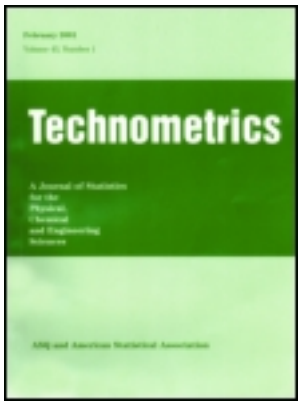

Technometrics
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/utch20

Assessing a Binary Measurement System With Varying Misclassification Rates When a Gold Standard Is Available

Oana Danila [a], Stefan H. Steiner [a] & R. Jock MacKay [a]
[a] Business and Industrial Statistics Research Group (BISRG), Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, N2L 3G1, Canada
Published online: 26 Aug 2013.

To cite this article: Oana Danila, Stefan H. Steiner & R. Jock MacKay (2013) Assessing a Binary Measurement System With Varying Misclassification Rates When a Gold Standard Is Available, Technometrics, 55:3, 335-345, DOI: 10.1080/00401706.2012.749653

To link to this article: http://dx.doi.org/10.1080/00401706.2012.749653

Assessing a Binary Measurement System With Varying Misclassification Rates When a Gold Standard Is Available

Oana DANILA, Stefan H. STEINER, and R. Jock MACKAY

Business and Industrial Statistics Research Group (BISRG)
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo N2L 3G1, Canada
(*omdanila@uwaterloo.ca*; *shsteiner@uwaterloo.ca*; *rjmackay@uwaterloo.ca*)

In manufacturing, we often use a binary measurement system (BMS) for 100% inspection to protect customers from receiving nonconforming product. We can assess the performance of a BMS by estimating the consumer's and producer's risks, the two misclassification rates. Here, we consider assessment plans and their analysis when a gold standard system (GSS) is available for the assessment study but is too expensive for everyday use. We propose a random-effects model to allow for variation in the misclassification rates within the populations of conforming and nonconforming parts. One possibility, here denoted the standard plan, is to randomly sample n parts and measure them once with the GSS and r times with the inspection system. We provide a simple analysis and planning advice for standard plans. In practice, the misclassification rates are often low and the underlying process has high capability. This combination of conditions makes the assessment of the BMS challenging. We show that we need a very large number of measurements with the standard plan in order to get precise estimators of the average misclassification rates and the true process performance. We consider an alternate design, here denoted the conditional assessment plan, where we select random samples from the sets of previously passed and failed parts. The sampled parts are measured once with the GSS and r times with the inspection system. When we augment the data from the conditional plans with available baseline information on the overall pass rate, we show that we can precisely estimate the parameters of interest with many fewer measurements. In the online supplementary materials, we provide R code to find maximum likelihood estimates and corresponding approximate standard errors, and to find the asymptotic standard deviation of the estimators with a selected plan size and assumed parameter values for both the standard and the conditional sampling plans.

KEY WORDS: Binary measurement systems; Gold standard measurement system; Likelihood methods; Random effects.

1. INTRODUCTION

Binary measurement systems (BMS) 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. Each part is conforming or not as indicated by the value of the random variable X, where

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

We can determine the value of X if we have an available gold standard system (GSS), a system with no measurement error.

If the part is measured 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 measurement system are then given by

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

Here, α represents the customer's risk, the proportion of nonconforming parts that pass the inspection and are presumably shipped to the customer. The parameter β represents the producer's risk, the proportion of conforming parts that fail the inspection and lead to unnecessary rework or scrap. We can also interpret α as the long-run proportion of times that a single nonconforming part passes repeated inspection by the BMS (and similarly for β). The parameter π_C is the proportion of parts that are conforming when measured by the GSS and depends on the underlying process and the GSS, not on the BMS. In the manufacturing context, we expect π_C to be large and α and β to be small. We focus on these conditions throughout the article. One consequence of this assumption is that we require the bias and standard deviation of any estimator to be small. We define, somewhat arbitrarily, an estimator to be useful if the relative bias is less than 0.1 and the relative standard deviation is less than 0.5.

Many other performance metrics describing both the BMS and the process are functions of the parameters α, β, and π_C.

TECHNOMETRICS, AUGUST 2013, VOL. 55, NO. 3
DOI: 10.1080/00401706.2012.749653

For example, the proportion of passed parts is

$$\pi_P = \Pr(Y = 1) = \alpha(1 - \pi_C) + (1 - \beta)\pi_C.$$

We can assess the BMS by measuring a randomly selected sample of n parts once each with the gold standard and $r \geq 1$ times with the BMS. We call this the standard plan. These plans with $r = 1$ have been studied by Danila, Steiner, and MacKay (2008), Farnum (1994), and Burke et al. (1995) in an industrial setting and by Pepe (2003) in the medical context.

To model the data from a standard plan, the simplest approach is to make the following assumptions:

- the misclassification rate α is the same for each nonconforming part (and similarly for β);
- measurements made on different parts are independent;
- 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

$$P(Y_1 = y_1, \ldots, Y_r = y_r | X = x) = \prod_{j=1}^{r} P(Y_j = y_j | X = x).$$

When we have a GSS, we know $X = x$, and so, for each part in the study

$$\begin{aligned} &P(Y_1 = y_1, \ldots, Y_r = y_r, X = x) \\ &\quad = [\alpha^s (1 - \alpha)^{r-s} \pi_c]^{1-x} [(1 - \beta)^s \beta^{r-s} (1 - \pi_c)]^x, \end{aligned} \quad (1)$$

where $s = \sum_{j=1}^{r} y_j$ is the number of times the part passes inspection. We refer to (1) and the corresponding assumptions as the fixed-effects model.

The assumptions underlying this model have been widely criticized. See, for instance, De Mast, Erdmann, and van Wierigen (2011), who discussed these issues in the situation when no GSS is available. In many cases, it may be unreasonable to assume that α and β are constant over all nonconforming and conforming parts, respectively. Some conforming (nonconforming) parts may be harder to classify correctly than others.

Suppose there is a vector of characteristics Z so that the probability that the BMS passes a part depends on the value of $Z = z$ as well as on $X = x$. If we measure Z for any part and assume that repeated measurements on the part are independent given $X = x$ and $Z = z$, then we can model $P(S = s | X = x, Z = z)$ using logistic regression, for example. The manual from AIAG (2010, p. 135) suggests a less formal approach. Now suppose that one or more of the components of Z are unidentified or not measurable. Each part has its own misclassification rate dependent on the unobserved Z. To better handle such situations, we make the following assumptions to model the data from a standard plan:

- the misclassification rate for part i is α_i if $X_i = 0$ and β_i if $X_i = 1$;
- measurements made on different parts are independent;
- given $X_i = 0$ and α_i, repeated measurements $Y_{i1}, \ldots, Y_{ir}$ on part i are (conditionally) independent, so

$$\begin{aligned} &P(Y_{i1} = y_1, \ldots., Y_{ir} = y_r | X_i = 0, \alpha_i) \\ &\quad = \prod_{j=1}^{r} P(Y_{ij} = y_j | X_i = 0, \alpha_i); \end{aligned}$$

and similarly given $X_i = 1$ and β_i.

- for each part i in the study, we have

$$\begin{aligned} &P(Y_{i1} = y_{i1}, \ldots, Y_{ir} = y_{ir}, X_i = x | \alpha_i, \beta_i) \\ &\quad = \left[\alpha_i^s (1 - \alpha_i)^{r-s} (1 - \pi_c)\right]^{1-x} \left[(1 - \beta_i)^s \beta_i^{r-s} \pi_c\right]^x, \\ &s_i = \sum\nolimits_{j=1}^{r} y_{ij}; \end{aligned}$$

- to model the variation of the misclassification rates in the populations of conforming and nonconforming parts, we specify the distributions of the α_i's and β_i's up to unknown parameters θ_0 and θ_1.

