Assessing a Binary Measurement System with Varying Misclassification Rates Using a Latent Class Random Effects Model

OANA DANILA, STEFAN H. STEINER, and R. JOCK MACKAY

University of Waterloo, Waterloo, N2L 3G1, Canada

When no gold standard measurement system is available, we can assess a binary measurement system by making repeated measurements on a random sample of parts and then using a latent class model for the analysis. However, there is widespread criticism of the model assumptions that, given the true state of the part, the repeated measurements are independent and have the same misclassification probability. We propose a latent class random effects model that relaxes these assumptions by modeling the distribution of the two misclassification rates with Beta distributions. We start by finding the likelihood, the maximum likelihood estimates (MLEs) and their approximate standard deviations with the standard assessment plan that selects parts at random from the process. However, to estimate the model parameters with reasonable precision, the standard plan requires extremely large sample sizes in the common industrial situation where the proportion of conforming parts is high and the misclassification probabilities are small. More realistic sample sizes are possible when we instead sample randomly from the population of previously failed parts and supplement the likelihood with baseline information on the overall pass rate. We show using simulation that, for feasible designs, the asymptotic standard deviation based on the expected information provides a reasonably close approximation to the simulated standard deviation. We then use these approximations to investigate how the properties of the MLEs for the unknown parameters depend on the baseline size, the number of parts in the sample, and the number of repeated measurements per part.

Key Words: Beta Binomial; Binary Measurement Systems; Latent Class Model; Misclassification Rates; Random Effects.

Introduction

Binary measurement systems (BMSs) are commonly used as diagnostic tools in medicine and inspection systems in industry. Understanding their properties is essential to making correct decisions with these systems. Here we adopt industrial language. To establish the notation, each part is conforming or not, as indicated by the value of the random variable X, where

$$X = \begin{cases} 1 & \text{if the part is conforming} \\ 0 & \text{if the part is nonconforming.} \end{cases}$$

If the part is measured once by the BMS under study, we use the random variable Y to indicate the result of that inspection, where

$$Y = \begin{cases} 1 & \text{if the part passes inspection} \\ 0 & \text{if the part fails inspection.} \end{cases}$$

The characteristics of the process and the measure-

Dr. Danila is a recent Ph.D. graduate from the Department of Statistics and Actuarial Science at the University of Waterloo. Her email address is omdanila@math.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 senior member 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.

ment system can be given by

$$\begin{aligned}\alpha &= P(Y = 1 \mid X = 0)\\ \beta &= P(Y = 0 \mid X = 1)\\ \pi_C &= P(X = 1). \qquad (1)\end{aligned}$$

In this simple model, α is the consumer's risk, the proportion of nonconforming parts that pass the inspection and are presumably shipped to the customer. The parameter β is the producer's risk, the proportion of conforming parts that fail the inspection and lead to unnecessary rework or scrap. The parameter π_C is the proportion of conforming parts, a property of the underlying process, not the BMS. Many other metrics describing both the BMS and the process can be constructed from this basic model. For example, we have the proportion of passed parts, π_P, that is a function of both the measurement and manufacturing processes, where

$$\begin{aligned}\pi_P &= P(Y = 1 \mid X = 0)P(X = 0)\\ &\quad + P(Y = 1 \mid X = 1)P(X = 1)\\ &= \alpha(1 - \pi_C) + (1 - \beta)\pi_C. \qquad (2)\end{aligned}$$

In the typical industrial context we have π_C large (>0.9) and α and β small (<0.1), so that π_P is close to one. The methods we present apply generally; however, our primary goal is to look for feasible assessment plans that are effective for this realistic narrow range of parameter values.

We assess the measurement system by conducting a study to estimate the parameters α and β. There are three distinct cases. In the gold standard case, we suppose that there is an alternate measurement system available that can classify parts as conforming or not without error. In terms of the above notation, using the gold standard measurement system, we can determine for any part if $X = 1$ or $X = 0$. In this context, the basic study design is to measure a randomly selected sample of n parts once each with the gold standard system and $r \geq 1$ times with the BMS. These designs and their analyses have been studied by Danila et al. (2008), Farnum (1994) and Burke et al. (1995) in an industrial setting and Pepe (2003) in the medical context. Boyles (2001) considers a second case in which there is no gold standard but instead there is a second BMS with known statistical properties. Boyles calls this an anchored measurement system and suggests study plans that involve measuring a sample of parts repeatedly with each system. In the third case, we assume no gold standard nor anchored measurement system is available. In this context, the standard study plan is to measure a randomly selected sample of parts $r \geq 3$ times with the BMS. See, for example, Danila et al. (2010), Van Wieringen and De Mast (2008), Van Weiringen and Van den Heuvel (2005), and Boyles (2001) in an industrial context and Pepe (2003) and Walter and Irwig (1988) in a medical setting.

In this paper, we consider only the third case. To model the data from the BMS assessment study in this context, the usual approach (Danila et al. (2010)) makes the following assumptions:

- the misclassification rate for every conforming (noconforming) part is the same, that is β (α) represents the misclassification rate for *each* conforming (nonconforming) part,
- measurements made on different parts are independent, and
- given the value of X, repeated measurements on the same part are (conditionally) independent. That is, if we make r measurements on the same part modeled by $Y_1, Y_2, \ldots, Y_r$, we have

$$\begin{aligned}&P(Y_1 = y_1, \ldots, Y_r = y_r \mid X = x)\\ &\quad = \prod_{j=1}^{r} P(Y_j = y_j \mid X = x).\end{aligned}$$

With no gold standard available, the value of X is unknowable, and so making the above assumptions we have the so-called latent class model for r repeated measurements on the same part:

$$\begin{aligned}&P(Y_1 = y_1, \ldots, Y_r = y_r)\\ &\quad = P(Y_1 = y_1, \ldots, Y_r = y_r \mid X = 0)P(X = 0)\\ &\qquad + P(Y_1 = y_1, \ldots, Y_r = y_r \mid X = 1)P(X = 1)\\ &\quad = \alpha^s(1 - \alpha)^{r-s}(1 - \pi_C) + (1 - \beta)^s\beta^{r-s}\pi_C,\end{aligned}$$

where $s = \sum_{j=1}^{r} y_j$ is the number of times the part passes inspection. Note that with this model, the random variables $Y_1, \ldots, Y_r$ are not independent marginally. Van Wieringen and Van den Heuvel (2005) show that we need to assume further that $\alpha < 1 - \beta$ and $r \geq 3$ for the parameters of the model to be identifiable. For the standard plan, using the assumption that measurements on different parts are independent, we can then build the likelihood function and estimate the unknown parameters (Danila et al. (2010)). Since we assume α and β are constant, we call this approach the fixed effects model.

The assumptions underlying this model have been widely criticized. See for instance De Mast et al. (2011), Van Wieringen and De Mast (2008), Pepe

(2003), Fujisawa and Izumi (2000), and Vacek (1983). In many cases, it may be unreasonable to assume that the misclassification rates are the same over all nonconforming and conforming parts. Some conforming (nonconforming) parts may be harder to classify correctly than others. Nonconstant misclassification probabilities may arise, for instance, if there is another characteristic Z associated with each part so that $P(Y = 1 \mid X = x, Z = z)$ is not equal to $P(Y = 1 \mid X = x)$ for some z, i.e., when the probability of passing inspection depends on the value of z as well as the true conforming/nonconforming state, x. In this instance, note that if we assume that repeated measurements on a single part are independent, given both $X = x$ and $Z = z$, it is easy to show that the repeated measurements on the part, given $X = x$, are now dependent, contrary to the basic assumption.

