Assessing Binary Measurement Systems: A Cost-Effective Alternative to Complete Verification

DANIEL E. SEVERN, STEFAN H. STEINER, and R. JOCK MACKAY

Business and Industrial Statistics Research Group, and
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1

Suppose we plan to assess a binary measurement system when the misclassification probabilities vary from part to part. We consider the estimation of the average error probabilities of such a system when a gold standard (error-free) system is available to verify the status of any part. We examine plans where we first measure a sample of n parts r times each with the binary measurement system. Then we study the impact on the precision and robustness of the estimates if we use the gold-standard system to verify the true status of none, some, or all of the sampled parts. We show that a partial verification plan has comparable performance to full verification in terms of the precision and robustness of the estimates while requiring as few as 10% of parts to be verified. When the gold-standard system is expensive or time consuming, eliminating the need to verify all parts dramatically reduces the cost of the assessment study.

Key Words: Beta Binomial; Gage Capability; Measurement System Analysis; Misclassification Probabilities; Pass–Fail Inspection; Random Effects.

1. Introduction

In this paper, we propose a new, efficient plan for assessing the average misclassification-error probabilities of a binary measurement system (BMS). We use the term "assessment" to replace "measurement system analysis" for a BMS. Many BMSs are used for 100% inspection to reduce the risk that a customer receives nonconforming product. For example, a vision system inspects fascias for ghosting, a surface defect that is either present or absent. In a second example, discussed in more detail later, an automated gauge measures a large number of continuous characteristics on a camshaft. A camshaft is conforming if all the characteristics meet specifications. A camshaft passes the inspection if the gauge determines that every measured characteristic is within specification; otherwise, the part fails the inspection. In both of these cases, the BMS is nondestructive and we can measure the same part repeatedly if we choose.

A BMS makes an error if it passes a non-conforming part or fails a conforming part. For any part, let X be 1 if the part is conforming and 0 otherwise. We denote by π_{C} the probability that a randomly selected part is conforming. Also, let Y be 1 if the part passes inspection and 0 otherwise. For many binary measurement systems, some parts are more difficult to classify correctly than others (De Mast et al. (2011)). For any particular part, we denote the part-specific error probability as α for a nonconforming part and β for a conforming part. We interpret α and β as the long-run error probabilities

Mr. Daniel Severn is a PhD student in the Department of Statistics and Actuarial Science at the University of Waterloo. His email address is dsevern@uwaterloo.ca.

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

Dr. MacKay is an Adjunct Professor in the Department of Statistics and Actuarial Science at the University of Waterloo. He is a member of ASQ. His email address is rjmackay@uwaterloo.ca.

for that particular part if it is measured repeatedly. Following Danila et al. (2012), we treat α and β as random effects, varying from part to part, and characterize the performance of the BMS by the average error probabilities $\mu_A = \mathrm{E}[\alpha]$ and $\mu_B = \mathrm{E}[\beta]$. Given that a part is nonconforming, we interpret $\mu_A = P(Y = 1 \mid X = 0)$ as the probability that a randomly selected part passes an inspection by the BMS and μ_B similarly. To assess the BMS, our goal is to estimate μ_A and μ_B. Throughout, we assume that μ_A and μ_B are small, i.e., the BMS is reasonably good, and that π_C is large to reflect the overall high performance of most industrial processes.

We assume that a gold standard-measurement system (GSS) is available so that the true conforming status of any part can be verified. The GSS is sometimes called a perfect reference system and the true status of a part is called the reference value. As noted by a referee, there are many BMSs where a GSS is not available. In that case, to estimate μ_A and μ_B, a sample of parts is measured repeatedly and the data are modeled treating the true status of each sampled part as a latent variable. See, for example, Danila et al. (2012), Van Wieringen and De Mast (2012), Beavers et al. (2011), and Boyles (2001). In the medical literature, repeated measurements by the BMS are often replaced by multiple diagnostic tests. See, for example, Qu et al. (1996), Albert and Dodd (2004), and Pepe and Janes (2007). When a GSS is available, the simplest assessment plan is to measure a sample of n parts $r \geq 1$ times and then determine the true status of each part with the GSS. We call this full verification. For example, see Danila et al. (2008), Farnum (1994), and Burke et al. (1995) in an industrial setting and Pepe (2003) in the medical context.

The GSS may be destructive or expensive and time consuming, as is the case in the camshaft example. To avoid the burden of measuring every part with the GSS, we propose a class of partial verification plans:

Phase I: Measure a randomly selected sample of n parts r times with the BMS under consideration.

Phase II: Use the GSS to determine the conforming status of a subsample of these parts deliberately selected based on the frequency of passes in phase I.

Partial verification of subjects in the assessment of diagnostic tests has been considered in the medical literature. The paper by De Groot et al. (2011) has many references. A major concern is bias in the estimates that can occur if the subjects verified are self-selected or chosen for ethical reasons. These considerations are not relevant in the industrial context where verification is strictly controlled by the investigator.

The remainder of the paper is laid out as follows. In the next section, we describe the model assumed for generating the data and develop the likelihood for the observed data for the general two-phase partial-verification plan. In Section 3, we provide an example and show numerically the benefits of employing a partial-verification plan. In Section 4, we show that, if the subsample of parts selected for verification in phase II is well chosen, the proposed plan provides estimates of μ_A and μ_B that are almost as precise as the estimates from the plan with complete verification and much more precise than the estimates from the plan with no verification. In addition, we discuss planning of both phase I and II and look at the robustness of the proposed plan. Finally, in Section 5, we summarize the proposed plan and discuss some additional issues.