For any part, we can determine the joint distribution of $(Y_1, \ldots, Y_r, X)$ with parameters $(\theta_0, \theta_1, \pi_C)$ using the conditional distributions and assumptions given above. Note that θ_0 and θ_1 may each represent a vector of parameters. In this model, $Y_1, \ldots, Y_r$ are not independent given $X = x$. We call this distribution and the accompanying assumptions the random-effects model. The random-effects model explicitly allows for the variation in the misclassification rates within the sets of conforming and nonconforming parts. In a sense, the distributions of the α_i's and β_i's capture the effects of all unknown or unmeasured variables on the properties of the BMS, given the value of X. The parameters α and β, the average misclassification rates, are functions of the parameters θ_0 and θ_1 in this model. That is,

$$\begin{aligned} \alpha &= P\left(Y_{ij} = 1 | X_i = 0\right) = E_{\theta_0}\left[P\left(Y_{ij} = 1 | X_i = 0, \alpha_i\right)\right] \\ &= E_{\theta_0}\left[\alpha_i\right] \end{aligned}$$

and

$$\begin{aligned} \beta &= P\left(Y_{ij} = 0 | X_i = 1\right) = E_{\theta_1}\left[P\left(Y_{ij} = 0 | X_i = 1, \beta_i\right)\right] \\ &= E_{\theta_1}\left[\beta_i\right]. \end{aligned}$$

As noted earlier, we expect π_C to be large and α and β to be small. We focus on the ranges $\pi_C \geq 0.8$, $0 < \alpha, \beta \leq 0.10$. The article is organized as follows. We first consider the standard plan and investigate the properties of the maximum likelihood estimates (MLEs) from the fixed-effects model and the random-effects model, where we assume the α_i's and β_i's follow a beta distribution. We find that even when the misclassification rates vary, the estimators of α, β, and π_C from the fixed-effects model perform relatively well. Next, we consider planning such an assessment by examining the effects of changing n and r on the properties of the estimators. We conclude that for the parameter ranges under investigation, we need a large sample of parts to get useful estimators of the primary parameters, especially α. This is not surprising because with π_C large, we need

a large sample to find a sufficient number of nonconforming parts.

In most inspection systems, parts are segregated into passes and fails and the pass rate is recorded over time. To deal with the issue of the large number of measurements required when π_C is large and α and β are small, we recommend a conditional assessment plan where we choose a random sample from the previously failed parts and measure each selected part once with the GSS and $r \geq 0$ times using the BMS. Additionally, we incorporate the recent recorded pass rate into the analysis. Danila, Steiner, and MacKay (2008) considered the case with $r = 0$ for the fixed-effects model. We use simulation to determine when the likelihood-based asymptotic approximations can be used. We show the marked improvement in the precision of the estimators corresponding to the MLEs using the conditional plan rather than the standard plan, and demonstrate that it is possible to assess the properties of the BMS with relatively small numbers of parts and repeated measurements.

We also investigate the sensitivity of the estimators corresponding to the MLEs based on a conditional plan if the GSS occasionally misclassifies parts; that is, it is really not a GSS. In this case, we show that the estimators based on the random-effects model can be severely biased. We offer an adhoc change to the estimation procedure that largely deals with the issue. We finish the article with a summary and some discussion.

2. STANDARD PLANS

For convenience, we assume that the α_i's for nonconforming parts ($X = 0$) and the β_i's for conforming parts ($X = 1$) follow beta distributions with densities

$$\frac{\alpha_i^{g_0-1}(1-\alpha_i)^{h_0-1}}{\text{Beta}(g_0,h_0)}, \quad 0 < \alpha_i < 1, \text{ and}$$
$$\frac{\beta_i^{g_1-1}(1-\beta_i)^{h_1-1}}{\text{Beta}(g_1,h_1)}, \quad 0 < \beta_i < 1,$$

respectively, where $\text{Beta}(g, h)$ is the beta function. Other distributional assumptions are possible, but the beta distribution is a natural choice as it is flexible and mathematically convenient. We reparameterize the distributions in terms of the means α, β, and measures of the variability γ_0, γ_1, where the subscript indicates if the part is conforming (subscript 1) or not (subscript 0). We have

$$\alpha = \frac{g_0}{g_0+h_0}, \quad \gamma_0 = \frac{1}{g_0+h_0} \quad \text{and}$$
$$\beta = \frac{g_1}{g_1+h_1}, \quad \gamma_1 = \frac{1}{g_1+h_1}$$

so that the variances are $\sigma_0^2 = (\gamma_0/(1+\gamma_0))\alpha(1-\alpha)$ and $\sigma_1^2 = (\gamma_1/(1+\gamma_1))\,\beta(1-\beta)$. In this model, α and β represent the consumer's and producer's risk, respectively, as in the fixed-effects model. Also, as γ_0 and γ_1 approach zero, we recover the fixed-effects model.

Suppose we select a random sample of n parts from the process and measure each part $r \geq 2$ times with the BMS and once with the GSS. The likelihood contribution (ignoring multiplicative constants) of any nonconforming part is

$$(1-\pi_C)\int_{\alpha=0}^{1}\alpha^s(1-\alpha)^{r-s}\frac{\alpha^{g_0-1}(1-\alpha)^{h_0-1}}{\text{Beta}(g_0,h_0)}d\alpha = \frac{\text{Beta}(g_0+s,h_0+r-s)}{\text{Beta}(g_0,h_0)}(1-\pi_C) \tag{2}$$

and, for any conforming part, is

$$\pi_C\int_{\beta=0}^{1}(1-\beta)^s\beta^{r-s}\frac{\beta^{g_1-1}(1-\beta)^{h_1-1}}{\text{Beta}(g_1,h_1)}d\beta = \frac{\text{Beta}(g_1+r-s,h_1+s)}{\text{Beta}(g_1,h_1)}\pi_C, \tag{3}$$

where s is the number of times the part passes inspection. These are beta binomial models as described by Griffiths (1973). If $r = 1$, we can simplify these contributions to $\alpha^s(1-\alpha)^{1-s}\pi_C$ and $\beta^{1-s}(1-\beta)^s(1-\pi_C)$ that do not depend on γ_0 and γ_1 and are, in fact, the likelihood contributions from the corresponding fixed-effects model. Thus, when $r = 1$, we cannot distinguish between the random- and fixed-effects models and hence we consider only standard assessment plans with $r \geq 2$.

Using (2) and (3), if there are t conforming parts and $n - t$ nonconforming parts in the random sample of parts selected for the measurement assessment study, we can write the log-likelihood up to an additive constant as the sum

$$l(\alpha,\gamma_0,\beta,\gamma_1,\pi_C) = l_{r0}(\alpha,\gamma_0) + l_{r1}(\beta,\gamma_1) + t\log(\pi_C) + (n-t)\log(1-\pi_C), \tag{4}$$

where $l_{r0}(\alpha,\gamma_0)$and $l_{r1}(\beta,\gamma_1)$ are the beta binomial log-likelihoods corresponding to the nonconforming and conforming parts in the random sample, respectively. Note that $l_{r0}(\alpha,\gamma_0)$ is the sum over all nonconforming parts of the log of the terms involving a ratio of Beta functions given by (2), and $l_{r1}(\beta,\gamma_1)$ is similarly defined based on (3). We then have $\hat{\pi}_C = t/n$, and we can maximize $l_{r0}(\alpha,\gamma_0)$ and $l_{r1}(\beta,\gamma_1)$ separately using the simplification provided by Griffiths (1973). This simplification also provides a route to the observed and expected information by first conditioning on $T = t$. See the Appendix for details. We maximize the likelihood numerically with R (R Core Team 2012) using the algorithm of Nelder and Mead (1965) that makes no use of the derivatives but conducts an organized search for the maximum over the parameter space. We denote the MLEs from the random-effects model by $\hat{\alpha}_{\text{re}}$ and $\hat{\beta}_{\text{re}}$. We also provide approximate standard errors for the MLEs based on substitution of the estimates into the inverse of the expected information matrix.

If we fit the fixed-effects model (1) to these data instead, then the corresponding MLEs are

$$\hat{\alpha}_{\text{fe}} = \frac{\sum_{\text{nonconforming}} s_i}{r(n-t)} = \frac{\sum_{\text{nonconforming}} \hat{\alpha}_i}{n-t}, \quad t \neq n,$$
$$\hat{\beta}_{\text{fe}} = \frac{\sum_{\text{conforming}} s_i}{rt} = \frac{\sum_{\text{conforming}} \hat{\beta}_i}{t}, \quad t \neq 0,$$
$$\hat{\pi}_c = \frac{t}{n}, \tag{5}$$

where s_i is the number of times part i passes inspection, and $\hat{\alpha}_i$ and $\hat{\beta}_i$ are the corresponding proportions of times part i is misclassified. The corresponding estimators are unbiased regardless of the distributions for the α's and β's. Furthermore,

suitably standardized, the sample variances of the $\hat{\alpha}_i$ and $\hat{\beta}_i$ are unbiased estimators of the variances of these simple estimators.

