Quality Engineering
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lqen20

GAUGE CAPABILITY AND DESIGNED EXPERIMENTS. PART I: BASIC METHODS

DOUGLAS C. MONTGOMERY [a] & GEORGE C. RUNGER [b]

[a] College of Engineering and Applied Sciences Arizona State University , Tempe, Arizona, 85287-5506

[b] School of Engineering Rensselaer Polytechnic Institute Troy , New York, 12180-3590
Published online: 03 Apr 2007.

Quality Engineering, 6(1), 115–135 (1993–94)

GAUGE CAPABILITY AND DESIGNED EXPERIMENTS. PART I: BASIC METHODS

DOUGLAS C. MONTGOMERY
College of Engineering and Applied Sciences
Arizona State University
Tempe, Arizona 85287-5506

GEORGE C. RUNGER
School of Engineering
Rensselaer Polytechnic Institute
Troy, New York 12180-3590

Key Words

Repeatability; Reproducibility; ANOVA; Variance components; Measurement error.

Introduction

Measurement plays a significant role in helping any organization improve quality. For example, the information obtained from manufacturing, and the continuous improvement of the production process based on that information, is an important competitive weapon in the global marketplace. The increased use of real-time sensors has increased the quantity of data, and tighter tolerances required for customer satisfaction over the life of the product have dramatically increased the requirements for information quality. Data, which in prior years were not essential, are now required for tuning a process to gain and maintain market share.

The quality of data used for the operation and management of our production processes depends on the gauges and other tools used for the measurement of process variables. An understanding of variability in our processes, a key component of the quality improvement strategies of Deming (2), requires an understanding of the variability due to the measurement procedure as well as product variability. The contribution of the complete measurement process to the overall variability of the process is crucial for projects such as evaluating process improvements, process control, and defect identification and isolation. Because of the much greater reliance on quantitative measurements in modern manufacturing operations, the criticality of successful gauge capability analysis has increased. A gauge capability analysis must be designed and developed with as much care as is devoted to other important experimental projects in the company.

The goal of a gauge capability analysis is the understanding and quantification of the sources of variability present in the measurement process. This quantification will provide the basis for projects such as improvements to the measurement process, compensation for the variability in evaluating the product, and the control of the measurement process. The quantification will provide inputs for assessing the capability, performance characteristics, control, defect isolation, etc., of the overall manufacturing process.

To understand the sources of variability in the measurement process, an empirical study of the sources of variability is required. The goal and procedure for a gauge capability analysis is essentially the paradigm of a designed experiment. We plan and analyze an experiment to quantify the sources of variability in the measurement process. Because of the random variations in the data, we require statistical methods for an objective analysis which provides valid conclusions. The tools of statistical experimental design are available for a comprehensive design and analysis of the measurement process. The fundamental principles of experimental design, replication, blocking, and randomization can be used to improve the validity and efficiency of our study. The understanding and use of experimental design will provide a much more comprehensive gauge capability analysis than ad hoc approaches. Furthermore, modern manufacturing operations require a thorough evaluation of the measurement process.

In this article, we present gauge capability analysis as an experimental design problem and discuss enhancements to the classical gauge repeatability and reproducibility (R&R) studies. We also provide examples of experimental design approaches and discuss statistical evaluation of the gauge capability parameters, expressed as variance components. In Part II of the paper, we will discuss confidence interval estimates of the variance components and make recommendations regarding experimental strategy, including sample sizes. These statistical criteria are typically not included in classical gauge R&R guidelines.

The product measured in these studies will be referred to as a part for simplicity in this discussion. The comments apply equally well to samples of product that might be selected from continuous or batch processes, such as is typical in the chemical industry.