We have organized the paper as follows. In the next section we propose a random effects model where we assume that the misclassification rates vary according to Beta distributions. We then show that with the standard sampling plan, in order to get reasonable estimates of the consumer's and producer's risks, we need an unrealistically large number of parts in the study. Next, we propose a conditional plan that samples heavily from the population of previously failed parts. With the conditional sampling plan, each of the n sampled parts is subsequently measured r times. In the analysis of the conditional plan, we supplement the resulting data using available baseline information on the pass rate. The conditional plan/analysis provides good estimates with feasible sample sizes. We use simulation to examine the performance of the Fisher information-based asymptotic approximations for the standard deviations of the estimates in the proposed plan. Then, we demonstrate how the precision of the parameter estimates changes as we vary the amount of baseline data, the number of parts selected and the number of repeated measurements on each part. We also examine the best choices for r and n when the total number of measurements $N = nr$ is fixed. Throughout, we compare the performance of the analysis assuming a fixed effects model, i.e., α and β do not vary across parts, when, in fact, the consumer's and producer's risks vary from part to part. Similarly, we also examine the performance of the analysis using the random effects model when α and β do not vary. Finally, we provide a brief discussion and a summary of the results in the paper.

Modeling the Varying Misclassification Rates

We adopt a random effects model to relax the assumption that α and β are constant for all nonconforming and conforming parts, respectively. That is, we suppose that for any randomly selected nonconforming part, the consumer's risk α has density $f(\alpha)$, $0 < \alpha < 1$. We also assume that, given $X = 0$ and α, repeated measurements on the part are independent so that

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

where $s = \sum_{j=1}^{r} y_j$. Similarly, for any conforming part, we assume that the producer's risk β has density $f(\beta)$, $0 < \beta < 1$ and, given $X = 1$ and β, repeated measurements on the part are independent so that

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

For any part, we can then determine the joint distribution of the observable $(Y_1, \ldots, Y_r)$ using the conditional distributions and assumptions given above. In this model, $Y_1, \ldots, Y_r$ given $X = x$ are not independent. The random effects model explicitly allows for variation in the producer's and consumer's risks within the sets of conforming and nonconforming parts.

As in Danila et al. (2011), who considered a similar model in the context of an available gold standard measurement system, we propose Beta distributions for α and β (Johnson et al. (1994)). We select Beta distributions for the convenience of the calculations and because they are highly flexible in shape. Assuming a Beta distribution, the probability density function (pdf) of α is

$$f(\alpha) = \frac{\alpha^{g_A - 1}(1-\alpha)^{h_A - 1}}{\mathrm{Beta}(g_A, h_A)}, \quad 0 < \alpha < 1, \quad (3)$$

where $\mathrm{Beta}(\cdot)$ is the Beta function. With this parameterization, the mean and variance of α are

$$E(\alpha) = \mu_A = \frac{g_A}{g_A + h_A},$$
$$\mathrm{Var}(\alpha) = \frac{\mu_A(1-\mu_A)}{g_A + h_A + 1} = \frac{\gamma_A}{\gamma_A + 1}\mu_A(1-\mu_A),$$

where $\gamma_A = (g_A + h_A)^{-1}$. Similarly we model the distribution of the βs, using a Beta distribution with parameters μ_B and γ_B. Note that, with this model, the consumer's and producer's risks are now $P(Y =$

$1 \mid X = 0) = \mu_A$ and $P(Y = 0 \mid X = 1) = \mu_B$, the mean misclassification rates. These are the parameters of primary interest in the BMS assessment.

Fujisawa and Izumi (2000) use a similar random effects model to address the issue of varying values of α and β over different units. However, in their model, they specify a value of both α and β for each part and assume that the joint distribution is Dirichlet. Their model is equivalent to ours with the additional constraint that $\gamma_A = \gamma_B$. In a somewhat more complex context where there are multiple measurement systems (or multiple operators), Qu et al. (1996) construct a random effects model to specify the joint distribution of $Y_1, \ldots, Y_r$. Using our notation, given $X = x$ and a latent variable $Z \sim N(0, 1)$, they assume $Y_1, \ldots, Y_r$ are conditionally independent with $P(Y_j = 1 \mid X = x, Z = z) = \Phi(a_x + b_x z)$ for $j = 1, \ldots, r$, where Φ is the distribution function of a standard normal random variable and a_x and b_x are additional parameters that need to be estimated. This formulation seems less direct than what we propose because of the introduction of the latent Z. Also, it is unclear how the normality assumption can be assessed. Finally, Dendukuri and Joseph (2001) use a fully Bayesian extension of the fixed effects model to deal with the case when $r < 3$, while Beavers et al. (2011) and Quinino et al. (2005) also use the Bayesian approach with the fixed effects model (i.e., when α and β do not vary from part to part).

The Beta distribution is flexible and allows a variety of shapes. However, some are unreasonable in this context. We eliminate u-shaped Beta pdfs where there is a relatively high probability of parts with misclassification probabilities close to 1. In terms of the parameters, we assume $h_A < 1$, $h_B < 1$ or equivalently $\mu_A + \gamma_A > 1$, $\mu_B + \gamma_B > 1$. We also assume that the chance that the misclassification rate for any conforming or nonconforming part is greater than 0.5 is small. Without an available gold standard measurement, having many sampled parts with misclassification rates greater than 0.5 will bias the estimates of the underlying parameters.

The Standard Plan

Suppose we employ the standard assessment plan in which we select a random sample of n parts from the process and measure each part r times with the BMS. Then, using the Beta pdfs as given in (3), for any part with s passes in the r repeated measurements, we have

$$\begin{aligned}
P(Y_1 = y_1, \ldots, Y_r = y_r) &\\
&= (1 - \pi_C) \int_{\alpha=0}^{1} \frac{\alpha^{s+g_A-1}(1-\alpha)^{r+h_A-s-1}}{\text{Beta}(g_A, h_A)} d\alpha \\
&\quad + \pi_C \int_{\beta=0}^{1} \frac{\beta^{r+g_B-s-1}(1-\beta)^{s+h_B-1}}{\text{Beta}(g_B, h_B)} d\beta \\
&= (1 - \pi_C) \frac{\text{Beta}(g_A + s, h_A + r - s)}{\text{Beta}(g_A, h_B)} \\
&\quad + \pi_C \frac{\text{Beta}(g_B + r - s, h_B + s)}{\text{Beta}(g_A, h_B)}, \qquad (4)
\end{aligned}$$

where $g_A = \mu_A/\gamma_A$, $h_A = (1-\mu_A)/\gamma_A$, $g_B = \mu_B/\gamma_B$, and $h_B = (1 - \mu_B)/\gamma_B$.

Note that the random effects model (4) depends only on the number of passes s in the r repeated measurements. There are $r + 1$ possible values for s, and the associated probabilities must add to one. The model has five parameters $(\mu_A, \mu_B, \gamma_A, \gamma_B, \pi_C)$, so to be identifiable we require $r \geq 5$. If $r = 4$, there are an infinite number of parameter values that give the same distribution for (Y_1, Y_2, Y_3, Y_4). From (4), it is clear that for identifiability we need the further constraint $\mu_A < 1 - \mu_B$. That is, we reasonably assume that the mean pass rate for nonconforming parts is less than the mean pass rate for conforming parts. This constraint is similar to the constraint $\alpha < 1 - \beta$ required for the fixed effects model. Note that if we hold μ_A and μ_B fixed and let γ_A and γ_B approach zero, $P(Y_1 = y_1, \ldots, Y_r = y_r)$ in (4) approaches $(1 - \pi_C)\mu_A^s(1 - \mu_A)^{r-s} + \pi_C(1 - \mu_B)^s \mu_B^{r-s}$, the fixed effects model with $\alpha = \mu_A$ and $\beta = \mu_B$. That is, the fixed effects model is a limiting case of the random effects model.

