This paper was presented at the Journal of Quality Technology Session at the 47th Annual Fall Technical Conference of the Chemical and Process Industries Division and Statistics Division of the American Society for Quality and the Section on Physical & Engineering Sciences of the American Statistical Association in El Paso, Texas, October 16–17, 2003.

A Review of Methods for Measurement Systems Capability Analysis

RICHARD K. BURDICK, CONNIE M. BORROR, and DOUGLAS C. MONTGOMERY

Arizona State University, Tempe, AZ 85287

We review methods for conducting and analyzing measurement systems capability studies, focusing on the analysis of variance approach. These studies are designed experiments involving crossed and possibly nested factors. The analysis of variance is an attractive method for analyzing the results of these experiments because it permits efficient point and interval estimation of the variance components associated with the sources of variability in the experiment. In this paper we demonstrate computations for the standard two-factor design, describe aspects of designing the experiment, and provide references for situations where the standard two-factor design is not applicable.

Introduction

DETERMINING the capability of a measurement system is an important aspect of most process and quality improvement efforts. Indeed, in any activity involving measurements, some of the observed variability will arise from the units that are measured and some variability will be due to the measuring instrument or gauge. The purposes of most measurement systems capability studies are to: (1) determine how much of the total observed variability is due to the gauge; (2) isolate the sources of variability in the system; and (3) assess whether the gauge is capable (that is, determine if it is suitable for use in the broader project or application). In many measurement systems capability studies, the gauge is used to obtain replicate measurements on units by several different operators, for different set-ups, or for different time periods. In these types of studies, two components of measurement systems variability are frequently generated: repeatability and reproducibility. Repeatability represents the variability from the gauge or measurement instrument when it is used to measure the same unit (with the same operator or set-up or in the same time period). Reproducibility reflects the variability arising from different operators, set-ups, or time periods. These studies are often referred to as gauge repeatability and reproducibility (GR&R) studies.

Two methods commonly used in the analysis of a GR&R study are: (1) an analysis of variance approach followed by estimation of the appropriate variance components; and (2) a tabular algorithm that relies on the range method to estimate the standard deviations of the components of gauge variability. We focus on the analysis of variance approach because the method is easy and widely available to practitioners, it can be adapted to deal with very complex experiments, and it admits confidence interval estimates of the important components of gauge variability. Furthermore, the properties of these confidence intervals are reasonably well understood. The tabular approach cannot be applied to any study other than the traditional two-factor design, and it

Dr. Burdick is a Professor in the W. P. Carey School of Business. He is a Member of ASQ. His email address is richard.burdick@asu.edu.

Dr. Borror is a Senior Lecturer in the Department of Industrial Engineering. She is a Senior Member of ASQ.

Dr. Montgomery is a Professor in the Department of Industrial Engineering. He is a Fellow of ASQ.

does not allow the analyst to obtain confidence intervals. Ballard, McCormack, Moore, Prins, Tobias, and Pore (1997) provided additional comparisons of these two methods.

In the following sections we review the standard GR&R experiment, provide a numerical example, discuss issues concerning design of a GR&R experiment, and cite additional references for more complex designs. Some aspects of measurement systems, such as calibration and assessing linearity, are beyond the scope of this paper. References that consider measurement systems analysis in a broader context include the manuals by the Automotive Industry Action Group (AIAG (1995, 2002), Horrell (1991), and Croarkin (2002, Chapter 2)).

Parameters of Interest

The purpose of a GR&R study is to determine if the variability of the measurement system is small relative to the variability of the monitored process. Several commonly reported ratios in GR&R studies are functions of the parameters in Table 1. These parameters describe the variation in the monitored process and the variation in the measurement system. We purposely avoid assignment of the terms "repeatability" and "reproducibility" to any of these parameters because they are defined differently by some authors (see, e.g., Vardeman and VanValkenburg (1999)), and such labels are not needed to address the questions of interest.

The precision-to-tolerance ratio (PTR) is a function of γ_M expressed as

$$\text{PTR} = \frac{k\sqrt{\gamma_M}}{USL - LSL} \times 100\%, \qquad (1)$$

where USL and LSL are specification limits (pass/fail) and k is either 5.15 or 6. The value $k = 6$ corresponds to the number of standard deviations between the "natural" tolerance limits of a normal process. The value $k = 5.15$ corresponds to the limiting value of the number of standard deviations between bounds of a 95% tolerance interval that contains at least 99% of a normal population. Montgomery and Runger (1993a) stated that PTR values of 10% or less indicated the measurement system is adequate. This recommendation is consistent with the recommendation of the AIAG Measurement Systems Analysis manual (1995, p. 60). Mader, Prins, and Lampe (1999) referenced Wheeler and Lyday (1989) for an "arbitrary" rule that states a measurement system is inadequate if the PTR exceeds 20%. Barrentine (1991, p. 10) provided a rule that states a measurement system is unacceptable if the PTR exceeds 30%. Montgomery and Runger (1993a) and Mader et al. (1999) noted that the PTR does not necessarily give a good indication of how well a measurement system performs for a particular process. This is because a process with a high capability can tolerate a measurement system with a higher PTR than a process that is not as capable. For this reason, the adequacy of a process is more often determined by some function of ρ_P (or, alternatively, ρ_M, since $\rho_M = 1 - \rho_P$). For example, the signal-to-noise ratio (SNR) defined by AIAG (1995, p. 32) can be written as a function of ρ_P. In particular,

$$SNR = \sqrt{\frac{2\rho_P}{1-\rho_P}}. \qquad (2)$$

AIAG (1995) defined SNR as the number of distinct levels of categories that can be reliably obtained from the data. A value of five or greater is recommended, and a value less than two indicates the measurement system is of no value in monitoring the process.

Another function of ρ_P defined by Mader et al. (1999) and Wheeler (1992) is the discrimination ratio

$$\text{DR} = \frac{1+\rho_P}{1-\rho_P}.$$

Mader et al. (1999) stated that DR must exceed four for the measurement system to be adequate.