Planning Gauge Capability Studies

Simple Example

A successful gauge capability study, one that provides good estimates of the variation in the measurement process and identifies the factors that are most influential to that variation, requires much more than just an accurate statistical analysis. The experimental details of conducting the gauge capability analysis are very important. Unfortunately, the result of errors in the design and analysis of these studies is often that the true variation in the measurement process is underestimated. This overly optimistic conclusion regarding gauge capability leads to inefficient operation of the overall process.

A simple gauge capability study is illustrated by the following example from Montgomery (8). Random selections of 20 parts are chosen from a process and each part is measured by one operator using a particular gauge (in random order) two times. The repeated measurements of each part are used to evaluate the capability of the measurement process. The data and the analysis are in Table 1.

It is well known that this experiment allows total variability σ^2_{total} to be decomposed into two variance components, say

$$\sigma^2_{\text{total}} = \sigma^2_{\text{product}} + \sigma^2_{\text{gauge}} \tag{1}$$

It is relatively simple to derive estimates of these variance components, as shown in the lower panel of Table 1. We strongly recommend that the data from a gauge capability study such as this be analyzed graphically via X-bar and R charts, as shown in Figure 1. The X-bar chart exhibits many out-of-control points. This is to be expected, as in this situation the X-bar chart has an interpretation that is somewhat different from the usual interpretation. The X-bar chart shows the discriminating power of the measurement instrument—literally, the ability of the gauge to distinguish between units of product. The R chart directly shows the magnitude of measurement error, or the gauge capability, often called gauge repeatability. In this example, the R chart is in control. This indicates that the operator is having no difficulty in making consistent

Table 1. A Simple Gauge Capability Study

PART NO.	MEASUREMENTS 1	2	$\bar{x}$	R
1	21	20	20.5	1
2	24	23	23.5	1
3	20	21	20.5	0
4	27	27	27.0	0
5	19	18	18.5	1
6	23	21	22.0	2
7	22	21	21.5	1
8	19	17	18.0	2
9	24	23	24.0	2
10	25	23	24.0	2
11	21	20	20.5	1
12	18	19	18.5	1
13	23	25	24.0	2
14	24	24	24.0	0
15	29	30	29.5	1
16	26	26	26.0	0
17	20	20	20.0	0
18	19	21	20.0	2
19	25	26	25.5	1
20	19	19	19.0	0
			$\bar{\bar{x}} = 22.3$	$\bar{R} = 1.0$

$$\sigma^2_{\text{total}} = \sigma^2_{\text{product}} + \sigma^2_{\text{gauge}}$$

$$\hat{\sigma}_{\text{gauge}} = \frac{\bar{R}}{d_2} = \frac{1.0}{1.128} = 0.887$$

$$\hat{\sigma}^2_{\text{total}} = 10.05, \quad \hat{\sigma}_{\text{total}} = 3.17$$

$$\hat{\sigma}^2_{\text{product}} = \hat{\sigma}^2_{\text{total}} - \hat{\sigma}^2_{\text{gauge}} = 10.05 - (0.887)^2 = 9.26$$

$$\hat{\sigma}_{\text{product}} = 3.04$$

measurements. Out-of-control points on the R chart would indicate that the operator is having difficulty using the instrument.

It has been our experience in practice that the X-bar and R chart analysis of gauge capability studies is frequently ignored. Below, we discuss a variation of the basic R chart that is also useful in gauge capability studies. Finally, we note that it is relatively standard practice to compare overall gauge capability (frequently taken as $6\hat{\sigma}_{\text{gauge}}$) to the specification or tolerance band (USL − LSL)

