Assessing a Binary Measurement System: A New Plan Using Targeted Verification with Conditional Sampling and Baseline Information

Daniel E. Severn, Stefan H. Steiner, R. Jock MacKay

PII: S0263-2241(19)30585-8
DOI: https://doi.org/10.1016/j.measurement.2019.06.019
Reference: MEASUR 6736

To appear in: *Measurement*

Received Date: 1 March 2018
Revised Date: 1 May 2019
Accepted Date: 10 June 2019

Please cite this article as: D.E. Severn, S.H. Steiner, R.J. MacKay, Assessing a Binary Measurement System: A New Plan Using Targeted Verification with Conditional Sampling and Baseline Information, *Measurement* (2019), doi: https://doi.org/10.1016/j.measurement.2019.06.019

Assessing a Binary Measurement System: A New Plan Using Targeted Verification with Conditional Sampling and Baseline Information

Daniel E. Severn, Stefan H. Steiner and R. Jock MacKay
Department of Statistics and Actuarial Science
University of Waterloo,
Waterloo, Ontario N2L 3G1 Canada

Abstract

We investigate efficient plans to assess the misclassification error rates of a binary measurement system used as an in-line inspection protocol. We assume that parts can be inspected repeatedly and that each part has its own (latent) misclassification rate. We propose a three-phase assessment plan. Phase I consists of data from recent inspection history. In Phase II, we select a sample of failed parts that we re-measure multiple times with the binary measurement system of interest. In Phase III, we verify a carefully selected subsample of the parts from Phase II with the aid of a binary gold standard measurement system. We show that the proposed plan is a substantial improvement over existing assessment plans in terms of cost and/or precision.

Key words: baseline data, binary measurement system assessment, conditional sampling, targeted verification

1. Introduction

Binary measurement systems (BMS) with two possible outcomes (pass or fail) are important components of inspection schemes designed to protect the customer by determining whether or not measured units, here called parts, meet specifications (conforming or non-conforming). In this article, we consider the situation where we assume both the underlying measurand and the realized quantity are binary. This differs from the commonly considered situation [1], [2], [3] where the measurand is continuous and compared to tolerance limits. We also assume that we have access to a gold standard measurement system, but that gold standard measurements are either expensive, destructive or both and so the gold standard is not used for regular production. There are many examples involving 100% final inspection where assuming a binary measurand makes sense. For instance, in the production of crankshafts, at the final inspection, we measure 50+ characteristics, including journal diameters, crank balance, various run-outs, etc. and each crankshaft either passes or fails the final inspection. In this case, the final decision on the conformance of each individual crankshaft is based on a large number of individual continuous characteristics and it is untenable to model the relationship between all the characteristics and the final binary assessment. Instead we proceed by considering a binary measurand. In the crankshaft example the gold standard measurement system is a coordinate measurement

machine. There are many other similar example situations. In Section 2.3, we consider the inspection of an electronic device. Here, as in the crankshaft example, each device passes the final inspection only if it passes a large number of individual performance tests. In this example, a gold standard measurement involves an engineer checking each functional test for a device one by one. Another example where the methodology presented in this paper is relevant arises in the inspection of train rails. In this application, the inspection process is looking for cracks and other types of defects using an ultrasound system. Here, the gold standard measurement is destructive as it involves sectioning the rail.

A BMS makes errors; specifically, a non-conforming part may pass inspection and a conforming part may fail inspection. We can quantify the performance of the measurement system by modeling and estimating the frequency of misclassifications. If we select a part at random from the process, let $Y_m = 1$ indicate that the part passes a single inspection and $Y_m = 0$ that it fails. Also, let the measurand $Y = 1$ indicate that the part is truly conforming and $Y = 0$ that it is not. Then,

$$R_C = P(Y_m = 1 \mid Y = 0) \text{ and } R_P = P(Y_m = 0 \mid Y = 1)$$

are the probabilities of misclassification. The parameters R_C and R_P are respectively the customer's and producer's risk. Note that in assessing a diagnostic test in a medical context with diseased replacing non-conforming, $1 - R_C$ and $1 - R_P$ are the sensitivity and specificity respectively. We assess the measurement system by estimating R_C and R_P. We also quantify the performance of the production process by estimating the probability that a randomly selected part is conforming, i.e. $P_C = P(Y = 1)$.

We start with several conditions:

- Any part can be measured repeatedly (i.e. the BMS is non-destructive) and the variability in the use of the BMS may result in different outcomes when measuring the same part.
- The measurand (*Y*) is binary, i.e. conforming/non-conforming
- The inspection system is automated so that there are no operator effects.
- Some parts are more difficult to classify correctly than others. That is, we suppose the misclassification rate is part specific. The parameters R_C and R_P are the mean misclassification rates over all the parts (also called the global consumer and producer's risks).
- The process quality P_C is close to 1 and the measurement system has high quality, i.e. R_C and R_P are small.
- There is a binary gold standard measurement system available that can verify without error whether any part is conforming or not. Gold standard measurements are expensive and/or time consuming.
- The production process has high volume and the inspection system is inline so that we have many parts segregated by passing or failing inspection. The records from recent inspections are called the baseline data and are freely available.

There are many assessment plans discussed in the literature for estimating the misclassification probabilities under some or all of these assumptions. All require a large number of parts and many measurements to get estimates of R_C and R_P (assumed small) with sufficient precision to be useful. Hence the interest is in developing more efficient assessment plans.

One way to categorize assessment plans is in terms of the use of the gold standard system. In a full verification plan, all parts in the study are measured with the gold standard. Full verification plans are widely recommended [2]. In contrast, for a no verification plan, we do not use a gold standard at all, while with a partial or targeted verification plan, the gold standard system is used to verify the status of a selected subset of the parts. Full verification requires the most effort, no-verification the least, and targeted verification is somewhere in between.