A discussion of the relationships between ρ_P, PTR, and the capability of the monitored process is

TABLE 1. Parameters of Interest in a GR&R Study

Parameter	Definition
γ_P	Variance of the monitored process
γ_M	Variance of the measurement system
$\gamma_T = \gamma_P + \gamma_M$	Total variance of the response variable
$\rho_P = \gamma_P/\gamma_T$	Proportion of total variance due to process
$\rho_M = \gamma_M/\gamma_T$	Proportion of total variance due to measurement system

provided by Majeske and Andrews (2002). Although we focus on the parameters in Table 1, we note that other authors have proposed alternative measures of measurement system performance (see, e.g., Vardeman and VanValkenburg (1999), van den Heuvel and Trip (2002), and Larsen (2002)).

Confidence Intervals

Montgomery and Runger (1993b), Conors, Merrill, and O'Donnell (1995), Burdick and Larsen (1997), Vardeman and VanValkenburg (1999), Hamada and Weerahandi (2000), and Chiang (2001) have noted the importance of computing confidence intervals in a GR&R study. We review two approaches for constructing confidence intervals in this paper.

The first approach is based primarily on the modified large-sample methods (MLS) first proposed by Graybill and Wang (1980) and summarized in the book by Burdick and Graybill (1992). This approach provides closed-form intervals and was applied to the standard two-factor design by Burdick and Larsen (1997).

The second approach is based on a computer-intensive method referred to as generalized intervals. Tsui and Weerahandi (1989) introduced the concept of generalized inference for testing hypotheses, and Weerahandi (1993) introduced generalized confidence intervals in situations where exact methods do not exist. This method was used in a GR&R two-factor study by Hamada and Weerahandi (2000). Chiang (2001) proposed a method called surrogate variables that produces the same intervals, and also applied this technique to the two-factor GR&R model. To compute a generalized confidence interval, one needs a generalized pivotal quantity (GPQ) with a distribution that is free of the parameters under study. Approximate confidence intervals are then constructed by computing required percentiles of the GPQ using either numerical integration or simulation. We demonstrate this process in the numerical example that follows later in this paper.

Empirical comparisons of the MLS and generalized intervals suggest the approaches provide comparable intervals in most cases. Thus, it is often a matter of preference as to which method an investigator employs. The fact that MLS intervals are written in closed-form makes them particularly amenable to computation in spreadsheet programs. An advantage of the generalized intervals approach is that it offers a general strategy for constructing confidence intervals. Unlike MLS intervals that are design specific, generalized intervals are easy to derive for complex designs that include crossed and nested factors, and for models with both fixed and random effects.

False Failures and Missed Faults

The quality of a measurement system can be defined by how well it discriminates between good and bad parts. To demonstrate, we consider the model

$$Y = X + \epsilon,$$

where X represents the true value of a randomly selected part, ϵ is the measurement error, and X and ϵ are independent normal random variables with means μ and 0, and variances γ_P and γ_M, respectively. Using the notation from Table 1, we have the following results:

$$\begin{aligned} E(Y) &= \mu, \\ \text{Var}(Y) &= \gamma_P + \gamma_M = \gamma_T, \text{ and} \\ \text{Cov}(X,Y) &= \gamma_P. \end{aligned}$$

The joint probability density function of the random vector $[Y \quad X]'$ is bivariate normal and represented as

$$f(y,x),$$

with mean vector $[\mu \quad \mu]'$ and variance-covariance matrix

$$\begin{bmatrix} \gamma_T & \gamma_P \\ \gamma_P & \gamma_P \end{bmatrix}.$$

A manufactured part is in conformance if

$$LSL \leq X \leq USL, \tag{3}$$

where LSL and USL are the lower and upper specification limits, respectively. A measurement system will "pass" a part if

$$LSL \leq Y \leq USL. \tag{4}$$

If Equation (3) is true, but Equation (4) is false, then a conforming part is incorrectly failed. This is called a false failure. Alternatively, if Equation (3) is false, but Equation (4) is true, a faulty part is incorrectly passed. This is called a missed fault. Of interest in a GR&R study are the producer's risk and the consumer's risk. The producer's risk, which we denote by δ, is the conditional probability that a measurement system will fail a part when the part conforms to specifications (false failure). The consumer's risk, denoted by β, is the conditional probability that a measurement system will pass a part when the part does not meet specifications (missed fault). Some authors such as Doganaksoy (2000) define these errors in terms of joint rather than conditional probabilities (e.g., the probability that a part is bad and it passes).

Mader et al. (1999) provide expressions based on the bivariate normal distribution for computing the conditional probabilities δ and β. Using the joint probability density function defined earlier, then

$$\delta = \frac{\int_{LSL}^{USL}\int_{-\infty}^{LSL} f(y,x)dydx + \int_{LSL}^{USL}\int_{USL}^{\infty} f(y,x)dydx}{\int_{LSL}^{USL} f(x)dx}, \quad (5)$$

and

$$\beta = \frac{\int_{-\infty}^{LSL}\int_{LSL}^{USL} f(y,x)dydx + \int_{USL}^{\infty}\int_{LSL}^{USL} f(y,x)dydx}{1-\int_{LSL}^{USL} f(x)dx}, \quad (6)$$

where $f(x)$ represents the marginal probability density function for X which is normal with mean μ and variance γ_P. In Figure 1 we illustrate the regions of false failures (FF) and missed faults (MF) on an equal density contour of the bivariate normal distribution. Thus, Equations (5) and (6) can be used to compute δ and β for given values of μ, γ_P, γ_T, LSL, and USL. The SAS code to perform this computation is provided in the Appendix.

In practice, we don't know the true values of μ, γ_P, and γ_T. If one uses only point estimates, the calculation does not account for the uncertainty in the estimates. Thus, we want to incorporate confidence intervals for these parameters in the calculation of δ and β. One way to proceed is to compute δ and β under different scenarios suggested by the confidence intervals. For example, a pessimistic scenario might consider the worst possible performance for the measurement system, and the worst possible capability for the manufacturing process. To do this, set γ_P equal to the upper bound of the confidence interval for γ_P and solve for the value of γ_T that provides the lower bound on ρ_P. Conversely, one might consider an optimistic scenario with the best possible performance for the measurement system combined with the best process capability. We will demonstrate this procedure with a numerical example later in the paper. Larsen (2003) provided an alternative scheme for using confidence intervals to estimate δ and β. The process of generalized inference can also be used to construct confidence intervals on δ and β. Engel and De Vries (1997) considered the relationship between PTR and δ and β for a model with one source of variation.