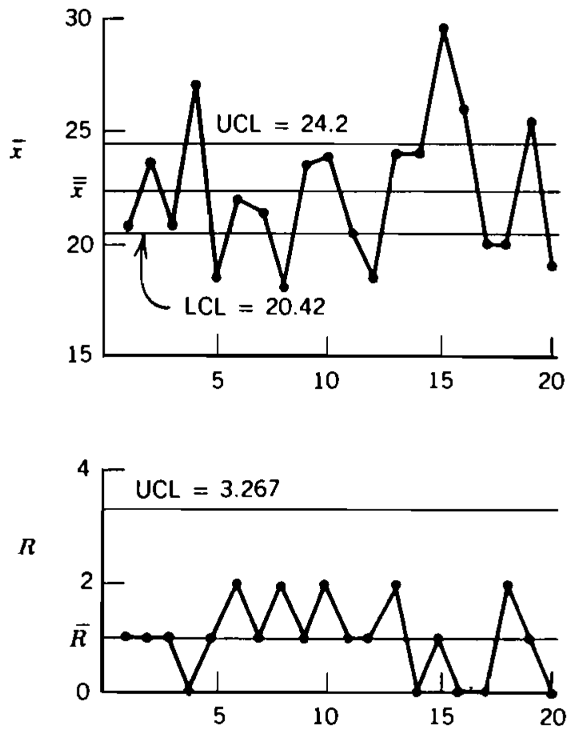

Figure 1. Control charts for the gauge capability analysis from Table 1.

for the part. This ratio of $6\hat{\sigma}_{\text{gauge}}$ to the total tolerance band is often called the precision-to-tolerance or P/T ratio:

$$\frac{P}{T} = \frac{6\hat{\sigma}_{\text{gauge}}}{\text{USL} - \text{LSL}}$$

The part used in this example has USL = 60 and LSL = 5. Therefore,

$$\frac{P}{T} = \frac{6(0.887)}{60 - 5} = \frac{5.32}{55} = 0.097$$

Values of P/T of 0.1 or less are frequently taken to imply adequate gauge capability. This is often justified on the widely used engineering rule that suggests a measurement device be calibrated in units one-tenth as large as the accuracy required in the final measurement. There are obvious dangers in relying too much on the P/T ratio: For example, the ratio may be made arbitrarily small by increasing the width of the specification band. It is more important that the gauge have sufficient capability to detect meaningful variation in the product.

The ratios of $\hat{\sigma}_{\text{gauge}}$ to $\hat{\sigma}_{\text{product}}$ or $\hat{\sigma}_{\text{gauge}}$ to $\hat{\sigma}_{\text{total}}$ are frequently more instructive measures than *P/T*. For the Table 1 data,

$$\frac{\hat{\sigma}_{\text{gauge}}}{\hat{\sigma}_{\text{product}}} \times 100 = \frac{0.877}{3.04} \times 100 = 29.2\%$$

$$\frac{\hat{\sigma}_{\text{gauge}}}{\hat{\sigma}_{\text{total}}} \times 100 = \frac{0.877}{3.17} \times 100 = 27.9\%$$

Thus, the gauge variability is approximately 29% of the product variability and approximately 28% of the total variability. These comparisons are frequently much more useful than the somewhat artificial *P/T*.

Although this is a very simple example, several aspects of conducting this type of gauge capability study are important. These aspects concern the number of parts and the number of measurements per part, the selection of the parts used in the analysis, and the experimental details of the replicated measurements on each part. How the study is conducted is as important as how many measurements are involved.

Number of Parts and Measurements per Part

Some individuals might prefer to use more measurements on fewer parts than in this example. There are at least three advantages to using many parts in the study, with few measurements on each, over the alternative of few parts with many measurements on each.

1. The parts are often selected from actual production and are representative of the material that the measurement system will encounter after the gauge capability analysis is completed and the instruments are certified for manufacturing use. A gauge might provide less variable results on a "standard" unit that is near the center of the manufacturing specifications than on product at the extremes of the manufacturing specifications. An obvious example is nonlinearity of the gauge causing unstable results beyond a linear operating region. Using many parts in the study increases the likelihood of detecting this problem. Also, actual product is often preferred to "standard" units in a gauge capability analysis because important product characteristics, which standards do not share, might produce unexpected measurement errors.
2. The variance of the measurements might not be constant; in particular, the variance might depend on the mean level of the measurements. This common phenomenon will not be detected if only a narrow range of product is used in the study. Obvious examples are any measurements proportional to the mean level of the product. Measurements over a range of products can be used to detect nonconstant measurement variance. Often, visual inspection of the data

is sufficient to detect this effect. A better approach is to construct an R chart for the data in which the ranges are ordered by the mean level of the part. Nonconstant variance is identified by a pattern such as a trend on this chart. This chart is a supplement to the usual R chart. This latter chart is sensitive to changes in the measurement system over the course of the study, but it is often ineffective in detecting nonconstant variance.

