fails the component} \end{cases}.$$

We model the characteristics of the binary measurement system and conforming rate of the process by

$$\alpha = P(Y = 1|X = 0), \quad \beta = P(Y = 0|X = 1), \quad \pi = P(X = 1).$$

Here, α represents the proportion of non-conforming components that are passed by the BMS and β represents the proportion of conforming components that are failed by the BMS. We can also interpret α as the long run proportion of times that a single non-conforming component passes repeated inspection by the BMS (and similarly for β). The parameter π is the proportion of conforming components produced by the production process and does not depend on the properties of the BMS or the inspection protocol. In a manufacturing context, we expect π to be large and α and β to be relatively small. We focus on these conditions throughout the article.

Table 1. Error rates for inspection protocols a (Double Fail) and B (Single Fail) as shown in Figure 1.

Protocol	θ_0	θ_1
A: Double Fail	$\frac{(1-(1-\alpha)^2)(1-\pi)}{1-((1-\alpha)^2(1-\pi)+\beta^2\pi)}$	$\frac{\beta^2\pi}{\beta^2\pi+(1-\alpha)^2(1-\pi)}$
B: Single Fail	$\frac{\alpha(1-\pi)}{\alpha(1-\pi)+(1-\beta)\pi}$	$\frac{\beta\pi}{\beta\pi+(1-\alpha)(1-\pi)}$

We characterize any inspection protocol by its error rates. That is:

$$\theta_0 = P(X = 0|\text{passed by the protocol}) \text{ and}$$
$$\theta_1 = P(X = 1|\text{failed by the protocol}). \quad (1)$$

Note that θ_0 and θ_1 are sometimes referred to as the consumer's risk and producer's risk, respectively. These error rates are of direct interest to the process managers. In the context of our example, if θ_0 is large, there is a high cost when the GSS detects a non-conforming component after assembly. If θ_1 is large, good components may be scrapped or operators may waste time searching for faults that do not exist. With good estimates of θ_0 and θ_1, managers can assess the costs associated with the protocol. If these costs are substantial, it may prove useful to improve the BMS (i.e., reduce α and or β) or to change the overall protocol.

In Table 1, we give the error rates (1) in terms of α, β, and π for both Protocols A and B, making the following assumptions.

- The BMS is non-destructive so that components can be repeatedly measured without changing their conforming status.
- The subsequent process steps do not change the conforming status of the component.
- The characteristics α and β of the BMS are the same for every non-conforming and conforming component, respectively.
- Given the conforming status of any component, repeated measurements by the BMS are (conditionally) independent.
- Measurements on different components are independent.

The main goal of this article is to design assessment plans to efficiently estimate the error rates of the inspection protocols A and B as shown in Figure 1. We choose assessment plans that first estimate α, β, and π. Then