2. Modeling and Likelihood

We assume that measurements made on different parts are independent. To describe the results of the repeated measurements of each part in phase 1, let $Y_1, \ldots, Y_r$ represent the output of the BMS for a single part. If the part is nonconforming with error probability α, we assume that $Y_1, \ldots, Y_r$ are conditionally independent so that

$$P(Y_1 = y_1, \ldots, Y_r = y_r \mid X = 0, \alpha) = (1-\alpha)^s \alpha^{r-s},$$

where $s = \sum_1^r y_i$ is the number of times the part passes inspection. Similarly, if the part is conforming with error rate β and the repeated measurements are conditionally independent, we have

$$P(Y_1 = y_1, \ldots, Y_r = y_r \mid X = 1, \beta) = (1-\beta)^s \beta^{r-s}.$$

In both cases, the number of passes S is a sufficient statistic. As in Danila et al. (2012), we suppose the part-to-part variation in the error probabilities α or β are described by beta distributions parameterized to have means μ_A and μ_B, the parameters of interest, and γ_A and γ_B, nuisance parameters related to the standard deviation of the random effects. The density function for the random effect α is

$$\frac{\alpha^{\frac{\mu_A}{\gamma_A}-1}(1-\alpha)^{\frac{1-\mu_A}{\gamma_A}-1}}{\text{Beta}\left(\frac{\mu_A}{\gamma_A},\frac{1-\mu_A}{\gamma_A}\right)}, \qquad 0<\alpha<1,$$

where $\text{Beta}(u,v)$ is the beta function. If γ_A approaches 0, the random effects becomes constant at μ_A. As γ_A increases, the standard deviation of the random effects increases to its maximum $\sqrt{\mu_A(1-\mu_A)}$. We make similar assumptions for β. Because we do not observe the random effects, α and β, we have, for any nonconforming part,

$$\begin{aligned}P(S=s\mid X=0)\\ &=\int_0^1\binom{r}{s}\alpha^s(1-\alpha)^{r-s}\frac{\alpha^{\frac{\mu_A}{\gamma_A}-1}(1-\alpha)^{\frac{1-\mu_A}{\gamma_A}-1}}{\text{Beta}\left(\frac{\mu_A}{\gamma_A},\frac{1-\mu_A}{\gamma_A}\right)}d\alpha\\ &=\binom{r}{s}\frac{\text{Beta}\left(s+\frac{\mu_A}{\gamma_A},r-s+\frac{1-\mu_A}{\gamma_A}\right)}{\text{Beta}\left(\frac{\mu_A}{\gamma_A},\frac{1-\mu_A}{\gamma_A}\right)},\end{aligned}$$

Similarly, for any conforming part,

$$\begin{aligned}P(S=s\mid X=1)\\ &=\binom{r}{s}\frac{\text{Beta}\left(r-s+\frac{\mu_B}{\gamma_B},s+\frac{1-\mu_B}{\gamma_B}\right)}{\text{Beta}\left(\frac{\mu_B}{\gamma_B},\frac{1-\mu_B}{\gamma_B}\right)}.\end{aligned}$$

In phase I, we observe the number of passes and not the conforming status of any part, so the probability that a phase I part passes s times is

$$\begin{aligned}P(S=s)&=P(S=s\mid X=0)(1-\pi_C)\\ &\quad+P(S=s\mid X=1)\pi_C\\ &=p_s+q_s, \qquad 0\le s\le r,\end{aligned}$$

where $p_s=P(S=s\mid X=0)(1-\pi_C)$ and $q_s=P(S=s\mid X=1)\pi_C$.

To construct the phase I likelihood, recall that we assume measurements on different parts are independent and, for each part, the number of passes S is sufficient. Let n_s be the number of phase I parts with s passes out of r repeated measurements. We say these parts are in bin s. If $\theta=(\mu_A,\mu_B,\pi_C,\gamma_A,\gamma_B)$, the phase I log likelihood based on a multinomial distribution for the observed data $(n_0,n_1,\ldots,n_r)$ is

$$l_{\text{I}}(\theta)=c+\sum_{s=0}^{r}n_s\ln(p_s+q_s), \tag{1}$$

where c is a constant not depending on θ.

In phase II, we select parts for verification by the GSS using stratified random sampling from the bins of parts with $s=0,1,\ldots,r$ passes. That is, we randomly select v_s parts for verification from bin s, where $0\le v_s\le n_s$. If $v_s=0$, no parts are verified from bin s, and, if $v_s=n_s$, all parts from bin s are verified. We discuss good choices for v_s, $s=0,1,\ldots,r$ later.

TABLE 1. Data Summary

Number of passes (bin #)	0	1	...	r
Number of parts (phase I)	n_0	n_1	...	n_r
Number verified	v_0	v_1	...	v_r
Number conforming among verified parts (phase II)	u_0	u_1	...	u_r

Because measurements on different parts are independent, we need only to record u_s the number of the v_s parts sampled from bin s that are conforming. The data and notation are summarized in Table 1. Note that the two-phase plan includes as a special case full verification where all parts are verified, i.e., v_s equals n_s for all s. The plan also includes the no-verification plan presented in Danila et al. (2012), where v_s is set to 0 for all s.

The probability of u_s conforming parts from the v_s parts in bin s is

$$\begin{aligned}P(U_s=u_s\mid S=s)\\ &=\binom{v_s}{u_s}P(x=1\mid S=s)^{u_s}P(x=0\mid S=s)^{v_s-u_s}\\ &=\binom{v_s}{u_s}\left(\frac{p_s}{p_s+q_s}\right)^{u_s}\left(\frac{q_s}{p_s+q_s}\right)^{v_s-u_s}, \quad 0\le u_s\le v_s.\end{aligned}$$

So, the phase II log likelihood is

$$\begin{aligned}l_{\text{II}}(\theta)=d+\sum_{s=0}^{r}\Bigg[u_s\log\left(\frac{p_s}{p_s+q_s}\right)\\ +(v_s-u_s)\log\left(\frac{q_s}{p_s+q_s}\right)\Bigg],\end{aligned} \tag{2}$$

where d is a constant. Combining phases I and II, the overall log likelihood, ignoring additive constants, is the sum of Eqs. (1) and (2),

$$\begin{aligned}\ell(\theta)=\sum_{s=0}^{r}(n_s-v_s)\log(p_s+q_s)+u_s\log p_s\\ +(v_s-u_s)\log q_s.\end{aligned} \tag{3}$$