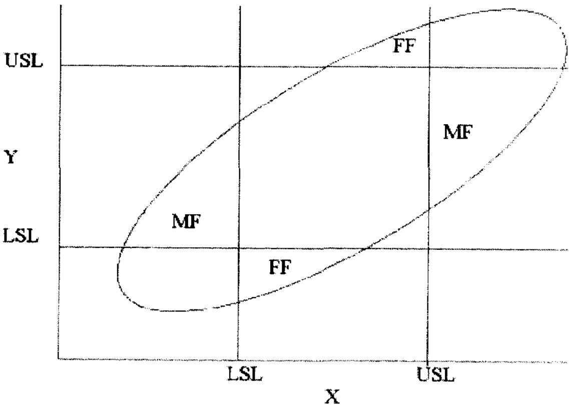

FIGURE 1. MF and FF Regions of an Equal Density Contour.

The Standard Experiment

The standard experiment employs a two-factor design with "parts" and "operators". The statistical model used to describe the response variable is the random two-factor model

$$Y_{ijk} = \mu + P_i + O_j + (PO)_{ij} + E_{ijk} \quad (7)$$
$$i = 1,\ldots,p; \quad j = 1,\ldots,o; \quad k = 1,\ldots,r;$$

where μ is a constant, and P_i, O_j, $(PO)_{ij}$, and E_{ijk} are jointly independent normal random variables with means of zero and variances σ_P^2, σ_O^2, σ_{PO}^2, and σ_E^2, respectively. The ANOVA table for the model in Equation (7) is shown in Table 2. Under the assumptions of Equation (7), $(p-1)S_P^2/\theta_P$, $(o-1)S_O^2/\theta_O$, $(p-1)(o-1)S_{PO}^2/\theta_{PO}$, and $po(r-1)S_E^2/\theta_E$ are jointly independent chi-squared random variables with $p-1$, $o-1$, $(p-1)(o-1)$, and $po(r-1)$ degrees of freedom, respectively.

In Table 3 we report the point estimators for the parameters of interest. The estimators for γ_P, γ_M, and γ_T are all minimum variance unbiased (MVU) estimators. The estimators for ρ_P and ρ_M are obtained by replacing each variance component with the corresponding MVU estimator. Table 4 contains upper and lower bounds for approximate $100(1-\alpha)\%$ MLS confidence intervals for these parameters. In Table 5 we list definitions of terms used in Table 4, where $F_{\alpha:df1,df2}$ represents the F-value with $df1$

TABLE 2. ANOVA for Model (7)

SV	DF	MS	EMS
Parts (P)	$p-1$	S_P^2	$\theta_P = \sigma_E^2 + r\sigma_{PO}^2 + or\sigma_P^2$
Operators (O)	$o-1$	S_O^2	$\theta_O = \sigma_E^2 + r\sigma_{PO}^2 + pr\sigma_O^2$
P×O	$(p-1)(o-1)$	S_{PO}^2	$\theta_{PO} = \sigma_E^2 + r\sigma_{PO}^2$
Replicates (E)	$po(r-1)$	S_E^2	$\theta_E = \sigma_E^2$

TABLE 3. Point Estimators

Parameter	Point Estimator
$\gamma_P = \sigma_P^2$	$\widehat{\gamma}_P = [S_P^2 - S_{PO}^2]/(or)$
$\gamma_M = \sigma_O^2 + \sigma_{PO}^2 + \sigma_E^2$	$\widehat{\gamma}_M = [S_O^2 + (p-1)S_{PO}^2 + p(r-1)S_E^2]/(pr)$
$\gamma_T = \sigma_P^2 + \sigma_O^2 + \sigma_{PO}^2 + \sigma_E^2$	$\widehat{\gamma}_T = [pS_P^2 + oS_O^2 + (po-p-o)S_{PO}^2 + po(r-1)S_E^2]/(por)$
$\rho_P = \gamma_P/\gamma_T$	$\widehat{\gamma}_P/\widehat{\gamma}_T$
$\rho_M = \gamma_M/\gamma_T = 1 - \rho_P$	$\widehat{\gamma}_M/\widehat{\gamma}_T$

and $df2$ degrees of freedom with area α to the left. References for the formulas in Table 4 can be found in Burdick and Larsen (1997), with one exception. As suggested by Chiang (2002), the intervals for ρ_P and ρ_M are based on a method by Leiva and Graybill (1986). This interval is easier to compute than the one reported by Burdick and Larsen (1997), and better maintains the stated confidence level. The numbers in the last columns of Tables 4 and 5 correspond to the numerical example that follows in the next section.

Generalized Confidence Intervals

In Table 6 we report the generalized pivotal quantity (GPQ) for each parameter in Table 1 where $U_1, \ldots, U_4$ are jointly independent chi-squared random variables with degrees of freedom $p-1$, $o-1$,

TABLE 4. $100(1-\alpha)\%$ MLS Confidence Intervals

Parameter	Lower Bound	Upper Bound	Example 95% Interval
γ_P	$\widehat{\gamma}_P - \frac{\sqrt{V_{LP}}}{or}$	$\widehat{\gamma}_P + \frac{\sqrt{V_{UP}}}{or}$	[22.69; 161.64]
γ_M	$\widehat{\gamma}_M - \frac{\sqrt{V_{LM}}}{pr}$	$\widehat{\gamma}_M + \frac{\sqrt{V_{UM}}}{pr}$	[1.20; 27.02]
γ_T	$\widehat{\gamma}_T - \frac{\sqrt{V_{LT}}}{por}$	$\widehat{\gamma}_T + \frac{\sqrt{V_{UT}}}{por}$	[24.48; 166.23]
ρ_P	$L_P = \frac{pL^*}{pL^*+o}$	$U_P = \frac{pU^*}{pU^*+o}$	[0.628; 0.991]
ρ_M	$1 - U_P$	$1 - L_P$	[0.009; 0.372]