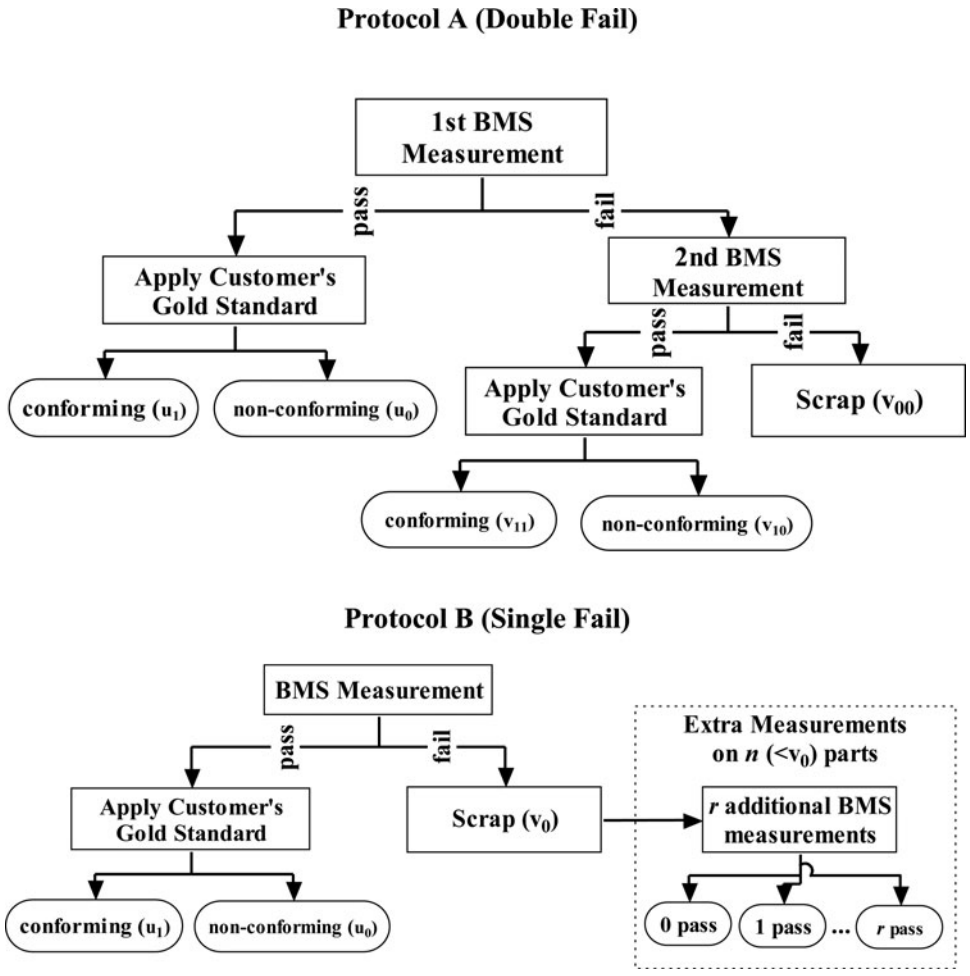

Figure 1. Flow charts for inspection Protocols A (top) and B (bottom).

we derive the estimates of the error rates θ_0 and θ_1 with corresponding measures of their precision.

This article is structured as follows. In the next section, we investigate the properties of Protocol A and assess its efficiency for estimating the error rates θ_0 and θ_1. We then repeat this exercise for Protocol B in the subsequent section. We end with a summary and discussion in the final section.

Protocol A (Double Fail)

With Protocol A, initial failures are re-inspected and, if they pass the second inspection, are shipped to the customer. Based on the motivating example, we assume that components are traceable, i.e., that components shipped after the second inspection are distinguishable from those that pass on the first inspection. With this assumption, we see from Figure 1 that there are five possible outcomes for any component. Of the m components inspected, let u_0 be the number of components that pass initially and are nonconforming, u_1 the number that pass initially and are conforming, v_{10} the number that pass the second inspection and are nonconforming, v_{11} the number that pass the second inspection and are conforming and finally, v_{00} the number that fail both inspections. The available data are $(u_o, u_1, v_{00}, v_{10}, v_{11})$ with $u_o + u_1 + v_{00} + v_{10} + v_{11} = m$. Based on the assumptions, we have a multinomial distribution with five possible outcomes for each part and three unknown parameters. Ignoring additive constants, the corresponding log-likelihood is

$$\begin{aligned} l_a(\alpha, \beta, \pi) = {} & u_0 \log[\alpha(1-\pi)] + u_1 \log[(1-\beta)\pi] \\ & + v_{10} \log[\alpha(1-\alpha)(1-\pi)] \\ & + v_{11} \log[\beta(1-\beta)\pi] + v_{00} \log[\beta^2\pi \\ & + (1-\alpha)^2(1-\pi)]. \end{aligned} \tag{2}$$

To estimate α, β and π, we numerically maximize the log-likelihood, given in (2), using the fmincon

Table 2. Results from Protocol A (Double Fail) for one day's production.