We maximize Eq. (3) numerically to find the maximum-likelihood estimates (MLEs). We derive approximate standard errors using the asymptotic

TABLE 2. Camshaft Data

Number of passes (s)	0	1	2	3	4	5
Number of camshafts (n_s)	29	9	7	33	132	290
Number verified (v_s)	0	0	7	33	0	0
Number conforming among verified camshafts (u_s)	0	0	5	33	0	0

properties of the log likelihood from the diagonal elements of the inverse of the observed information matrix. Matlab code (2008) to find the MLEs and approximate standard errors is available upon request from the authors. See the Appendix for a justification of the asymptotic approximations based on the expected information matrix.

3. Camshaft Example

The context is real, but we constructed the data to be realistic. An automated gauge determines whether or not the lobes on a camshaft are within specification with respect to their geometry. Each of the 12 lobes is checked for six continuous critical characteristics. If one or more of these characteristics are out of specification for any lobe, the camshaft is rejected for scrap or rework. The binary measurement system is used for 100% inspection. Individual gauge R&R studies on specific continuous characteristics are conducted by lobe on a regular basis. It is known that these characteristics are correlated. To assess the overall performance of the BMS, 500 camshafts were measured 5 times each with the automated gauge and the number of times that each passed was recorded. The geometry of each of the 40 problematic camshafts that passed 2 or 3 times out of the 5 repeated measurement made in phase I was also measured using a high-precision coordinate-measuring machine, here taken to be the gold standard. Five of the seven camshafts that passed twice in phase I were found to be nonconforming on one or more of the 72 characteristics. None of the 33 camshafts with three initial passes were out-of-specification for any characteristic. The data are given in Table 2.

We show the maximum-likelihood estimates and their associated asymptotic standard errors in Table 3. For comparison, Table 3 also gives the estimates obtained without using the verification data (i.e., maximizing the phase I likelihood of Eq. (1)). In Table 3, we also quantify the benefits of using the verification data in terms of reduced standard errors. With the partial verification plan, the primary parameters μ_A, μ_B, and π_C are estimated with good precision but the nuisance parameters γ_A and γ_B are poorly estimated.

Compared with the no-verification plan, using partial verification (here the GSS measured only 8% of the phase I parts) results in a large reduction in the standard errors of all the parameter estimates, particularly those relating to nonconforming parts, i.e., μ_A and γ_A. Of the three primary quantities of interest, μ_A is estimated with the least precision. Note that μ_A affects the customer, making it perhaps the most important characteristic of the BMS.

4. Proposed Two-Phase Partial-Verification Plan

We have many options in defining a two-phase plan. We do not look for an optimal choice that would depend on the relative costs of making measurements with the BMS and GSS and the values of

TABLE 3. Camshaft Example Estimation Summary

Parameter	μ_A	μ_B	π_C	γ_A	γ_B
Without verification					
Estimate	0.0661	0.0935	0.9208	0.0483	0.0301
Std. error	0.0690	0.0093	0.0181	0.3032	0.0336
With Partial verification					
Estimate	0.0902	0.0896	0.9141	0.0886	0.0103
Std. error	0.0239	0.0061	0.0126	0.1081	0.0177
Reduction of std. err. with partial verification	65.4%	33.8%	30.4%	64.3%	47.2%

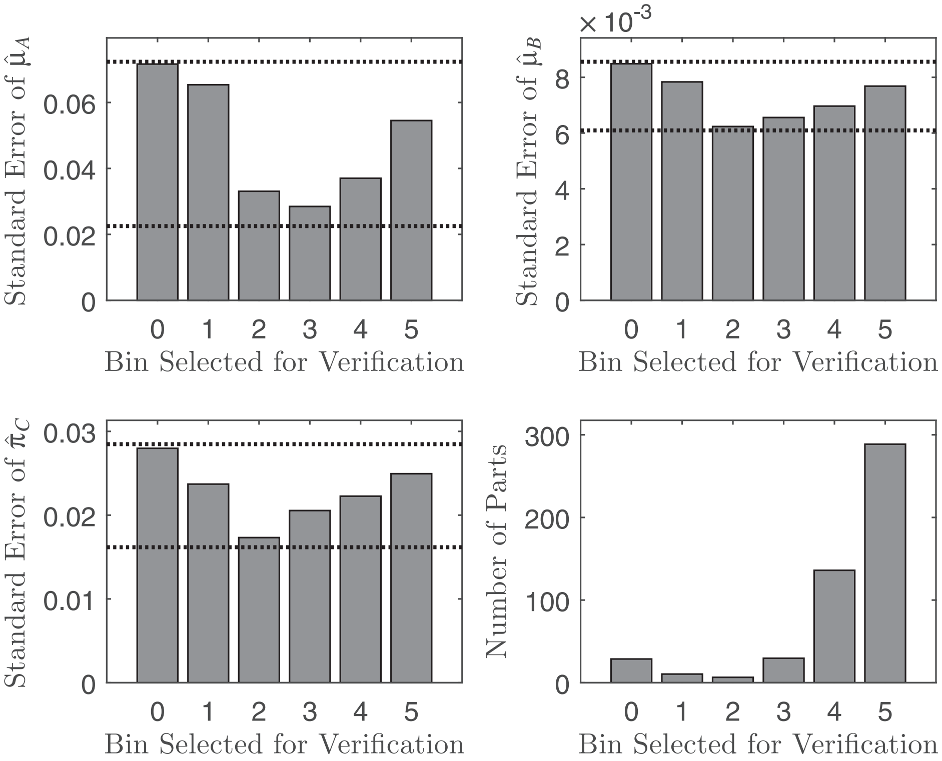

FIGURE 1. One Bin Verification Example: $n = 500$, $r = 5$, $\mu_A = 0.0902$, $\mu_B = 0.0896$, $\pi_C = 0.9141$, $\gamma_A = 0.0886$, $\gamma_B = 0.0103$. Dashed lines represent the standard errors of full verification (lower) and no verification (higher) plans.

the unknown model parameters. Instead we propose a generic plan that we show to have good properties over a wide range of situations. The recommended two phase plan is

Phase I: Measure a random sample of n parts $r = 5$ times each with the BMS. Based on the number of times passed, separate the parts into six bins to determine $n_0, n_1, \ldots, n_5$.