Combining the data across parts using the independence assumption, we have the log-likelihood

$$\begin{aligned}
l(\mu_A, \mu_B, \gamma_A, \gamma_B, \pi_C) &\\
&= \sum_{i=1}^{n} \ln \left[(1 - \pi_C) \frac{\text{Beta}(g_A + s_i, h_A + r - s_i)}{\text{Beta}(g_A, h_B)} \right. \\
&\qquad \left. + \pi_C \frac{\text{Beta}(g_B + r - s_i, h_B + s_i)}{\text{Beta}(g_A, h_B)} \right]. \qquad (5)
\end{aligned}$$

We can estimate the five parameters by maximizing (5) using a standard approach such as that used in Nelder and Mead (1965) that does not require gradients.

For high-quality processes with π_C close to 1, even when n is large, it is possible that the maximum likelihood estimate of π_C is $\hat{\pi}_C = 1$. That is, based on the likelihood, it appears that all parts are conforming. When this happens, the estimates of the five parameters are not unique. There is no information about μ_A and γ_A. Also, with the same data, we can set $\hat{\pi}_C = 0$ and then get the same maximum value for the likelihood function by varying μ_A, γ_A and ignoring μ_B, γ_B. In any application, if the output of the maximization routine is $\hat{\pi}_C = 1$, we suggest using the estimates $\hat{\pi}_C = 1$, $\hat{\mu}_B$, and $\hat{\gamma}_B$ and accept that there is no information about μ_A and γ_A. The same problem and suggested resolution apply to the fixed effects model.

We determined the Fisher (expected) information matrix from the second derivatives of the log-likelihood function (5) and used the square root of the diagonal elements of the inverse of the information matrix evaluated at the parameter estimates to get the approximate standard deviations of the estimates. We used Maple (2009) for the symbolic calculations and Matlab (2008) for the numerical calculations, and we will provide the Matlab code, on request, to fit the random effects model and to determine the approximate standard deviations.

We demonstrate the poor performance of the standard plan when π_C is close to one through a simulation study. We consider the plan parameters $r = 10$ and $n = 500$, 1000, 2000. We recognize that these sample sizes are impractical but note that the performance of smaller plans will be worse. For the model parameters, we use a factorial structure with μ_A, $\mu_B = 0.02$, 0.1, γ_A, $\gamma_B = 0.01$, 0.1, and $\pi_C = 0.8$, 0.9, 0.95. Each simulation run consists of 5000 trials. We exclude the estimates from any run in which $\hat{\pi}_C = 1$ in the calculation of the simulated standard deviations (this happens infrequently with the sample sizes and parameter values used in the simulation). In each run, we estimate the parameters from both the fixed and random effects models.

The complete results (not shown) suggest that, with the standard plan and large sample sizes, we can estimate μ_B and π_C well and γ_B reasonably well, with negligible bias and small standard deviations. However, estimating μ_A, a key parameter, and γ_A is problematic. Figure 1 shows the biases for $\hat{\mu}_A$ and $\hat{\alpha}$ (the estimate from the fixed effects model) stratified by the true values of μ_A, π_C, and n, while Figure 2 similarly shows the standard deviations. For $\hat{\mu}_A$, we see relatively large biases and standard deviations compared to the true small values of μ_A, especially with π_C large. The results for $\hat{\gamma}_A$ (not shown) are much worse. We also see in Figure 1 very large biases in $\hat{\alpha}$, the consumer's risk estimate from the fixed effects model. While not shown, there are also similar biases in the estimates of β and π_C with the fixed effects model.

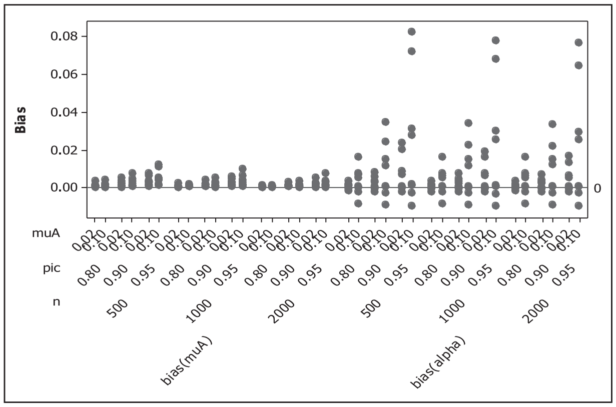

FIGURE 1. Simulated Bias of μ_A from the Random Effects Model and α from the Fixed Effects Model Across all Simulation Runs with the Standard Plan.

We draw two conclusions if the misclassification rates vary from part to part. First, for high-quality processes with π_C close to one, the standard assessment plan is not practical. We require huge sample sizes to produce useful and reliable estimates of the primary parameter μ_A. The problem occurs because we need n very large in order to get sufficient non-

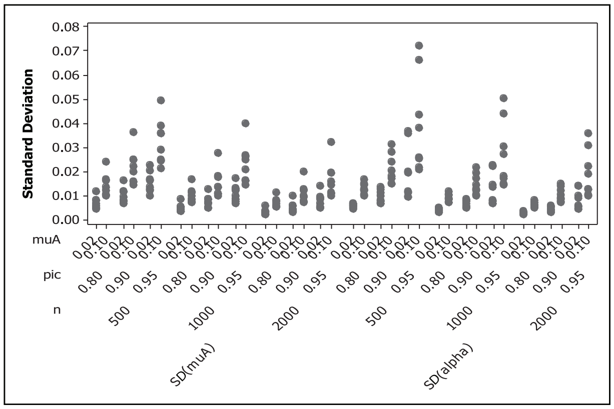

FIGURE 2. Simulated Standard Deviation of μ_A from the Random Effects Model and α from the Fixed Effects Model Across all Simulation Runs with the Standard Plan.

conforming parts in the sample. Second, we conclude that using the fixed effects model with the standard plan can produce badly biased estimates of the consumer's and producer's risks.

Conditional Sampling and Baseline Data

To address the sample size issue, we explore conditional sampling plans (Danila et al. (2008, 2010, 2011)) in which we sample parts randomly from the populations of previously passed and/or failed parts. We develop general results for sampling from both the population of previously failed and the population of previously passed parts. However, in our context with π_C close to 1, it is best to sample only from the population of previously failed parts. Conveniently, parts that fail inspection are often readily available as they are segregated for scrap or rework. In situations in which conditional sampling is feasible, (i.e., when we have populations of previously passed and failed parts) the pass rate is also usually recorded by hour, shift, or some other fixed time period. We assume that such pass rate baseline data are available, and we include this data in the likelihood and, thus, the analysis.