Models used for data from plans with no verification, treat Y as a latent variable for each part in the study. See, for example, Danila et al. [4], Van Wieringen and de Mast [5], and Boyles [6]. However, as shown by Akkerhuis et al. [7] and Albert and Dodd [8] among other, the estimates of R_C and R_P from the latent class model are highly sensitive to untestable underlying assumptions. Unless necessary, we cannot recommend a no-verification plan because of the lack of robustness. Boyles [6] provides an alternative plan that uses an imperfect reference measurement system in lieu of a gold standard system. Full verification plans in an industrial context have been studied by Danila et al. [9] and Burke et al. [10] and many authors in the diagnostic testing literature. See Pepe [11]. Note that, to the authors' knowledge, existing standards, such as [2], [3], do not consider the case where we assume both the measurand and the measured characteristics are binary. Instead, they assume an observable continuous measurand that can be discretized. Assuming a continuous measurand changes the problem considerably. De Mast et al. [12] consider the case of assessing a BMS under four scenarios defined by whether a gold standard measurement system is available or not and whether the underlying measurand is binary or continuous. They discuss various plans and the possibility of bias in estimation. For the gold standard available and binary measurand case we consider in this paper, they present three plans. The plan proposed by Farnum [13] that requires parts sampled at random from populations of conforming/non-conforming parts is rejected as too difficult and expensive to implement. The other two plans discussed in de Mast et al. [12] are plans from Danila et al. [4] and an extension of what is proposed in Danila et al. [14]. We compare the proposed targeted verification plan to these plans in Section 3.

There is little literature regarding targeted verification where only a sample of parts in the study are verified with the gold standard measurement system. Severn et al. [15] showed that a targeted verification plan using a random sample of parts could have performance close to that of the corresponding full verification plan with effort close to the no verification plan. Albert and Dodd [16] explored a similar concept in the medical context that they called over-sampling for assessing a diagnostic test with multiple raters. When only some of the units in the study are verified, there is discussion of bias in the estimates of R_C and R_P from naïve analysis that does not account for the selection mechanism. See Begg and Greenes [17].

We can also categorize plans by the use or not of the available baseline data. For example, a standard R&R study, see Burdick et al. [18], for the assessment of a continuous measurement system makes no use of available data from recently (once) measured parts. In this paper we have assumed that there are pass/fail data (baseline data) from the BMS freely available on a large number of parts. Danila et al. [14] showed the substantial gain in efficiency using these freely available data in the full and no-verification plan cases.

We can also use the baseline to select parts for inclusion in the assessment study. Because we are dealing with a high quality process (P_C near 1), a random sample of parts from the process will contain relatively few non-conforming parts making the precise estimation of R_C difficult. We can increase the number of non-conforming parts in the study by sampling from the population of parts that failed initial inspection. We call this approach conditional sampling. Figure 1 shows the probability of *selecting* (with conditional and random sampling) a non-conforming part as the proportion of conforming parts (p_C) increases. The concept of conditional sampling was first proposed in Haitovsky and Rapp [19] as an extension of Tenenbein [20]. Both papers focus on using the BMS to improve the estimation of P_C , the process quality.

The goal of this paper is to extend the work of Severn et al. [15] by showing the advantages of an assessment plan that uses all three ideas: targeted verification, conditional sampling and the inclusion of baseline data. The plan has three phases. First, we specify a baseline set of parts that have already been inspected. Second, we select a random sample of parts from those that initially failed the inspection and measure each selected part r times with the BMS. Finally, we verify the conforming/nonconforming status of a stratified sample of these parts where the strata are defined by the number times a part passed in Phase II.

In the next section, we fully describe the three-phase plan and associated data. We derive the likelihood and explain how to get the maximum likelihood estimates and approximate standard errors. We use an example to show the value of conditional sampling, baseline data and targeted verification. Subsequently, we recommend a specific targeted verification plan. Then, in Section 3, we demonstrate the advantages of the recommended plan compared to full or no-verification plans (as proposed by Danila et al. [4, 9]), to population sampling plans (i.e. plans based on a random sample of parts in Phase II with no baseline data, as proposed by Severn et al. [15]) and to a naïve plan without baseline data and repeated measurements as described by Pepe [11]). We end with a brief discussion in Section 4 and a conclusion in Section 5.

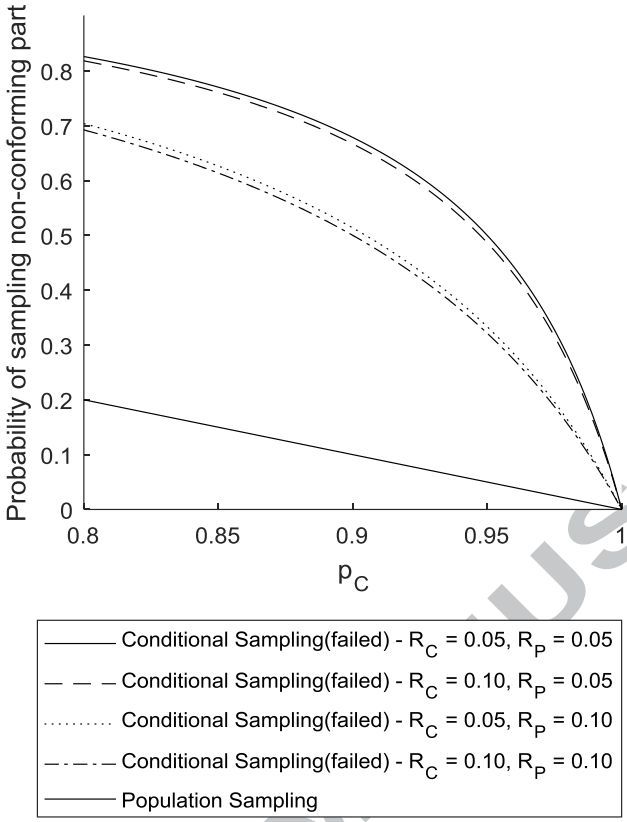

Figure 1 – Comparison of the Probability of Sampling a Nonconforming Part as a Function of the Probability a Randomly Selected Part is Conforming and the Sampling Approach

2. Proposed Three-Phase BMS Assessment Plan

This section details how to assess a BMS with our proposed three-phased plan that allows for targeted verification. It also provides an example and gives a recommended plan with some justification.

2.1 Plan Overview and Notation

In the baseline (Phase I), n_B parts are measured once with the BMS. Failed parts are set aside for future use. Let y_B denote the number of parts that passed inspection. In the repeated measurement phase (Phase II), n_{RM} parts are selected randomly from the $n_B - y_B$ parts that failed initial inspection and each are measured r additional times. These parts are then separated into bins, indexed by $s \varepsilon \{0,1,\ldots r\}$, which represents the number of times parts in the bin passed inspection. Let n_s denote the number of parts in bin s. The proposed plan is summarized in Table 1 and Section 2.4.