Phase II: Verify with the gold standard the conforming status of all $n_2 + n_3$ parts in the bins 2 and 3 and five randomly selected parts from each of the other bins (where possible).

So we have $v_s = \min(5, n_s)$ for $s = 0, 1, 4, 5$ and $v_2 = n_2$, $v_3 = n_3$. We denote bins 2 and 3 as the "middle" bins and the others as the "outside" bins. We investigate the following issues for the recommended plan:

- Selection of parts to verify in phase II.
- Sample size n and the number of replicated measurements r in phase I.
- Performance relative to the complete ($v_s = n_s$, $s = 0, 1, \ldots, r$) and no verification plans ($v_s = 0$, $s = 0, 1, \ldots, r$).
- Robustness against model misspecification.

More details on the justification and performance of this recommended plan and some extensions are provided in Severn (2016).

4.1. Planning Phase II

After phase I, parts are selected for verification based on the number of times they passed inspection. Given the assumption that the error probabilities are small for all (or almost all) parts, we expect parts that always pass to be conforming and those that always fail to be nonconforming. This suggests that there is little value in verifying such parts. The example also suggests that we can get a large improvement over the no-verification plan using only a few verifications in phase II from the middle bins.

To demonstrate that selecting from the middle bins is an effective strategy and results in precise parameter estimates, we carried out a study where we verify all parts in one and only one bin. We present the results in Figure 1 for the case where the parameter values ($\mu_A = 0.0902$, $\mu_B = 0.0896$, $\pi_C = 0.9141$, $\gamma_A = 0.0886$, $\gamma_B = 0.0103$), sample size ($n = 500$) and number of repeated measurements ($r = 5$) are based on the estimates and plan in the camshaft example. The results are based on the asymptotic standard errors derived from the Fisher information matrix with the phase I data given by the expected values of n_s for this set of parameter values. The number of parts in each bin in phase I data are displayed in the lower right subplot of Figure 1. The remaining three plots show the standard errors of $\hat{\mu}_A$, $\hat{\mu}_B$, and $\hat{\pi}_C$ when all the parts from only one bin are verified. Each of these plots has two dashed lines for reference; the higher line represents the standard error of the corresponding no-verification plan while the lower line represents the standard error for the full-verification plan.

We see in Figure 1 that verifying the relatively few parts with two passes reduces the standard errors of the estimates for all primary parameters more than verifying the 300+ parts that always passed inspection. We also see in Figure 1 that verifying all parts with two passes in phase I gives standard errors almost as small as full verification.

Figure 1 shows results for only one set of parameter values. To obtain more general conclusions, we conducted a factorial experiment for each combination of parameter values in Table 4. We chose the levels of the parameters to correspond to likely ranges for a BMS in practice. We repeated the calculations as in the previous paragraph and chose the optimal bin for each of the three model parameters of primary interest. For example, for the set of parameters used in Figure 1, bin 3 was optimal for μ_A because verifying all parts in bin 3 resulted in the lowest standard error for $\hat{\mu}_A$. Similarly, bin 2 was optimal for μ_B and π_C. A summary of the results is displayed in Table 5. As in Figure 1, standard errors are calculated using the asymptotic results as described in the Appendix.

Table 5 shows that, with five repeated measurements verifying all parts in either bin 2 or bin 3 is best in all 32 cases tested. Bin 2 seems the best overall when considering all three parameters. It is clear that selecting from the middle bins is the good strategy for verification over the selected range of parameter values that were chosen to represent typical values for a BMS in industry. Through further investigation (details not given here), we found that choosing parts from the middle bin(s) is effective with other values of r. Note also that the results from Table 5 do not depend on the sample size (n) because they are based on the Fisher information matrix and the log-likelihood function is approximately linear in n.

TABLE 4. Factorial Experiment Levels

Factor	μ_A	μ_B	π_C	γ_A	γ_B
Levels	0.05	0.05	0.90	0.05	0.05
	0.10	0.10	0.95	0.20	0.20

TABLE 5. Optimal Bin Factorial Experiment with $r = 5$. Percentage of time a bin is optimal for reducing standard error by parameters μ_A, μ_B = 0.05, 0.1; π_C = 0.9, 0.95; γ_A, γ_B = 0.05, 0.2 (see Table 4)

bin #	μ_A	μ_B	π_C
0	0	0	0
1	0	0	0
2	43.8	100	87.5
3	56.2	0	12.5
4	0	0	0
5	0	0	0

Verifying from the middle bins is optimal for many extreme scenarios as well. The only exception we found is when the parameters γ_A and γ_B are large enough so that the distribution of misclassification probabilities is U shaped and clustered around 0% and 100% as opposed to the average misclassification rate. This situation seems unrealistic.

In the recommended plan with $r = 5$, we select all parts from bins 2 and 3 for verification. We explain the additional verifications from the other bins when we discuss robustness of the estimates.

To see how increasing the number of parts verified from the middle bins reduces the standard errors, we again use the phase I plan and parameter values as estimated in the camshaft example. We start with a single verification from bin 2 and increase the number of verifications from that bin until it is exhausted. Then, we additionally sample from bin 3 until it is exhausted and continue to verify parts in increasingly outlying bins. We see the results in Figure 2. The vertical axis is the ratio of the standard error for the partial-verification plan over the standard error of the no-verification plan. The standard errors in Figure 2 are calculated using the asymptotic expressions.

