Improved Measurement-System Assessment for Processes with 100% Inspection

RYAN P. BROWNE, R. JOCK MACKAY, and STEFAN H. STEINER

Business and Industrial Statistics Research Group, Department of Statistics and Actuarial Sciences
University of Waterloo, Wateroo, N2L 3G1, Canada

We consider the assessment of an automated continuous measurement system used for 100% inspection in a high- volume manufacturing process. Because of the automation, we assume that there are no operator effects. If the system stores the measured values, we effectively know the current process mean and standard deviation. Because of the high volume, we have parts available with values spread across the whole distribution.

The standard plan for measurement-system assessment is to select k parts at random from the process and measure each of the selected parts n times. We then estimate the repeatability of the system using ANOVA. We propose two improvements. First, we demonstrate the substantial value of using the known process characteristics in the analysis. Second, we describe an alternative sampling plan where we deliberately select parts with extreme values from the population of measured parts to remeasure. We call this selection leveraging. We discuss the analysis of the leveraged plan and show that it is more efficient than the standard plan. We also discuss the planning and implementation of a leveraged assessment study and some associated issues and extensions.

Key Words: Components of Variation; Gauge R&R; Leveraging; Measurement System Assessment; Sample Size Determination.

Verifying the quality of a measurement system is important to any manufacturing process because all measurements are subject to error. As discussed in Shrout and Fleiss (1979), measurement errors can seriously affect process control and decisions about product quality. It is important to quantify such errors by assessing the measurement system.

Consider a measurement system used in a high-volume process for 100% inspection of a key product characteristic. Suppose that the measurement system is automated so that there are no operator effects and that the measured values are stored and can be traced back to the individual parts. Throughout the article, we refer to the stored measurements as the initial values.

One such measurement system is used to inspect journal diameters and several other characteristics of finished crankshafts in an engine-assembly plant. Another involves the measurement of many functions of a circuit board assembled for use in a hand-held electronic device. In both of these examples and many others, quality systems (such as ISO/TC 16949 or QS 9000) adopted by the manufacturer require periodic assessments of the measurement system to ensure that the current measurement variability is relatively small compared with the underlying process variation.

In a standard measurement system assessment plan (SP), we measure a sample of k parts n times each. The parts may be selected at random or, more commonly, haphazardly from the process. A common

Mr. Browne is a PhD candidate in the Department of Statistics and Actuarial Science at the University of Waterloo. His email address is rpbrowne@uwaterloo.ca.

Dr. MacKay is an Associate Professor in the Department of Statistics and Actuarial Science and Director of the Business and Industrial Statistics Research Group. He is a member of ASQ. His email address is rjmackay@uwaterloo.ca.

Dr. Steiner is an Associate Professor in the Department of Statistics and Actuarial Science. He is a senior member of ASQ. His email address is shsteiner@uwaterloo.ca.

statistical model for this plan is

$$Y_{ij} = P_i + E_{ij}, \quad i = 1, 2, \ldots, k \text{ and } j = 1, 2, \ldots, n \tag{1}$$

where P_i is a random variable representing the possible values for the true dimension of part i and E_{ij} is a random variable representing the measurement error. We assume that the part effects $\{P_i\}$ are independent and identically distributed normal random variables with mean μ and standard deviation σ_p, the measurement errors $\{E_{ij}\}$ are independent and identically distributed normal random variables with mean zero and standard deviation σ_m, and $\{P_i\}$ and $\{E_{ij}\}$ are mutually independent. Under the independence assumption, the standard deviation of Y_{ij}, called the total variation, is $\sigma_t = \sqrt{\sigma_p^2 + \sigma_m^2}$.

In a gauge R&R study (Automotive Industry Action Group (2002)), operators are an additional source of variation and one purpose of the study is to separate the contributions of the operators (reproducibility) from the other sources of variation (repeatability) in the measurement system. In the context described here, there are no operator effects and σ_m represents only the repeatability.

Bias, stability, and linearity are other important characteristics of a measurement system. To assess bias, we must measure parts or standards with known true values. To assess stability, we must look at the behavior of the measurement system over time. To assess linearity (i.e., measures of changes in the bias and variability over true part size), we must measure parts with a wide range of true values. In the plans that we propose, we can see if measurement variability changes with part size (one aspect of linearity) but assess neither bias nor stability. This is also the case with the standard plan that includes one or several operators.

To quantify the contribution of the measurement system to the total variation, the most commonly used performance measure in a manufacturing context is the gauge repeatability, $\theta = \sigma_m/\sigma_t$. An equivalent measure, more frequently used in a medical context, is the intraclass correlation coefficient, $\rho = \sigma_p^2/\sigma_t^2$, the ratio of process variation to the total variation. We note that ρ is the correlation between two measurements on the same part under the assumptions of model (1). Smaller values of θ or, equivalently, larger values of ρ correspond to a (relatively) less variable measurement system. In this paper, we discuss assessment in terms of θ.

To assess the measurement system, we can test the hypothesis

$$H_0 : \theta \geq \theta_0 \text{ versus } H_A : \theta < \theta_0 \tag{2}$$

to determine whether the measurement system is acceptable. Two commonly used values for θ_0 are 0.10 and 0.30. That is, the measurement system contributes 10% or 30% of the total variability seen in the process. To compare assessment plans, we can compare the power of the corresponding hypothesis tests over all values of θ when each test has equivalent size and the same total number of measurements.

In the context described here, we can calculate the current process mean μ and standard deviation σ_t with negligible error from recently stored values. If there is bias in the measurement system (i.e., the mean of the measurement errors $\{E_{ij}\}$ is not 0), then this bias is subsumed in the known value of μ and has no effect on the estimation of σ_m or θ.

Practitioners often do not recognize that there is substantial value in making use of the known process characteristics μ and σ_t. In the next section, we demonstrate the considerable value of this information and show that it should not be ignored. We then introduce a new plan that we call a leveraged measurement system assessment plan (LP). In an LP, k parts are deliberately selected based on their initial measurement from the population of measured parts and each selected part is then remeasured n times. We use the term leveraged because we recommend choosing parts with extreme initial values relative to the process mean μ. We can consider such a purposeful selection because there is no need to estimate the overall variation σ_t. Next, we compare the standard and leveraged plans. We then discuss how to design an LP to meet prespecified size and power requirements. Finally, we summarize the conclusions and discuss some other issues and extensions to leveraged plans.