Outcome	1st Pass & Nonconforming (u_0)	1st Pass & Conforming (u_1)	Fail, Pass & Nonconforming (v_{10})	Fail, Pass & Conforming (v_{11})	Fail Twice (v_{00})
Observed Frequency	23	1892	26	256	253

function in Matlab (2008). We also derive approximate standard errors using the asymptotic properties of the log-likelihood. Using Maple (2009) we find the Fisher Information matrix, the expected value of the matrix of second partial derivatives of the log-likelihood (2). The approximate standard errors for the three parameters are then given by the diagonals of the inverse after substitution of the estimates. In the Appendix, using simulation, we demonstrate the adequacy of this approximation for the smallest sample size ($m = 1000$) we consider over a range of the model parameters (α, β, and π). We estimate θ_0 and θ_1 by substituting the estimates of the model parameters into the expressions given in Table 1. We use the delta method (Lehmann and Casella 1998) to find their approximate standard errors. Matlab code is available at http://www.bisrg.uwaterloo.ca/software/.

Numerical example

The context and process are real; the data have been constructed to be realistic. For the example described in the Introduction, one day's production was $m = 2450$ units. The resulting data are given in Table 2.

Table 3 gives the results of maximizing the log-likelihood given by (2) and applying the analysis procedure described earlier for obtaining approximate standard errors. We see that in this example all the parameters of interest are precisely estimated without any effort other than organizing the available data and applying the estimation procedure.

To improve the inspection protocol, i.e., to simultaneously reduce θ_0 and θ_1, we need to improve the BMS, i.e., reduce α and β. If Protocol B (single fail) were followed (as the control plan indicated), then the estimates of θ_0 and θ_1 are 0.0133 and 0.5559 with standard errors 0.0041 and 0.1525 respectively, based on the estimates of α and β given in Table 3. These estimates suggest that by following Protocol B, more than half the rejected components would in fact be conforming. This may explain the behaviour of the operators who adopted Protocol A contrary to the control plan.

Table 3. Parameter estimates and standard errors for Protocol A (Double Fail) data given in Table 2.

Parameter	Estimate	Standard Error
α	0.0978	0.0137
β	0.1352	0.0090
π	0.8931	0.0071
θ_0	0.0222	0.0031
θ_1	0.1580	0.0239

Note that when comparing Protocols A (Double Fail) and B (Single Fail), there is always a tradeoff. For fixed values of α, β and π, θ_0 is larger for Protocol A and θ_1 is larger for Protocol B. Over the realistic range of parameter values for α, β and π considered in this article, θ_0 is generally twice as large and θ_1 is about 4 times smaller with Protocol A compared to Protocol B.

With Protocol A, the design of the assessment study is determined by m, the total number of components screened by the inspection system over the assessment period. Since we assume measurements on all components are independent, the Fisher information matrix is m times the Fisher information available from a single component. As a result, the standard errors of the estimates of θ_0 and θ_1 (and of α, β, and π) are proportional to $1/\sqrt{m}$ and so increasing m increases the precision of the estimates in a predictable way. We suggest choosing m as large as possible, noting that we are assuming no changes in the properties of the BMS (α, β) nor in the quality of the process (π) while the data are being collected.

To help in the choice of m, we show in Table 4 the proportionality constants for the approximate standard deviations for the parameters of interest for all combinations of the parameters given in Table 5. We can use the results provided in Table 4 to determine the approximate standard deviation for any sample size m by dividing the values in the table by $\sqrt{m}$. We show in the Appendix that the asymptotic approximations are reasonable with sample sizes as small as $m = 1000$. So, for example, from the first row of Table 5, if $(\alpha, \beta, \pi) = (0.01, 0.01, 0.9)$ and $m = 1000$, the standard errors for estimating θ_0 and θ_1 are $0.049/\sqrt{1000} = 0.0015$ and $0.02/\sqrt{1000} = 0.0006$ respectively. Since these standard errors are large relative to the corresponding values $\theta_0 = 0.002$ and $\theta_1 = 0.001$ (given in columns four and five of Table 4), we need to increase m accordingly. For combinations of parameter values not given in Table 5, we interpolate.