3. With multiple measurements on the same part, we are less likely to perform complete replications of the measurement process. Without complete replications, we omit important sources of variability and our estimate of measurement variability is overly optimistic. A new part often provides more opportunity for sources of variation to occur in the measurement process than repeat measurements of the same part and encourages complete replications of the measurement process. An example is the variation caused by fixturing parts for measurement. Each new part will require fixturing and the variation in the measurement process due to fixturing will be included in our study. Slightly different parts might affect the consistency and stability of our fixturing. By measuring the same part many times, we might not even change fixturing, and we might omit an important source of variation in the measurement process. Further comments are provided in the discussion of experimental details for replicated measurements below.

Additional criteria for choosing the number of parts and the number of measurements per part are to consider the length of confidence interval estimates of the relevant variance components. These confidence intervals are discussed further in Part II of the paper.

Selection of Parts

The selection of parts used in the study should, as much as possible, span the range of measurements that are anticipated from the manufacturing process. This might include parts that are out of specifications, if production will send those type of parts through the measurement process.

A random sample from current production might result in acceptable parts for the study. However, a manufacturing process might produce unusual parts at a low frequency and the measurement process might not be effective nor sensitive to parts that are very unusual. This might occur from nonlinearity of transducers beyond a certain distance from the process target. A sample of 20 parts obtained within a particular time period might not include any of these unusual parts.

Also, for an in-control process, it is unlikely that a sample of 20 parts will include product at or beyond three standard deviations from the process mean; in fact, even product beyond two standard deviations from the process mean will not be selected often. Therefore, it might be valuable to supplement random samples with historical parts from the extremes of the manufacturing process. If historical part collections are not available, as will be the case in new processes, it will be valuable to supplement random samples with standards at the extremes of the expected manufacturing process. This approach can be useful, but it must be verified that current production is not substantially different from the era in which the historical set was chosen.

Experimental Details of Replicated Measurements

The experimental details of the replicated measurements on each part is probably the most important aspect of a gauge capability study and the area where errors will most likely occur. Unfortunately, errors in this phase of the experiment can result in underestimates of the true measurement variability.

Surface Flatness

A replicated measurement in a gauge capability analysis should be a complete replicate of the measurement process. For example, consider a surface flatness measurement. Assume a part is fixtured under a particular probe tip, the measurements are taken on the same location on the part, and the measurements are taken over a short enough time interval that there are no ambient environmental changes, electrical surges, or gauge wear-out. The measurements will most likely be very repeatable. However, the measurement process that will be used as part of the manufacturing process will be quite different. An operator will select a part. There might be more than one gauge to be used, possibly from different manufacturers. Possibly there are many probes that could be selected for use with this gauge. The part must be fixtured. A location on the part is selected. The probe might have worn since the time a calibration was performed. Environmental conditions, power supplies, and other gauge operating conditions might have changed since the last calibration.

Chemical Processes—Sample Preparation and Batch Homogeneity

In chemical processes, a complete replication of the measurement process entails a new sample preparation, including all requisite heating, stirring, mixing, reagents, purifying, etc., possibly performed by different technicians. For measurements of a batch's characteristics in a batch process, a complete replication of the measurement process will often require a new sample from the batch. A new sample is a particularly controversial recommendation for laboratory personnel because it extends the factors being included in the measurement process from the confines of the lab to batch homogeneity in the manufacturing operation. However, if a sample is taken to characterize a batch and if a batch might not be homogeneous, then sample-to-sample variation is an additional component of measurement variability and this component should be included in a gauge capability analysis.