In the verification (Phase III), the experimenter decides how many parts to verify (i.e. measure with the gold standard system) from each bin. Let v_s denote the number of parts verified from bin s where $0 \le v_s \le n_s$. Setting v_s equal to zero indicates no parts are verified from bin s, whereas setting v_s equal to n_s indicates all parts from that bin are verified. For other choice of v_s, parts are selected at random from bin s. After determining the parts to verify, the experimenter measures those parts with the gold

standard and records u_s, the number from each bin that conform to specification. If no parts are verified in bin s, then u_s is set to zero.

We summarize the plan phases and the resulting data as in Table 1.

Table 1 – Three Phase Plan Data and Notation Summary

Phase I: Baseline ($0 \le y_B \le n_B$)

Select n_B parts (at random) from the production process and measure them each once	
Number of Parts Measured in Baseline	n_B
Number of Parts Passing Inspection in Baseline	y_B

Phase II: Repeated Measurements ($n_0 + n_1 + \ldots + n_r = n_{RM}$)

Select n_{RM} parts (at random) from the baseline rejects and measure each *r* times				
Number of Passes (bin #, *s*)	0	1	…	*r*
Number of Parts	n_0	n_1	…	n_r

Phase III: Verification ($0 \le v_s \le n_s$, $0 \le u_s \le v_s$)

Select v_s parts (at random) from bin *s* and measure them with the gold standard				
Bin #	0	1	…	*r*
Number of Parts Verified from Each Bin	v_0	v_1	…	v_r
Number of Conforming Parts among those Verified	u_0	u_1	…	u_r

2.2 Beta Binomial Model and Likelihood Derivation

Below we present the likelihoods for each of the three phases. In the baseline phase, we model the data using a binomial distribution where we give the probability of y_B passes from n_B parts each measured once. In the repeated measurement phase, we model the data using a multinomial distribution that gives the probability of obtaining n_i, $i = 0,1,\ldots,r$, parts with *i* passes from the n_{RM} parts each measured *r* times. Finally, the verification phase data are modeled using a series of binomial distributions, one for each of the $r+1$ bins. The likelihood expressions for the three phases are

$$\mathcal{L}_I(\theta) = \binom{n_B}{y_B} P(Y_m = 1)^{y_B} P(Y_m = 0)^{n_B - y_B},$$

$$\mathcal{L}_{II}(\theta) = \binom{n_{RM}}{n_0\ n_1 \ldots n_r}\left(\prod_{s=0}^{r} P(S=s \mid Y_m=0)^{n_s}\right), \quad (1)$$

$$\mathcal{L}_{III}(\theta) = \prod_{s=0}^{r}\binom{v_s}{u_s} P(Y=1 \mid S=s)^{u_s} P(Y=0 \mid S=s)^{v_s-u_s}\ .$$

where θ represents the model parameters. The overall likelihood is the product of the likelihoods for each phase.

We use a beta-binomial based model developed in Danila et al. [4] to give expressions for each factor in $\mathcal{L}_I$, $\mathcal{L}_{II}$ and $\mathcal{L}_{III}$ in terms of five model parameters. The model parameters are the attributes of interest R_C, R_P and P_C as well as two nuisance parameters, denoted by γ_C, γ_P. In the model, each part may have a different rate of misclassification if measured repeatedly. The model we propose assumes the misclassification rates for non-conforming parts are independently and identically distributed according to a Beta distribution with parameters R_C/γ_C and $(1-R_C)/\gamma_C$. Similarly, the misclassification rates for conforming parts follow a Beta distribution with parameters R_P/γ_P and $(1-R_P)/\gamma_P$. The parameters R_C and R_P represent the mean of the misclassification rates over all non-conforming and conforming parts. The parameters γ_C and γ_P control the variation of the misclassification rates. These assumptions can be used to derive expressions for $P(S=s, Y_m=0 \mid Y=0)$ and $P(S=s, Y_m=0 \mid Y=1)$ in terms of the five model parameters. For non-conforming parts (i.e. $Y=0$), we have,

$$\begin{aligned}
&P(S=s, Y_m=0 \mid Y=0) = \int_0^1 P(S=s, Y_m=0 \mid \alpha)\,\mathrm{f}_A(\alpha)\,d\alpha \\
&= \int_0^1 P(S=s \mid \alpha) P(Y_m=0 \mid \alpha)\,\mathrm{f}_A(\alpha)\,d\alpha \\
&= \int_0^1 \left(\binom{r}{s}(\alpha)^s(1-\alpha)^{r-s}\right)(1-\alpha)\left(\frac{\alpha^{-1+R_C/\gamma_C}(1-\alpha)^{-1+(1-R_C)/\gamma_C}}{\mathrm{Beta}\left(\frac{R_C}{\gamma_C}, \frac{1-R_C}{\gamma_C}\right)}\right) d\alpha \\
&= \binom{r}{s}\frac{\mathrm{Beta}\left(s+\frac{R_C}{\gamma_C}, r+1-s+\frac{1-R_C}{\gamma_C}\right)}{\mathrm{Beta}\left(\frac{R_C}{\gamma_C}, \frac{1-R_C}{\gamma_C}\right)}.
\end{aligned}$$

where α represents the part specific misclassification rate and f_A is the corresponding Beta density function. Similarly, for conforming parts,

$$P(S=s, Y_m=0 \mid Y=1) = \binom{r}{s}\frac{\mathrm{Beta}\left(r+1-s+\frac{R_P}{\gamma_P}, s+\frac{1-R_P}{\gamma_P}\right)}{\mathrm{Beta}\left(\frac{R_P}{\gamma_P}, \frac{1-R_P}{\gamma_P}\right)}.$$

We used these expressions along with the definitions of the attributes of interest and applications of Bayes' rule to derive expressions for all of the factors in the likelihoods from each of the three phases. Doing so and combining the likelihoods in Equation (1) from the different phases, yields the following overall log-likelihood function,

$$\begin{aligned}\ell(\theta) = k &+ (n_B - y_B - n_{RM})\log\left(p_C R_P + (1-p_C)(1-R_C)\right) \\ &+ (y_B)\log\left(p_C(1-R_P) + (1-p_C)R_C\right) \\ &+ \sum_{s=0}^{r}(n_s - v_s)\log(f_s + g_s) + u_s \log f_s + (v_s - u_s)\log g_s\end{aligned} \tag{2}$$

where $g_s = P(S=s, Y_m = 0 \mid Y = 0)(1-p_C)$, $f_s = P(S=s, Y_m = 0 \mid Y = 1)p_C$, and k is a constant which does not depend on any of the five model parameters.