Table 4. Protocol a (Double Fail) asymptotic standard deviation for $m = 1$.

Parameter Values			Error Rates		Asymptotic Standard Deviation $m = 1$				
α	β	π	θ_0	θ_1	$SD(\alpha)$	$SD(\beta)$	$SD(\pi)$	$SD(\theta_0)$	$SD(\theta_1)$
0.01	0.01	0.9	0.002	0.001	0.223	0.106	0.300	0.049	0.020
0.01	0.01	0.95	0.001	0.002	0.315	0.104	0.218	0.033	0.041
0.01	0.01	0.99	0.000	0.010	0.705	0.102	0.100	0.014	0.227
0.01	0.05	0.9	0.002	0.022	0.223	0.248	0.305	0.050	0.235
0.01	0.05	0.95	0.001	0.046	0.316	0.241	0.225	0.033	0.494
0.01	0.05	0.99	0.000	0.202	0.707	0.236	0.114	0.014	2.646
0.01	0.1	0.9	0.002	0.084	0.223	0.369	0.322	0.050	0.683
0.01	0.1	0.95	0.001	0.162	0.316	0.359	0.249	0.033	1.365
0.01	0.1	0.99	0.000	0.503	0.716	0.351	0.158	0.014	5.079
0.05	0.01	0.9	0.011	0.001	0.494	0.106	0.300	0.108	0.021
0.05	0.01	0.95	0.005	0.002	0.698	0.104	0.218	0.073	0.045
0.05	0.01	0.99	0.001	0.011	1.562	0.102	0.100	0.031	0.248
0.05	0.05	0.9	0.011	0.024	0.494	0.248	0.305	0.108	0.256
0.05	0.05	0.95	0.005	0.050	0.700	0.241	0.225	0.073	0.538
0.05	0.05	0.99	0.001	0.215	1.586	0.236	0.114	0.032	2.869
0.05	0.1	0.9	0.011	0.091	0.497	0.369	0.322	0.109	0.740
0.05	0.1	0.95	0.005	0.174	0.709	0.359	0.249	0.074	1.474
0.05	0.1	0.99	0.001	0.523	1.682	0.351	0.158	0.032	5.341
0.1	0.01	0.9	0.021	0.001	0.688	0.106	0.300	0.148	0.024
0.1	0.01	0.95	0.010	0.002	0.974	0.104	0.218	0.101	0.050
0.1	0.01	0.99	0.002	0.012	2.179	0.102	0.100	0.044	0.279
0.1	0.05	0.9	0.021	0.027	0.690	0.248	0.305	0.149	0.286
0.1	0.05	0.95	0.010	0.055	0.980	0.241	0.225	0.101	0.601
0.1	0.05	0.99	0.002	0.234	2.252	0.236	0.114	0.044	3.185
0.1	0.1	0.9	0.021	0.100	0.699	0.369	0.322	0.150	0.822
0.1	0.1	0.95	0.010	0.190	1.005	0.359	0.249	0.102	1.629
0.1	0.1	0.99	0.002	0.550	2.528	0.351	0.158	0.044	5.684

Protocol B (Single Fail)

Now suppose that Protocol B, as shown in the bottom panel of Figure 1, is in use and that the result of each inspection is recorded. Of the m components inspected, we have u_0 components that pass inspection and are not conforming, u_1 components that pass inspection and are conforming and v_0 components that fail inspection. Note that m is under our control and is one element of the design of the assessment study. The data (u_0, u_1, v_o) with $u_0 + u_1 + v_o = m$ are available from the process under normal usage. For Protocol B, the log-likelihood for the available data is given by [3]:

$$l_B(\alpha, \beta, \pi) = u_1 \log[(1-\beta)\pi] + u_0 \log[\alpha(1-\pi)] + v_0 \log[\beta\pi + (1-\alpha)(1-\pi)] \quad (3)$$

Here we have a three cell multinomial model and three unknown parameters α, β, and π to estimate. Since the multinomial probabilities add to 1, the three parameters are not identifiable using the available data.