Many factors that influence the measurement—many sources of variation—can cause the measurement *now* to be much different from the identical part measured at another time. These sources of variations determine the variation in our measurement process and, to the extent possible, they should be included in a gauge capability analysis. A complete replication in a gauge capability analysis will include as many sources of variation as possible by starting at the beginning of the measurement process for every replicate in the study. For the flatness example, this might entail measuring the same part on another gauge with different fixturing. The location on the part measured should only be identified to the extent it is described in the operating procedures for the measurement process. Extra emphasis to measure the same location of the part will cause measurement variability to be underestimated.

Accelerating Long-Term Variability

Even with efforts to perform a complete replication in the measurement study, we might not normally be able to include long-term sources of variability over the course of the gauge capability analysis. We might have to be satisfied with a study that only estimates short-term gauge capability. However, long-term variability can be estimated if important long-term factors in the measurement process can be accelerated for the purpose of the study. For example, different gauges can be used in different locations possibly to include factors such as equipment, environmental conditions, operating instructions, etc. Also, chemical analyses can use different lots or batches of reagents for different replications to accelerate reagent changes that will be encountered when the manufacturing and measurement process is operating. Even environmental chambers can be used for environmental factors, and heating might be an appropriate substitute for aging of measurement equipment. Many ingenious approaches can be used to accelerate changes to the long-term factors in the measurement process for the purposes of the study. The resources devoted to including these factors will result in an improved estimate of gauge capability.

Recommendations

Based on the discussion above, we now recommend some approaches for beginning a gauge capability study. A simple approach for a gauge capability analysis is to conduct a preliminary analysis by including as many sources of variability as possible in the study; that is, vary as many factors from the measurement process as possible between replicated measurements of the same part. Use accelerated methods when possible to simulate long-term sources of variability. The same part might be measured by different equipment in different locations on different days. A second part might be measured by equipment different from any that was used for the replicates on the first part. This is a simple, unstructured approach for a gauge capability analysis that should provide representative data for the subsequent statistical analysis.

The simple approach is useful as an easy method to obtain preliminary information on the measurement process. It represents an attempt to obtain a worst-case evaluation of measurement capability. If the results of the preliminary study are satisfactory, the measurement process can be considered to be acceptable.

However, the simple preliminary approach above is limited when the results from the gauge capability analysis indicate that the measurement process is not capable of meeting the requirements of the manufacturing process. The unstructured inclusion of factors which might affect variability prevents isolation of the factors (the sources of variation) that are responsible for the unacceptable performance of the measurement system. It might be difficult or impossible to determine if the measurement variability in the unstructured study is due to operators, equipment, environmental conditions, etc.

Experimental Designs

A better approach is to design a gauge capability analysis systematically to include all factors that might contribute variation to the measurement system. By systematically including many factors in a designed study, the effect of different factors on the measurement variability can be determined. The basic principles used in designing experiments will be very effective in a gauge capability analysis. A factorial experiment is useful for investigating many factors simultaneously in an experiment and will allow us to estimate easily the contribution of each factor to measurement variation. Thus, the analysis takes the form of a variance components study.

A general recommendation for designing experiments is that 25% or less of total resources be allocated to the initial study. Based on this guideline, a gauge capability analysis should use only two or three different settings (levels) for each factor in the experiment. For example, we can include two operators and two different gauges as two levels for each factor in the study. Each part in this example would be measured two times by each of the four combinations of operators and gauges. After the important factors have been identified, further experiments can use more levels of the factors (if necessary) to quantify their effect on measurement variability.

Fractional factorial designs could also be used in the initial phases of a gauge capability study to screen out the important sources of variability. A screening experiment would then be followed by a more elaborate experiment. This strategy is a relatively standard part of experimental design practice (1,9).