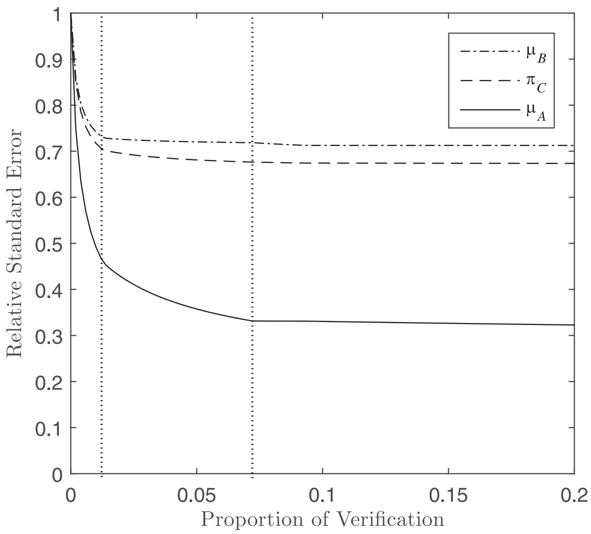

FIGURE 2. Verification Proportion Plot: $n = 500$, $r = 5$, $\mu_A = 0.0902$, $\mu_B = 0.0896$, $\pi_C = 0.9141$, $\gamma_A = 0.0886$, $\gamma_B = 0.0103$. Relative standard error is SE (partial verification)/SE (no verification). Vertical dotted lines show when bins 2 and 3 are exhausted.

Figure 2 shows that the majority of the benefit of selective verification occurs very quickly. Once the parts in the middle two bins have been verified (the vertical dotted lines show when bins 2 and 3 are exhausted), the improvement in standard errors is negligible.

In summary, we propose in the recommended plan that, in phase II, we verify all parts in the middle bins from phase I corresponding to two and three passes.

4.2. Planning Phase I

To plan Phase I, we specify the sample size n and the number of repeated measurements r.

We start by considering the choice of n. The recommended plan has $r = 5$ and, in phase II, $v_2 = n_2$ and $v_3 = n_3$. Based on the results in Section 4.1, the contribution to the information matrix of the 20 verifications in the four outside bins is small. If we ignore this contribution, then the phase II log-likelihood function of Eq. (3) becomes

$$l_{\mathrm{II}}^*(\theta) = d + \sum_{s=2}^{3}\left[u_s \log\left(\frac{p_s}{p_s+q_s}\right) + (n_s - u_s)\log\left(\frac{q_s}{p_s+q_s}\right)\right].$$

The modified log likelihood $l_{\mathrm{I}}(\theta) + l_{\mathrm{II}}^*(\theta)$ is linear in $n_0, \ldots, n_5$, u_2 and u_3. For the modified likelihood, the expected information matrix is n times a 5×5 matrix $J(\theta)$ that depends only on θ. The approximate standard deviation of the MLEs is the corresponding diagonal element of $J^{-1}(\theta)$ divided by $\sqrt{n}$. We can investigate various values of n for a reasonable range of values for the unknown parameters. For any particular value of θ, we provide software to calculate the diagonal elements of $J^{-1}(\theta)$.

If we hold n fixed, given the assumptions that μ_A and μ_B are small and that the variability of the part-specific error probabilities for conforming and nonconforming are not too large, we can decide whether or not each part is conforming or not with a small chance of error as r increases. So the information in the phase I plan with r large corresponds to complete verification, i.e., the additional information from phase II becomes negligible. There is then no need to use the gold-standard system.

To compare different r values, we conducted an experiment where we fixed nr, the total number of measurements in phase I. The verifications are allocated using the proposed rule described at the beginning of Section 4 that includes verifying five parts in each bin. For an even value of r, the two middle bins are defined as the middle bin and the bin with one pass less in phase I.

The experiment is first conducted with the parameter values taken from the camshaft example. For each value of r between 3 and 9, we calculated the asymptotic standard deviation of the estimates of the primary parameters. The results are found in Table 6.

For the set of parameter values used in Table 6, $r = 5$ is best for estimating μ_A, while $r = 4$ is best for estimating μ_B and π_C. Note that μ_A is the most poorly estimated parameter and the most important to the customer. Table 6 shows the results for μ_A, μ_B, and π_C but not the nuisance parameters, γ_A and γ_B, because they are poorly estimated in all cases. The precision of the estimates of the nuisance parameters improves when we increase r.

More generally, we conducted the same experiment over the grid of parameter values in Table 4. For each value of r, Figure 3 shows a boxplot of the standard deviation of the estimate for μ_A relative to the optimal value of r for each of the 32 combinations of parameter values. For example, if the results in Table 6 were included in Figure 3, the value for $r = 5$ would be $0.0239/0.0239 = 1$, whereas the value for $r = 6$ would be $0.0241/0.0239 = 1.0086$. In Figure 3, we do not give the results for $r = 3$ because

TABLE 6. Optimal n Experiment. Asymptotic standard deviations as n and r vary with $nr = 2{,}500$ fixed
$\mu_A = 0.0902$, $\mu_B = 0.0896$, $\pi_C = 0.9141$, $\gamma_A = 0.0886$, $\gamma_B = 0.0103$

	$r = 3$	$r = 4$	$r = 5$	$r = 6$	$r = 7$	$r = 8$	$r = 8$
	$n = 833$	$n = 625$	$n = 500$	$n = 416$	$n = 357$	$n = 312$	$n = 277$
$\mathrm{SE}(\hat{\mu}_A)$	0.0596	0.0265	**0.0239**	0.0241	0.0244	0.0248	0.253
$\mathrm{SE}(\hat{\mu}_B)$	0.0063	**0.0061**	0.0061	0.0061	0.0062	0.0062	0.0062
$\mathrm{SE}(\hat{\pi}_C)$	0.0122	**0.0114**	0.0126	0.0138	0.0148	0.0159	0.016

they are so far from optimal that they would make differences between the other choices for r difficult to observe.