Table 5. Levels of parameters.

Parameter	α	β	π
Levels	0.01, 0.05, 0.10	0.01, 0.05, 0.10	0.90, 0.95, 0.99

To allow estimation of α, β, and π, we supplement the available data by re-measuring components (with the BMS) that fail the initial inspection. That is, we select a random sample of n components from the v_0 failures (with $n \le v_0$) and re-measure each of these components $r \ge 1$ times. For each component sampled from the initial failures, using the BMS we observe $t = 0, 1, \ldots, r$, the number of failures in r additional measurements. We assume conditional independence. That is, given the true status of any component (conforming or non-conforming), repeated measurements of the component by the BMS are independent. So, for any failed component re-measured r times we have

$$P(T = t) = \frac{\binom{r}{t}\left[\beta^{t+1}(1-\beta)^{r-t}\pi + (1-\alpha)^{t+1}\alpha^{r-t}(1-\pi)\right]}{\beta\pi + (1-\alpha)(1-\pi)}, \quad t = 0, \ldots, r$$

and the log-likelihood for the supplemental data, ignoring additive constants, is

$$l_S(\alpha, \beta, \pi) = \sum_{i=1}^{n} \log\left(\frac{\beta^{t_i+1}(1-\beta)^{r-t_i}\pi + (1-\alpha)^{t_i+1}\alpha^{r-t_i}(1-\pi)}{\beta\pi + (1-\alpha)(1-\pi)}\right), \quad (4)$$

where t_i is the observed number of failures in the r additional measurements on the i$^{\text{th}}$ $i = 1, \ldots, n$ component selected from the initial failures. Combining the available and supplementary data the overall log-likelihood is the sum of l_B and l_S as given by (3) and (4). Now there are $r + 3$, $r \ge 1$ possible outcomes and the three parameters are estimable. As with Protocol A, to estimate α, β and π, we numerically maximize the overall log-likelihood using the fmincon function in Matlab (2008). We also derive approximate standard errors using the asymptotic properties of the log-likelihood based on the Fisher information. For Protocol B, Matlab code is available at http://www.bisrg.uwaterloo.ca/software/

In planning the assessment study, we have more choices to make with Protocol B (Single Fail) than with Protocol A (Double Fail). The study is determined by m, n, and r. In what follows, we focus on plans that have $r = 1$. Increasing r results in more precise estimates. However, for estimating θ_0, increasing r has very little impact while for estimating θ_1 it is more effective to increase n than r. In addition using r larger than 1 results in additional logistical work since we have to keep track of which components pass/fail each of the

Table 6. Protocol B (Single Fail) asymptotic standard deviation $m = 1$.