Standard Plans

Suppose we measure k randomly selected parts n times each to get the data $\{y_{ij}, i = 1, 2, \ldots, k; j = 1, 2, \ldots, n\}$. To analyze these data, AIAG Automotive Industry Action Group (2002) recommends an analysis of variance (ANOVA) (see also Burdick et al. (2003)) that ignores the known values of μ and σ_t. This analysis is briefly outlined in the next section. Then we make use of the known parameter values in the analysis with one method based on ANOVA and a second using maximum-likelihood estimation (MLE). Finally, we show that the ANOVA method

is efficient and we demonstrate the value of using the information about the known process characteristics.

Analysis of the Standard-Plan Data

If we follow common practice and ignore the known process characteristics and the initial values $\{y_{i0}, i = 1, 2, \ldots, k\}$ for the selected parts, we estimate θ using the appropriate mean squares from an ANOVA. We get

$$\hat{\theta}_s = \sqrt{\frac{\text{MSW}}{\text{MSA} + (1 - 1/n)\text{MSW}}}, \tag{3}$$

where $\text{MSW} = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \overline{y}_{i.})^2/k(n-1)$ is the mean square among the repeated measured values within parts and $\text{MSA} = \sum_{i=1}^{k} n(\overline{y}_{i.} - \overline{y}_{..})^2/(k-1)$ is the mean square among the part averages.

In this context, to test the hypothesis (2), we use a standard F-test with size α and with power at θ_1 given by

$$P(F \geq c\, F_\alpha; v_1, v_2), \tag{4}$$

where

$$c = \left(1 + n\frac{1 - \theta_0^2}{\theta_0^2}\right) \Big/ \left(1 + n\frac{1 - \theta_1^2}{\theta_1^2}\right),$$

with $v_1 = k - 1$ and $v_2 = k(n-1)$.

Details of the derivation, expressed in terms of the intraclass correlation coefficient ρ, can be found in Donner and Eliasziw (1987), who use Equation (4) to determine appropriate sample sizes for a measurement-system assessment. They first rearrange the equation, expressing ρ_1 as a function of n and k when testing Equation (2) with $\alpha = 0.05$ and power 0.80. Then they plot contours of ρ_1 as a function of n and k. In the section on Selecting a Leveraged Plan, we use the same type of display to calculate sample sizes for a leveraged plan.

The standard analysis is appropriate only if the parts are selected at random because we are using the data to estimate σ_t. In this instance, we can incorporate the initial values into the analysis, which increases n by 1.

We expect the estimate (3) and the corresponding hypothesis test to perform relatively poorly because we are using only the data from the repeatedly measured parts to estimate σ_t.

Analysis Using the Known Process Characteristics

We look at two ways of incorporating the known values σ_t and μ into the estimation of θ. First, using ANOVA, we estimate θ by

$$\widehat{\theta}_a = \sqrt{\frac{\text{MSW}}{\sigma_t^2}}. \tag{5}$$

The corresponding estimator $\widetilde{\theta}_a$ of the estimate $\widehat{\theta}_a$ is distributed as the square root of a Chi-squared random variable. Note that we use a circumflex (ˆ) to overscore a parameter to denote the estimate (a number) and a overscore tilde (˜) to denote the corresponding estimator (a random variable). We have

$$E(\widetilde{\theta}_a) = \theta\sqrt{\frac{2}{v_2}}\frac{\Gamma((v_2+1)/2)}{\Gamma(v_2/2)}$$

and

$$\text{Var}(\widetilde{\theta}_a) = \theta^2\left[1 - \left(\sqrt{\frac{2}{v_2}}\frac{\Gamma((v_2+1)/2)}{\Gamma(v_2/2)}\right)^2\right]. \tag{6}$$

We use a standard χ^2-test with size α for testing Equation (2) with power at θ_1 given by

$$P\left(\chi^2_{k(n-1)} \leq \left[\frac{\theta_0}{\theta_1}\right]^2 \chi^2_{\alpha, k(n-1)}\right). \tag{7}$$

One benefit of this plan is that the properties of the estimate and test do not depend on how the parts are selected from the process because we are using the repeated measurements to estimate only σ_m. This means the parts do not have to be randomly selected. If we have selected the parts at random, we can include the initial values of the selected parts in the estimate, increasing the degrees of freedom from $k(n-1)$ to kn.

Maximum Likelihood

The n measurements ($n+1$ if we include the initial measurement) on the same part have joint distribution

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} \backsim N\left(\mu \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \sigma_t{}^2 \begin{bmatrix} 1 & 1-\theta^2 & \ldots & 1-\theta^2 \\ 1-\theta^2 & 1 & & \\ \vdots & & \ddots & \vdots \\ 1-\theta^2 & & \ldots & 1 \end{bmatrix}\right) \tag{8}$$

because the covariance between two measurements on the same part is σ_p^2.

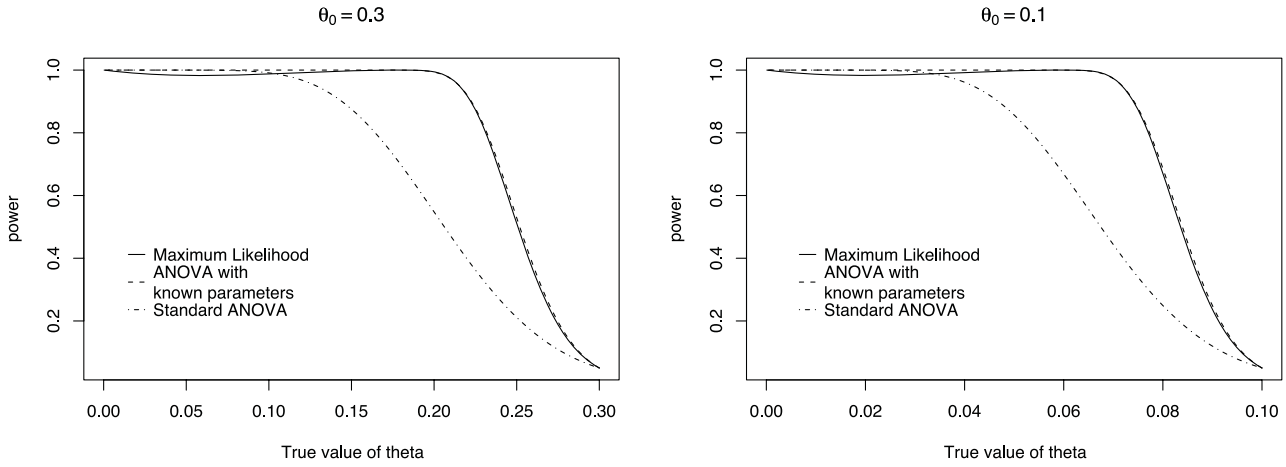

FIGURE 1. Power Curves for Testing the Hypothesis (2) When $\rho_0 = 0.80$ and $\rho_0 = 0.91$.