The Classical Gauge Repeatability and Reproducibility Study

A classical gauge repeatability and reproducibility study is illustrated by an example from Montgomery (8). A sample of 3 operators each measure 20 parts 2 times each, in random order. The data for this example and the analysis is shown in Table 2.

The classical analysis in Table 2 uses sample ranges to obtain estimates of the $\sigma_{\text{repeatability}}$ and $\sigma_{\text{reproducibility}}$. We now describe an experimental design approach for the analysis of this gauge study in which we estimate the variance components of an experimental design model. A similar approach has been used previously by Tsai (12) and more recently by Deutler (3).

The measurement by operator i on part j at replication k is denoted as X_{ijk}. We assume the two-factor factorial model

$$X_{\text{ijk}} = \mu + O_i + P_j + (OP)_{ij} + R_{k(ij)} \qquad \begin{cases} i = 1, 2, \ldots, o; \\ j = 1, 2, \ldots, p; \\ k = 1, 2, \ldots, n \end{cases} \tag{2}$$

where O_i, P_j, $(OP)_{ij}$, and $R_{k(ij)}$ are random variables representing the effects of the Operators, Parts, Operator by Part interaction, and replications on the measurement, respectively, and μ is an overall mean common to all observations. In gauge capability studies, we usually assume that the Operator, Part, Operator by Part interaction, and Replication factors are independent, random effects that are normally distributed with means 0 and variances σ_O^2, σ_P^2, σ_{OP}^2, and σ_R^2, respectively. The parts in gauge capability studies would almost always be treated as random effects. In some cases, the operators may be the only operators that are available; thus, it could be argued that operators would be treated as a fixed effect. However, we feel that it is usually desirable to regard the operators as representative of a larger population of potential operators from which they were randomly selected. Thus, a random effects viewpoint for operators is usually appropriate. The Operator and Part factors are factorial variables because all 20 parts are measured by all three operators and the Repeatability factor is nested within Part and Operator.

Although the model above is easily described and analyzed, a number of implications of this model are important: (1) the assumption that the terms in the model associated with Part, Operator, Operator by Part, and Replication have mean zero implies that the measurement system is unbiased (the measurement variability under investigation is due to precision) and (2) the variance components are assumed constant. For example, variability does not increase with the magnitude of the measurement. Transformations to logarithms of the response or to percent changes can be used if the variability is proportional to the magnitude of the measurement. (3) The model assumes the measurement is

Table 2. Classical Gauge Repeatability and Reproducibility Study from Montgomery (8)

PART NO.	OPERATOR 1 MEASUREMENTS 1	2	$\bar{x}$	R	OPERATOR 2 MEASUREMENTS 1	2	$\bar{x}$	R	OPERATOR 3 MEASUREMENTS 1	2	$\bar{x}$	R
1	21	20	20.5	1	20	20	20.0	0	19	21	20.0	2
2	24	23	23.5	1	24	24	24.0	0	23	24	23.5	1
3	20	21	20.5	1	19	21	20.0	2	20	22	21.0	2
4	27	27	27.0	0	28	26	27.0	2	27	28	27.5	1
5	19	18	18.5	1	19	18	18.5	1	18	21	19.5	3
6	23	21	22.0	2	24	21	22.5	3	23	22	22.5	1
7	22	21	21.5	1	22	24	23.0	2	22	20	21.0	2
8	19	17	18.0	2	18	20	19.0	2	19	18	18.5	1
9	24	23	23.5	1	25	23	24.0	2	24	24	24.0	0
10	25	23	24.0	2	26	25	25.5	1	24	25	24.5	1
11	21	20	20.5	1	20	20	20.0	0	21	20	20.5	1
12	18	19	18.5	1	17	19	18.0	2	18	19	18.5	1
13	23	25	24.0	2	25	25	25.0	0	25	25	25.0	0
14	24	24	24.0	0	23	25	24.0	2	24	25	24.5	1
15	29	30	29.5	1	30	28	29.0	2	31	30	30.5	1
16	26	26	26.0	0	25	26	25.5	1	25	27	26.0	2
17	20	20	20.0	0	19	20	19.5	1	20	20	20.0	0
18	19	21	20.0	2	19	19	19.0	0	21	23	22.0	2
19	25	26	25.5	1	25	24	24.5	1	25	25	25.0	0
20	19	19	19.0	0	18	17	17.5	1	19	17	18.0	2
			$\bar{\bar{x}}_1 = 22.30$	$\bar{R}_1 = 1.00$			$\bar{\bar{x}}_2 = 22.28$	$\bar{R}_2 = 1.25$			$\bar{\bar{x}}_3 = 22.60$	$\bar{R}_3 = 1.20$