TABLE 5. Definition of Terms in Table 4

Term	Definition	Value in Example
V_{LP}	$G_1^2S_P^4 + H_3^2S_{PO}^4 + G_{13}S_P^2S_{PO}^2$	53,076.17
V_{UP}	$H_1^2S_P^4 + G_3^2S_{PO}^4 + H_{13}S_P^2S_{PO}^2$	1,040,643.4
V_{LM}	$G_2^2S_O^4 + G_3^2(p-1)^2S_{PO}^4 + G_4^2p^2(r-1)^2S_E^4$	321.282
V_{UM}	$H_2^2S_O^4 + H_3^2(p-1)^2S_{PO}^4 + H_4^2p^2(r-1)^2S_E^4$	572,150.12
V_{LT}	$G_1^2p^2S_P^4 + G_2^2o^2S_O^4 + G_3^2(po-p-o)^2S_{PO}^4 + G_4^2(po)^2(r-1)^2S_E^4$	5,311,676.7
V_{UT}	$H_1^2p^2S_P^4 + H_2^2o^2S_O^4 + H_3^2(po-p-o)^2S_{PO}^4 + H_4^2(po)^2(r-1)^2S_E^4$	109,230,276
G_1	$1 - 1/F_{1-\alpha/2:p-1,\infty}$	0.5269
G_2	$1 - 1/F_{1-\alpha/2:o-1,\infty}$	0.7289
G_3	$1 - 1/F_{1-\alpha/2:(p-1)(o-1),\infty}$	0.4290
G_4	$1 - 1/F_{1-\alpha/2:po(r-1),\infty}$	0.2797
H_1	$1/F_{\alpha/2:p-1,\infty} - 1$	2.3329
H_2	$1/F_{\alpha/2:o-1,\infty} - 1$	38.4979
H_3	$1/F_{\alpha/2:(p-1)(o-1),\infty} - 1$	1.1869
H_4	$1/F_{\alpha/2:po(r-1),\infty} - 1$	0.4821
G_{13}	$\frac{(F_{1-\alpha/2:p-1,(p-1)(o-1)}-1)^2-G_1^2F_{1-\alpha/2:p-1,(p-1)(o-1)}^2-H_3^2}{F_{1-\alpha/2:p-1,(p-1)(o-1)}}$	-0.0236
H_{13}	$\frac{(1-F_{\alpha/2:p-1,(p-1)(o-1)})^2-H_1^2F_{\alpha/2:p-1,(p-1)(o-1)}^2-G_3^2}{F_{\alpha/2:p-1,(p-1)(o-1)}}$	-0.1800
L^*	$\frac{S_P^2-F_{1-\alpha/2:p-1,(p-1)(o-1)}S_{PO}^2}{p(r-1)F_{1-\alpha/2:p-1,\infty}S_E^2+F_{1-\alpha/2:p-1,o-1}S_O^2+(p-1)F_{1-\alpha/2:p-1,\infty}S_{PO}^2}$	0.5075
U^*	$\frac{S_P^2-F_{\alpha/2:p-1,(p-1)(o-1)}S_{PO}^2}{p(r-1)F_{\alpha/2:p-1,\infty}S_E^2+F_{\alpha/2:p-1,o-1}S_O^2+(p-1)F_{\alpha/2:p-1,\infty}S_{PO}^2}$	31.6827

$(p-1)(o-1)$, and $po(r-1)$, respectively. The terms s_P^2, s_O^2, s_{PO}^2, and s_E^2 are the realized values of the mean squares for a particular data set.

The quantities shown in Table 6 were recommended by Hamada and Weerahandi (2000) and correspond to the tailored variables recommended by Chiang (2001).

One can use either numerical integration or computer simulation to compute the generalized intervals. We will apply simulation because it is easier computationally. To construct a generalized interval in this manner, we use the following process:

1. Compute S_P^2, S_O^2, S_{PO}^2, and S_E^2 for the collected data set and denote the realized values by s_P^2, s_O^2, s_{PO}^2, and s_E^2, respectively.
2. Simulate 10,000 values of the appropriate GPQ

TABLE 6. GPQs for the Two-Factor With Interaction Model

Parameter	GPQ	Example 95% Interval
γ_P	$\left[\frac{(p-1)s_P^2}{U_1} - \frac{(p-1)(o-1)s_{PO}^2}{U_3}\right]/(or)$	[22.22; 164.92]
γ_M	$\left[\frac{(o-1)s_O^2}{U_2} + \frac{(p-1)^2(o-1)s_{PO}^2}{U_3} + \frac{p^2o(r-1)^2s_E^2}{U_4}\right]/(pr)$	[1.18; 27.50]
γ_T	$\left[\frac{p(p-1)s_P^2}{U_1} + \frac{o(o-1)s_O^2}{U_2} + \frac{(po-p-o)(p-1)(o-1)s_{PO}^2}{U_3} + \frac{p^2o^2(r-1)^2s_E^2}{U_4}\right]/(por)$	[25.14; 181.76]
ρ_P	$\mathrm{GPQ}(\gamma_P)/\mathrm{GPQ}(\gamma_T)$	[0.630; 0.989]

TABLE 7. ANOVA for Numerical Example

SV	DF	MS
Parts (P)	$p-1=9$	$S_P^2 = 437.3284$
Operators (O)	$o-1=2$	$S_O^2 = 19.6333$
P×O	$(p-1)(o-1)=18$	$S_{PO}^2 = 2.6951$
Replicates (E)	$po(r-1)=60$	$S_E^2 = 0.5111$

by simulating 10,000 independent values each of $U_1, \ldots, U_4$.

3. Order the resulting 10,000 GPQ values from least to greatest.
4. Define the lower bound for a $100(1-\alpha)\%$ interval as the value in position $10,000 \times (\alpha/2)$ of the ordered set of simulated GPQs. Define the upper bound as the value in position $10,000 \times (1-\alpha/2)$ of this same ordered set.

The numerical example in the next section demonstrates this process. The Appendix provides SAS code that performs this computation.

Numerical Example

We demonstrate the formulas in this section by analyzing a data set reported by Houf and Berman (1988). These data consist of measurements taken on a power module for an induction motor starter. The units of measurement are degrees C per watt. The experiment consists of a two-factor crossed design with $p = 10$ parts, $o = 3$ operators, and $r = 3$ replicates. The resulting ANOVA after multiplying each response by 100 for convenience of scale is shown in Table 7. The mean of all the observations is 35.8. The last columns in Tables 4 and 6 contain the 95% MLS and generalized confidence intervals, respectively, for these data. In the last column of Table 5 we list values used to compute the intervals in Table 4. Here, we use $\alpha = 0.05$, $p = 10$ parts, $o = 3$ operators, and $r = 3$ replicates. All lower bounds have been rounded down to the reported number of decimals and all upper bounds have been rounded up.