We see from Figure 3, using $r = 5$ is optimal or close to optimal for estimating μ_A in all cases considered. Using five repeated measurements is optimal in 53% of cases tested, within 2% of optimal in 97% of cases, and is always within 3% of optimal. Using 6 repeated measurements also works well but, with 4 repeated measurements, we lose quite a lot of precision for estimating compared with the optimal choice. Looking at similar results for the other parameters suggests that, for estimating μ_B and π_C, $r = 4$ is best but $r = 5$ also works well. Generally, the optimal choice of r depends on the unknown parameter values. Because we feel μ_A is the most important parameter and typically the worst estimated, we recommend $r = 5$. We see similar results for other values of nr.

4.3. Performance

To test the performance of the proposed plan as described in Section 4, we conducted another factorial experiment with the levels of the parameters given in Table 4. For each combination of model parameter values, the standard deviations were calculated for the full-verification plan, the no-verification plan, and the proposed plan. From these three quantities, two performance measures are calculated as detailed in Figure 4. One compares the proposed partial-verification plan with the no-verification plan and the second compares the proposed plan with the full-verification plan. The performance for estimating μ_A is shown in Figure 5 using boxplots calculated over the 32 different combinations of parameter values. Asymptotic standard deviations were used for the full-verification plan and the proposed plan while standard deviations for the no-verification plan were estimated using simulation. In the simulation, for each combination of the parameter values, we generated 1000 phase I data sets and found the MLEs from the log likelihood $l_{\mathrm{I}}(\theta)$ of Eq. (1). Then we calculated the sample standard deviations of the 1000 values of each estimate. In Figure 5, we stratify the results by γ_A because changing the level of that parameter makes the most difference.

We see from Figure 5 that the proposed plan offers a large percentage reduction in standard deviation of $\hat{\mu}_A$ compared with the no-verification plan (left panel) and that it attains the majority of the potential improvement we could achieve by verifying *all* parts with the gold standard (right panel). The improvement in standard deviation with the proposed plan compared with the no-verification case is dramatic, with most cases seeing more than the 65% improvement obtained in the camshaft example. The standard deviation of $\hat{\mu}_A$ under the proposed plan is typically reduced to less than a third of that same standard deviation under the no-verification plan. Furthermore, on average, 97% of the possible reduction available through verification was attained using the proposed plan. The number of parts verified us-

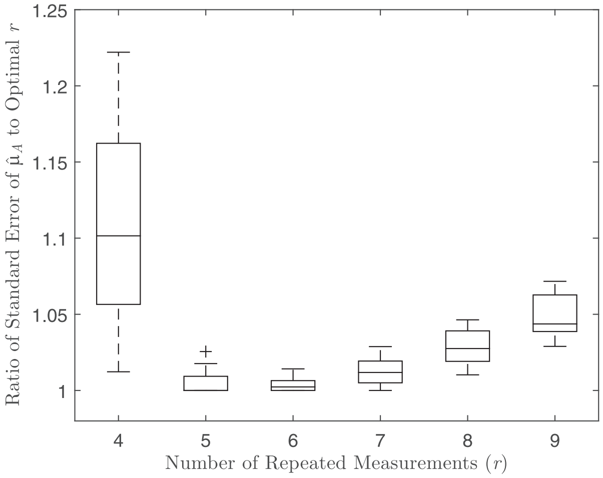

FIGURE 3. Optimal r Experiment: Factorial Experiment $nr = 2{,}500$ Run at All Combinations of μ_A, $\mu_B = 0.05$, 0.1; $\pi_C = 0.9$, 0.95; γ_A, $\gamma_B = 0.05$, 0.2 (see Table 4).

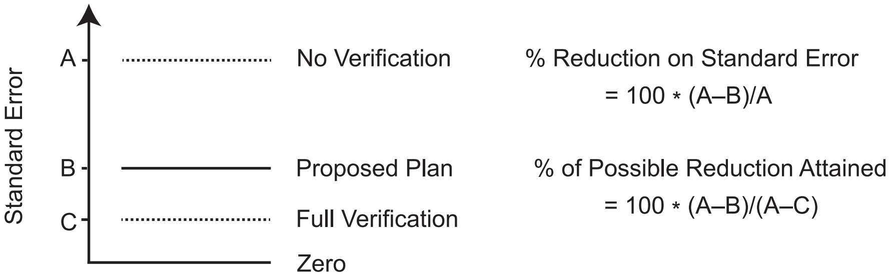

FIGURE 4. Performance Measures Summary.

ing the proposed plan varies and is affected by all five model parameters. However, because π_C is assumed large, μ_B plays the dominant role in determining how many parts fall into the two middle bins to be verified. For the grid of parameters values specified in Table 4, the proportion of parts verified averaged 9% when μ_B was 0.05 and 15% when μ_B was 0.1.

For the other parameters μ_B and π_C, the improvement is not as large but still dramatic. For the same experiment, the average improvement for estimating μ_B was 32%, which represented on average 96% of the possible improvement available through complete verification and for π_C the average improvement was 50%, which represented on average 99% of the improvement possible through full verification.

To summarize, the proposed plan gives large gains in precision over the no-verification plan for little additional cost. At the same time, it attains comparable performance to the full-verification plan while eliminating the majority of the cost and effort associated with using the gold standard. We expect these conclusions to be somewhat better as n increases because then relatively less effort is devoted to verifying parts from outside bins.

4.4. Robustness

One of the problems with the no-verification plan is that it is not robust to model misspecification. See Albert and Dodd (2008), who showed that, under two different random effects models, the plan with no verification had significant bias in estimating μ_A and μ_B.