$$\sigma^2_{\text{total}} = \sigma^2_{\text{product}} + \sigma^2_{\text{gauge}}$$

Repeatability

$$\bar{\bar{R}} = \tfrac{1}{3}(\bar{R}_1 + \bar{R}_2 + \bar{R}_3)$$
$$= \tfrac{1}{3}(1.00 + 1.25 + 1.20)$$
$$= 1.15$$

$$\hat{\sigma}_{\text{repeatability}} = \frac{\bar{\bar{R}}}{d_2} = \frac{1.15}{1.128} = 1.02$$

Reproducibility

$$\bar{\bar{x}}_{\max} = \max(\bar{\bar{x}}_1, \bar{\bar{x}}_2, \bar{\bar{x}}_3) = 22.60$$
$$\bar{\bar{x}}_{\min} = \min(\bar{\bar{x}}_1, \bar{\bar{x}}_2, \bar{\bar{x}}_3) = 22.28$$
$$R_{\bar{\bar{x}}} = \bar{\bar{x}}_{\max} - \bar{\bar{x}}_{\min} = 22.60 - 22.28 = 0.32$$

$$\hat{\sigma}_{\text{reproducibility}} = \frac{R_{\bar{\bar{x}}}}{d_2} = \frac{0.32}{1.693} = 0.19$$

$$\hat{\sigma}^2_{\text{gauge}} = \hat{\sigma}^2_{\text{repeatability}} + \hat{\sigma}_{\text{reproducibility}}$$
$$= (1.02)^2 + (0.19)^2$$
$$= 1.08$$

linear in the Part, Operator, Operator by Part, and Replication factors. Prior study of the measurement system would evaluate the range of linearity of the measurment process. A calibration study is appropriate, before the measurement capability study, to assess adequately these assumptions. In this paper, we restrict attention to measurement capability studies and assume the model above. Standard methods for assessing departures from a model in an experimental design contact should be applied to gauge capability experiments. These methods include normal probability plotting of residuals and plots of residuals versus the predicted response and the design factors. Refer to the works of Box, Hunter, and Hunter (1) and Montgomery (9) for details.

The analysis of variance for this model using the data from Table 2 is presented in Table 3. The repeatability of the gauge is estimated by $\hat{\sigma}^2_{\text{repeatability}} = 0.99$. This estimate of $\sigma^2_{\text{repeatability}}$ is similar to the results obtained from the classical R&R analysis.

The estimate of the reproducibility of the gauge requires a quantitative description of reproducibility applicable to the analysis of variance model.