The log-likelihood expression in Equation (2) is used with data recorded as in Table 1 to calculate maximum likelihood estimates and asymptotic standard errors based Fisher's [21] asymptotic theory. In the appendix we provide evidence through a simulation study that the Fisher information based approximation to the standard errors is reasonable in this context. MATLAB code is available upon request. Equation (2) is the log-likelihood function for the general three phase plan regardless of the outcomes of the various phases and the experimenter's choices for v_s, $s = 0,1,...,r$.

2.3 Example

In order to give a tangible example, we calculate estimates from the example in Danila et al. [9]. That paper used a data set from a hypothetical but realistic situation involving a functional test stand in a 100% inspection scheme for an electronic device. The test stand simultaneously evaluates a number of characteristics of the device. As a result, we assume the measurand is binary. Many of the features depend on underlying latent variables, and thus we expect that the misclassification rates will vary from device to device. The gold standard inspection is an exhaustive manual check of all the device characteristics.

The test stand was used to measure 100 parts that previously failed inspection. The parts were re-measured five times each and separated into six bins according to the number of passed inspections. All parts were then verified with the gold standard as conforming or not. This plan is a full verification plan. From the data set, we calculate estimates of the parameters of interest for the full, targeted and no-verification plans. To calculate targeted verification plan estimates, we discard the verification information from all bins except 2 and 3. For the no-verification plan estimates, we discard all verification information from all bins except 2 and 3. For the no-verification plan estimates, we discard all verification information. Table 2 gives the data and the corresponding estimates are given in Table 3.

Table 2 – Electronics Testing Example Data

Phase I: Baseline

Number of Parts Measured in Baseline (n_B)	1243
Number of Parts Passing Inspection in Baseline (y_B)	960

Phase II: Repeated Measurements

Select n_{RM} = 100 parts from rejects and measure each *r* = 5 times						
Number of Passes (bin #, *s*)	0	1	2	3	4	5
Number of Parts (n_s)	41	18	5	9	5	22

Phase III: Verification

Bin #	0	1	2	3	4	5
Full Verification Plan						
Number Verified (ν_s)	41	18	5	9	5	22
Number Conforming among Verified (u_s)	0	0	0	5	5	22
Targeted Verification Plan						
Number Verified (ν_s)	0	0	5	9	0	0
Number Conforming among Verified (u_s)	0	0	0	5	0	0

To help understand the data, consider the following. In Table 2, the data from Phase II show that of the 100 previously failed parts 41 passed zero times when measured five additional times with the BMS. In Phase III with the full verification plan these 41 parts were measured with the gold standard and all were found to be conforming. With the targeted verification plan, the Phase III likelihood uses only the 14 gold standard measurements from bins 2 and 3.

Table 3 – Parameter Estimates for the Example from Each of the Three Plans: Full Verification, Targeted Verification and No Verification

Parameter	R_C	R_P	P_C
Full Verification Plan			
Estimate	0.134	0.086	0.820
Standard Error	0.029	0.013	0.016
Targeted Verification Plan			
Estimate	0.146	0.085	0.816
Standard Error	0.040	0.013	0.019
No Verification Plan (no Phase III)			
Estimate	0.235	0.072	0.778
Standard Error	0.128	0.0162	0.052

Table 3 shows that the standard errors for $\hat{R}_C, \hat{R}_P$ and $\hat{P}_C$ in the targeted verification case increased by 38%, 0% and 19% compared to the full verification plan. However, the targeted verification plan only uses 14 gold standard measurements as compared to 100 for the full verification plan. The no-

verification plan has much worse performance with standard errors for $\hat{R}_C, \hat{R}_P$ and $\hat{P}_C$ being 3-4 times greater than those of the full verification plan.

2.4 Recommended Three Phase BMS Assessment Plan

The Three-Phase Plan as outlined in Section 2.1 leaves many design decisions open, including the number of parts in the repeated measurement phase (n_{RM}), how many repeated measurements are used (r) as well as how many and which parts are verified ($\nu_0, \nu_1, \ldots, \nu_r$). In Table 4 we describe a recommended three-phase plan that uses targeted verification and specific choices for the general design parameters. A brief justification of the design choices is given in Section 2.5. The recommended plan is not optimal for all parameter combinations but is close to optimal for parameter values thought to be plausible in practice. There are two design parameters left open. The first, n_B , is not specified, but rather the experimenter should use all relevant baseline data available. The second, n_{RM} , the number of parts selected for phase II and is left open to the experimenter to adjust in order to meet the precision requirements of the study.

Table 4: Recommended Targeted Verification Plan with Suggested Design Parameters

Baseline Phase	• Measure n_B parts once and record the number of parts that passed inspection as y_B .
Repeated Measurement Phase	• Randomly select n_{RM} parts that failed inspection in the baseline phase and measure them (*r* =) **7** additional times. Separate the parts into eight bins based upon the total number of times each part passed inspection • Record the total number of parts in each bin as n_s , *s* = 0, 1 ,…,7
Verification Phase	• Measure all parts with the gold standard in the bins representing **3** or **4** passes, i.e. select $\nu_0 = \nu_1 = \nu_2 = \nu_5 = \nu_6 = \nu_7 = 0$, $\nu_3 = n_3$ and $\nu_4 = n_4$ • For each bin, record the number of parts that were measured by the gold standard to be conforming as u_s

2.5 Justification for the Recommended Plan

Notice that in the example from Section 2.3, the targeted verification plan only verifies parts in the middle bins, and yet obtains performance not much worse than that of the full verification plan. To visualize how standard errors are reduced as the number of parts verified increases, we present an illustration where parts are verified one by one in a given order, and standard errors are calculated at each point using Fisher's asymptotic theory. The illustration uses seven repeated measurements as we recommend in Section 2.4. We start by verifying parts one-by-one from bin 4. When all parts in bin 4 are

verified, we next start verifying parts from bin 3. When bin 3 is exhausted we start with parts in bin 5, then bin 2, then 6, then 1, then 7, and finally bin 0.