We can use the intervals on γ_M and ρ_P to compute confidence intervals for PTR and SNR, respectively. To demonstrate, for this example the specification limits are $LSL = 18$ and $USL = 58$. Using Equation (1), the bounds for γ_M in Table 4, and $k = 5.15$, a 95% confidence interval for the PTR has the limits

$$L = \frac{5.15\sqrt{1.20}}{58-18} = 14.1\%, \quad \text{and}$$

$$U = \frac{5.15\sqrt{27.02}}{58-18} = 67.0\%.$$

A 95% confidence interval for the SNR based on Equation (2) and the bounds for ρ_P in Table 4 is

$$L = \sqrt{\frac{2 \times 0.628}{1-0.628}} = 1.8, \quad \text{and}$$

$$U = \sqrt{\frac{2 \times 0.991}{1-0.991}} = 15.$$

Since not all values in the interval for SNR exceed five, there is not sufficient evidence to claim this measurement system is adequate for monitoring the process.

Table 8 shows the calculation of producer's risk (δ) and consumer's risk (β) using Equations (5) and (6) under two scenarios. The scenario labeled "Pessimistic" is computed assuming the worst possible performance for both the production process and the measurement system. This is done by computing δ and β using the upper bound on γ_P and the lower bound on ρ_P. We used the sample mean of 35.8 for the value of μ, the computed confidence bounds in Table 4, and solved for γ_T using the relation $\gamma_T = \gamma_P/\rho_P$. The SAS code shown in the Appendix was used to make this calculation. The scenario labeled "Optimistic" uses the best conditions for both the process and the measurement system. In particular, we use the lower bound of γ_P and the upper bound of ρ_P. As with the first scenario, we

TABLE 8. Error Rates for Two Scenarios

Scenario	γ_P	ρ_P	δ	β
Pessimistic	161.64	0.628	15.2%	31.0%
Optimistic	22.69	0.991	0.002%	12.3%

use the point estimate of 35.8 for μ. The range for the producer's risk is from 0.002% to 15.2% and for consumer's risk from 12.3% to 31.0%.

All of the intervals in this example are relatively wide. The reason for the wide intervals is that there are only three operators in the experimental design. This provides only two degrees of freedom for estimation of the operator effect, and this impacts the interval length for any parameter that is a function of σ_O^2. Hence, it is necessary to increase the number of operators beyond the traditional size of three. The next section provides additional guidelines for designing a GR&R experiment.

Finally, as this example suggests, there is not much practical difference between the closed-form intervals and the generalized confidence intervals. Empirical comparisons of these two sets of intervals suggest they provide comparable intervals for the model in Equation (7). Thus, for this particular model, an investigator can use either method.

Designing a GR&R Experiment

A successful gauge capability study is one that provides reliable estimates of the components of variation in the measurement process and identifies the factors that are most influential. The study should also provide information about the potential effectiveness of the gauge as a measurement tool. Consequently, the design of the experiment is very important. Poor statistical design of the experiment can lead to a situation where the true variation in the measurement process is underestimated, and this results in an overly optimistic conclusion regarding gauge capability. Some important statistical design issues include the number of parts to be used in the study, the number of measurements per part, how the parts are selected, and ensuring that true replicates are actually obtained as opposed to repeat measurements. Very few sources in the literature provide guidance on these experimental design aspects of a measurement systems capability analysis study, so we include some discussion here.

A good general practice is to use many parts in the experiment with relatively few measurements each, as opposed to few parts with many measurements per part. There are several reasons for this recommendation. First, parts are typically selected from actual production and are representative of the material that the measurement system will encounter during routine operation. The gauge may exhibit less variability on a production unit that is near the center of the manufacturing specifications than on product at the extremes of this specification. An extreme example of this is non-linearity of the gauge, which results in unstable or unreliable results beyond a certain operating region. Using a relatively large number of parts in the study increases the likelihood of detecting this problem. Some analysts like to use "golden" or "standard" parts in a measurement systems capability study as opposed to production parts. We do not recommend the exclusive use of standard parts because standard units may not share important product characteristics with the typical production units, and they might produce unexpected measurement errors. Alternatively, standard units typically exhibit less variability than production units with respect to key quality characteristics.

Second, it is not unusual to find that the variance of the measurements is not constant, and often depends on the mean level of the product characteristic. This is unlikely to be detected if only a narrow range of good production parts or standard parts are used in the study. Sometimes visual inspection of the data can reveal this problem, but a better approach is to carefully analyze the residuals from a gauge capability experiment, using the same residual plots typically employed in any designed experiment. In particular, the plots of residuals versus the predicted response, residuals versus parts, residuals versus operators, and residuals versus time order all convey very specific diagnostic information. For example, an outward-opening funnel pattern on the plot of residuals versus time order suggests that variability in the measurement process is increasing with time, perhaps due to operator fatigue, an instrument that does not hold calibration, or environmental factors such as temperature that may change over time and affect the performance of the gauge.

Finally, when many measurements are to be made on the same part, our experience has been that operating personnel are less likely to perform complete replications of the measurement process or to completely randomize the order of the trials. Sometimes all measurements on a part will be taken successively without any change in measurement system setup. Some analysts refer to the repeatability component obtained in this manner as "static repeatability" while if complete replicates are performed the repeatability component is called "dynamic repeatability." We feel that it is important to use complete replicates and to conduct the trials in random order.

Without complete replicates and randomization, we omit important sources of variability due to such factors as fixturing and positioning of the part, measurement tool alignment, batches of reagent in a chemical assay, or sources of variability that are associated with time. Consequently, the "static" estimate of the repeatability component of measurement variability is overly optimistic. Using a relatively large number of parts and making few measurements on each part encourages true or complete replication as opposed to simply making repeat measurements.