2.1 Example

The context is real; the data are not. A functional test stand is used for 100% inspection of a component of an electronic device. The test stand examines a number of features and characteristics of the device. With many of these features we believe there are underlying latent variables that differentiate the devices, so we expect the misclassification rates to vary from device to device. To assess the performance of the stand, a sample of 100 parts was selected haphazardly from one shift's production and measured five times with the inspection system and then offline with a GSS. The GSS is an expensive exhaustive examination of the device by a human operator. The sample contained $t = 78$ conforming devices. We give the data in Table 1.

The fixed-effects model estimates from (5) and their standard errors (within parentheses) are $\hat{\pi}_C = 0.78\ (0.041)$, $\hat{\alpha}_{\text{fe}} = 0.127\ (0.038)$, and $\hat{\beta}_{\text{fe}} = 0.087\ (0.015)$. The MLEs $\hat{\pi}_C$, $\hat{\alpha}_{\text{re}}$, and $\hat{\beta}_{\text{re}}$ from the random-effects model log-likelihood (4) and their corresponding standard errors based on the asymptotic approximation are the same (to three decimal places). We also have $\hat{\gamma}_0 = 0.131\ (0.135)$ and $\hat{\gamma}_1 = 0.035\ (0.047)$. Even with 500 measurements, we have relatively poor precision for estimating α, β, and especially γ_0 and γ_1. For these latter parameters, we do not expect the asymptotic normal approximations to work because of the boundary issues, that is, the possibility that $\gamma_0 = 0$ and/or $\gamma_1 = 0$. To test the hypothesis that $\gamma_0 = 0$ and/or $\gamma_1 = 0$, we followed the suggestions of Self and Liang (1987) to adjust the distribution of the likelihood ratio statistic to deal with the boundary issue. Through simulation, we found that the distribution of the test statistic under the null hypothesis was far from the equal mixture of a discrete random variable with probability 1 at the origin and a χ^2_1 random variable suggested by the asymptotic results, especially in the tail. A simpler approach is to test the fit of the observed s_i to a binomial distribution. In the example, there is no evidence against the fit of the two binomial models corresponding to $\gamma_0 = \gamma_1 = 0$.

2.2 Performance of the Estimators in the Standard Plan

Here, we examine the properties of the estimators defined above. Because of the way the parameters separate in the log-likelihood (4), the Hessian matrix is block diagonal and we can investigate the asymptotic behavior of $(\hat{\alpha}_{\text{re}}, \hat{\gamma}_0)$, $(\hat{\beta}_{\text{re}}, \hat{\gamma}_1)$ and $\hat{\pi}_C$ separately. Since the asymptotic results for $\hat{\pi}_C$ are driven by the binomial distribution of T that depends solely on n and π_C, we do not consider its properties further. We focus on the estimators $\hat{\alpha}_{\text{re}}$ and $\hat{\gamma}_0$ since, with π_C large, we expect fewer nonconforming than conforming parts in the sample. Thus, in this case, the performance of the estimators $\hat{\alpha}_{\text{re}}$ and $\hat{\gamma}_0$ will be worse than that of $\hat{\beta}_{\text{re}}$ and $\hat{\gamma}_1$.

Table 1. Standard plan assessment data

	$s=5$	$s=4$	$s=3$	$s=2$	$s=1$	$s=0$	Total
$x=1$	51	21	5	1	0	0	78
$x=0$	0	0	1	3	5	13	22

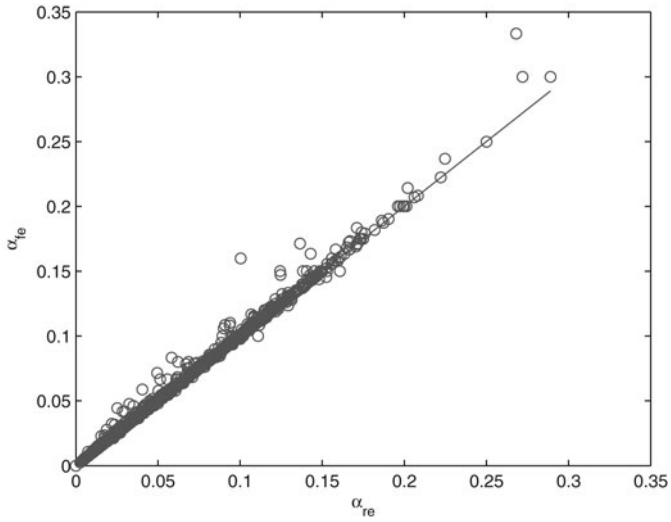

Figure 1. Plot of $\hat{\alpha}_{\text{fe}}$ versus $\hat{\alpha}_{\text{re}}$ for varying design and model parameters.

We start by numerically comparing the estimators $\hat{\alpha}_{\text{fe}}$ and $\hat{\alpha}_{\text{re}}$ over a wide range of design and model parameters. For each combination of $n = 100, 200, 500, 1000$; $r = 2, 5, 10$; $\alpha, \beta = 0.02, 0.05, 0.10$; $\gamma_0, \gamma_1 = 0.02, 0.10, 0.20$; and $\pi_C = 0.8, 0.9, 0.95$, we generated one sample and calculated $\hat{\alpha}_{\text{fe}}$ and $\hat{\alpha}_{\text{re}}$. We see in Figure 1 that the two estimates are almost always nearly equal. The most divergent cases correspond to n and r small with π_C large. In these cases, neither estimator is precise because there are few nonconforming parts in the sample. Based on the plot, we assume that we can assess the properties of estimator $\hat{\alpha}_{\text{re}}$ using the properties of simple estimator $\hat{\alpha}_{\text{fe}}$.

Conditioning on $T = t$, we can easily show, for any distribution of the random effects, that

$$\text{var}[\hat{\alpha}_{\text{fe}}] = \frac{1}{n-t}\left[\frac{\alpha(1-\alpha)}{r} + \sigma_0^2\frac{r-1}{r}\right],$$

where σ_0^2 is the variance of the α_i's in the population of nonconforming parts. And so, for n large, we have

$$\text{var}[\hat{\alpha}_{\text{fe}}] \approx \frac{\alpha(1-\alpha) + \sigma_0^2(r-1)}{nr(1-\pi_C)}. \tag{6}$$

If $\pi_C = 0.95$, we see that the standard deviation of the estimator $\hat{\alpha}_{\text{fe}}$ is relatively large unless n is extreme. For example, with $n = 1000$, $r = 5$, $\alpha = 0.05$, $\gamma_0 = 0.1$, we have $\sigma_0 = 0.065$ and the standard deviation of $\hat{\alpha}_{\text{fe}}$ is 0.017, about one-third of α. An estimator with such small relative precision may not be useful, even though the study involved 5000 measurements with the inspection system.

To assess the properties of the MLE for γ_0, we carried out a simulation where we varied the design specifications n, r over the ranges $100 \le n \le 1000$, $r = 2, 5, 10$ and the parameter values over the ranges $0.02 \le \alpha \le 0.10$, $0 < \gamma_0 \le 0.20$, and $0.8 \le \pi_C \le 0.95$. The number of conforming parts t varies from run to run. We used 10,000 simulation runs for each combination of the parameters. We found that $\hat{\gamma}_0$ performed poorly over most of the range. For example, with $n = 1000$, $r = 5$, $\alpha =$

0.05, $\gamma_0 = 0.1$ and $\pi_C = 0.95$, the bias and the standard deviation of $\hat{\gamma}_0$ are 0.007 and 0.111, respectively. The standard deviation is so large that the estimator is not useful.

In summary, we draw the following conclusions for standard plans:

- We can safely use the fixed-effects estimates of α, β, and π_C even if we suspect that the misclassification probabilities vary within the sets of conforming and nonconforming parts. We get standard errors, for example, by substituting $\hat{\sigma}_0^2$, the sample standard deviation of the $\hat{\alpha}_i$'s, and $\hat{\alpha}_{\text{fe}}$ (for α) into (6).
- If π_C is close to 1 (i.e., a high-quality process), then we need large n and moderately large r for the estimator $\hat{\alpha}_{\text{fe}}$ to have small relative precision.
- The MLEs $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are badly behaved unless n is extremely large.
- As r increases, the standard deviations of $\hat{\alpha}_{\text{fe}}$ and $\hat{\beta}_{\text{fe}}$ do not approach zero.