					"Asymptotic" StDev with $m = 1$, $r = 1$ and n/m = 0.01,0.02,0.05				
α	β	π	θ_0	θ_1	$SD(\alpha)$	$SD(\beta)$	$SD(\pi)$	$SD(\theta_0)$	$SD(\theta_1)$
0.01	0.01	0.9	0.001	0.08	0.31	0.35, 0.25, 0.16	0.43, 0.37, 0.32	0.04	2.96, 2.10, 1.35
0.01	0.01	0.95	0.001	0.16	0.44	0.24, 0.17, 0.12	0.30, 0.26, 0.22	0.02	3.83, 2.72, 1.75
0.01	0.01	0.99	0.000	0.50	0.99	0.12, 0.10, 0.09	0.12, 0.10, 0.08	0.01	5.13, 3.64, 2.34
0.01	0.05	0.9	0.001	0.31	0.32	0.79, 0.57, 0.38	0.77, 0.57, 0.41	0.04	5.10, 3.61, 2.29
0.01	0.05	0.95	0.001	0.49	0.45	0.56, 0.42, 0.30	0.56, 0.41, 0.28	0.02	5.54, 3.92, 2.49
0.01	0.05	0.99	0.000	0.83	1.02	0.33, 0.28, 0.24	0.27, 0.19, 0.12	0.01	4.49, 3.18, 2.02
0.01	0.1	0.9	0.001	0.48	0.33	1.19, 0.86, 0.58	1.17, 0.84, 0.55	0.04	6.12, 4.33, 2.75
0.01	0.1	0.95	0.001	0.66	0.47	0.88, 0.65, 0.47	0.88, 0.63, 0.40	0.03	6.07, 4.30, 2.73
0.01	0.1	0.99	0.000	0.91	1.11	0.57, 0.46, 0.38	0.52, 0.37, 0.23	0.01	4.77, 3.38, 2.15
0.05	0.01	0.9	0.006	0.09	0.68	0.40, 0.29, 0.20	0.46, 0.39, 0.34	0.08	3.51, 2.53, 1.67
0.05	0.01	0.95	0.003	0.17	0.95	0.26, 0.19, 0.13	0.31, 0.27, 0.23	0.05	4.22, 3.04, 2.03
0.05	0.01	0.99	0.001	0.51	2.11	0.13, 0.11, 0.09	0.13, 0.10, 0.09	0.02	5.22, 3.77, 2.53
0.05	0.05	0.9	0.006	0.32	0.74	0.80, 0.58, 0.39	0.78, 0.58, 0.42	0.08	5.29, 3.76, 2.41
0.05	0.05	0.95	0.003	0.50	1.05	0.56, 0.42, 0.30	0.55, 0.41, 0.29	0.06	5.58, 3.97, 2.55
0.05	0.05	0.99	0.001	0.84	2.40	0.33, 0.28, 0.24	0.26, 0.19, 0.13	0.02	4.42, 3.14, 2.02
0.05	0.1	0.9	0.006	0.49	0.86	1.18, 0.86, 0.58	1.16, 0.84, 0.55	0.09	6.19, 4.39, 2.80
0.05	0.1	0.95	0.003	0.67	1.24	0.87, 0.65, 0.47	0.87, 0.62, 0.41	0.06	6.04, 4.28, 2.73
0.05	0.1	0.99	0.001	0.91	3.24	0.57, 0.46, 0.38	0.51, 0.36, 0.23	0.03	4.72, 3.35, 2.14
0.1	0.01	0.9	0.011	0.09	0.94	0.45, 0.33, 0.22	0.50, 0.41, 0.35	0.11	4.07, 2.94, 1.98
0.1	0.01	0.95	0.005	0.17	1.30	0.27, 0.20, 0.14	0.32, 0.27, 0.24	0.07	4.63, 3.37, 2.31
0.1	0.01	0.99	0.001	0.52	2.82	0.13, 0.11, 0.09	0.13, 0.11, 0.09	0.03	5.31, 3.89, 2.70
0.1	0.05	0.9	0.012	0.33	1.11	0.80, 0.58, 0.40	0.78, 0.59, 0.42	0.11	5.49, 3.92, 2.54
0.1	0.05	0.95	0.006	0.51	1.56	0.55, 0.42, 0.30	0.55, 0.41, 0.29	0.08	5.64, 4.02, 2.61
0.1	0.05	0.99	0.001	0.85	3.60	0.33, 0.28, 0.24	0.26, 0.19, 0.13	0.03	4.35, 3.10, 2.01
0.1	0.1	0.9	0.012	0.50	1.39	1.16, 0.85, 0.59	1.15, 0.83, 0.56	0.12	6.28, 4.46, 2.86
0.1	0.1	0.95	0.006	0.68	2.03	0.85, 0.64, 0.47	0.85, 0.61, 0.41	0.08	6.02, 4.28, 2.74
0.1	0.1	0.99	0.001	0.92	5.63	0.56, 0.46, 0.39	0.51, 0.36, 0.23	0.04	4.68, 3.33, 2.13

additional measurements so that we can determine the t_i s. Using $r = 1$, on the other hand, we only need to count the number of failures in the n repeated measurements.