The number of parts and the number of operators to choose is an important consideration. A useful approach to these decisions is to consider the length of the confidence interval estimates of the relevant parameters that will result. As demonstrated in the numerical example, a two-factor random design with only three operators provides very wide intervals. Unfortunately, there are no closed-form solutions for sample sizes that will tell us how many parts and how many operators to use to produce confidence intervals of a specified length at a stated confidence. However, a trial-and-error approach using preliminary estimates of the quantities in these equations and simulated data can be used to obtain reasonably good estimates of the required sample sizes. For example, simulations by Burdick and Larsen (1997) demonstrated that increasing the number of operators from three to six in a random two-factor model greatly decreases interval length.

Other Measurement System Models

In this section, we cite a few references for situations where the standard model in Equation (7) is not appropriate. Although we do not discuss these papers in detail, they provide starting points for investigators who encounter these situations.

Mixed ANOVA Models

Although it is customary to assume that all effects in Equation (7) are random, such an assumption is not always warranted. In particular, operators are often more properly considered as fixed effects. This is the situation when the set of operators used in the experiment is also used to monitor the process. The assumption that operators are fixed changes the distributional assumptions associated with the standard experiment. If operators are fixed, then $(o-1)S_O^2/\theta_O$ no longer has a chi-squared distribution, but $(o-1)S_O^2/\theta_{PO}$ has a non-central chi-squared distribution. Dolezal, Burdick, and Birch (1998) proposed a simple modification to the operator degrees of freedom that allows one to use the closed-form intervals in Table 4 to compute intervals for this mixed model. Generalized confidence intervals can also be computed for this mixed model.

More Complex ANOVA Designs

Designs more complex than the two-factor crossed design are needed in some applications. Such designs require more than two factors and involve both nested and crossed effects. Adamec and Burdick (2003) provided a closed-form interval for ρ_M in a three-factor crossed random effects design. Examples of other designs in the literature include a completely nested design provided by John (1994, p. 12) and an example with both nested and crossed factors by Borror, Montgomery, and Runger (1997). Deutler (1991) demonstrated that nested designs are also very common in interlaboratory test studies. Various MLS closed-form confidence intervals that can be computed for parameters in these models are provided by Burdick and Graybill (1992). Additionally, the generalized confidence interval method can be applied to all such balanced designs, with both fixed and random effects.

Comparison of Two Measurement Systems

It is often of interest to examine changes in γ_T after attempts to improve a measurement system. Comparison of γ_T for two different locations as described by Morchower (1999) is also of interest in many applications. Details of how such a comparison can be performed using closed-form intervals for the two-factor design are provided by Burdick, Allen, and Larsen (2002). Generalized confidence intervals can be used in the same manner as the closed-form intervals.

Attribute (Pass-Fail) Data

Attribute data results from measurement variables that assign only a finite set of values. The most common type of attribute data is pass-fail data. Boyles (2001) proposed a method for such data and derived maximum likelihood estimators and confidence intervals for the misclassification rates under a model where a standard is given, and one where no standard is given. AIAG (2002, pp. 125-140) suggested a cross-tabulation approach for this process.

Truncated Data

Lai and Chew (2000) stated that distributions of measurements from automated instruments are likely

TABLE 9. Summary of Software Features

Package	Q1	Q2	Q3
Minitab	Yes	No	No
Statgraphics	Yes	No	No
JMP	Yes	No	No
NCSS	Yes	Yes	Yes
Statistica	Yes	Yes*	No
SAS	No	No	No

*Requires the Industrial Statistics and Six Sigma Module

to be truncated. To address this situation, they proposed a non-parametric approach that employs a two-factor model with no interaction. They used simulation to demonstrate the improvement in point estimates of the GR&R parameters under different underlying truncated distributions.

Destructive Testing

When a measurement process destroys parts, it is not possible to estimate repeatability with the traditional two-factor model. Mitchell, Hegemann, and Liu (1997) presented a method for estimating repeatability and assessing gauge capability for destructive testing. Bergeret, Maubert, Sourd, and Puel (2001) applied this method to three case studies and suggested a slight modification. Phillips, Jeffries, Schneider, and Frankoski (1997) described a case study where two phases were used to conduct a GR&R study involving fiberglass shingles.

Two-Dimensional Data

Voelkel (2003) proposed a method for a study in which the data are two-dimensional and associated with measurements of a circle. The examples in the paper concerned balancing rotors in centrifugal pumps. Several two-dimensional summary measures were proposed and compared in the paper.

Statistical Software

We performed a review of Minitab 13 (2001), Statgraphics Plus 5.0 (2000), JMP 4.0.4 (2001), NCSS (2001), Statistica 6.0 (2001), and SAS 8.2 (2001). Table 9 reports the features of each program in response to the following three questions:

1. Does the program have a separate measurement systems capability module?
2. Does the program construct a confidence interval for the PTR?
3. Does the program construct a confidence interval for the SNR?

These software packages don't compute all of the confidence intervals we have recommended in this paper. Given the computational ease of the closed-form intervals and the SAS code for computing generalized intervals shown in the Appendix, we recommend implementation of these methods instead of the software packages. An Excel program that computes the closed-form intervals shown in Table 4 is available from the first author.

Concluding Remarks

We have reviewed statistical methods for conducting measurement systems capability studies, with emphasis on the ANOVA procedure to analyze the results. We have emphasized constructing confidence intervals on relevant parameters in the ANOVA model, and provided a brief review of the literature.

Based on our review, it appears that while the ANOVA method is becoming more widely used, there is still much reliance on the range method. For example, we note that the range method is included in four of the six software packages represented in Table 9. This is unfortunate because the range method does not support computation of confidence intervals. Point estimates alone do not convey a complete picture of the capability of the measuring instrument, just as point estimates of a process capability ratio do not completely describe process capability. In both situations, confidence intervals are of considerable practical value and should become part of any standard GR&R study report. There is still much reliance on precision-to-tolerance ratios and other indices to summarize the capability of the gauge. Much of the guidance regarding interpretation of these indices is arbitrary, and often does not provide direct information on the ability of the gauge to discriminate between good and bad parts. For this reason, it might be beneficial to consider the intervals we have constructed for δ and β in the decision process.