In the recommended plan, five parts are verified from the four outside bins in addition to all parts in bins 2 and 3. These extra verifications were added to remove oddities in the likelihood surface that sometime occurred in our simulation when we did not include the additional parts from outside bins. When no verifications were taken from the other groups,

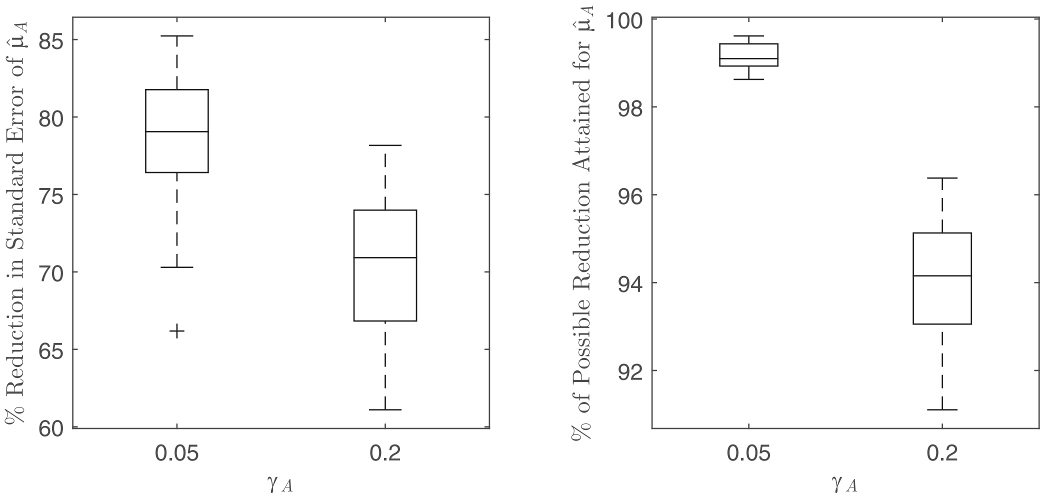

FIGURE 5. Proposed Plan Performance. Factorial experiment run at all combinations with $n = 500$ using proposed plan μ_A, $\mu_B = 0.05, 0.1$; $\pi_C = 0.9, 0.95$; γ_A, $\gamma_B = 0.05, 0.2$ (see Table 4). Left panel compares with no-verification plan, right panel compares with full-verification plan.

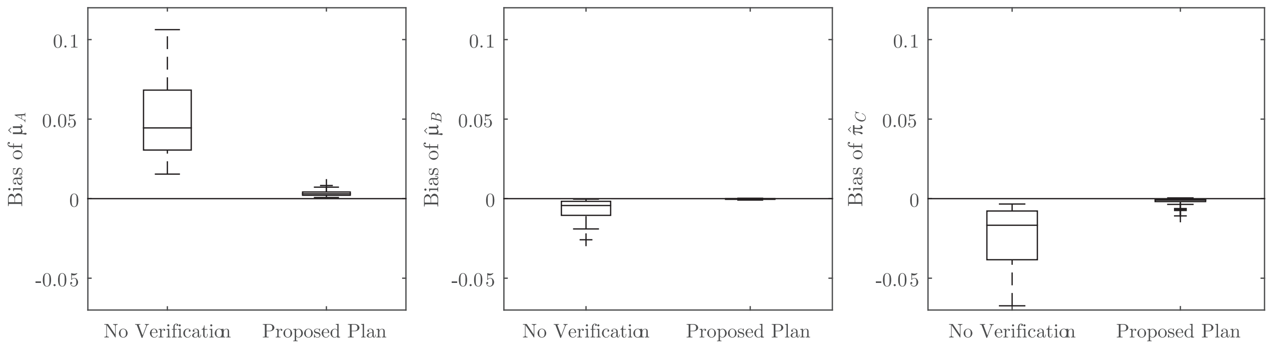

FIGURE 6. Gaussian Random Effect Model Bias Experiment Results. Factorial experiment run at all combinations with $n = 500$, μ_A, $\mu_B = 0.05, 0.1$; $\pi_C = 0.9, 0.95$; γ_A, $\gamma_B = 0.05, 0.2$ (see Table 4).

sometimes the MLE estimates would correspond to U-shaped beta distributions when this was not appropriate. Verifying a few observations from each of the bins makes the likelihood surface better behaved and eliminates these undesirable situations. This also improves the robustness properties of the recommended plan.

We conducted another simulation to demonstrate that the proposed plan has robustness properties similar to that of the full verification plan. Rather than use the beta-binomial random-effects model for the misclassification probabilities, we generated phase I and II data from the proposed plan using the Gaussian random effects (GRE) model developed by Qu (1996) over all combinations of the parameters in Table 4. The parameter values for the GRE were chosen to match the mean and variance of the corresponding beta distributions for each combination of parameter values. We calculated the MLEs from the log likelihood of Eq. (3) for 1000 replicates and the bias for each estimate. We also looked at the MLEs from the phase I data corresponding to the no-verification plan. The results are summarized in Figure 6.

We see large bias for the no-verification plan and negligible bias for the proposed plan. The proposed plan performs similarly to the full-verification plan that provides unbiased estimates of μ_A, μ_B, and π_C. At least for the GRE model, the proposed partial-verification plan is much less sensitive to model misspecification than the no-verification plan.

5. Summary and Discussion

The proposed plan has comparable performance and robustness to the full verification plan while eliminating the majority of the cost inherent in using the gold standard. This is possible because the amount of information gained in verifying different parts is not the same. It is better to repeatedly measure parts with the BMS and then verify only parts that have roughly equal number of passes and failures. The proposed plan is effective for a wide range of different parameter values.