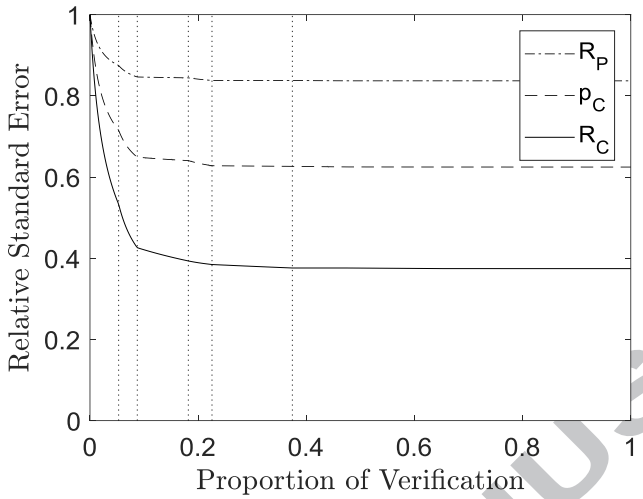

Figure 2 – Reduction in Standard Error as the Proportion of Verification (starting in middle bins) Increases

$n_{RM} = 500, r = 7, n_B = 10000, R_C = 0.075, R_P = 0.075, p_C = 0.925, \gamma_C = 0.125, \gamma_P = 0.125$

Vertical dashed lines represent when each of the first 5 bins are exhausted (Bin order: 4,3,5,2,6,1,7,0)

Figure 2 shows that the standard errors (displayed relative to the standard error of the no-verification plan) decrease rapidly and then flatten out. Additionally, we see that verifying parts in the middle bins provides tremendous benefit while verifying parts in non-central bins provides almost no benefit. The dashed lines represent when each of the first 5 bins is exhausted. We note that the reduction in standard error dramatically slows down after all the parts in bins 3 and 4 are verified. We investigated other parameter values combinations and found similar results.

We used a factorial experiment with two levels (see Table 5) for each of the five parameters to assess properties of the recommended plan. We chose the levels to represent realistic values for inspection systems used in industry.

Table 5 - Factorial Experiment Parameter Levels

Factor	R_C	R_P	p_C	γ_C	γ_P
Levels	0.05 0.10	0.05 0.10	0.90 0.95	0.05 0.20	0.05 0.20

Chapter 4 of the PhD thesis Severn [22] gives a more detailed justification for the recommended plan. Next, we summarize some of the results from Severn [22] without providing all the details.

To justify the choice $r = 7$, the total number of BMS measurements was fixed at 2500, i.e. $r * n_{RM} = 2500$, and r was varied from 3 through 11. Then, using the same order of verification used to create Figure 2, we calculated the trajectory of standard errors as the proportion of verification

increased, for each value of r and all 32 combinations of factor levels found in Table 5. We then plotted the average trajectory over the 32 combinations for each value of *r*. No trajectory was uniformly the lowest; rather the best value of r depended on the proportion of verifications. However at the point in the trajectory where most of the decrease had been obtained, and the reduction slowed down thereafter, $r = 7$ had the lowest standard error.

For $r = 7$, to give a fuller justification for the choice of verifying the two central bins, we consider verifying whole bins one after the other. We find that the reduction in standard error, computed as an average over the 32 parameter combinations in Table 5, is greatest when bin 4 is verified. Next, we repeat the calculation for the remaining unverified bins and find that verifying bin 3 reduces the standard errors the most. Third, repeating the calculation for the remaining six bins, we find that the reductions in the standard errors are negligible and the number of verifications is greatly increased. Thus, we conclude that verifying parts from bins 3 and 4 alone was ideal.

3. Comparisons of Plans

3.1 Comparative Study of Targeted Verification Performance

We now assess the performance of the recommended plan in comparison to the full and no-verification plans as proposed by Danila et al. [4, 9]. Note that all plans have the same baseline and repeated measurement phases. The performance comparisons make use of the results from a full factorial experiment with factors defined by the five parameters using the levels given in Table 5. Each experiment has 32 runs and the results are summarized by box plots. The design parameters for the experiment are $n_B = 10000, n_{RM} = 500, r = 7$. For each run, ten thousand full verification data sets were generated, then full verification, targeted verification and no-verification estimates were calculated. The full verification plan used the full data set, while the targeted and no verification plans used the same data set but with the appropriate verification information removed.

We know that the full verification must result in the smallest standard errors, no verification the largest and targeted verification somewhere in-between. Thus, to compare the three plans, we use two performance measures that quantify how close the standard error from targeted verification is to the two extremes. The performance measures are the percent reduction in standard error and percent possible reduction attained. Figure 3 provides a pictorial explanation of these two performance measures where we denote the standard errors for the no-verification, the proposed plan and the full verification as A, B and C ($A \geq B \geq C$) respectively.

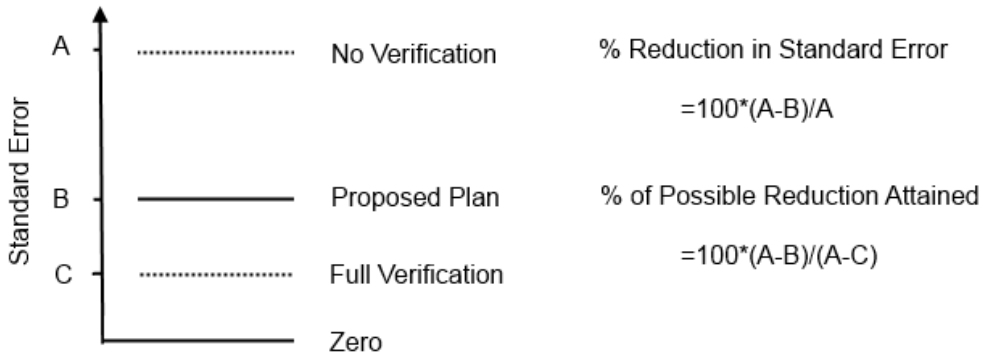

Figure 3 – Definition of the Two Performance Measures

Figure 4 shows the great improvement obtained by verifying bins 3 and 4, which, on average, represent only 8.4% of the parts repeatedly measured. The reduction in the standard error of $\hat{R}_C$ compared to the no verification plan is, on average, about 60% and represents 90% of the reduction possible through verification. This is a dramatic improvement for very little work. Targeted verification also reduces the standard errors of $\hat{R}_P$ and $\hat{P}_C$ by 18% and 41% respectively, which in both cases, represents, on average, 90% of the reduction possible.