3. CONDITIONAL SAMPLING PLANS

In many inspection processes, passed and failed parts are segregated after an initial measurement by the BMS. Let $Y_0 = y_0$ indicate this measurement for any inspected part. As well, the pass rate is recorded by hour, shift, or some other fixed time period. Here, we use the idea of Danila, Steiner, and MacKay (2008, 2010) to select the parts for the assessment by separately sampling randomly from the previously passed and failed parts. We call this the conditional sampling approach, and when π_C is large, we expect to increase the number of nonconforming parts in the sample by oversampling the failed parts. We also propose to include the pass rate data from a recent fixed time period in the likelihood. Both of these measures help to reduce the large sample sizes required by the standard plan to produce useful estimators.

For the measurement system assessment study, we sample n_0 and n_1 parts at random from the previously failed and passed parts. Then, we measure each selected part once with the GSS and r times with the BMS. With the conditional sampling plan, it is not possible to use the simple average misclassification rate estimates given by (5) because they do not take into account the conditional sampling nor can we easily add the additional information given by the baseline data. Instead, we resort to maximum likelihood estimation. The contribution to the likelihood of any part that passes s times in the assessment study is

$$\begin{aligned} &P(S = s,\ X = x|Y_0 = y_0) \\ &\quad = P(Y_0 = y_0, S = s|X = x)P(X = x)/P(Y_0 = y_0). \end{aligned}$$

We evaluate the first factor on the right using the appropriate beta distribution to model the random effects. For any nonconforming part, the contribution to the likelihood is

$$\begin{aligned} &P(S = s, X = 0|Y_0 = y_0) \\ &\quad = P(Y_0 = y_0, S = s|X = 0)P(X = 0)/P(Y_0 = y_0) \\ &\quad = \binom{r}{s} \int_0^1 \frac{\alpha^{s+y_0}\alpha^{r-s+1-y_0}\alpha^{g_0-1}(1-\alpha)^{h_0-1}}{\text{Beta}(g_0, h_0)} d\alpha \frac{P(X = 0)}{P(Y_0 = y_0)} \\ &\quad = \binom{r}{s} \frac{\text{Beta}(g_0 + s + y_0, h_0 + r - s + 1 - y_0)}{\text{Beta}(g_0, h_0)} \\ &\qquad \times \frac{(1-\pi_C)}{\pi_P^{y_0}(1-\pi_P)^{1-y_0}}, \end{aligned} \tag{7}$$

and for any conforming part, is

$$\begin{aligned} &P(S = s, X = 1|Y_0 = y_0) \\ &\quad = P(Y_0 = y_0, S = s|X = 1)P(X = 1)/P(Y_0 = y_0) \\ &\quad = \binom{r}{s} \int_0^1 \frac{(1-\beta)^{s+y_0}\beta^{r-s+1-y_0}\beta^{g_1-1}(1-\beta)^{h_1-1}}{\text{Beta}(g_1, h_1)} \\ &\qquad \times d\alpha \ \frac{P(X = 1)}{P(Y_0 = y_0)} \\ &\quad = \binom{r}{s} \frac{\text{Beta}(g_1 + r - s + 1 - y_0, h_1 + s + y_0)}{\text{Beta}(g_1, h_1)} \\ &\qquad \times \frac{\pi_C}{\pi_P^{y_0}(1-\pi_P)^{1-y_0}}. \end{aligned} \tag{8}$$

Other than the additional divisor, (7) and (8) correspond to (2) and (3) with r replaced by $r + 1$ and s replaced by $s + y_0$. We can again use Griffith's (1973) simplification to evaluate the likelihood.

As noted above, most inspection systems record the pass rate over some fixed time period. Suppose the baseline data showed that u of the last m parts were passed by the BMS. Under the random-effects model, these data contribute the additional term $u\log(\pi_P) + (m-u)\log(1-\pi_P)$ to the log-likelihood. So for the conditional sampling plan, the log-likelihood is given by

$$\begin{aligned} &l_{c0}(\alpha, \gamma_0) + l_{c1}(\beta, \gamma_1) + t\log(\pi_C) + (n-t)\log(1-\pi_C) \\ &\quad + (u - n_1)\log(\pi_P) + (m - u - n_0)\log(1-\pi_P). \end{aligned} \tag{9}$$

Here t is the total number of conforming parts in the two samples. The terms $l_{c0}(\alpha, \gamma_0)$ and $l_{c1}(\beta, \gamma_1)$ are based on the log-ratio of the Beta functions on the right sides of (7) and (8), respectively, and each is the sum of two terms ($y_0 = 0, 1$) determined by the frequencies of the numbers of passes in the r inspections.

We maximize the overall conditional sampling likelihood using the Nelder–Mead algorithm. Since π_P depends on α, β, and π_C, the information matrix is no longer block diagonal, as it was with the standard plan. With conditional sampling, the fixed-effects estimators are severely biased (see below), and so for convenience, we drop the subscript "fe" and "re" in the notation for the MLEs of α and β as we only look in detail at the properties of the random-effects model-based estimators.

3.1 Example

For the functional test stand described earlier, in the previous three shifts before the assessment, there were 960 units passed in 1243 inspections. Over time, failed parts were set aside until there were 100 available for the assessment study. Each previously failed part was remeasured five times with the BMS and once with the GSS. The results are given in Table 2.

The MLEs from (9) and their approximate standard errors (within parentheses) are

$$\hat{\alpha} = 0.134\,(0.029),\ \hat{\beta} = 0.086\,(0.013),\ \ \hat{\gamma}_0 = 0.141\,(0.098),$$
$$\hat{\gamma}_1 = 0.020\,(0.030),\ \hat{\pi}_C = 0.82\ (.016).$$

Table 2. Conditional sampling plan assessment data

	$s=5$	$s=4$	$s=3$	$s=2$	$s=1$	$s=0$	Total
$x=1$	22	5	5	0	0	0	32
$x=0$	0	0	4	5	18	41	68

Note the improved precision from the earlier example due to the greater number of conforming parts in the sample and the additional information from the baseline. We show below that the asymptotic approximations for the standard deviations of the estimators for α, β, and π_C work well for this design with the selected range of parameter values. However, both $\hat{\gamma}_0$ and $\hat{\gamma}_1$ have significant positive bias, and asymptotic approximations badly underestimate standard deviations evaluated by simulation. We must increase the number of parts in the sample to get good estimators of γ_0 and γ_1. As with the standard plan, we can test the hypothesis(is) $\gamma_0 = 0$ and/or $\gamma_1 = 0$ using a goodness-of-fit test on the appropriate distributions of the rows of Table 2.

3.2 Performance of the Asymptotic Standard Deviation Approximations for the MLEs

In this section, we assessed how well the asymptotic standard deviation approximations for the conditional sampling plan match the standard errors of the MLEs using simulation. We focus on the case $n_1 = 0$ and $n_0 = n$, where we sample only failed parts. If π_C is large, as expected, then sampling failed parts increases the number of nonconforming parts in the study. In practice, it is convenient to use failed parts in the assessment study as these are not shipped to customers and typically set aside as scrap or for rework. Depending on the parameter values, we may end up with more than half the parts being nonconforming, which suggests that β and γ_1 would become the more difficult parameters to estimate. However, the baseline data provide significantly more information about β than about α. We discuss why this is true at the end of Section 4. As noted above, the information matrix for a plan augmented by baseline data is not block diagonal. Hence, we expect the properties of the estimator of any one of the parameters to depend on all of the others. Accordingly, we report the results for all of α, β, γ_0, γ_1, and π_C.

We start with a small plan $n_0 = n = 100$, $n_1 = 0$; that is, we sample 100 parts from the population of previously failed parts, $r = 5$ repeated measurements on each part with the BMS, and $m = 1000$ baseline observations. We expect that increasing any of the design characteristics n, r, or m will improve the performance of the estimators and the asymptotic approximations. We selected two levels for each of the model parameters $\alpha = 0.02, 0.10$, $\beta = 0.02, 0.10$, $\gamma_0 = 0.05, 0.20$, $\gamma_1 = 0.05, 0.20$, and $\pi_C = 0.90, 0.95$. For each of the 32 combinations, we estimate the bias and standard deviation of the estimator for each parameter based on the results of the 5000 simulation runs. We also calculate the ratio of this standard deviation to the asymptotic approximation based on the Fisher information at the known parameter values. The results are summarized in Table 3.