We have briefly introduced several additional areas of research and applications. In addition to mixed ANOVA models, attribute data, truncated data, and destructive testing presented in this paper, there are several areas where we think further research is needed. Automated measurement systems frequently collect data at multiple locations on the same unit or on several different variables simultaneously. These types of gauges also routinely result in censored and/or missing data. Furthermore, the methods that we have discussed are based on the assumption that the response variable from the experiment is normally distributed. Situations exist in industrial settings where the measurement error is not normally distributed. Methodology for analyzing these experiments needs to be developed, evaluated, and carried into practice.

Appendix-SAS Code

In this appendix we provide SAS code for computing the misclassification rates and generalized confidence intervals for the standard two-factor random crossed experiment. Comments from a reviewer were helpful in improving our code.

δ and β Errors

The SAS function PROBBNRM computes probabilities from a standardized bivariate normal by computing

$$\text{PROBBNRM}(a, b, \rho) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(y, x) dy dx,$$

where $f(y, x)$ is a standardized bivariate normal distribution with Y and X having a correlation of ρ. The SAS function PROBNORM computes

$$\text{PROBNORM}(a) = \int_{-\infty}^{a} f(x) dx,$$

where $f(x)$ is a standardized normal distribution. The following SAS code uses these functions to compute δ and β as defined by Equations (5) and (6) for the numerical example. The computation for mf2 in the code is based on the assumption that there are no parts where $Y < LSL$ and $X > USL$, and the computation for ff2 is based on the assumption that there are no parts where $Y > USL$ and $X < LSL$.

```
data misclass;
input mu lsl usl rhop gammap;
gammat=gammap/rhop;
cov=gammap;
corr=cov/(sqrt(gammap)*sqrt(gammat));
uslstdy=(usl-mu)/sqrt(gammat);
uslstdx=(usl-mu)/sqrt(gammap);
lslstdy=(lsl-mu)/sqrt(gammat);
lslstdx=(lsl-mu)/sqrt(gammap);

ff1=probbnrm(uslstdx,lslstdy,corr)-
probbnrm(lslstdx,lslstdy,corr);
ff2=probnorm(uslstdx)-probbnrm(uslstdx,uslstdy,corr);
mf1=probbnrm(lslstdx,uslstdy,corr)-
probbnrm(lslstdx,lslstdy,corr);
mf2=probnorm(uslstdy)-probbnrm(uslstdx,uslstdy,corr);
delta=(ff1+ff2)/(probnorm(uslstdx)-
probnorm(lslstdx));
beta=(mf1+mf2)/(1-(probnorm(uslstdx)-
probnorm(lslstdx)));

keep mu lsl usl rhop gammap gammat delta beta;
datalines;
35.8 18 58 .628 161.64
;
proc print data=misclass;
run;
```

Generalized Confidence Intervals

The following SAS code can be used to compute 95% generalized confidence intervals for the example problem.

```
data gci;
input parts operators reps msparts msoperators msint
mse iter
seed1 seed2 seed3 seed4;
df1=parts-1;
df2=operators-1;
df3=df1*df2;
df4=parts*operators*(reps-1);
do i=1 to iter;
u1s=2*rangam(seed1,df1/2);
u2s=2*rangam(seed2,df2/2);
u3s=2*rangam(seed3,df3/2);
u4s=2*rangam(seed4,df4/2);
*GPQ values from Table 6 are shown below;
qp=(df1*msparts/u1s-df3*msint/u3s)/(operators*reps);
qm=(df2*msoperators/u2s+df3*(parts-1)*msint/u3s
+parts**2*operators*(reps-
1)**2*mse/u4s)/(parts*reps);
qt=(parts*df1*msparts/u1s+operators*df2*msoperators
/u2s
+(parts*operators-parts-operators)*df1*df2*msint/u3s
+parts**2*operators**2*(reps-1)**2*mse/u4s)
/(parts*operators*reps);
qr=qp/qt;
output;
end;

datalines;
10 3 3 437.328 19.633 2.695 .5111 10000
1684625530 1646374628 1688825530 1876453274
;
proc univariate data=gci;
var qp qm qt qr;
ods select quantiles;
output out=percentiles pctlpre=P_p  P_m P_t P_r
pctlpts=2.5  97.5;
run;

proc print data=percentiles;run;
```

References

Adamec, E. and Burdick, R. K. (2003). "Confidence Intervals for a Discrimination Ratio in a Gauge R&R Study with Three Random Factors". *Quality Engineering* 15, pp. 383–389.

Automotive Industry Action Group (1995). *Measurement Systems Analysis*, 2nd ed. Detroit, MI.

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

BALLARD, D. H.; MCCORMACK, JR., D. W.; MOORE, T. L.; PRINS, J.; TOBIAS, P. A.; and PORE, M. (1997). "A Comparison of Gauge Study Practices". *ASA Proceedings of the Section on Quality and Productivity,* pp. 31–36.

BARRENTINE, L. B. (1991). *Concepts for R&R Studies.* ASQC Quality Press, Milwaukee, WI.

BERGERET, F.; MAUBERT, S.; SOURD, P.; and PUEL, F. (2001). "Improving and Applying Destructive Gauge Capability". *Quality Engineering* 14, pp. 59–66.

BORROR, C. M.; MONTGOMERY, D. C.; and RUNGER, G. C. (1997). "Confidence Intervals for Variance Components from Gauge Capability Studies". *Quality and Reliability Engineering International* 13, pp. 361–369.

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

BURDICK, R. K.; ALLEN, A. E.; and LARSEN, G. A. (2002). "Comparing Variability of Two Measurement Processes Using R&R Studies". *Journal of Quality Technology* 34, pp. 97–105.

BURDICK, R. K. and GRAYBILL, F. A. (1992). *Confidence Intervals on Variance Components.* Marcel Dekker, New York, NY.

BURDICK, R. K. and LARSEN, G. A. (1997). "Confidence Intervals on Measures of Variability in R&R Studies". *Journal of Quality Technology* 29, pp. 261–273.

CHIANG, A. K. L. (2001). "A Simple General Method for Constructing Confidence Intervals for Functions of Variance Components". *Technometrics* 43, pp. 356–367.

CHIANG, A. K. L. (2002). "Improved Confidence Intervals for a Ratio in an R&R Study". *Communications in Statistics—Simulation and Computation* 31, pp. 329–344.