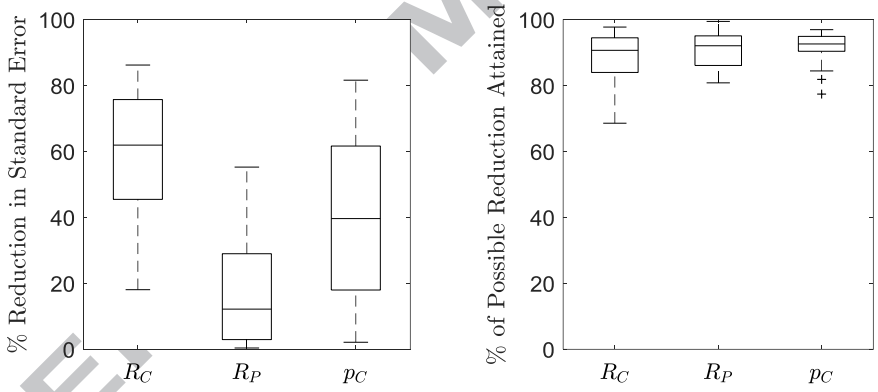

Figure 4 – Comparison of Full, Targeted and No Verification on the Standard Error of the Estimates

Results of Factorial Experiment (See Table 5) using performance measures defined in Figure 4

Left panel shows the % reduction, right panel shows the % of possible reduction

In Figure 5 we compare the bias of the estimates from the recommended targeted verification plan to that of the full and no verification plans. Relative bias is the bias of the estimate for a parameter divided by the true parameter value.

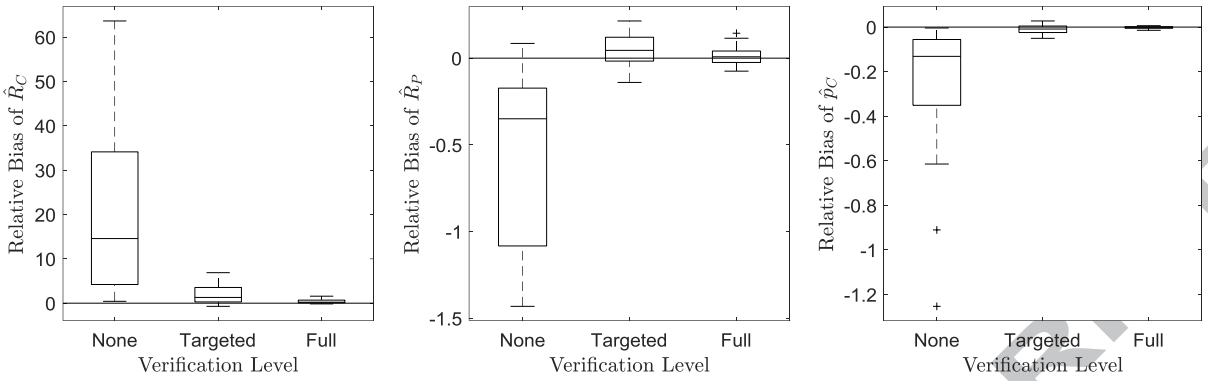

Figure 5 – Relative Bias of the Conditional Plan with Full, Targeted and No Verification

Results show bias divided by true parameter value (times 100) for the factorial experiment (See Table 5)

Figure 5 shows a dramatic reduction in the bias of the estimates from the targeted verification plan for all three quantities of interest when compared to the no verification plan. The reduction in biases of the targeted verification plan compared to the no verification plan for $\hat{R}_C$, $\hat{R}_P$ and $\hat{P}_C$ are 84%, 70% and 83% on average respectively, which represents around 90% of the possible reduction possible through verification. We attain this improvement through verifying, on average, only 8.4% of the parts in the repeated measurement phase.

We conclude that when using conditional sampling, the targeted verification plan has superior performance compared to the no-verification plan. The performance is close to full verification while requiring only a small number of gold standard measurements.

3.2 Comparison of Conditional and Population Sampling Plans

In Phase II, we select n_{RM} parts for repeated measurement. In this subsection we compare a targeted verification plan that uses a random sample of parts (population sampling) and no baseline measurements (as proposed by Severn et al. [15]) versus a plan that uses parts selected at random from those that fail initial inspection (conditional sampling – as recommended in Section 2). Sampling parts that failed inspection is easy because these parts are collected for scrap or rework. Sampling rejected parts is also less intrusive to the manufacturing process because it does not interfere with production goals. To make a quantitative comparison, we conducted a comparison using ten thousand simulated data sets generated for each of the 32 sets of parameter values found in Table 5. The design parameters for the experiment are $n_B = 10000$ (used for the conditional sampling plan), $n_{RM} = 500$ and $r = 7$. The simulations were run for the population sampling plan and separately for the conditional plan. The number of repeated measurements in Phase II and the number of verifications in Phase III is the same. To keep the number of verifications the same, we used a fixed number of verifications for each set of parameter values. That number was equal to the expected number of parts (rounded to the nearest integer) that would fall in the two central bins in the population sampling case, assuming the beta-binomial model. This slightly favours the population sampling approach because the number of

verifications is better adapted to that plan. The verifications were done in the order described earlier, starting with middle bins and, as they are exhausted moving outwards. Figures 6 and 7 are based on the results of these simulations. These figures use relative measures, meaning that they are expressed as a percentage of the parameter value being estimated.

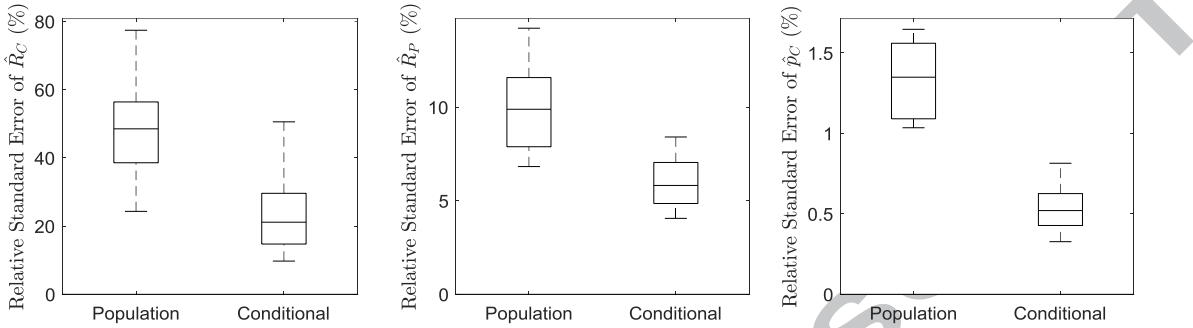

Figure 6 – Relative Standard Error Comparison of Population and Conditional Sampling

Results show standard errors divided by true parameter value (times 100) for the factorial experiment (See Table 5)

Figure 6 shows dramatic improvements in the estimates for all three quantities of interest when using the conditional plan rather than the population plan. Using the conditional sampling plan instead of the population plan yields reductions in the standard errors of $\hat{R}_C$, $\hat{R}_P$ and $\hat{p}_C$ of 53%, 39% and 60% on average respectively. Note that although the boxplots for population and conditional sampling overlap somewhat in Figure 6, the conditional sampling plan always provides an improvement over the population sampling plan in each of the 32 parameter value combinations.