Table 3. Bias, standard deviation, and ratio for the conditional sampling plan: minimum and maximum values over the 32 combinations given in parenthesis with $n_0 = 100$, $n_1 = 0$, $r = 5$, $m = 1000$

Estimator	Bias	Standard deviation	Ratio
$\hat{\alpha}$	(0.000, 0.004)	(0.008, 0.049)	(0.973, 1.089)
$\hat{\gamma}_0$	(0.004, 0.108)	(0.055, 0.377)	(0.951, 2.048)
$\hat{\beta}$	(0.000, 0.004)	(0.004, 0.012)	(0.981, 1.034)
$\hat{\gamma}_1$	(−0.000, 0.004)	(0.026, 0.086)	(0.990, 1.081)
$\hat{\pi}_C$	(−0.000, 0.000)	(0.007, 0.012)	(0.988, 1.028)

There is negligible bias in all of the estimators except $\hat{\gamma}_0$. There is very little information about γ_0 available using the plan. We found through further simulation that to estimate γ_0 without bias and reasonable standard deviation, we need $n \geq 200$, $r \geq 5$, with even larger samples required as π_C gets closer to 1. The standard deviations of the estimators are sensitive to the underlying parameter values. The standard deviation of $\hat{\alpha}$ increases as α, β, and γ_0 increase with no indication of two-factor interactions. The standard deviation of $\hat{\beta}$ increases as β increases. The standard deviation of $\hat{\gamma}_1$ increases as γ_1 increases, and decreases as β and π_C increase. Finally, the standard deviation of $\hat{\pi}_C$ increases as β increases, and decreases as π_C increases. With the exception of $\hat{\gamma}_0$, the asymptotic approximations are close to the standard deviations over the whole range of parameter values that we investigated. We can then safely compare various plans as large as or larger than the plan $n_0 = 100$, $n_1 = 0$, $r = 5$, $m = 1000$ using the asymptotic approximations. The standard deviations of the estimators of the key parameters α, β, and π_C (especially α) are just small enough so that this plan provides useful information.

In the above experiment, we also examined the MLEs for α and β assuming the fixed-effects model. Unlike the situation under the standard plan, the estimators from the fixed-effects likelihood that take into account the conditional sampling and baseline are severely biased. The bias ranges between (–0.017, –0.001) and (0.004, 0.042) for the estimators of α and β, respectively, and between (–0.005, 0.010) for the fixed-effects estimator of π_C.

In summary, we conclude:

- When $\pi_C > 0.9$, with conditional sampling and a reasonable size baseline ($m > 1000$), we need a sample with at least $n = 100$, $r = 5$ to estimate α and β with useful precision. We also find the smaller α and β are, the larger is the sample size that is required. We cannot estimate γ_0 well unless we have a much larger plan.
- We can substitute the estimates into the approximate asymptotic standard deviations to get standard errors that are close to correct.
- If the misclassifications error rates vary, we cannot use the fixed-effects model with the conditional sampling plan without the risk of large bias in the estimators of α and β.

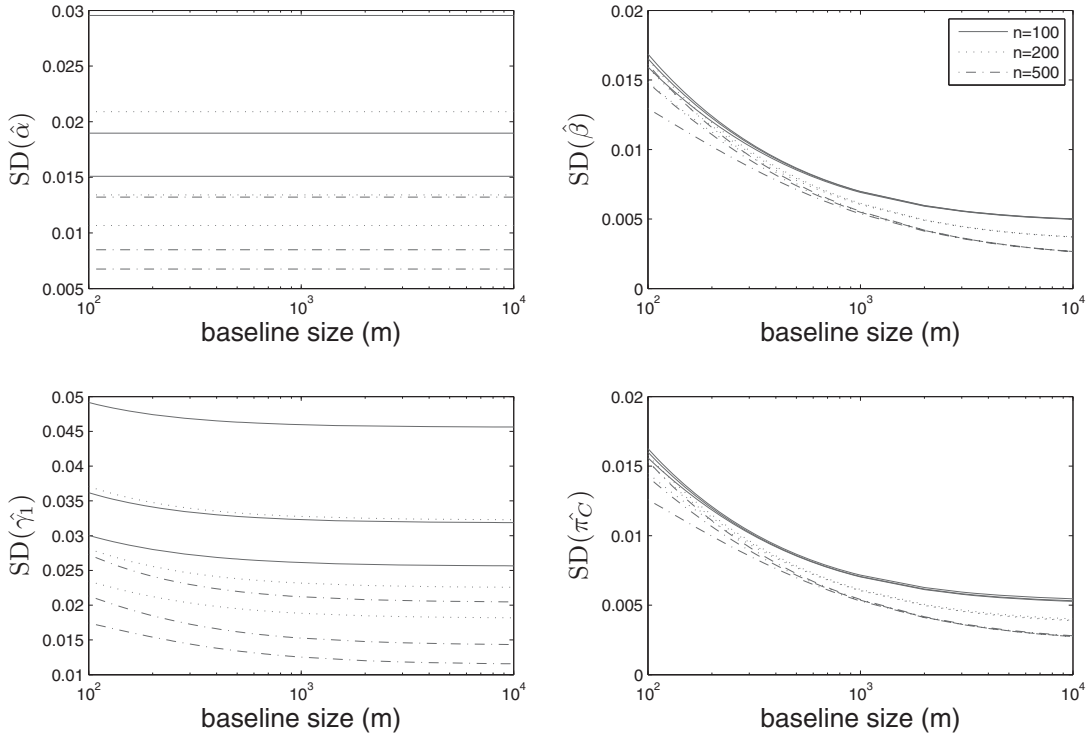

Figure 2. Effect of increasing m on the asymptotic standard deviations of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}_1$, and $\hat{\pi}_C$ for $r = 2, 5, 10, n = 100$ (solid lines), $n = 200$ (dotted lines), $n = 500$ (dot-dashed lines), $\pi_C = 0.95$, $\alpha = \beta = 0.05$, and $\gamma_0 = \gamma_1 = 0.10$. Note that lower lines of the same type correspond to increasing values of r.

3.3 Effect of Changing the Design Characteristics

With conditional sampling and baseline data, the investigator can choose the four design characteristics n_0, n_1, r, and m. Here, we consider only the case $n_1 = 0$ since we are assuming π_C is large. We investigate the effects of changing $n_0 = n$, r, and m using the asymptotic standard deviations of the MLEs as the basis of comparison. We present the results in Figures 2, 3, and 4, with $\pi_C = 0.95$, $\alpha = \beta = 0.05$, and $\gamma_0 = \gamma_1 = 0.10$. In each plot, we show how the asymptotic standard deviations for $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}_1$, and $\hat{\pi}_C$ change. We do not show the plots for $\hat{\gamma}_0$ because, for the smaller sample sizes, the estimator is badly biased and the asymptotic approximations perform poorly.

Note that the horizontal axis in Figure 2 is on a log scale and the scale of the vertical axis varies from plot to plot. To create the plots, we considered a wide range of values of m for each of the nine combinations of $r = 2, 5, 10$ and $n = 100, 200, 500$. We see that increasing the baseline size has no effect on the asymptotic standard deviations of $\hat{\alpha}$, a small positive effect on the standard deviation of $\hat{\gamma}_1$, and a dramatic effect on the standard deviations of $\hat{\beta}$ and $\hat{\pi}_C$ for all values of n and r, with reductions in the standard deviation of more than 80% 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 parameters. Since the baseline data are freely available, we strongly recommend their inclusion in the analysis.

In Figure 3, we examine the effects of increasing n, the number of parts, on the asymptotic standard deviations with $m = 100, 1000, 10000$ and $r = 2, 5, 10$. We use a log scale for the horizontal axis. For $\hat{\alpha}$, the standard deviations decrease at a rate $1/\sqrt{n}$ and m has no effect. For $\hat{\beta}$ and $\hat{\pi}_C$, when $m = 100$, the effect of increasing n depends strongly on the value of r. In the cases with the baseline size $m \geq 1000$, the standard deviations are weakly dependent on r and the effect of increasing n is greater for $\hat{\pi}_C$. For $\hat{\gamma}_1$, increasing either n or m substantially improves the precision, but baseline sizes larger than 1000 provide little benefit.