Regardless of the value of r, for a fixed ratio, n/m, the asymptotic information is a multiple of m, the total number of components screened by the inspection system. Thus, given the ratio n/m the asymptotic standard deviation decreases at the rate $1\sqrt{m}$. Similar to Table 4 for Protocol A, Table 6 gives the approximate standard deviations times $m = 1$ for the three parameter estimates and the two derived estimated error rates $\hat{\theta}_0$ and $\hat{\theta}_1$ for different combinations of (α, β, π) for Protocol B. We again use the parameter values given in Table 5. The standard deviations for $\hat{\alpha}$ and $\hat{\theta}_0$ depend weakly on the ratio n/m so we give only a single value in those two columns. We can use these results as with Table 4. For any reasonable m, we can get the expected standard deviation of the estimators from Protocol B by dividing the results in Table 6 by $\sqrt{m}$. So for instance, suppose we believe $(\alpha, \beta, \pi) = (0.1, 0.05, 0.9)$ approximately and we select $m = 1000$ and $n = 20$ (i.e. $n/m = 0.02$). Then we expect standard deviations for $\hat{\theta}_0$ and $\hat{\theta}_1$ to be $0.11/\sqrt{1000} = 0.0034$ and $3.92/\sqrt{1000} = 0.012$, respectively. The standard errors are small relative to $\theta_0 = 0.012$ and $\theta_1 = 0.33$ as given in Table 6.

From Table 6, we make the following observations. For $\hat{\theta}_0$, increasing m reduces the *relative* standard deviation $SE(\hat{\theta}_0)/\theta_0$. The large relative standard deviations correspond to low values of α and high values of π. For $\hat{\theta}_1$, increasing n with m fixed substantially reduces the standard deviation.

Summary and discussion

In this article, we investigate two inspection protocols that are common in industry. Both protocols are used to decide which components to ship to the customer and which to scrap or rework. We examine a situation where shipped components are measured with a gold standard measurement to determine if they are conforming or not. Protocol A ships a component to the customer if it passes the BMS the first time or after being given a second chance. Protocol B is simpler and only ships components to the customer if they pass the first measurement.

Note that in this context it is not possible to measure each component with the gold standard since only components that go through the whole assembly process can be tested with the gold standard. With gold standard measurements on all components the assessment of the protocol is straightforward.

We can supplement Protocol A (Double Fail) with additional measurements. However, as long as m is reasonably large, the added information from the additional measurements is relatively small. We also explored a version of Protocol A where we did not assume that conforming and nonconforming components found by the customer could be linked to the pool of first or second time passes. We found that the tracing information was valuable; without it, the plan did not perform well.

We made the simplifying assumption that the misclassification rates of the BMS are the same for every component. For some BMSs, this assumption is unrealistic since some components are easier to correctly classify than others. We are concerned that the estimates from the assumed model may not be robust to varying misclassification rates. This was shown to be a problem when assessing a BMS without any gold standard measurements. See Albert and Dodd (2004) and Danila et al. (2012). However, by assumption, with Protocols A and B a large proportion of the components shipped to the customer is measured by the gold standard. Albert (2007), Albert and Dodd (2008) and Danila et al. (2013) found that when the true status of *every* component is known, the maximum likelihood estimates based on a model with constant misclassification rates are robust against the random effects model. However, in a small simulation study, using Beta distributions to model the varying misclassification rates, we found for both protocols that the proposed estimate of θ_1 was significantly biased. There does not appear to be a simple remedy. Using a random effects model, where the misclassification rates vary from component to component, requires a much more complex assessment study for either protocol and is left to further work.

We select m to be as large as possible, assuming the process and inspection protocol are within a stable period. We assume that the process has high volume to make this possible. For Protocol B (Single Fail), we also need m to be large to produce a sufficient number of failures so that we can select n components to be repeatedly measured.

References

Albert, P. S. 2007. Imputation approaches for estimating diagnostic accuracy for multiple tests from partially verified designs. *Biometrics* 63:947–957.