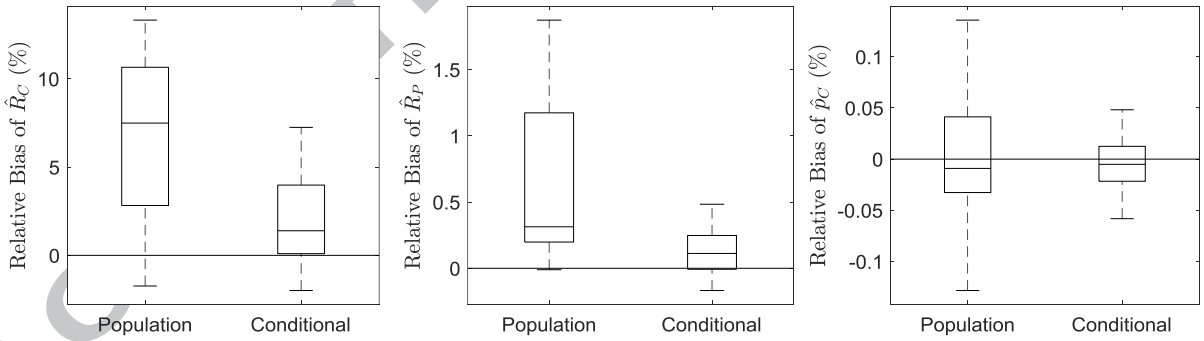

Figure 7 – Relative Bias Comparison of Population and Conditional Sampling

Results show bias divided by true parameter value (times 100) for the factorial experiment (See Table 5)

Figure 7 shows that the bias is also reduced for each of the quantities of interest. The reduction in the average bias for $\hat{R}_C$, $\hat{R}_P$ and $\hat{p}_C$ is 65%, 75% and 59%, respectively. In conclusion, the targeted verification conditional sampling plan greatly outperforms the comparable targeted verification population sampling plan. Therefore, wherever possible, we recommend using the conditional sampling plan. This is particularly easy to do when a BMS is currently used in a manufacturing environment.

In summary, in Phase II, conditional sampling is substantially more efficient than selecting the parts to be repeatedly measured at random from the population.

3.3 Comparison to a Naïve Plan

Under the assumptions given in the bulleted list in Section 1, we require hundreds of parts and measurements to assess reliably the misclassification rates of a binary measurement system. Consider using a naïve plan where we sample at random from the stream of parts and then measure each selected part with the BMS (*r=1*) and the gold standard. Pepe [11] calls this a cohort study. In a medical context, it is likely that baseline data are not available and that subjects cannot be measured repeatedly. In a cohort study we ignore baseline data and repeated measurement of the selected parts in Phase II and use full verification in Phase III. Suppose we want to obtain a relative standard error of, at most, 0.25 for each of the three parameters of interest for each of the 32 combination of parameters given in Table 5. The average sample size needed for the naïve plan is 3,360 parts with a range of 1,440 to 6,080. Compare this to the recommended plan where on average only 243 failed parts are needed for repeated measurement ($r = 7$) with a range of 52 to 740. In this case the number of BMS measurements is on average 1701 with range 364 to 5189. Furthermore, in the recommended plan on average only 22 parts need to be measured with the gold standard with a range of 2 to 115. The naïve plan not only requires more measurements with the BMS, but many more measurements with the gold standard. The standard error calculations are based on Fisher's asymptotic theory; they follow an inverse squared relationship with sample size. The recommended plan was run with baseline size equal to twenty times n_{RM} the Phase II sample size. In summary, the proposed plan is far superior to the naïve plan under the assumed conditions listed in Section 1.

4. Discussion

We have assumed throughout that we are dealing with an inspection system currently in use so that baseline data are available. If we are dealing with a new system we can use targeted verification without the baseline and conditional sampling.

The results in this paper are based on the assumption that the part specific misclassification rates follow beta distributions. Estimates from no- verification plans are sensitive to this assumption; see Akkerhuis et al. [7], Albert and Dodd [13]. The recommended plan is more robust to model misspecification, and although we have not investigated thoroughly, verifying a few additional parts, say 5, from each of the non-central bins improves robustness further. Additionally, with these extra verifications, we can construct estimates that are nearly as efficient as those from the beta-binomial model and that make no assumptions about the underlying distribution of misclassification rates. See Severn [22].

There are other possible applications of targeted verification. For example, we may want to compare two BMSs using the same set of parts or to include operator effects in the analysis.

5. Conclusions

We have assumed throughout that we are dealing with an inspection system currently in use so that baseline data are available. The BMS is not destructive and parts can be measured more than once.

Assessing a BMS requires much larger sample sizes than assessing a continuous measurement system. To control the cost, we need sophisticated plans and analysis. In general, we recommend

- Use as much baseline data as possible, subject to the assumption that the inspection system has been stable over the collection period.
- Sample several hundred parts that failed an initial inspection. With larger P_C , i.e. a higher quality process, we need more parts in Phase II.
- Measure each selected failed part $r = 7$ times.
- Verify all of the parts in the middle bins, i.e. for $r = 7$ from bins corresponding to $s = 3$ or 4.

We showed that the precision and bias of the estimated mean error rates obtained using the recommended three-phase plan are much better than those without the Phase III data, i.e. when we do not use gold standard measurements. We also showed that the performance of the recommended plan is similar to much more costly plans that measure all parts in Phase III with the gold standard. In addition, we quantified the substantial improvement in precision for estimating R_C and R_P when using the proposed plan with conditional sampling rather than population sampling, as recommended in earlier work. Finally, we showed that the proposed assessment plan is vastly superior to a cohort study that ignores baseline data and does not repeatedly measure parts with the BMS.