In Figure 4, we see the effect of increasing r, the number of repeated measurements, on the asymptotic standard deviations of the estimators when we consider the cases $m = 100, 1000, 10000$ and $n = 100, 200, 500$. Increasing r has a similar positive effect to increasing n on the standard deviation of $\hat{\alpha}$. When $m = 100$, increasing r strongly reduces the standard deviations of $\hat{\beta}$ and $\hat{\pi}_C$. However, when m is much larger, we see that increasing r has very little effect on the standard deviations. For $\hat{\gamma}_1$, increasing r reduces the standard deviation, but most of the gains have been achieved when $r = 5$ or 6; larger values are only marginally better.

We see similar patterns for other values of the parameters. In summary, we conclude that for $\hat{\alpha}$, the baseline data has little impact on the precision of the estimator, which depends largely on n and r. For $\hat{\beta}$ and $\hat{\pi}_C$, even a small baseline sample can have a large impact. Once the baseline reaches a size near 1000, there is little value in increasing r in estimating the corresponding parameters. In this case, increasing n, the number of parts, has diminishing returns, with the asymptotic standard deviations decreasing more slowly than the usual $1/\sqrt{n}$ rate.

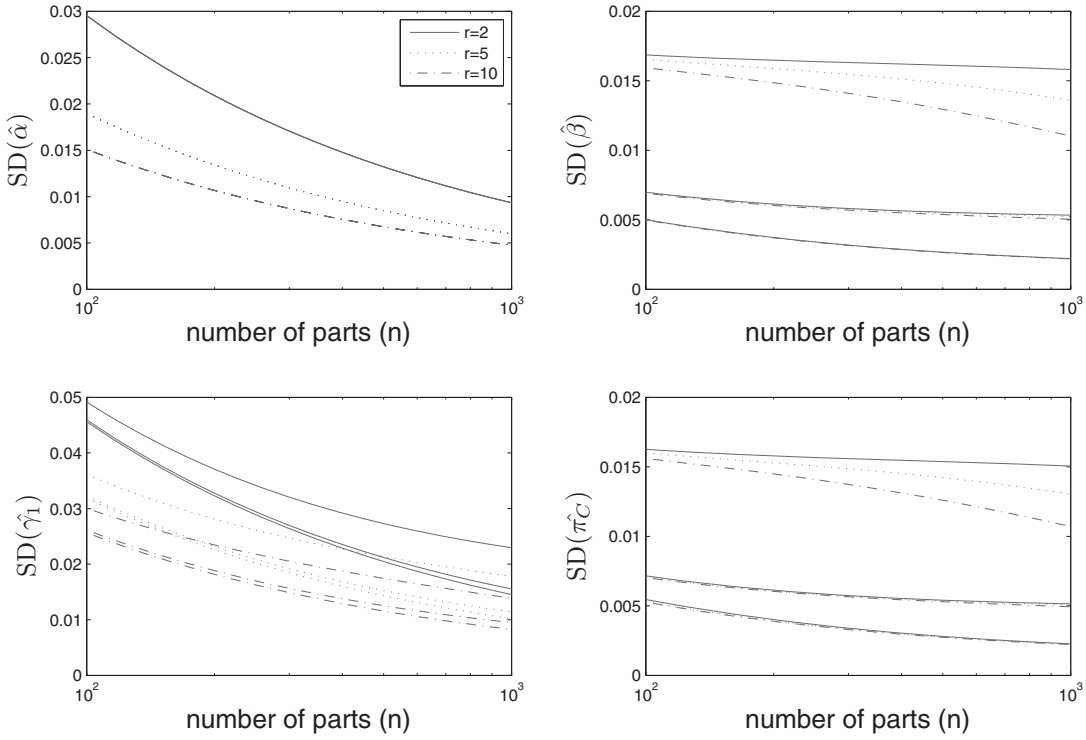

Figure 3. Effect of increasing n on the asymptotic standard deviations of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}_1$, and $\hat{\pi}_C$ for $m = 100, 1000, 10{,}000$, $r = 2$ (solid lines), $r = 5$ (dotted lines), $r = 10$ (dot-dashed lines), $\pi_C = 0.95$, $\alpha = \beta = 0.05$, and $\gamma_0 = \gamma_1 = 0.10$. Note that lower lines of the same type correspond to increasing values of m.

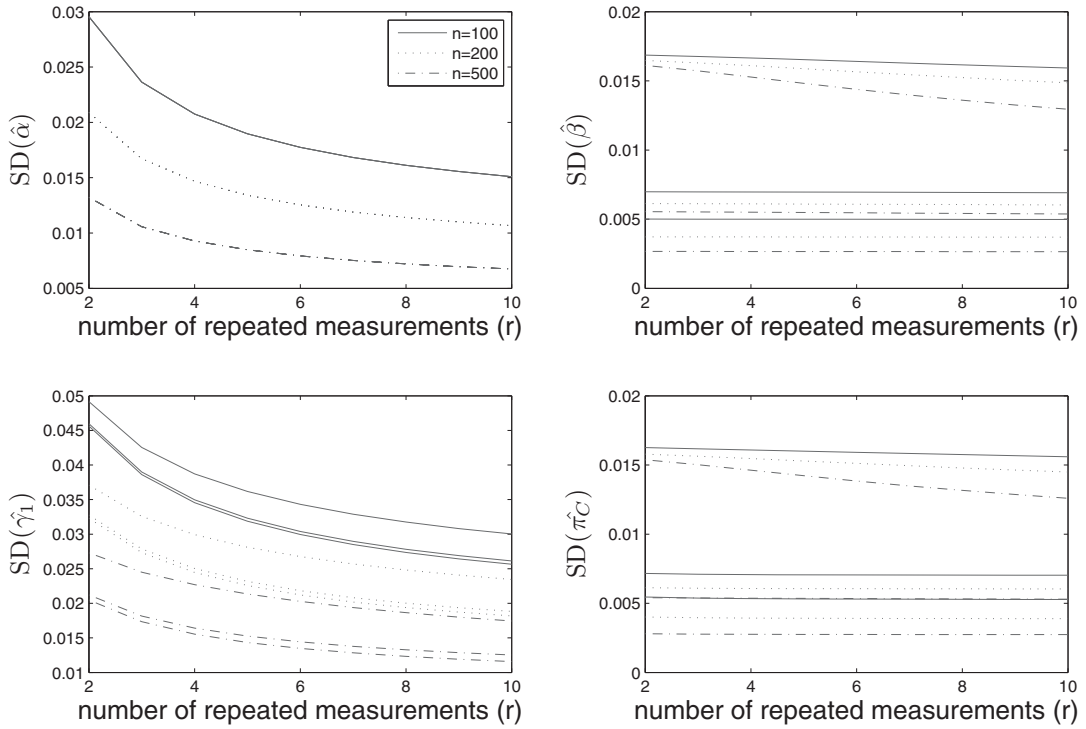

Figure 4. Effect of increasing r on the asymptotic standard deviations of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}_1$, and $\hat{\pi}_C$ for $m = 100, 1000, 10{,}000$, $n = 100$ (solid lines), $n = 200$ (dotted lines), $n = 500$ (dot-dashed lines), $\pi_C = 0.95$, $\alpha = \beta = 0.05$, and $\gamma_0 = \gamma_1 = 0.10$. Note that lower lines of the same type correspond to increasing values of m.

3.4 Errors in the GSS

A referee raised the valid point that a gold standard is an idealization and, in practice, unlikely to be error free. In our example, we have no evidence that if the same or different inspector tore down the same device, that the same result (i.e., value of X) would occur. When we calibrate a continuous measurement system, we are faced with a similar problem in that the standards are to some extent inaccurate. Note that if the GSS is perfectly repeatable (i.e., even if in error, it consistently makes the same error), then we can redefine "conforming" using the gold standard and the properties of the inspection system are defined relative to the gold standard.

However, another possibility is that the GSS results are uncertain. For the conditional plan, we investigated by simulation the effect of a GSS that occasionally misclassifies parts at a rate of one-fifth of α and β (as given by the properties of the BMS). For the cases we considered, only the properties of $\hat{\alpha}$ are materially affected. For example, if $\alpha = \beta = 0.10$, $\gamma_0 = \gamma_1 = 0.10$, $\pi_C = 0.95$, the bias in estimating α is 0.064 when $n = 100$, $r = 5$, $m = 1000$ and the standard deviation is inflated by a factor of 1.76 relative to the results with a true GSS. For a larger design, the effect of the misclassifications with the assumed GSS is even greater. With the same parameter values as given above but with $n = 500$, $r = 5$ and $m = 1000$, the corresponding bias and inflation factors are 0.071 and 2.34. We see similar distortions for other designs and parameter values. One immediate consequence of the above finding is that it is essential to examine the data carefully, looking for anomalies. For example, a table like Table 2 could be useful.