With the conditional sampling plan, we sample n_0 and n_1 parts at random from the previously failed and passed parts, respectively, where $n_0 + n_1 = n$. We let f denote the sampling proportion of previously passed parts, that is $n_1 = fn$, so $f = 0$ implies that we sample only from the previously failed parts. Then we measure each selected part r times with the BMS. Let $Y_0 = y_0$ indicate the (initial) measurement for any inspected part. With conditional sampling, the contribution to the likelihood of any part that passes s times in the assessment study is

$$
\begin{aligned}
&P(S = s \mid Y_0 = y_0) \\
&\quad = P(S = s, Y_0 = y_0)/P(Y_0 = y_0) \\
&\quad = \left[(1-\pi_C) \int_{\alpha=0}^{1} \frac{\alpha^{s+y_0+g_A-1}(1-\alpha)^{r+h_A-s-y_0-1}}{\mathrm{Beta}(g_A, h_A)} d\alpha + \pi_C \int_{\beta=0}^{1} \frac{\beta^{r+g_B-s-y_0-1}(1-\beta)^{s+y_0+h_B-1}}{\mathrm{Beta}(g_A, h_A)} d\beta \right] \\
&\qquad \div P(Y_0 - y_0) \\
&\quad = \left[(1-\pi_C) \frac{\mathrm{Beta}(g_A + s + y_0, h_A + r - s - y_0)}{\mathrm{Beta}(g_A, h_A)} + \pi_C \frac{\mathrm{Beta}(g_B + r - s - y_0, h_B + s + y_0)}{\mathrm{Beta}(g_B, h_B)} \right] \\
&\qquad \div \left(\pi_P^{y_0} (1-\pi_P)^{1-y_0} \right) \\
&\quad = W(s, y_0) / \left(\pi_P^{y_0} (1-\pi_P)^{1-y_0} \right), \qquad (6)
\end{aligned}
$$

where $W(s, y_0)$ is the factor in the square brackets, a function of r and the model parameters. With the random effects model, similar to (2), $\pi_P = \mu_A(1 - \pi_C) + (1 - \mu_B)\pi_C$ is the marginal probability that a randomly selected part passes inspection.

Since measurements on different parts are independent, the log-likelihood for the measurements made in the assessment study is the sum of the logarithm of terms like (6) for each selected part. Suppose we also have independent baseline data where in m inspections, there are u passed parts. These data contribute the additional term $u\log(\pi_P) + (m-u)\log(1-\pi_P)$ to the log-likelihood, if we assume π_P is constant over the time of the collection of the baseline data. The overall log-likelihood for the conditional sampling plan augmented with baseline data is thus

$$
\sum_{i=1}^{n_0} \ln[W(s_i, 0)] + \sum_{i=1}^{n_1} \ln[W(s_i, 1)] + (u - n_1)\ln[\pi_P] + (m - u - n_0)\ln[1 - \pi_P] \quad (7)
$$

where s_i is the number of passes out of r measurements for the ith part sampled from either the population of previously passed or failed parts. The maximum likelihood estimates of the parameters in (7) must be found numerically. Again, we determine the approximate standard errors using the Fisher information matrix. Upon request we will provide Matlab (2008) code to produce the estimates and their approximate standard errors for the conditional sampling plan.

Example

The context is real; the data are not. An automated inspection system for credit card blanks rejects cards for many reasons, so we expect the misclassification rates to vary from blank to blank. Onerous manual inspection provides a gold standard sys-

TABLE 1. Number of Parts with s Passes in Sample

s	0	1	2	3	4	5	6	7	8	9	10
Frequency	37	26	3	3	2	1	6	14	11	42	55

tem. Here, to see how the proposed methodology works and to avoid using the gold standard, $n = 200$ cards were selected haphazardly from the recent rejects and re-inspected $r = 10$ times each, with the results given in Table 1. As well, $u = 1734$ of the last $m = 2000$ cards checked passed the inspection.

Fitting the log-likelihood (7) gives the following MLEs and corresponding asymptotic standard errors (in parentheses): $\hat{\mu}_A = 0.069$ (0.0125), $\hat{\gamma}_A = 0.033$ (0.0337), $\hat{\mu}_B = 0.084$ (0.0063), $\hat{\gamma}_B = 0.038$ (0.0136), and $\hat{\pi}_C = 0.95$ (0.0056). Based on these estimates we get $\hat{\pi}_P = 0.874$, which closely matches the estimate available from the baseline data (1734/2000). Note that sampling only previously failed parts gives a proportion of conforming parts in the sample equal to $P(X = 1 \mid Y = 0) = \mu_B \pi_C / (1 - \pi_P)$. Plugging in the estimates $\hat{\mu}_B$, $\hat{\pi}_C$, and $\hat{\pi}_P$, this proportion is estimated as 0.63. This seems reasonable given the data in Table 1 since, with small misclassification rates, we expect that any part that yielded more than five passes when measured 10 times is conforming and the proportion of such parts in the sample is 0.645 (129/200).

More generally, to examine how well the conditional sampling plan works, we conducted another simulation study. We use $r = 10$ since this is the number of repeated measurements recommended in a subsequent section. We investigate large but feasible plans with $n = 100$, 250, and 500 parts sampled from the set of failed parts, i.e., $f = 0$. We used a baseline sample of $m = 1000$ parts. We varied the model parameters in a factorial structure with μ_A, $\mu_B = 0.02$, 0.1, γ_A, $\gamma_B = 0.01$, 0.1, $\pi_C = 0.9$, 0.95. Each simulation run consists of 5000 trials, and in each run we estimate the parameters of the fixed and random effects models. To fit the fixed effect model we use the model and methods described in Danila et al. (2010).

Figure 3 shows the bias of the estimates from both the random and fixed effects model using the conditional sampling plan. The estimates from the fixed effects model are badly biased and should not be used if there is a suspicion that the misclassification rates vary. We found in another simulation (results not shown) that if the data are generated from a fixed effects model, there is little loss of efficiency using the random effects estimates for μ_A and μ_B. We also see in Figure 3 that, with the random effects model, we get unbiased estimates for μ_B, π_C, and γ_B if the sample size is large. There are, however, relatively large biases in the estimates of μ_A and, especially, γ_A. More detailed exploration of the causes of the observed large biases for μ_A and γ_A reveals that for some runs of the simulation, $\hat{\gamma}_A$ was surprisingly large. Recall that we imposed the constraint $\mu_A + \gamma_A < 1$ to avoid u-shaped Beta distributions for the random effects. To deal with the large values of $\hat{\gamma}_A$, we tried strengthening the constraint to $\mu_A + \gamma_A \leq 0.5$ which reduces the probability of getting a nonconforming part with a very large misclassification rate. However, we did not solve the bias issue with this stronger constraint. We still had large values for $\hat{\gamma}_A$ and in many runs the estimates fell on the boundary $\hat{\mu}_A + \hat{\gamma}_A = 0.5$. We propose to handle these cases by refitting the random effects model with the further constraint that $\gamma_A = \gamma_B = \gamma$. The result of this proposal is that, when needed, we borrow strength from the better performing estimate $\hat{\gamma}_B$. Using a model with a common γ is similar to the so-

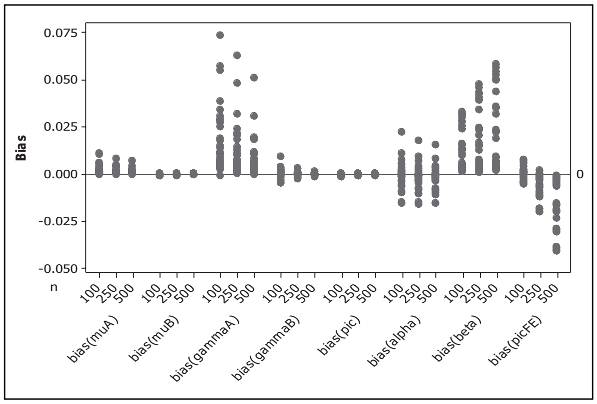

FIGURE 3. Biases of the Estimates from the Random and Fixed Effects Models with the Conditional Plan pic, picFE (estimates of π_C from random and fixed effects models, respectively).