To perform the test, we calculate the MLE (see Appendix A) and then reject the hypothesis $\theta \geq \theta_0$ if the calculated value of $W(\theta_0; \hat{\theta})$ is less than Z_α, the α^{th} quantile of the standard normal distribution. The approximate power of the Wald test, when $\theta = \theta_1$, is

$$P\left\{Z \leq \left(Z_\alpha \left[J_1(\theta_0)\right]^{-1/2} + \theta_0 - \theta_1\right)\left[J_1(\theta_1)\right]^{1/2}\right\}. \quad (9)$$

We compare the three power functions (4), (7), and (9) in the next section.

Value of Using the Known Process Characteristics

We use power to compare the three hypothesis tests. All have the same size, $\alpha = 0.05$. To make the tests comparable, we do not include the initial values of the selected parts in the analysis. Figure 1 shows power curves, calculated from formulas (4), (7), and (9) when the sampling plan has ten parts with six repeated measurements on each part, a common choice in the SP (see Automotive Industry Action Group (2002)). We selected the values for θ_0 to match the standard cut-off values used in R&R studies. Similar pictures emerge for other plausible values of n, k, and θ_0.

Figure 1 indicates, not surprisingly, that maximum likelihood and ANOVA using the known characteristics are significantly more powerful than the standard ANOVA analysis. This strongly suggests that, when μ and σ_t are known, we should use this information.

Maximum likelihood should be the most efficient method of estimation. However, Figure 1 shows that ANOVA using the known process characteristics has a slightly higher power curve than the MLE. This likely happened because the power of the Wald test used in Figure 1 is based on an approximation using the asymptotic distribution of the MLE. We verified using simulation for a range of sample sizes and θ_0 that the Wald and ANOVA tests are virtually equivalent in terms of power. Given the simplicity of the calculation, the ANOVA test is a clear winner.

In summary, when the process parameters are known (or well estimated using the available data), we recommend the analysis using ANOVA with known parameters. This approach has a closed-form estimate for θ and performs as well as the Wald test when θ_0 is less than 0.30, the typical values of interest. In addition, because with the ANOVA and known process characteristics we only estimate σ_m (and not σ_p), we do not require that the selected parts be representative of the process. As well, there is no need to select as many as 10 parts. In line with this observation, Steiner and Mackay (2005, chapter 7), suggests using only three parts having initial values spread across the known distribution of measured values.

Leveraged Plans

For the LP, we sample parts based on their initial measurements as observed and stored in regular production. In particular, we sample parts that are extreme relative to the known process mean μ. Because we are dealing with a high-volume process, we can quickly find parts with initial values in the tails of the distribution. Again, we present three approaches for testing the hypothesis of interest. The first method uses maximum likelihood, the second is

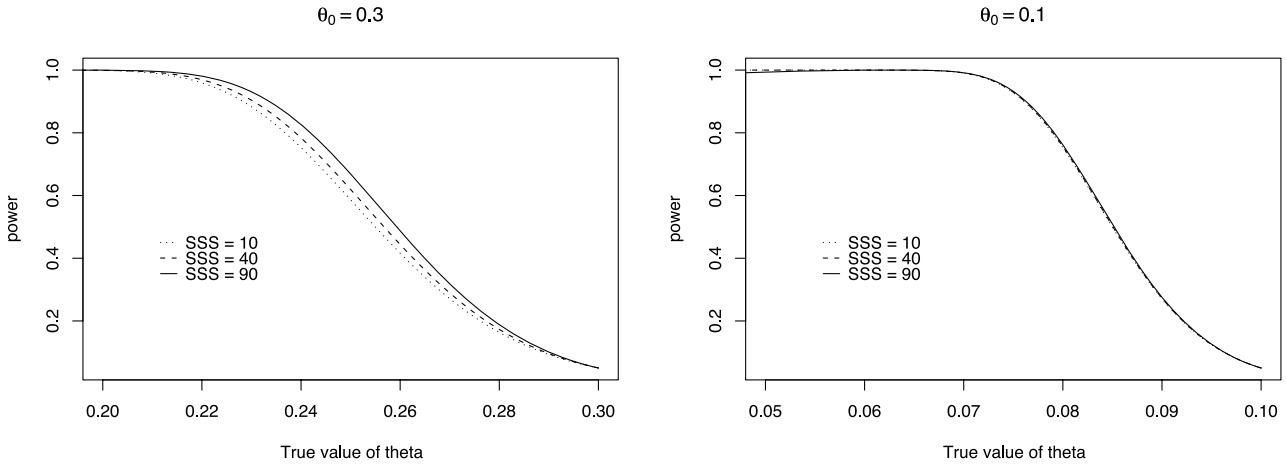

FIGURE 2. Power Curves for Testing the Hypothesis (2) When $\theta_0 = 0.3$ and 0.1 at $\alpha = 0.05$, with Six Repeated Measurements and Ten Parts Having $SSS = 10$, 40, and 90.

based on a regression estimate, and the third is a weighted average of the regression estimate and the ANOVA estimate with known parameter values.

Maximum Likelihood and Fisher Information

For a single part, the joint distribution of the initial value Y_0 and the n repeated measurements is given by Equation (8). The conditional distribution of the repeated measurements $\{Y_1, \ldots, Y_n\}$, given that the selected part has initial measured value $Y_0 = y_0$, is then

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \Bigg| Y_0 = y_0 \Bigg) \backsim N\left(\mu + \left(1-\theta^2\right)(y_0 - \mu)\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \sigma_t^2\theta^2 \begin{bmatrix} (2-\theta^2) & & (1-\theta^2) \\ & \ddots & \\ (1-\theta^2) & & (2-\theta^2) \end{bmatrix}\right). \quad (10)$$

We use the Wald test to test the hypothesis in Equation (2). The power is given by

$$P\left\{Z \leq \left(Z_\alpha \left[J_2(\theta_0)\right]^{-1/2} + \theta_0 - \theta_1\right)\left[J_2(\theta_1)\right]^{1/2}\right\}, \quad (11)$$

where $J_2(\theta_1)$ is given in Appendix B. The Fisher information (26) depends on the initial values of the chosen parts through the sum of the squared standardized values (SSS) of the initial measurements,

$$SSS = \sum_{i=1}^{k} z_{i0}^2 = \sum_{i=1}^{k}\left[\frac{y_{i0} - \mu}{\sigma_t}\right]^2. \quad (12)$$

Selecting the parts to increase the z_{i0}^2's, i.e., the deviation of the intial measurements from the mean μ, will increase the Fisher information. Figure 2 shows the effect of changing SSS on the power of the Wald test when 10 parts are selected and measured 6 times each. Note that $SSS = 10$ corresponds to the average value of SSS if 10 parts are selected at random.