Appendix: Asymptotic Variance Justification

It is not possible to find an analytic expression for the maximum likelihood estimates or the associated standard errors. Therefore in Sections 2 and 3, asymptotic variance results due to Fisher [22] are used to estimate the standard errors. Here we assess the accuracy of these estimates for different sets of parameter values using a factorial experiment structure. There are 32 different treatments which are made up by varying the five model parameters, see Table 5. For each treatment, ten thousand datasets were simulated from the beta-binomial model, parts were selected and verified according the recommended conditional sampling plan and the maximum likelihood estimates were calculated. The standard errors of these estimates were calculated over the 10,000 simulation runs. We also calculated the asymptotic standard error approximation for each set of parameter values. Figure 8 shows the ratio of the simulated and the asymptotic standard errors for each combination of parameter values. The design parameters were kept the same for all treatments with $n_B = 10000$, $n_F = 500$ and $r = 7$.

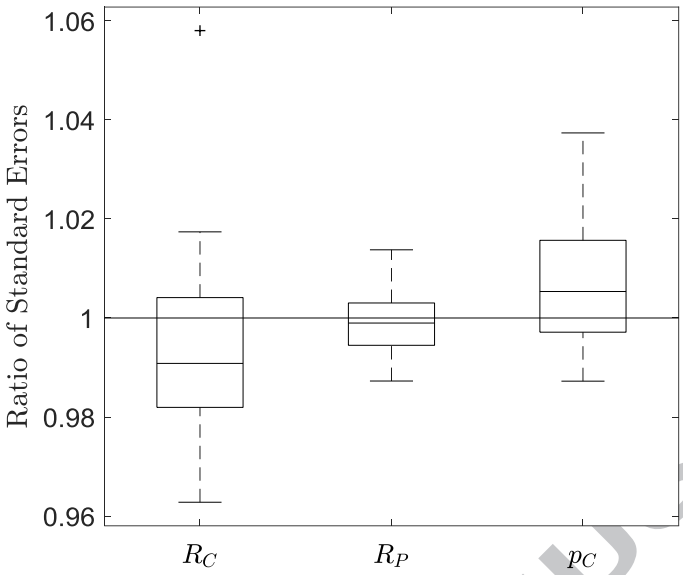

Figure 8 – Boxplots of the Ratio of Simulated and Asymptotic Standard Errors
Results of Factorial Experiment (See Table 5)

Figure 8 shows that the ratios of simulated standard errors and asymptotic standard errors are very close to one, and thus the asymptotic approximation are sufficiently accurate for the manner in which they are used.

Acknowledgements

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) discovery grant #105240. We also thank the referees for helpful comments.

References

[1] ISO Guide GUM (1995) Guide to the expression of uncertainty in measurement, 2nd edn. International Organisation for Standardisation, Geneva

[2] BIPM JCGM 2012 Evaluation of measurement data—the role of measurement uncertainty in conformity assessment JCGM 106:2012 (Paris: BIPM) (www.bipm.org/utils/common/documents/jcgm/JCGM_106_2012_E.pdf)

[3] ISO/FDIS 10576-1 (2002) Statistical methods—Guidelines for the evaluation of conformity with specified requirements Part 1: General principles, ISO/TC69/SC 6.

[4] O. Danila, S.H. Steiner, R.J. MacKay, Assessing a binary measurement system with varying misclassification rates using a latent class random effects model, Journal of Quality Technology 44, no. 3 (2012): 179.

[5] W.N. Van Wieringen, J. De Mast, Measurement system analysis for binary data, Technometrics 50, no. 4 (2008): 468-478.

[6] R.A. Boyles, Gauge Capability for Pass—Fail Inspection, Technometrics 43, no. 2 (2001): 223-229.

[7] T. Akkerhuis, J. de Mast, T. Erdmann, The statistical evaluation of binary tests without gold standard: Robustness of latent variable approaches, Measurement 95 (2017): 473-479.

[8] 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, no. 2 (2004): 427-435.

[9] 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, no. 3 (2013): 335-345.

[10] R.J. Burke, R.D. Davis, F.C. Kaminsky, A.E.P. Roberts, The effect of inspector errors on the true fraction non-conforming: an industrial experiment, Quality Engineering 7, no. 3 (1995): 543-550.

[11] M.S. Pepe, The Statistical Evaluation of Medical Tests for Classification and Prediction, first ed., Oxford University Press Inc., New York, 2003.

[12] J .De Mast, T.P. Erdmann & W.N. Van Wieringen, Measurement System Analysis for Binary Inspection: Continuous Versus Dichotomous Measurands, Journal of Quality Technology, (2011), 43:2, 99-112.

[13] N. R. Farnum, *Modern Statistical Quality Control and Improvement*. (1994). Belmont, CA: Duxbury Press.

[14] O. Danila, S.H. Steiner, R.J. MacKay, Assessment of a binary measurement system in current use, Journal of Quality Technology, (2010): 42, no. 2, 152.

[15] D.E. Severn, S.H. Steiner, R.J. MacKay, Assessing Binary Measurement Systems: A Cost-Effective Alternative to Complete Verification." Journal of Quality Technology 48, no. 2 (2016): 128.

[16] P.S. Albert, L.E. Dodd, On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation, Journal of the American Statistical Association 103, no. 481 (2008): 61-73.

[17] C.B. Begg, R.A. Greenes, Assessment of diagnostic tests when disease verification is subject to selection bias, Biometrics (1983): 207-215.

[18] R.K. Burdick, C.M. Borror, D.C. Montgomery, Design and analysis of gauge R&R studies: making decisions with confidence intervals in random and mixed ANOVA models, Society for Industrial and Applied Mathematics, 2005.

[19] Y. Haitovsky, J. Rapp, Conditional resampling for misclassified multinomial data with applications to sampling inspection, Technometrics (1992): 34, no. 4, 473-483.

[20] A. Tenenbein, A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection, Technometrics 14, no. 1 (1972): 187-202.

[21] R.A. Fisher, Theory of statistical estimation, In Mathematical Proceedings of the Cambridge Philosophical Society (1925): vol. 22, no. 5, pp. 700-725.

[22] D.E. Severn, Assessing Binary Measurement Systems Using Targeted Verification with a Gold Standard, PhD Thesis, University of Waterloo (2017) *UWSpace. http://hdl.handle.net/10012/11930*

Highlights

- Estimate misclassification rates for non-destructive pass/fail inspection system
- Efficient plans using available data, repeated measurement & conditional sampling
- Demonstrated improvement to existing assessment plans