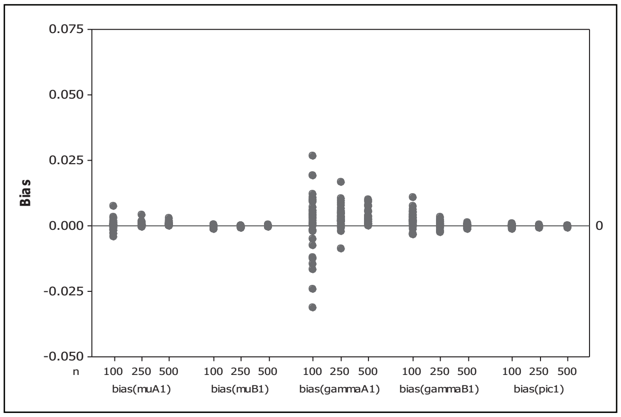

FIGURE 4. Biases of the Ad Hoc Estimates from the Random Effects Model.

called 2LCR1 model described by Qu et al. (1996) and equivalent to the Direchlet model of Fujisawa and Izumi (2000).

Figure 4 gives simulated bias results using the proposed ad hoc method in which, if needed (i.e., when the standard MLEs hit the constraint $\hat{\mu}_A + \hat{\gamma}_A = 0.5$), we fit the random effects model with $\gamma_A = \gamma_B = \gamma$. Comparing Figures 3 and 4 we see that, with the ad hoc procedure, the biases are much smaller and better centered on zero than are those of the standard MLEs. In addition, unlike in Figure 3, we see clear reductions in the biases as the sample size increases. In Figure 5, we show the standard deviations for the estimates from both the MLEs and the ad hoc procedure (denoted by an additional 1 in their labels). We see that the ad hoc procedure improves the precision of the estimates for μ_A and γ_A and has virtually no effect on the precision of the other three estimates.

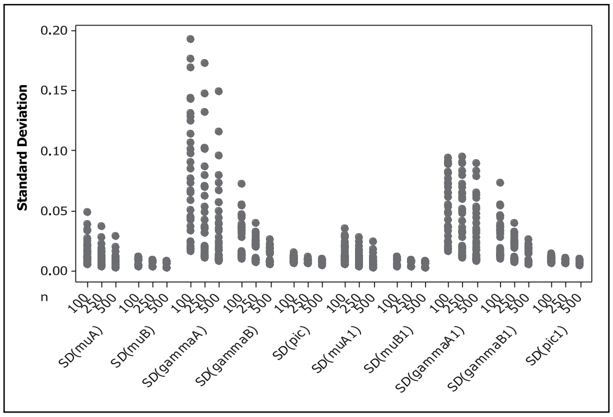

FIGURE 5. Standard Deviations for the Estimates from the Random Effects Model and Conditional Plan.

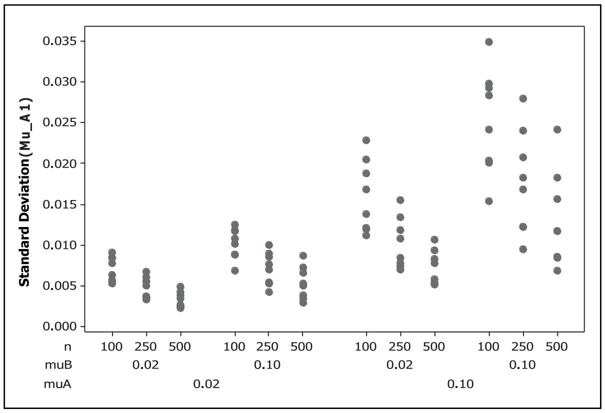

FIGURE 6. Standard Deviations of $\hat{\mu}_A$ with the Ad Hoc Procedure.

It is interesting to note that, despite sampling only from the population of previously failed parts, estimating μ_A is still more difficult than estimating μ_B, even though for some combinations of the parameters more nonconforming than conforming parts are selected. This occurs because with large values of π_C the baseline data provides much more information about μ_B than μ_A.

With the common γ as needed (i.e., ad hoc) approach and reasonable sample sizes, we can estimate the primary parameters μ_A, μ_B, and π_C fairly well. However, the biases and standard deviations for the measures of variability in the misclassification rates, i.e., $\hat{\gamma}_A$ and $\hat{\gamma}_B$, are still too large for these estimates to be reliable. Even with the conditional plan and the ad hoc estimation procedure, we need larger sample sizes to estimate these two parameters, especially γ_A.

Next, we focus further investigation on $\hat{\mu}_A$, the least precise estimate of the three primary parameters. In Figure 6, we show that with the ad hoc procedure the standard deviations for $\hat{\mu}_A$ are larger when μ_B and, especially, when μ_A are larger. We need a sample size of at least $n = 250$, so that the standard deviation of $\hat{\mu}_A$ is less than half the actual parameter value in the worst case.

We hope to use standard likelihood results to get approximate standard deviations of the estimates as derived from the expected (Fisher) information matrix, both for analysis of a particular application and for planning purposes. In Figure 7, we plot the ratio of simulated over asymptotic approximate standard

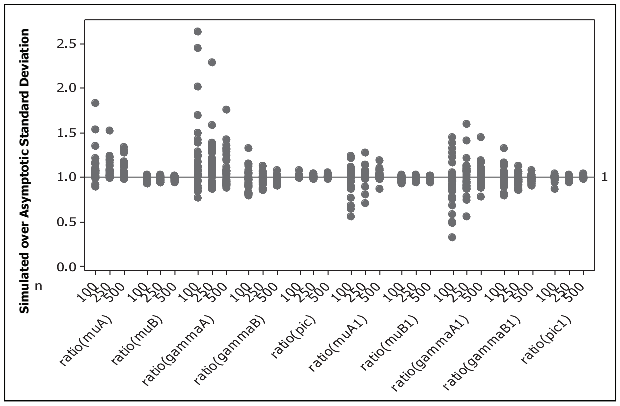

FIGURE 7. Ratios of Simulated over Asymptotic Standard Deviations with Conditional Sampling Plan and Baseline.

deviation for both the standard MLEs and the ad hoc maximum likelihood estimates. The ratios are close to one for μ_B and π_C. There is, however, considerably more variation for μ_A. The ratios are worst for small μ_A and large μ_B. While there is not perfect agreement between the asymptotic and simulated standard deviations, the asymptotic approximations can be used to get approximate confidence intervals for the parameters and to provide a guide to compare plans.