To get $SSS = 40$, we can select 10 parts with initial values equal to $\mu \pm 2\sigma_t$ or any other set of $\{y_{10}, \ldots, y_{k0}\}$ where

$$\sum_{i=1}^{k}\left[\frac{y_{i0} - \mu}{\sigma_t}\right]^2 = \sum_{i=1}^{k} z_{i0}^2 = 40.$$

In the analysis, the properties of the test will depend on the actual value of SSS achieved.

Regression Estimate

Maximum likelihood is an efficient method of estimation but, when an explicit expression of the MLE cannot be found, it can be useful to look for simple, efficient estimators to avoid complex calculations. The distribution of the average of the repeated measurements on single part i, given the initial mea-

surement y_{i0}, is

$$\overline{Y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij} \backsim N\left(\mu + (1-\theta^2)(y_{i0}-\mu), \sigma_t^2\theta^2\left[\frac{n+1}{n} - \theta^2\right]\right). \quad (13)$$

Notice that the $\overline{Y}_{i.}$'s are independent and have constant variance. The conditional mean is a linear function of $y_{i0}-\mu$ and so we can estimate $\beta = 1-\theta^2$ and hence θ using regression. See Appendix A for details. The estimate of β is

$$\hat{\beta} = \frac{\sum_{i=1}^{k} r_i z_{i0}}{\sum_{i=1}^{k} z_{i0}^2}, \quad (14)$$

where $r_i = (\overline{y}_{i.} - \mu)/\sigma_t$. Then the estimate of θ is $\hat{\theta}_r = \sqrt{1-\widehat{\beta}}$. If $\widehat{\beta}$ is negative, we set $\hat{\theta}_r$ to zero.

The regression estimator $\widetilde{\theta_r}$ is approximately unbiased with

$$\mathrm{Var}\left(\widetilde{\theta_r}\right) \approx \frac{\theta\left[\frac{n+1}{n} - \theta^2\right]}{4\sum_{i=1}^{k} z_{i0}^2}. \quad (15)$$

We see from the denominator of Equation (15) that this estimator has smaller variance when we choose parts that increase *SSS*, the sum of squares of the standardized initial measurements (12).

Using the regression estimate, we reject the hypothesis $\theta \geq \theta_0$ if

$$\frac{\widehat{\theta}_r - \theta_0}{\mathrm{Var}(\tilde{\theta}_r; \theta = \theta_0)^{1/2}} \leq Z_\alpha. \quad (16)$$

The approximate power of the test when $\theta = \theta_1$ is given by

$$P\left\{ Z \leq \left(Z_\alpha \left[\mathrm{Var}(\tilde{\theta}_r; \theta = \theta_0)\right]^{1/2} + \theta_0 - \theta_1 \right) \times \left[\mathrm{Var}(\tilde{\theta}_r; \theta = \theta_1)\right]^{-1/2} \right\}. \quad (17)$$

Note that this test uses the initial values but not the information from the variability of the repeated measurements.

Combining the ANOVA and Regression Estimates

As shown in the next subsection, the test based on the regression estimator does not perform well when compared with the MLE. To improve the power, we propose estimating θ using a weighted average of the regression and ANOVA estimates. The ANOVA estimator, as described above, uses only the repeated measurements for each selected part and not the initial values. It is easy to see that the two estimators are statistically independent, so we would expect a linear combination to be better.

The variances of the estimators (6) and (15) depend on the unknown θ, which makes finding optimal weights impossible. We can, however, find optimal weights at $\theta = \theta_0$, the hypothesized value in Equation (2). See Appendix D for details. To simplify the algebra, we find the optimal weighted average for estimating θ^2. The combined estimate is

$$\widehat{\theta}_c = \sqrt{w_1\left(1 - \frac{\sum_{i=1}^{k} r_i z_{i0}}{\sum_{i=1}^{k} z_{i0}^2}\right) + w_2\left(\frac{\mathrm{MSW}}{\sigma_t^2}\right)}, \quad (18)$$

with MSW defined as in Equation (3) and

$$w_1 = \frac{2n\theta_0^2\sum_{i=1}^{k} z_{i0}^2}{2n\theta_0^2\sum_{i=1}^{k} z_{i0}^2 + k(n-1)\left(\left[1-\theta_0^2\right]n+1\right)}$$

$$w_2 = \frac{k(n-1)\left(\left[1-\theta_0^2\right]n+1\right)}{2n\theta_0^2\sum_{i=1}^{k} z_{i0}^2 + k(n-1)\left(\left[1-\theta_0^2\right]n+1\right)}.$$

From Appendix B, the variance of $\tilde{\theta}_c$ with the true value θ is

$$\mathrm{Var}\left(\tilde{\theta}_c; \theta\right) \approx \frac{1}{2}\frac{\theta^2\left(\left[1-\theta^2\right]n+1\right)}{2n\theta^2\sum_{i=1}^{k} z_{i0}^2 + k(n-1)\left(\left[1-\theta^2\right]n+1\right)}. \quad (19)$$

We construct the test of the hypothesis (2) using the normal approximation for the estimator. The approximate power of the test is

$$P\left\{ Z \leq \left(Z_\alpha \left[\mathrm{Var}(\tilde{\theta}_c; \theta = \theta_0)\right]^{1/2} + \theta_0 - \theta_1 \right) \times \left[\mathrm{Var}(\tilde{\theta}_c; \theta = \theta_1)\right]^{-1/2} \right\}. \quad (20)$$

To derive the expression for the power, we use a normal approximation to the distribution of MSW. Because the numerator of MSW has a chi-square distribution with $k(n-1)$ degrees of freedom, a normal approximation is reasonable if $k(n-1)$ is larger than 30. The common choice for an SP is $k = 10$ and $n = 6$, which means $k(n-1) = 50$.