While this plan stands on its own, there is room for further study and possible extensions. One possible extension is to include baseline information and use conditional sampling as in Danila et al. (2012). This can be an especially useful alteration to the plan when the conforming rate π_C is very high.

Appendix: Justification of Asymptotic Results

The purpose of this appendix is to assess the reliability of the asymptotic results for the proposed plan at different sample sizes and sets of parameter values. Determining a closed-form expression for the maximum-likelihood estimates or their standard deviations is not feasible. This paper uses asymptotic variance results due to Fisher (1925). We conducted a factorial experiment with six factors: sample size, n, and all 5 model parameters. For each of the $3 \cdot 2^5$ combinations, we generated 1000 data sets (phase I and phase II) using the beta-binomial model described in Section 2. Parts were selected and verified according to the proposed plan described in Section 4. For each data set, the parameters were estimated using maximum likelihood. The bias and standard deviations of the 1000 values of the estimates were calculated and recorded for comparison with the asymptotic approximation of the standard deviations based on the inverse of the expected information matrix. Figure A1 shows the ratio of the simulated standard deviations to the asymptotic approximation of the standard deviation for each combination of parame-

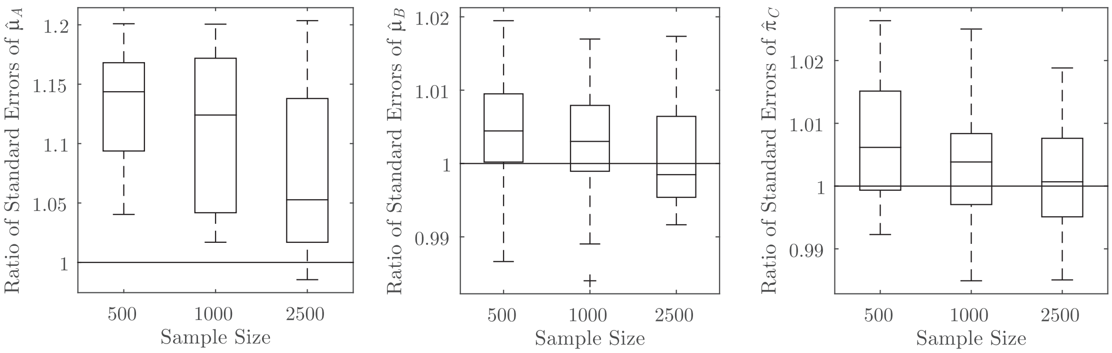

FIGURE A1. Ratio of Simulated and Asymptotic Standard Deviations for $\hat{\mu}_A$ (left), $\hat{\mu}_B$ (centre) and $\hat{\pi}_C$ (right). μ_A, μ_B = 0.05, 0.1; π_C = 0.9, 0.95; γ_A, γ_B = 0.05, 0.2 (see Table 4).

ter values. The results are separated by sample size, n. Thus, each box represents 32 combinations of parameter values.

The ratios are typically close to 1, indicating that the asymptotic variance is a reasonable approximation. Only in the case of $\hat{\mu}_A$ does the asymptotic approximation underestimate the simulated standard deviations. The approximations are sufficiently accurate for the manner in which they are used; that is, to make broad conclusions, not to calculate precise standard error estimates.

Acknowledgments

The authors would like to thank the Editor, an Associate Editor, and an anonymous reviewer for making insightful comments that improved the paper. This research was supported, in part, by a research grant (#105240) from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Albert, P. S. and Dodd L. E. (2004). "A Cautionary Note on the Robustness of Latent Class Models for Estimating Diagnostic Error Without a Gold Standard". *Biometrics* 60, pp. 427–435.

Albert, P. S. and Dodd L. E. (2008). "On Estimating Diagnostic Accuracy from Studies with Multiple Raters and Partial Gold Standard Evaluation". *Journal of the American Statistical Association* 103, pp. 61–73.

Beavers, D. P.; Stamey, J. D.; and Bekele, B. N. (2011). "A Bayesian Model to Assess a Binary Measurement System When No Gold Standard System Is Available". *Journal of Quality Technology* 43, pp. 16–27.

Begg, C. B. and Greenes, R. A. (1983). "Assessment of Diagnostic Tests When Disease Verification Is Subject to Selection Bias". *Biometrics* 39, pp. 207–215.

Boyles R. A. (2001). "Gauge Capability for Pass–Fail Inspection". *Technometrics* 43(2), pp. 223–229.

Burke, R. J.; Davis, R. D.; Kaminsky, F. C.; and Roberts, A. E. P. (1995). "The Effect of Inspector Errors on the True Fraction Non-Conforming: An Industrial Experiment". *Quality Engineering* 7, pp. 543–550.

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

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

Danila, O.; Steiner, S. H.; and MacKay, R. J. (2013). "Assessing a Binary Measurement System with Varying Misclassification Rates When a Gold Standard Is Available". *Technometrics* 55, pp. 335–345.

De Groot, J. A. H.; Janssen, K. J. M.; Zwinderman, A. H.; Bossuyt, P. M. M.; Reitsma, J. B.; and Moons, K. G. M., (2011). "Correcting for Partial Verification Bias: A Comparison of Methods". *Annals of Epidemiology* 21, pp. 139–148.

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

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

Fisher, R. A. (1925). "Theory of Statistical Estimation". *Proceedings of the Cambridge Philosophical Society* 22, pp. 700–725.

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 edition. New York, NY: Oxford University Press Inc.

Pepe, M. S. and Janes, H. (2007). "Insights into Latent Class Analysis of Diagnostic Test Performance". *Biostatistics* 8, pp. 474–484.

Qu, Y.; Tan, M.; and Kutner, M. H. (1996). "Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests". *Biometrics* 52, pp. 797–810.

Severn D. (2016). "Assessing Binary Measurement Systems Using Partial Verification with a Gold Standard". Ph.D. thesis (in progress), Department of Statistics and Actuarial Science, University of Waterloo.

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