Planning a Conditional Sampling BMS Assessment Study

In this section, we consider planning an assessment study. Since we are assuming π_C is large, we look only at the case $f = 0$ (i.e., we sample only from previously failed parts). To investigate the effects of changing n, r, and m, we consider the cases $m = 100$, 1000, 10,000; $r = 5$, 10, 15; and $n = 100$, 250, 500 for a range of values of the model parameters. We present the results, based on asymptotic standard deviations from the Fisher information calculated from the log-likelihood (7), in Figures 8, 9, and 10 with $\pi_C = 0.95$, $\mu_A = 0.05$, $\mu_B = 0.05$, $\gamma_A = 0.10$, $\gamma_B = 0.10$. In each plot, we show how the asymptotic standard deviations for $\hat{\mu}_A$, $\hat{\mu}_B$, $\hat{\gamma}_B$, and $\hat{\pi}_C$ change. We do not show the plots for $\hat{\gamma}_A$ because for the smaller sample sizes the estimate

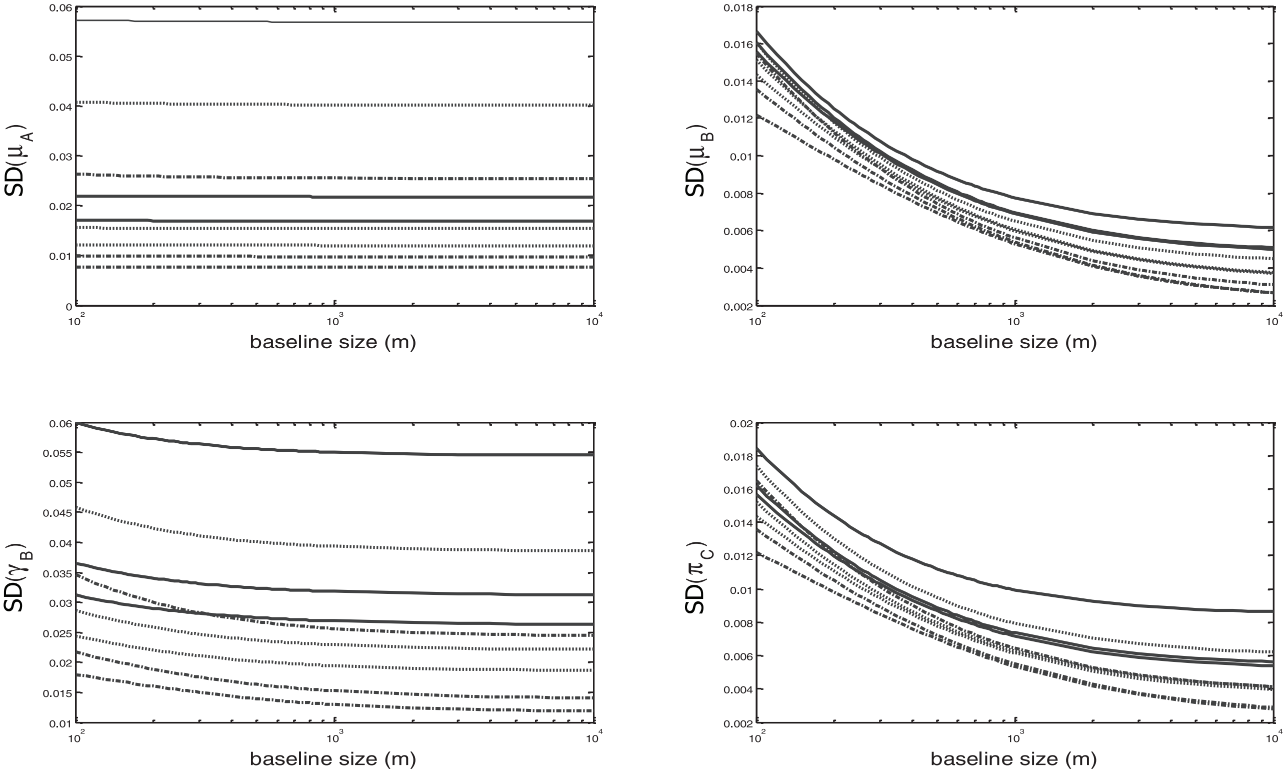

FIGURE 8. Effect of Increasing m on the Asymptotic Standard Deviations of $\hat{\mu}_A$, $\hat{\mu}_B$, $\hat{\gamma}_B$, and $\hat{\pi}_C$ for $r = 5$, 10, 15; $n = 100$ (solid lines), $n = 250$ (dotted lines), $n = 500$ (dot-dashed lines); $\pi_C = 0.95$, $\mu_A = 0.05$, $\mu_B = 0.05$, $\gamma_A = 0.10$, $\gamma_B = 0.10$. Note that lower lines of fixed type correspond to increasing values of r.

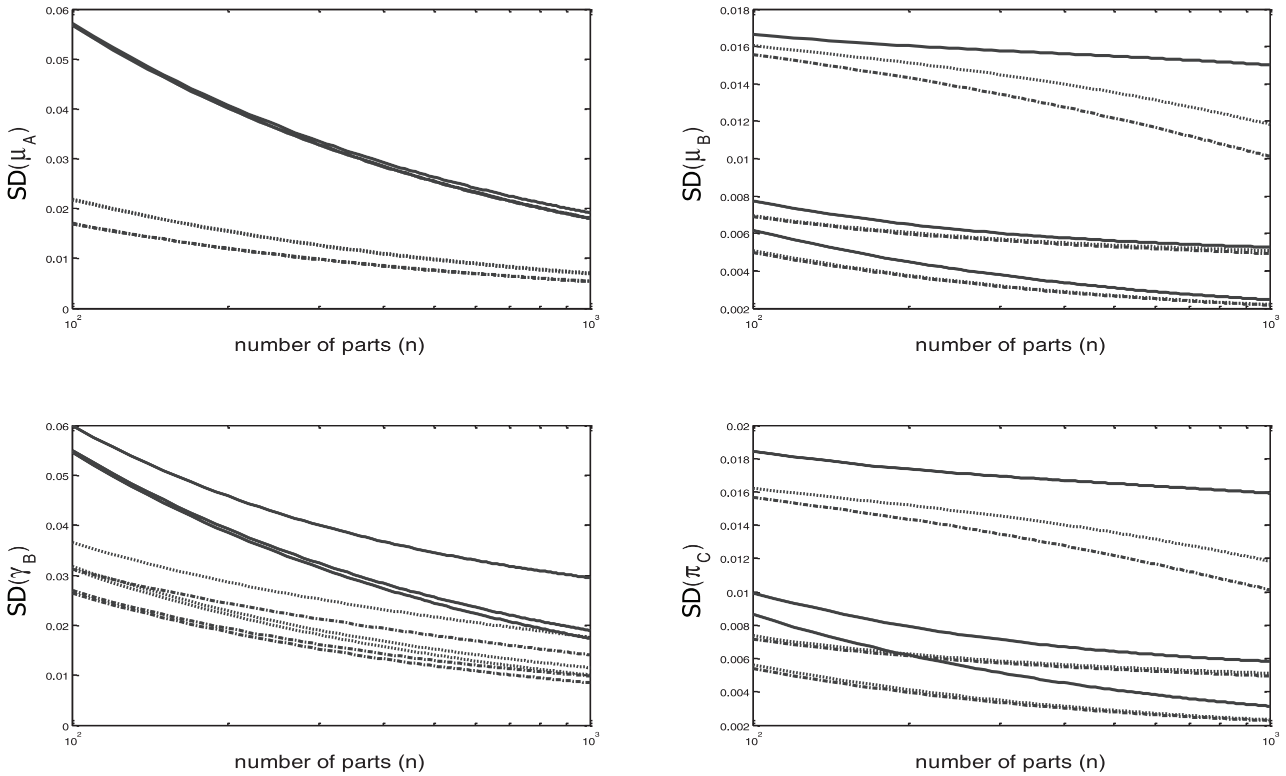

FIGURE 9. Effect of Increasing n on the Asymptotic Standard Deviations of $\hat{\mu}_A$, $\hat{\mu}_B$, $\hat{\gamma}_B$, and $\hat{\pi}_C$ for $m = 100$, 1000, 10,000; $r = 5$ (solid lines), $r = 10$ (dotted lines), $r = 15$ (dot-dashed lines); $\pi_C = 0.95$, $\mu_A = 0.05$, $\mu_B = 0.05$, $\gamma_A = 0.10$, $\gamma_B = 0.10$. Lower lines of the same type correspond to increasing values of m.