TABLE 1. Example LP Data

Part	y_0	Repeats 1	2	3	4	5	Average	Standard deviation
1	1.42	1.36	1.30	1.42	1.35	1.33	1.352	0.0444
2	1.96	1.89	1.97	1.82	1.84	1.94	1.892	0.0638
3	2.26	2.11	2.19	2.26	2.16	2.21	2.186	0.0559
4	7.76	7.81	7.83	7.72	7.66	7.80	7.764	0.0716
5	7.98	7.97	8.06	7.82	7.83	7.88	7.912	0.1018
6	8.78	8.67	8.59	8.62	8.52	8.62	8.604	0.0550

LP Analysis Example

We present an artifical example to illustrate the numerical calculations. Using our process knowledge, we have $\mu = 5$ and $\sigma_t^2 = 2$, and we wish to test Equation (2) when $\theta_0 = 0.1$. We selected six parts and remeasured them five times each. The six parts selected had three small and three large initial values, with $SSS = 30.175$. The results are shown in Table 1.

The estimates and standard errors for the MLE, ANOVA with known process characteristics, and ANOVA and regression combined are shown in Table 2. The weights for the combined estimate are $w_1 = 0.022$ and $w_2 = 0.978$. The standard error for each estimation method was evaluated at each estimate and under the null hypothesis that $\theta_0 = 0.1$. We also give the result of the hypothesis tests when the size of the test is $\alpha = 0.05$.

We can check some of the model assumptions using the residuals from the repeated measurements. A QQ plot of these residuals will check the normality assumption and a plot of the residuals against the intial values can indicate if the measurement variation depends on the true value.

Comparison of the LP Analysis Methods

We use power to compare the three methods of testing the hypothesis (2). For the MLE, we use the Wald test with power given by Equation (11). For the regression and combined estimators, the power is given by Equations (17) and (20), respectively. In all the comparisons, we use a sample of 10 parts with $SSS = \sum_{i=1}^{k} z_{i0}^2 = 40$ and six repeated measurements on each part.

Figure 3 shows that the test based on the regression estimator performs poorly and that the test based on the combined estimator performs almost as well as the Wald test based on the MLEs, with the advantage of a closed-form solution. We see similar results for other values of n, k, and SSS.

The Value of Leveraging

We can use leveraging whenever there is a supply of parts with measured initial values so that we can select extremes. We have such a supply if the measurement system is used routinely in production as described in the Introduction. There is a small cost to find the extreme parts but this cost is likely comparable with that incurred if random selection is

TABLE 2. Example LP Analysis When Testing $\theta_0 = 0.1$

Method	Estimate $\widehat{\theta}$	Standard error at $\theta = \widehat{\theta}$	Standard error at $\theta = \theta_0$	Test result
Maximum likelihood	0.0500	0.00644	0.0128	Reject
Regression	0.0571	0.02377	0.0314	Accept
ANOVA with known process characteristics	0.0478	0.00687	0.0144	Reject
ANOVA and regression combined	0.0481	0.00692	0.0071	Reject

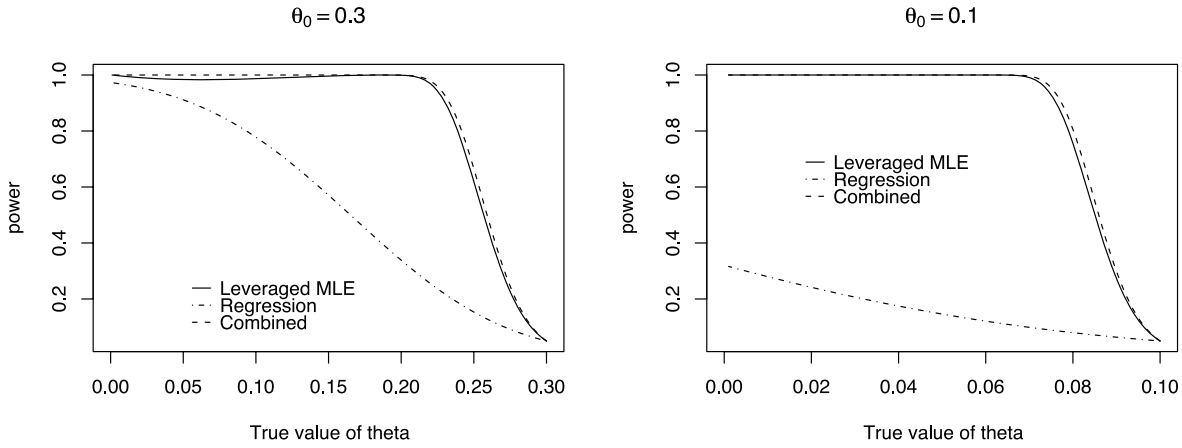

FIGURE 3. Power Curves for Testing the Hypothesis (2) When $\theta_0 = 0.3$, $\theta_0 = 0.1$, $\sum z_{i,0}^2 = 40$, and $k = 10$, $n = 6$.

properly implemented. What is the gain? To address this question, we compare power curves for leveraged and standard plans with the same number of parts and repeated measurements.

A common sample size used in a SP is 10 parts and 6 repeated measurements. For this plan, we select the parts at random and use the known μ and σ_t. We consider both the case when the initial values are included in the calculations and the more common case when they are not. For the LP, we select the 10 parts so that $\sum_{i=1}^{k} z_{i0}^2 = 40$. Note that about one part in 20 will have z_0 greater than 2. We generated the power curves for testing the hypothesis (2) using the Wald test for each plan. The power of the Wald test for the SP is found in Equation (9). We increase n by 1 for the case that includes the initial values. The power of the Wald test for the LP uses Equation (11). We used maximum likelihood for the power curves because this way, the three plans can be compared fairly.

Figure 4 shows that the LP is more powerful than the SP when the 10 parts are used with $SSS = 40$ and there are six repeated measurements on each part. The gain is greater when the initial values are not used in the analysis of the SP. We see similar results for other scenarios.