A few errors in the GSS have a surprisingly (to us) large effect. We can suggest two possible remedies. If r is reasonably large ($r \geq 3$), we can examine the data and identify possible discrepancies, for example, parts with s close to r and $x = 0$ or s close to 0 and $x = 1$. These correspond to parts where the GSS has likely made a mistake (since we assume α and β are small). We propose to flip the value of x for these discrepant parts. In the simulation described above, this adhoc procedure (with swaps occurring for parts with either $x = 0$ and $s \geq r - 1$ or $x = 1$ and $s \leq 1$) virtually eliminated the bias in the estimators of α but produced estimated standard deviations that were deflated by a factor of about 0.8. We did not search for the best swapping rule. An alternative is to remeasure the discrepant parts with the GSS.

There are other possible remedies to an imperfect GSS. If it is known that the GSS is imperfect, then at the cost of a much larger sample size, we can use the latent class model described by Danila, Steiner, and MacKay (2012) that does not require GSS data to estimate the model parameters. Alternatively, if the GSS has known misclassification rates, we can extend the methods of Boyles (2001) using a so-called anchored model.

4. COMPARISON OF THE STANDARD AND CONDITIONAL SAMPLING PLANS

In this section, we compare the performance of the conditional plan with baseline data to the standard plan using the asymptotic approximations to the standard deviations of the MLEs for α, γ_0, β, γ_1, and π_C. We used a factorial structure in the study where we looked at all possible combinations of $n = 100, 500$, $r = 5, 10$, $\alpha = 0.02, 0.10$, $\beta = 0.02, 0.10$, $\gamma_0 = 0.05, 0.20$, $\gamma_1 = 0.05, 0.20$, and $\pi_C = 0.9, 0.95$. For conditional sampling, we selected parts only from the population of previously failed parts (i.e., $n_0 = n$ and $n_1 = 0$) and we assumed 1000 baseline measurements (i.e., $m = 1000$). We feel that with this setup, the standard and conditional plans are equivalent in terms of effort because we assume the baseline measurements are already available from routine inspection. To summarize the results, Figure 5 shows the ratio of the asymptotic standard deviation for the standard plan over the conditional sampling plan with baseline for all the 32 combinations of the model parameters for each combination of n and r. As noted earlier, the asymptotic results do not work well for $\hat{\gamma}_0$ when n is small for the standard plan. We include results for $\hat{\gamma}_0$ in our conclusions because we found that the standard deviations of $\hat{\gamma}_0$ in the simulation were larger than the asymptotic approximations, so the conclusions for $\hat{\gamma}_0$ are conservative.

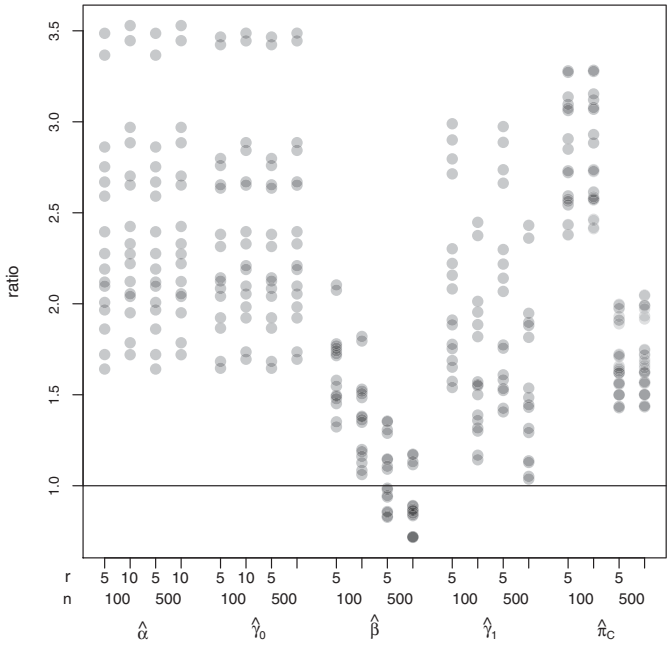

Figure 5. Ratios of the asymptotic standard deviations (standard over conditional plan with baseline).

In Figure 5, we see that all the ratios are greater than 1 except for $\hat{\beta}$ when $n = 500$. Based on the study, we draw the following conclusions:

- The conditional sampling plans gives more precise estimators of all the parameters except β. In many cases, the ratio of standard deviations exceeds 2.
- As α increases from 0.02 to 0.10, the ratios decrease for $\hat{\alpha}$, $\hat{\gamma}_0$, and $\hat{\pi}_C$ and increase for $\hat{\gamma}_1$. For $\hat{\beta}$, the ratios increase slightly for $n = 100$ and are essentially unchanged when $n = 500$.
- As β increases from 0.02 to 0.10, all the ratios decrease except for $\hat{\pi}_C$, which decreases when $n = 100$ and increases when $n = 500$.

- As γ_0 increases from 0.05 to 0.20, the ratios for $\hat{\alpha}$ and $\hat{\gamma}_0$ decrease, for $\hat{\pi}_C$ decrease slightly, and are unchanged for $\hat{\beta}$ and $\hat{\gamma}_1$.
- As γ_1 increases from 0.05 to 0.20, the ratios for $\hat{\alpha}$ and $\hat{\gamma}_0$ are unchanged, for $\hat{\beta}$ and $\hat{\gamma}_1$ increase, and for $\hat{\pi}_C$ decrease slightly.
- As π_C increases, all the ratios increase.
- Increasing r decreases the ratios for $\hat{\beta}$ and $\hat{\gamma}_1$, leaving the others the same.
- Increasing n decreases the ratios for $\hat{\beta}$, $\hat{\gamma}_1$, and $\hat{\pi}_C$, leaving the others the same.

When comparing the standard and conditional plans, we expect to get more information about α and γ_0 with conditional sampling since we will likely select more nonconforming parts. The increased precision for β and γ_1 is perhaps surprising. Here, it is the baseline measurements that help, as shown in Figures 2 and 3. The baseline provides an estimate of $\pi_P = \alpha(1-\pi_C) + (1-\beta)\pi_C$, a function of α, β, and π_C. Since we are considering situations where π_C is large and α and β are small, π_P is strongly influenced by β and π_C. For most cases in the simulation, the additional information about β and π_C from the baseline outweighs the lost information due to fewer conforming parts in the sample.

To summarize, in almost all cases, the conditional sampling plan with the baseline data provides substantially greater precision (for the same sample size) than the standard plan. The only exception occurs when the number of parts n is large, where the standard plan produces slightly smaller standard deviations for $\hat{\beta}$ (see Figure 5). In this case, however, $\hat{\beta}$ is very well estimated and so the loss is minor. Especially when π_C is close to 1, as we might expect with high-quality processes, we strongly recommend the conditional sampling plan.

5. DISCUSSION AND CONCLUSIONS

In this article, we consider the assessment of a BMS when a GSS is available. We concentrate on the industrial context where the BMS is used for inspection. In this case, it is likely that the misclassification probabilities are small and the overall conforming rate is close to 1. We investigate a random-effects model that relaxes the assumption that the misclassification probabilities for all conforming (and separately all nonconforming) parts are the same.

We apply the model to the standard assessment plan where each part in a random sample of parts is measured a number of times with the BMS and once with the GSS. We show that to get reasonable precision for the estimators of the average misclassification probabilities, we need a large sample of parts and a large number of repeated measurements on each part. Even with large samples, it is difficult to estimate the measures of variation in the misclassification rates.

As an alternative, we recommend using a conditional sampling plan where we sample at random from previously failed parts. Then, each selected part is measured a number of times using the BMS and once with the GSS. We augment the data from the assessment study with baseline data available from the inspection records of the BMS. With this plan, we show that there are large gains in efficiency of estimation using the same number of parts and repeated measurements. Or put differently, we can use much smaller sample sizes with conditional sampling and baseline data and get the same precision as with the standard plan.

The MLEs of the parameters in the random-effects model must be found numerically. We provide R code (R Core Team 2012) to produce the estimates and their approximate standard errors for both the standard and the conditional assessment plans in the online supplementary materials. We show that with the standard plan, if the misclassification rates vary from part to part, we can use the simpler estimates from the fixed-effects model to estimate the average misclassification rates. However, with conditional sampling plan, the fixed-effects model estimators are not appropriate since they have significant bias.