is somewhat biased, the standard deviations are large, and the asymptotic approximation can be poor. Note the log scale for the horizontal axis in Figure 8 and that the scale of the vertical axis varies from plot to plot. To create Figure 8, we consider a wide range of values of m, for each of the nine combinations of the other two design parameters n and r. We see that increasing the baseline size has no effect on the asymptotic standard deviations of $\hat{\mu}_A$, a small positive effect on the standard deviation of $\hat{\gamma}_B$, and a dramatic positive effect on the standard deviations of $\hat{\mu}_B$ and $\hat{\pi}_C$, for all values of n and r, with reductions in standard deviation of more than 70% possible. As the baseline size, m, goes to infinity, the standard deviations approach a positive limit, corresponding to the situation when the pass rate π_P is known. We see similar results for other values of the model parameters. Since the baseline data are freely available, we strongly recommend their inclusion in the plan and analysis of the BMS assessment study.

Figure 9 explores the effect of increasing the sample size, n. Note that, while each part in the study contributes the same information, the asymptotic standard deviations are generally not simply $1/\sqrt{n}$ times a function of r, due to the effect of the baseline data. However, for $\hat{\mu}_A$ and $\hat{\gamma}_A$, the $1/\sqrt{n}$ rule works as a close approximation since the baseline information has very little effect on estimates for these two parameters. However, for $\hat{\mu}_B$ and $\hat{\pi}_C$, the standard deviations decrease only slowly with n; while, for $\hat{\gamma}_B$, increasing either n or m substantially improves the precision, baseline sizes larger than 1000 provide little benefit.

Figure 10 shows the effect of changing r, the number of repeated measurements on each selected part. There are large gains in efficiency available by increasing r from its minimum value of five when estimating μ_A or γ_B, but not so much for μ_B or π_C. Also note that, as r approaches infinity with n fixed, unlike with the fixed effects model (Danila et al. (2010)), we do not get consistent estimates of μ_A and μ_B (i.e., their standard errors do not go to zero) since there is a finite number of parts in the investigation.

Next, we use the asymptotic results to look at the

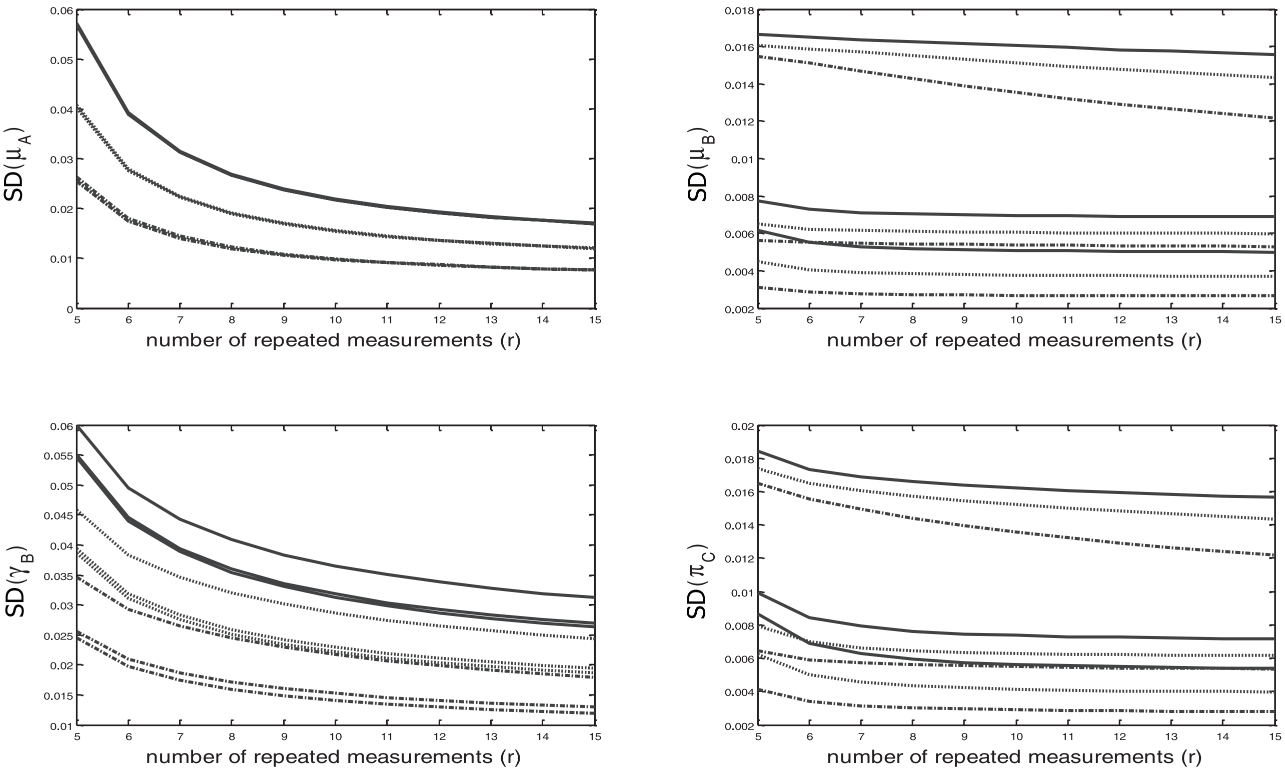

FIGURE 10. Effect of Increasing r on the Asymptotic Standard Deviations of $\hat{\mu}_A$, $\hat{\mu}_B$, $\hat{\gamma}_B$, and $\hat{\pi}_C$ for $m = 100$, 1000, 10,000; $n = 100$ (solid lines), $n = 250$ (dotted lines), $n = 500$ (dot-dashed lines); $\pi_C = 0.95$, $\mu_A = 0.05$, $\mu_B = 0.05$, $\gamma_A = 0.10$, $\gamma_B = 0.10$. Note that lower lines of fixed type correspond to increasing values of m.

relative importance of n and r for estimating μ_A and μ_B if the total number of measurements $N = nr$ is fixed. Here we assume the baseline sample size m is determined separately. We see in Table 2 that, when $N = 2500$ (and $m = 1000$), there are significant gains for estimating μ_A if we increase r at the expense of n, especially when μ_B is large. However, for estimating μ_B, usually the minimum number of repeated measurements is optimal, except when μ_B is large and γ_B is small; even in this case, $r = 5$ is close to optimal. In general, since the parameters are unknown, we recommend $r = 10$ as a reasonable choice. Table 2 also includes columns showing the ratio of the standard deviations for $r = 10$ and the optimal value of r. Qualitatively, we see similar results to those in Table 2 for other values of γ_A and π_C.