We can also quantify the effects of leveraging by comparing sample sizes. For example, suppose we select five parts with initial measured values about two standard deviations from the mean and measure each

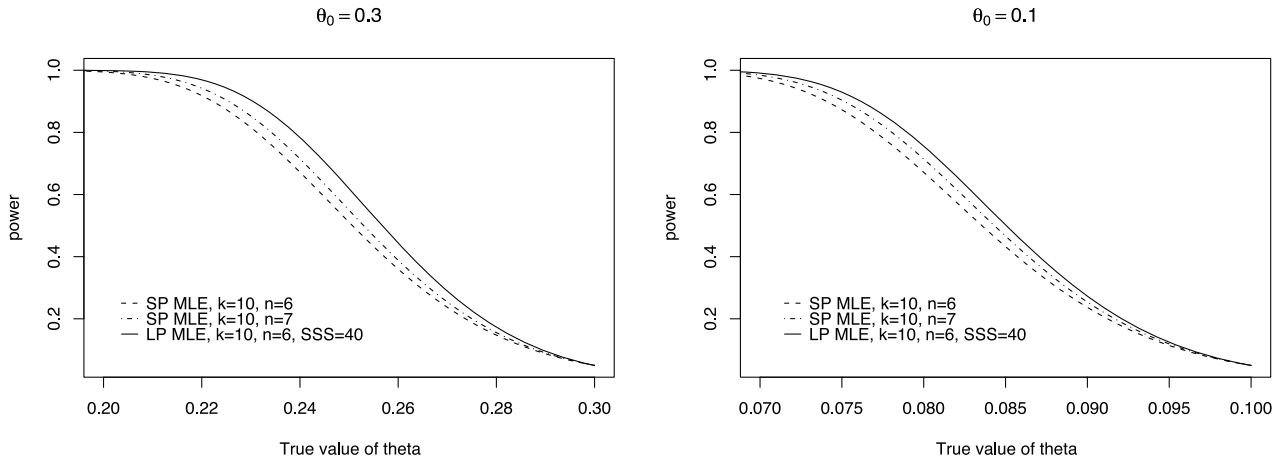

FIGURE 4. Power Curve for Leveraged and Standard Plans from Testing the Hyppothesis (2) When $\theta_0 = 0.3$ and 0.1, Ten Parts Having $SSS = \sum z_{i,0}^2 = 40$ and 6 Repeated Measurements.

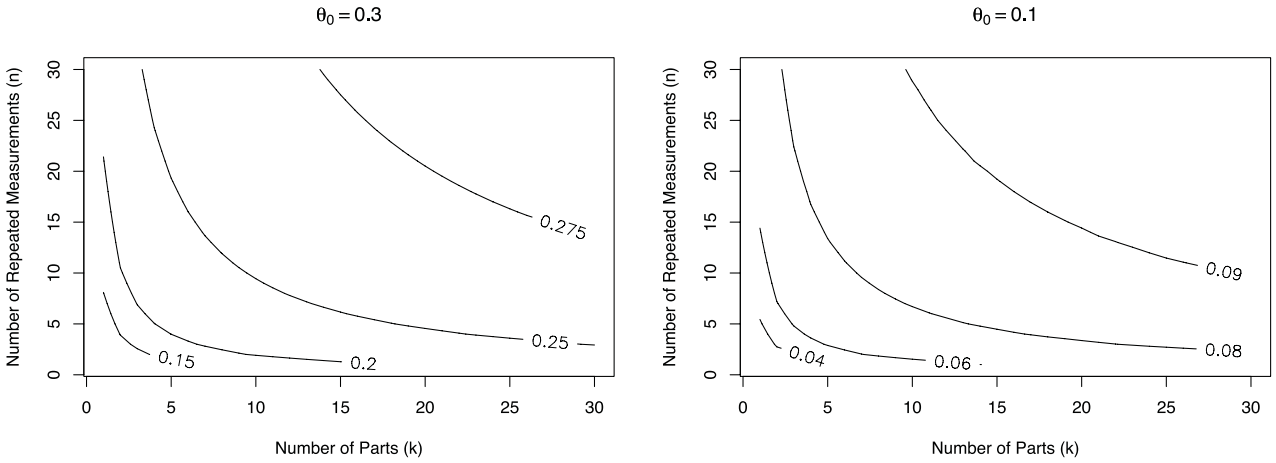

FIGURE 5. Contours of θ for Testing the Hypothesis (2) When $\theta_0 = 0.3$ and 0.1 with Size 0.05, Power 0.80, and $\sum_{i=1}^{k} z_{i,0}^2 = 4k$.

five times. Using the test based on the combined estimate, the power of the test for $\theta_0 = 0.3$ at $\theta_1 = 0.209$ is about 0.80. To get the same power with the SP and the test based on the ANOVA estimate, we would need to measure five randomly selected parts 6.4 (actually 7) times each.

The conclusion is that leveraging increases power with little or no increase in cost. Although not presented here, we see corresponding improvements in the precision of the estimate of θ when we use leveraging.

Selecting a Leveraging Plan

To plan a measurement system assessment to carry out the test of hypothesis (2), we need to specify the size and θ_0 as well as the desired power and θ_1. For size 0.05, $\theta_0 = .1$, and $\theta_0 = .3$, Figure 5 shows the values of θ_1 that have 0.80 power when testing with various values of k and n. These contours show the (n, k) combinations that satisfy the required size and power for the assumed values of θ_1. Again we assume that each selected part contributes, on average, 4 to the sum SSS given in Equation (12).

To obtain the contours in Figure 5, we determined, for all discrete points (n, k), the values of θ that have power 0.80 using the asymptotic Wald test when testing θ_0 in Equation (2). To determine the power of the Wald test for an LP, we use Equation (11).

Figure 5 can be used to determine the necessary sample sizes for testing Equation (2) when $\theta_0 = 0.30$ and $\theta_0 = 0.10$ with size 0.05 and power 0.80. For example, suppose we wish to determine if the gauge repeatability is 30% or less of the total variation (i.e., $\theta_0 \leq 0.30$) and it is currently thought that θ is around 0.20. The point on Figure 5, with $\theta_0 = 0.30$, that corresponds to five parts with $\sum_{i=1}^{5} z_{i0}^2 = 20$ and five repeated measurements on each part satisfies the needs of this investigation because it is above the $\theta = 0.20$ contour. Similar figures for other values of SSS are available from the authors on request.

The properties of the test associated with an LP depend on the part selection only through SSS. We could, for example, choose only large parts so that all z_{i0} are positive. We recommend instead that a balanced sample is selected with roughly equal numbers of large and small parts. With this plan, we can check if the measurement variability σ_m is constant across the range of true part dimensions.