To use conditional sampling, we need to choose n_0, n_1, and r as we assume the baseline size m is given. We recommend $n_1 = 0$, that is, 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 even having more nonconforming than conforming parts results in precise estimators for β since the baseline helps substantially. To choose appropriate values for the number of parts n and the number of repeated measurements r, we provide R code (R Core Team 2012) that determines the asymptotic standard deviations of all five parameters. This code can be used to meet any desired precision goals so long as the numbers of parts is not very small (say $n > 100$ for all parameters other than γ_0). One constraint is that there needs to be enough parts so that we get a reasonable number of both conforming and nonconforming parts, otherwise the asymptotic results are not valid.

APPENDIX

For the standard plan, the log-likelihood is

$$\sum_{x,s} f_x(s) \log\left(P(S=s, X=x)\right) = \sum_{x,s} f_x(s) \log\left(P(S=s|X=x)\right) + \sum_{x} n_x \log\left(P(X=x)\right),$$

where $f_x(s)$ is the number of parts in the sample with $S = s$ passes in the r repeated measurements, $X = x$, and n_x is the number of parts with $X = x$. Considering the case $X = 1$ (i.e., a conforming part), then from the beta binomial model (Griffiths 1973), we have for $s = 0$

$$\begin{aligned}\log P(S=0|X=1) &= \log\binom{r}{0} + \log\left(\text{Beta}(g_1, h_1+r)/\text{Beta}(g_1, h_1)\right)\\ &= \sum_{j=0}^{r-1} \log(1-\beta+\gamma_1 j) - \sum_{j=0}^{r-1} \log(1+\gamma_1 j),\end{aligned}$$

and for $1 \le s \le r$, the recursion

$$\begin{aligned}\log\left(\frac{P(S=s|X=1)}{P(S=s-1|X=1)}\right) &= \log\left(\frac{r-s+1}{s}\right) + \log\left((\beta+\gamma_1(s-1))\right)\\ &\quad - \log\left(1-\beta+\gamma_1(r-s)\right).\end{aligned}$$

Table A1. Probabilities for $X = 1$ for conditional sampling

y_0	$P(S = 0, y_0 \mid X = 1)$	$\frac{P(S=s,y_0\mid X=1)}{P(S=s-1,y_0\mid X=1)},\ 1 \le s \le r$
0	$\sum_{j=0}^{r}\log(\beta + j\gamma_1) - \sum_{j=0}^{r}\log(1 + j\gamma_1)$	$\log(\frac{r-s+1}{s}) + \log((1-\beta+\gamma_1(s-1))) - \log(\beta + \gamma_1(r+1-s))$
1	$\log(1-\beta) + \sum_{j=0}^{r-1}\log(\beta + j\gamma_1) - \sum_{j=0}^{r}\log(1 + j\gamma_1)$	$\log(\frac{r-s+1}{s}) + \log((1-\beta+\gamma_1 s)) - \log(\beta + \gamma_1(r-s))$

Any series of the form $\sum_{s=0}^{r} u_s v_s$ can be rewritten as $U_0 v_0 + U_1(v_1 - v_0) + \cdots + U_r(v_r - v_{r-1})$, where $U_s = u_s + \cdots + u_r$. Applying this result to the first term in the log-likelihood with $u_s = f_1(s)$ and $v_s = \log(P(S = s|X = 1))$ and ignoring the additive constants, the contribution for parts with $X = 1$ to the log-likelihood is

$$F_1(0)\left(\sum_{j=0}^{r-1}\log(1-\beta+j\gamma_1) - \log(1+j\gamma_1)\right) + \sum_{j=1}^{r} F_1(j)(\log(\beta+(j-1)\gamma_1) - \log(1-\beta+(r-j)\gamma_1)), \quad \text{(A.1)}$$

where $F_1(j) = f_1(j) + f_1(j+1) + \cdots + f_1(r)$ for $j = 1, 2, \ldots, r$. For parts with $X = 0$, we have a similar expression with $1 - \beta$ replaced by α, γ_1 by γ_0, and $F_1(j)$ by $F_0(j)$. Albeit somewhat messy, it is easy to derive the first and second derivatives of the log-likelihood. We used Maple 13 (2009) to obtain the symbolic expressions. To calculate the expected information, we see that the second derivatives of the log-likelihood are linear in $F_x(s)$ and n_x, so we can evaluate $E(F_x(s))$, that is, $nP(X = x)\sum_{j=s}^{r} P(S = j|X = x)$, using the recursions to calculate the probabilities.

For conditional sampling, we proceed in a similar manner except that the likelihoods (7) and (8) contain the additional divisor $P(Y_0 = y_0)$, and we need to consider four cases defined by $x = 0, 1$ and $y_0 = 0, 1$. Table A1 gives the expressions for conforming parts ($X = 1$) selected from either the previously passed or failed parts.

As with the standard plan, we can use the four expressions in Table A1 to write down the log-likelihood ratio contribution for parts with $X = 1$. With conditional sampling, there are now two sets of F and f, where $f_1(s, y_0)$ is the number of parts in the sample with $S = s$ and $Y_0 = y_0$. For parts with $X = 0$, we have the same expressions with $1 - \beta$ replaced by α, γ_1 by γ_0, and $F_1(j, y_0)$ by $F_0(j, y_0)$. To determine the expected information from the second derivatives, we need to determine the expected value of $f_x(s, y_0)$ in each of the four cases. These are easy to derive from the results given in Table A1.

SUPPLEMENTARY MATERIALS

R code: A zip archive containing four R code files (R Core Team 2012) listed below:

- The first two files provide code to find the random-effects model MLEs and corresponding approximate standard errors for the standard assessment plan and the conditional sampling plan when we assume a gold standard measurement is available. Each file provides an example of how to specify the assessment study data:
  - MLE_se_section2_SP.R - standard (random) sampling plan
  - MLE_se_section3_CS.R - conditional sampling plan
- The second set of two R files provide code to find the asymptotic standard errors (from the Fisher information) for all the random-effect model estimators given assumed true values when we assume a gold standard measurement is available:
  - Asympt_sd_section2_SP.R - standard (random) sampling plan
  - Asympt_sd_section3_CS.R - conditional sampling plan

ACKNOWLEDGMENTS

The authors thank the Associate Editor for suggesting the simple analysis in the standard plan, and a referee for the suggestion to investigate the effects of misclassification errors by the assumed GSS.

[Received August 2011. Revised October 2012.]

REFERENCES

Automotive Industry Action Group (AIAG) (2010), *Measurement Systems Analysis* (4th ed.), Southfield, MI: AIAG. [336]

Boyles, R. A. (2001), "Gage Capability for Pass-Fail Inspection," *Technometrics*, 43, 223–229. [343]

Burke, J. R., 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, 543–550. [336]

Danila, O., Steiner, S. H., and MacKay, R. J. (2008), "Assessing a Binary Measurement System," *Journal of Quality Technology*, 40, 310–318. [336,337,339]

——— (2010), "Assessment of a Binary Measurement System in Current Use," *Journal of Quality Technology*, 42, 152–164. [339]

——— (2012), "Assessing a Binary Measurement System With Varying Misclassification Rates Using a Latent Class Random Effects Model," *Journal of Quality Technology*, 44, 179–191. [343]

De Mast, J., Erdmann, T. P., and van Wierigen, W. N. (2011), "Measurement System Analysis for Binary Inspection: Continuous Versus Dichotomous Measurands," *Journal of Quality Technology*, 43, 99–112. [336]

Farnum, N. R. (1994), *Modern Statistical Quality Control and Improvement*, Belmont, CA: Duxbury Press. [336]

Griffiths, D. A. (1973), "Maximum Likelihood Estimation for the Beta-Binomial Distribution and an Application to the Household Distribution of the Total Number of Cases of a Disease," *Biometrics*, 29, 637–648. [337,339,344]

Maple 13 (2009), Waterloo, Ontario: Maplesoft. Available at *www.maplesoft.com*. [345]

Nelder, J. A., and Mead, R. (1965), "A Simplex Method for Function Minimization," *Computer Journal*, 7, 308–313. [337]

Pepe, M. S. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction* (1st ed.), New York: Oxford University Press. [336]

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing. Available at *www.R-project.org*. [337,344,345]

Self, S. G., and Liang, K. Y. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605–610. [338]