Albert, P. S., and L. E. Dodd. 2004. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 60:427–435.

Albert, P. S., and L. E. Dodd. 2008. On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *Journal of American Statistical Association* 103:61–73.

Automotive Industry Action Group (AIAG). 2003. *Measurement systems analysis*. 3rd ed. Southfield, Michigan: Automotive Industry Action Group.

Burdick, R. K., C. M. Borror, and D. C. Montgomery. 2005. *Design and analysis of gauge r&r studies: making decisions with confidence intervals in random and mixed effects models*. Philadelphia, PA: American Statistical Association - Society for Industrial and Applied Mathematics. Series on Statistics and Applied Probability.

Burke, J. R., R. D. Davis, F. C. Kaminsky, and A. E. P. Roberts. 1995. The effect of inspector errors on the true fraction nonconforming: an industrial experiment. *Quality Engineering* 7 (4):543–550.

Danila, O., S. H. Steiner, and R. J. MacKay. 2008. Assessing a binary measurement System. *Journal of Quality Technology* 40:310–318.

Danila, O., S. H. Steiner, and R. J. MacKay. 2010. Assessment of a binary measurement system in current use. *Journal of Quality Technology* 42:152–164.

Danila, O., S. H. Steiner, and R. J. MacKay. 2012. Assessing a binary measurement system with varying misclassification rates using a latent class random effects model. *Journal of Quality Technology* 44:179–191.

Danila, O., S. H. Steiner, and MacKay R. J. 2013. Assessing a binary measurement system with varying misclassification rates when a gold standard is available. *Technometrics* 55:335–345.

De Mast, J., T. P. Erdmann, and W. N. Van Wieringen. 2011. Measurement system analysis for binary inspection: Continuous versus dichotomous measurands. *Journal of Quality Technology* 43:99–112.

Farnum, N. R. 1994. *Modern statistical quality control and improvement*. Belmont, CA: Duxbury Press.

Lehmann, E. L., and G. Casella. 1998. *Theory of point estimation*. 2nd ed. New York, NY: Springer.

Maple 13. 2009. Maplesoft, Toronto: Maplesoft, a division of Waterloo Maple Inc., www.maplesoft.com.

Matlab 7.7.0. 2008. The MathWorks, Inc. Natick, Massachusetts, www.mathworks.com.

Pepe, M. S. 2003. *The statistical evaluation of medical tests for classification and prediction*. 1st ed. New York: Oxford University Press Inc.

Appendix

In this appendix, we report simulation results to justify the use of the asymptotic standard deviation based on the Fisher information for reasonable sample sizes.

For large sample sizes for both Protocols A and B, the asymptotic results derived from the Fisher information will provide good approximations of the standard deviation of the MLEs. The question is for how small a sample size are the results based on the asymptotics appropriate. To explore this question, for both Protocol A and B we ran a simulation study at the worst case, i.e., when the sample size is the smallest we would reasonably recommend.

We look at $m = 1000$ for both protocols and $n = 20$ for protocol B (Single Fail) and all combinations of $\alpha = (0.02, 0.10)$, $\beta = (0.02, 0.10)$, and $\pi = (0.90, 0.95)$. For each set of parameter values and each protocol, we conduct 50,000 simulation runs. We evaluate how well the asymptotic standard deviation of each estimate works by determining the ratio of the simulated standard deviation (i.e. the sample standard deviation of the 50,000 estimates) divided by the asymptotic standard deviation. For protocol A (Double Fail), the ratios fall in the range 0.99—1.03 for all estimates except for $\hat{\theta}_1$ where the ratio varies from 1.03—1.09. For Protocol B (Single Fail), the ratio varies between 0.98 and 1.01 for all estimates except $\hat{\alpha}$ where the ratio exceeds 1.40 when $\beta = 0.10$. We need to increase n, say $n \geq 50$ so that the asymptotic approximation for $\hat{\alpha}$ is adequate.