We may occasionally find a part with a very large or small initial value, with $|z_{i0}| \geq 3$, say. We do not recommend using such a part in an LP because the reason for the large value may be due to some special cause acting in the measurement system. This unusual outcome should be investigated separately.

The contour plots in Figure 5 are useful in the planning of an LP. In executing the plan, we do not expect to find k parts with initial measurements so that SSS exactly meets the specified value. The analysis depends only on the realized value of SSS. We can use Equation (19) to estimate θ and, substituting the estimate, Equation (19) to get an approximate standard error leading to a Z-test or an approximate confidence interval.

Conclusions and Discussion

We compared two classes of plans for assessing a measurement system: the standard plan, in which parts are selected at random from the process, and a leveraged plan, where the parts are selected based on relatively large or small initial values. The leveraged plan is feasible in the context we have described, where an automated gauge (without an operator source of variation) is used for 100% inspection and the measured values are recorded. In this instance, we have shown that the LP can produce modest gains in efficiency at little or no extra cost or complexity. In cases where leveraging is not used, we have shown the somewhat obvious result that, if the overall process mean and variance are known, this information should be included in the analysis of any measurement assessment plan. What may be surprising is the increase in efficiency due to this information. We have also noted that we can avoid random selection of the parts to be remeasured when we have known process mean and variance.

All of our results are given in terms of power comparisons. We see similar gains in efficiency in using the known parameter values and leveraging if we make the comparisons in terms of standard errors of the estimates of θ. For an LP, we can use Equation (19) to estimate θ and, substituting the estimate of Equation (19), get an approximate standard error. This can then be used to get an approximate confidence interval for θ.

One issue that we have skirted is the choice of time period over which we use the data records to determine μ and σ_t. These parameters must describe the process behavior at the time we plan to assess the measurement system. This problem arises with both the LP and SP, and, in fact, whether initial values are available or not. We can only select a random sample of parts by first specifying a particular production period so that μ and σ_t have meaning. One possibility is to apply a run chart to the stored data and look for a substantial recent period of stability.

When implementing an LP, we can avoid searching for previously measured parts by first estimating μ and σ_t using some period of production. Then we can flag and save a set of extreme parts from current production for the remeasurement.

Leveraging can be applied to other variance component problems where one source of variation can be held fixed (e.g., the process) and the other can be varied (e.g., the measurement system). One example is an assembly–disassembly experiment (see Steiner and Mackay (2005), 10.4) where we have an assembled product with several components. The goal is to determine which is the greater source of output variation, the components or the assembly operation. We select $k = 2$ products for the study. Each product is disassembled, reassembled, and then remeasured. These three steps are repeated several times. Performing the experiment in this way ensures the variation from the components is held fixed while the assembly contribution varies in each run. Again, we assume that the distribution of the output of interest is known or estimated from production records with little error. We use leveraging by selecting the two products for the experiment to have extreme and opposite performance when initially measured.

We can also exploit known process characteristics and purposeful sampling when assessing binary measurement systems used for 100% inspection. See Danila et al. (2008) for details.

In this work, we have assumed that the process mean, μ, and variance, σ_t^2, are known. For the context we have described, this is a reasonable assumption. In many other situations, we can use leveraging by first randomly selecting a baseline sample of b parts to estimate μ and σ_t^2. Second, we select k extreme parts from the baseline sample and remeasure each n times. We expect leveraging to provide substantial benefit. A key issue is how to allocate resources between the two stages of the study. See Browne et al. (2009) for details.

Acknowledgments

Ryan Browne acknowledges the financial support of the National Science and Engineering Research Council of Canada (NSERC), Research in Motion Limited, and the Mathematics of Information Technology and Complex Systems (MITACS). Stefan Steiner and Jock MacKay acknowledge the financial support of NSERC. We thank the editor and the referees for numerous suggestions that greatly improved the paper.

Appendix A
Analysis Using the Known Process Characteristics

The covariance matrix in Equation (8) has a special form with the following well-known properties (Dillon and Goldstein (1984)). If $\mathbf{W} = a\mathbf{I}_n + b\mathbf{J}_n$, where $\mathbf{I}_n$ is an n by n identity matrix and $\mathbf{J}_n$ is an n by n matrix of ones, then

$$\det(\mathbf{W}) = a^{n-1}\left[a + bn\right] \quad \text{and} \quad \mathbf{W}^{-1} = \frac{1}{a}\mathbf{I}_n - \frac{b}{a(a+bn)}\mathbf{J}_n. \tag{21}$$

We use these results to write the log -likelihood for each part and, because observations from the k different parts are independent, we add the individual components to get the overall log likelihood,

$$\begin{aligned} l_1(\theta) = &-\frac{k}{2}\left\{n\log\sigma_t^2 + (n-1)\log\theta^2 + \log\left[n - \theta^2(n-1)\right]\right\} \\ &- \frac{1}{2\sigma_t^2\theta^2\left[n-\theta^2(n-1)\right]}\left\{\left[n-\theta^2(n-1)\right]\mathrm{SSW}^* - (1-\theta^2)n^S\mathrm{SA}^*\right\}, \end{aligned} \tag{22}$$

where $\mathrm{SSW}^* = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij}-\mu\right)^2$ and $\mathrm{SSA}^* = \sum_{i=1}^{k} n\left(\overline{y}_{i.}-\mu\right)^2$.

We must numerically calculate the maximum likelihood estimate $\hat{\theta}$. We can, however, examine the asymptotic properties of the estimator using the Fisher information,

$$J_1(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2}\, l_1(\theta)\right] = 2\frac{nk(n-1)\left[(1-\theta^2)^2(n-1)+1\right]}{\theta^2\left[n-\theta^2(n-1)\right]^2}. \tag{23}$$

The Fisher information is the inverse of the asymptotic variance of the maximum likelihood estimator. The larger the information, the better the precision of the estimate.

To test the hypothesis (2), we use a Wald test based on the asymptotic distribution of the MLE. Suppose that $\tilde{\theta}$ is the maximum likelihood estimator of θ. Then, approximately, we have

$$W(\theta_0;\tilde{\theta}) = (\tilde{\theta}-\theta_0)\left[J_1(\theta_0)\right]^{1/2} \sim N(0,1). \tag{24}$$

Appendix B
Leveraged Plan Maximum Likelihood and Fisher Information

As in Equation (8), the covariance matrix in Equation (10) has a special form that allows us to easily write down an expression for its inverse and determinant.