CONORS, M.; MERRILL, K.; and O'DONNELL, B. (1995). "A Comprehensive Approach to Measurement System Evaluation". *ASA Proceedings of the Section on Physical and Engineering Sciences,* pp. 136–138.

CROARKIN, C., EDITOR (2002). "Gauge R&R Studies". Section 2.4 of the *Beta Version of the NIST/SEMATECH Engineering Statistics Internet Handbook.* Located at http://www.itl.nist.gov/div898/handbook/.

DEUTLER, T. (1991). "Grubbs-Type Estimators for Reproducibility Variances in an Interlaboratory Test Study". *Journal of Quality Technology* 23, pp. 324–335.

DOGANAKSOY, N. (2000). "Assessment of Impact of Measurement Variability in the Presence of Multiple Sources of Product Variability". *Quality Engineering* 13, pp. 83–89.

DOLEZAL, K. K.; BURDICK, R. K.; and BIRCH, N. J. (1998). "Analysis of a Two-Factor R&R Study With Fixed Operators". *Journal of Quality Technology* 30, pp. 163–170.

ENGEL, J. and DE VRIES, B. (1997). "Evaluating a Well-Known Criterion for Measurement Precision". *Journal of Quality Technology* 29, pp. 469–476.

GRAYBILL, F. A. and WANG, C. M. (1980). "Confidence Intervals on Nonnegative Linear Combinations of Variances". *Journal of the American Statistical Association* 75, pp. 869–873.

HAMADA, M. and WEERAHANDI, S. (2000). "Measurement System Assessment Via Generalized Inference". *Journal of Quality Technology* 32, pp. 241–253.

HORRELL, K. (1991). *Introduction to Measurement Capability Analysis.* SEMATECH report 91090709A-ENG.

HOUF, R. E. and BERMAN, D. B. (1988). "Statistical Analysis of Power Module Thermal Test Equipment Performance". *IEEE Transactions on Components, Hybrids, and Manufacturing Technology* 11, pp. 516–520.

JMP 4.0.4 (2001). SAS Institute, Inc. Cary, NC.

JOHN, P. (1994). *Alternative Models for Gauge Studies.* SEMATECH report 93081755A-TR.

LAI, Y. W. and CHEW, E. P. (2000). "Gauge Capability Assessment for High-Yield Manufacturing Processes with Truncated Distribution". *Quality Engineering* 13, pp. 203–210.

LARSEN, G. (2002). "Measurement System Analysis: The Usual Metrics Can Be Noninformative". *Quality Engineering* 15, pp. 293–298.

LARSEN, G. (2003). "Measurement System Analysis in a Production Test Environment with Multiple Test Parameters". *Quality Engineering,* to appear.

LEIVA, R. A. and GRAYBILL, F. A. (1986). "Confidence Intervals for Variance Components in the Balanced Two-Way Model With Interaction". *Communications in Statistics—Simulation and Computation* 15, pp. 301–322.

MAJESKE, K. D. and ANDREWS, R. W. (2002). "Evaluating Measurement Systems and Manufacturing Processes Using Three Quality Measures". *Quality Engineering* 15, pp. 243–251.

MADER, D. P.; PRINS, J.; and LAMPE, R. E. (1999). "The Economic Impact of Measurement Error". *Quality Engineering* 11, pp. 563–574.

MINITAB 13 (2001). Minitab Inc., State College, PA.

MITCHELL, T.; HEGEMANN, V.; and LIU, K. C. (1997). "GRR Methodology for Destructive Testing and Quantitative Assessment of Gauge Capability for One-Side Specifications" in *Statistical Case Studies for Industrial Process Improvement* edited by V. Czitrom and P. D. Spagon, SIAM, Philadelphia, PA, pp. 47–59.

MONTGOMERY, D. C. and RUNGER, G. C. (1993a). "Gauge Capability and Designed Experiments. Part I: Basic Methods". *Quality Engineering* 6, pp. 115–135.

MONTGOMERY, D. C. and RUNGER, G. C. (1993b). "Gauge Capability Analysis and Designed Experiments. Part II: Experimental Design Models and Variance Component Estimation". *Quality Engineering* 6, pp. 289–305.

MORCHOWER, N. D. (1999). "Two-Location Gauge Evaluation". *Quality Progress* 32/4, pp. 79–86.

NCSS (2001). NCSS Statistical Software, Kaysville, UT.

PHILLIPS, A. R.; JEFFRIES, R.; SCHNEIDER, J.; and FRANKOSKI, S. P. (1997). "Using Repeatability and Reproducibility Studies to Evaluate a Destructive Test Method". *Quality Engineering* 10, pp. 283–290.

SAS 8.2 (2001). SAS Institute, Inc., Cary, NC.

STATGRAPHICS PLUS 5.0 (2000). Manugistics, Inc. Rockville, MD.

STATISTICA 6.0 (2001). Statsoft, Inc., Tulsa, OK.

TSUI, K. and WEERAHANDI, S. (1989). "Generalized p-values in Significance Testing of Hypotheses in the Presence of Nuisance Parameters". *Journal of the American Statistical Association* 84, pp. 602–607.

VAN DEN HEUVEL, E. R. and TRIP, A. (2002). "Evaluation of Measurement Systems with a Small Number of Observers". *Quality Engineering* 15, pp. 323–331.

VARDEMAN, S. B. and VAN VALKENBURG, E. S. (1999). "Two-Way Random-Effects Analyses and Gauge R&R Studies". *Technometrics* 41, pp. 202–211.

VOELKEL, J. G. (2003). "Gauge R&R Analysis for Two-Dimensional Data with Circular Tolerances". *Journal of Quality Technology* 35, pp. 153–167.

WEERAHANDI, S. (1993). "Generalized Confidence Intervals". *Journal of the American Statistical Association* 88, pp. 899–905.

WHEELER, D. J. (1992). "Problems With Gauge R&R Studies". *ASQC Quality Congress Transactions,* pp. 179–185.

WHEELER, D. J. and LYDAY, R. W. (1989). *Evaluating the Measurement Process.* SPC Press, Knoxville, TN.

Key Words: *Analysis of Variance, Confidence Intervals, Gauge Studies, Repeatability, Reproducibility, Variance Components.*