Discussion and Conclusions

In this paper, we consider the assessment of a binary measurement system when no gold standard system is available. We concentrate on the industrial context in which it is likely that the misclassification probabilities are small and the overall conforming rate is close to one. We investigate a random effects model that relaxes the assumption that the misclassification probability is the same for all conforming (nonconforming) parts. We make the important observation that, if the data are generated from the random effects model, the estimates of the (average) misclassification probabilities obtained by fitting the standard fixed effects model can be seriously biased. Furthermore, the loss of precision in fitting the random effects model to fixed effects data is small.

We first apply the model to the standard assessment plan, where each part in a random sample of parts is measured a number of times. We show that to estimate the average misclassification probabilities with reasonable precision we need a sample of parts that is so large as to be impractical. We also need substantially more repeated measurements on each part than required to make the model identifiable (i.e., $r > 5$). To reduce the sample size, we propose using conditional sampling from the population of previously failed parts, supplemented by baseline data. For some samples for which the estimates seem unreasonable, we adopt an ad hoc pro-

TABLE 2. Optimal Value of r for Estimating μ_A and μ_B When $N = nr = 2500$, $\gamma_A = 0.05$, $\pi_C = 0.95$, $m = 1000$, $f = 0$

μ_A	μ_B	γ_B	Best r for μ_A	Best SD($\hat{\mu}_A$)	SD($\hat{\mu}_A$) at best r over $r = 10$	Best r for μ_B	Best SD($\hat{\mu}_B$)	SD($\hat{\mu}_B$) at best r over $r = 10$
0.02	0.02	0.01	7	0.0042	0.9758	5	0.0028	0.9015
0.02	0.02	0.10	11	0.0047	0.9973	5	0.0029	0.9051
0.02	0.05	0.01	9	0.0051	0.9940	5	0.0054	0.9678
0.02	0.05	0.10	14	0.0060	0.9658	5	0.0055	0.9446
0.02	0.10	0.01	11	0.0066	0.9980	16	0.0076	0.9934
0.02	0.10	0.10	17	0.0079	0.8667	6	0.0083	0.9740
0.05	0.02	0.01	7	0.0066	0.9798	5	0.0028	0.9054
0.05	0.02	0.10	11	0.0074	0.9968	5	0.0029	0.9139
0.05	0.05	0.01	9	0.0082	0.9997	5	0.0054	0.9702
0.05	0.05	0.10	14	0.0095	0.9633	5	0.0056	0.9511
0.05	0.10	0.01	12	0.0105	0.9959	17	0.0076	0.9902
0.05	0.10	0.10	17	0.0128	0.8720	6	0.0084	0.9785
0.10	0.02	0.01	8	0.0094	0.9920	5	0.0029	0.9127
0.10	0.02	0.10	13	0.0108	0.9850	5	0.0030	0.9386
0.10	0.05	0.01	10	0.0118	1	5	0.0055	0.9748
0.10	0.05	0.10	16	0.0140	0.9345	6	0.0057	0.9671
0.10	0.10	0.01	13	0.0153	0.9720	17	0.0076	0.9850
0.10	0.10	0.10	21	0.0194	0.8250	7	0.0085	0.9888

cedure that assumes a common variance multiplier γ for both conforming and nonconforming parts. This procedure reduces the number of parameters in the model and borrows strength from the baseline data. Conditional sampling and the corresponding analysis provide reasonable parameter estimates with a moderate sample size and a feasible number of repeated measurements on each selected part.

We can think of several other ad hoc procedures to deal with the difficulty of estimating μ_A and γ_A. If the estimates are unreasonable, we can collect additional data by sampling additional parts or by remeasuring parts with observed pass rates close to 0.5. If a gold standard is available but expensive, as in the credit card example, we can instead use the gold standard to classify these problematic parts. The likelihood (7) requires adjustment in this case. We have not investigated either of these alternatives.

To use the conditional sampling assessment plan we need to choose n, r, and f, as we assume the baseline size, m, is given. We recommend $f = 0$ (i.e., sampling only from the failed parts) since this increases the expected number of nonconforming parts in the study, and failed parts are usually readily available. Our results suggest that it is desirable to have more nonconforming than conforming parts in the sample. We still get precise estimates for μ_B due to the baseline information. To choose appropriate values for the number of parts n and the number of repeated measurements r, we provide Matlab (2008) code that determines the asymptotic standard errors for the five parameters, given initial guesses. This code can be used to select a conditional plan to meet any desired precision goals for all the estimates other than $\hat{\gamma}_A$, as long as the number of parts is not very small (say $n > 250$) so that the asymptotic approximations are applicable.

In the analysis, we suppose that both the BMS and the underlying process are stable, that is, the model parameters do not change during the measurement system assessment study. This begs the question "How much baseline data should we use?" More is better for estimating μ_B and π_C, as shown in Figure 8, but we need to be careful. We do not want to use baseline data from a time period in which any of the parameters were substantially different. To protect against this we recommend examining the stability of the baseline data using statistical process control techniques (Montgomery (1996)).

References

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.

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

Burke, J. R.; Davis, R. D.; Kaminsky, F. C.; and Roberts, A. E. P. (1995). "The Effect of Inspector Errors on the True Fraction Nonconforming: An Industrial Experiment". *Quality Engineering* 7(4), pp. 543–550.

Danila, O., Steiner, S. H.; and MacKay, R. J. (2008). "Assessing a Binary Measurement System". *Journal of Quality Technology* 40(3), pp. 310–318.

Danila, O., Steiner, S. H.; and MacKay, R. J. (2010). "Assessment of a Binary Measurement System in Current Use". *Journal of Quality Technology* 42, pp. 152–164.

Danila, O., Steiner, S. H.; and MacKay, R. J. (2011). "Assessing a Binary Measurement System with Varying Misclassification Rates When a Gold Standard is Available". *Technometrics*, submitted.

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.

Dendukuri, N. and Joseph, L. (2001). "Bayesian Approaches to Modeling the Conditional Dependence Between Multiple Diagnostic Tests". *Biometrics* 57, pp. 158–167.

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

Fujisawa, H. and Izumi, S. (2000). "Inference about Misclassification Probabilities from Repeated Binary Responses". *Biometrics* 56(3), pp. 706–711.

Johnson, N. L.; Kotz, S.; and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. 1, 2nd ed. New York, NY: John Wiley and Sons.

Maple 13 (2009). Maplesoft, Waterloo Maple Inc. Waterloo, Ontario, www.maplesoft.com.

Matlab 7.7.0 (2008). The MathWorks Inc. Natick, MA, www.mathworks.com.

Montgomery, D. C. (1996). *Introduction to Statistical Quality Control*, 3rd ed., New York, NY: John Wiley and Sons.

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

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, 1st ed., New York, NY: Oxford University Press.

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.

Quinino, R.; Ho, L. L.; and Trindale, A. L. G. (2005). "Bayesian Judgment of a Dichotomous Inspection System When the True State of an Evaluated Item Is Unknown". *Computers and Industrial Engineering* 49, pp. 591–599.

Vacek, P. (1983). "The Effect of Conditional Dependence on the Evaluation of Diagnostic Tests". *Biometrics* 41, pp. 959–968.

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

Van Wieringen, W. N. and Van den Heuvel, E. R. (2005). "A Comparison of Methods for the Evaluation of Binary Measurement Systems". *Quality Engineering* 17, pp. 495–507.

Walter, S. D. and Irwig, L. M. (1988). "Estimation of Test Error Rates, Disease Prevalence and Relative Risk for Misclassified Data: A Review". *Journal of Clinical Epidemiology* 41, pp. 923–937.