Using these results, we can find the conditional likelihood (conditional on y_0) for the n repeated measurements on a single part. Because the repeated measurements for one part are independent of the repeated measurements from another part, the conditional log likelihood for k parts, each with n measurements, is the sum of their log likelihoods. Given the initial values, the log likelihood for n repeated measurements on k parts is then

$$\begin{aligned} l_2\left(\theta\,|y_{10},\ldots,y_{k0}\right) = &-\frac{nk}{2}\log\sigma_t^2\theta^2 - \frac{k}{2}\log\left[1+n(1-\theta^2)\right] - \frac{1}{2}\frac{1}{\sigma_t^2\theta^2\left[1+n(1-\theta^2)\right]} \\ &\times\left\{\left[1+n(1-\theta^2)\right]SSW + n\sum_{i=1}^{k}\left[\overline{y}_{i.}-\mu-(1-\theta^2)(y_{i0}-\mu)\right]^2\right\}, \end{aligned} \tag{25}$$

Note that this likelihood does not depend on the method we use to select the parts, only the initial values of these parts.

To estimate θ, we maximize Equation (25) numerically. We calculate the Fisher information directly from the log-likelihood function where the expectation is conditioned on the initial values.

$$J_2(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2} l_2(\theta)\right]$$

$$= \frac{1}{2}\frac{4\theta^2 kn^2}{\left[1+n(1-\theta^2)\right]^2} + \frac{4kn(1-\theta^2)(n+1)}{\left[1+n(1-\theta^2)\right]\theta^2} - \frac{2kn}{\theta^2} + \frac{4n}{\left[1+n(1-\theta^2)\right]}\frac{\sum (y_{i0}-\mu)^2}{\sigma_t^2}. \quad (26)$$

Appendix C
Regression Estimate

The distribution of the average of the repeated measurements on single part i, given the initial measurement y_{i0}, is

$$\overline{Y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij} \backsim N\left(\mu + (1-\theta^2)(y_{i0}-\mu),\ \ \sigma_t^2\theta^2\left[\frac{n+1}{n} - \theta^2\right]\right). \quad (27)$$

Notice that the $\overline{Y}_{i.}$'s are independent and have constant variance. The conditional mean is a linear function of $y_{i0} - \mu$ and so we can estimate $1 - \theta^2$ and hence θ using regression. Letting $\beta = 1 - \theta^2$, $R_i = (\overline{Y}_{i.} - \mu)/\sigma_t$, and $z_{i0} = (y_{i0} - \mu)/\sigma_t$ and rearranging Equation (27), we have

$$R_i = \beta z_{i0} + \epsilon_i \quad \text{where } \epsilon_i \sim N\left(0, \sigma_r^2\right) \quad (28)$$

and $\sigma_r^2 = \theta^2[(n+1)/n - \theta^2]$.

The estimate of β (Montgomery et al. (2001)) is

$$\hat{\beta} = \left(\sum_{i=1}^{k} r_i z_{i0}\right) \Big/ \left(\sum_{i=1}^{k} z_{i0}^2\right), \quad (29)$$

where $r_i = (\overline{y}_{i.} - \mu)/\sigma_t$. Then the estimate of θ is $\hat{\theta}_r = \sqrt{1-\widehat{\beta}}$. If $\widehat{\beta}$ is negative, we set $\hat{\theta}_r$ to zero.

Appendix D
Combining the ANOVA and Regression Estimates

If we have two unbiased independent estimators of θ, $\tilde{\theta_1}$ and $\tilde{\theta_2}$ with known variances σ_1^2 and σ_2^2, the optimal linear combination is

$$\tilde{\theta}^2 = w_1\tilde{\theta_1}^2 + w_2\tilde{\theta_2}^2 = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2}\tilde{\theta_1}^2 + \frac{1/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}\tilde{\theta_2}^2. \quad (30)$$

The variance of the combined estimator is

$$\text{Var}\left(\tilde{\theta}^2\right) = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}. \quad (31)$$

Now, if we obtain $\tilde{\theta_1}^2$ using Equation (14) and $\tilde{\theta_2}^2$ from Equation (5), we can easily show that the two estimators are independent. Applying Equation (30), we obtain the combined estimate

$$\widehat{\theta}_c = \sqrt{w_1\left(\frac{\sum_{i=1}^{k} r_i z_{i0}}{\sum_{i=1}^{k} z_{i0}^2}\right) + w_2\left(1 - \frac{\text{MSW}}{\sigma_t^2}\right)} \quad (32)$$

with MSW defined as in Equation (3) and

$$w_1 = \frac{2n\theta_0^2\sum_{i=1}^{k} z_{i0}^2}{2n\theta_0^2\sum_{i=1}^{k} z_{i0}^2 + k(n-1)\left(\left[1-\theta_0^2\right]n+1\right)}$$

$$w_2 = \frac{k(n-1)\left(\left[1-\theta_0^2\right]n+1\right)}{2n\theta_0^2\sum_{i=1}^{k} z_{i0}^2 + k(n-1)\left(\left[1-\theta_0^2\right]n+1\right)}.$$

References

Automotive Industry Action Group (2002). *Measurement Systems Analysis*, 3rd edition. Southfield, MI.

Browne, R. P.; Steiner, S. H.; and Mackay, R. J. (2009). "Two Stage Leverged Measurement System Assessment". *Technometrics*, to appear.

Burdick, R. K.; Borror, C. M.; and Montgomery, D. C. (2003). "A Review of Methods for Measurement Systems Capability Analysis". *Journal of Quality Technology* 35, pp. 342–354.

Danila, O.; Steiner, S.; and Mackay, R. J. (2008). "Assessing a Binary Measurement System". *Journal of Quality Technology* 40, pp. 312–320.

Dillon, W. and Goldstein, M. (1984). *Multivariate Analysis Methods and Applications*. John Wiley & Sons.

Donner, A. and Eliasziw, M. (1987). "Sample Size Requirements for Reliability Studies". *Statistics in Medicine* 6, pp. 441–448.

Montgomery, D. C.; Peck, E. A.; and Vining, G. G. (2001). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Shrout, P. E. and Fleiss, J. L. (1979). "Intraclass Correlations: Uses in Assessing Rater Reliability". *Psychological Bulletin* 86, pp. 420–428.

Steiner, S. H. and Mackay, R. J. (2005). *Statistical Engineering: An Algorithm for Reducing Variation in Manufacturing Processes*. ASQ Quality Press.