Table 3. Analysis of Variance for the Gauge Capability Study in Table 2

SOURCE OF VARIABILITY	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F_0	P-value
Operators	2.62	2	1.31	1.85	0.1711
Parts	1185.43	19	62.39	87.63	0.0001
Operator by Part	27.05	38	0.71	0.72	0.4909
Repeatability	59.50	60	0.99		
Total	1274.60	119			

Expected Mean Squares:

$$E(MS_O) = \sigma_R^2 + n\sigma_O^2 + pn\sigma_O^2$$
$$E(MS_P) = \sigma_R^2 + n\sigma_{OP}^2 + on\sigma_P^2$$
$$E(MS_{OP}) = \sigma_R^2 + n\sigma_{OP}^2$$
$$E(MS_R) = \sigma_R^2$$

Estimates of Variance Components:

$$\hat{\sigma}_O^2 = (MS_O - MS_{OP})/pn$$
$$\hat{\sigma}_P^2 = (MS_P - MS_{OP})/on$$
$$\hat{\sigma}_{OP}^2 = (MS_{OP} - MS_R)/n$$
$$\hat{\sigma}_R^2 = MS_R$$

Estimates of Variance Components:

$$\sigma^2_{\text{repeatability}} = \sigma_R^2 = 0.99$$
$$\hat{\sigma}_{OP}^2 = (0.71 - 0.99)/2 = -0.14(\text{estimate} < 0)$$
$$\sigma_P^2 = (62.39 - 0.71)/6 = 10.28$$
$$\sigma_O^2 = (1.31 - 0.71)/40 = 0.015$$
$$\sigma^2_{\text{reproducibility}} = \hat{\sigma}_O^2 + \hat{\sigma}_{OP}^2 = 0.015 - 0.14 = -0.125(\text{estimate} < 0)$$

Because Operator is the only factor in the study other than Part, a reasonable quantification of reproducibility is σ_O^2. The value of σ_O^2 describes the amount of variability in the study attributed to different operators. However, in the gauge capability analysis, there could be a substantial Operator by Part interaction. This could result from operators having difficulty with part fixturing or, for a chemical measurement, sample preparations, resulting in differences between operators' measurements for some parts, but not for others. However, if a selected part is measured by two different operators, there can be substantial differences in the readings. Thus, there is measurement error due to Operator by Part interaction.

For these reasons, we prefer to define reproducibility as

$$\sigma_{\text{reproducibility}}^2 = \sigma_O^2 + \sigma_{OP}^2 \tag{3}$$

By obtaining estimates of the variance components σ_O^2 and σ_{OP}^2 we may obtain an estimate of $\sigma_{\text{reproducibility}}^2$. The analysis-of-variance method of variance component estimation, in which expected mean squares are equated to their observed values and the resulting equations solved for estimates of the variance components, is an obvious choice. The technical details of this method, which is really a moment estimator, are provided by Searle (11). The bottom panel of Table 3 presents the equations that result for $\hat{\sigma}_O^2$ and $\hat{\sigma}_{OP}^2$.

Table 3 also presents the analysis-of-variance-method estimators of $\hat{\sigma}_O^2 = 0.015$ and $\hat{\sigma}_{OP}^2 = -0.14$. Note that because the mean square for Repeatability is larger than the mean square for Operator by Part interaction, a negative estimate of σ_{OP}^2 is obtained. Furthermore, from Eq. (3),

$$\begin{aligned}\hat{\sigma}_{\text{reproducibility}}^2 &= \hat{\sigma}_O^2 + \hat{\sigma}_{OP}^2 \\ &= 0.015 - 0.14 \\ &= -0.125\end{aligned}$$

Thus, our estimate $\hat{\sigma}_{\text{reproducibility}}^2$ is negative. The analysis-of-variance method can produce negative estimates of variance components. This is considered to be one of its major drawbacks. Because Operator by Part interaction may be small in gauge capability studies, negative estimates of $\hat{\sigma}_{OP}^2$ and $\hat{\sigma}_{\text{reproducibility}}^2$ will not be unusual.

One approach to a negative estimate of a variance component is to replace the negative estimate by zero. However, as Searle (11) points out, replacing a negative estimate of a variance component by zero results in biased estimates of other model parameters. We feel that the modified analysis-of-variance method is a better alternative [see Milliken and Johnson (7)]. In this procedure, a model-building approach is used to determine which variance components should be included in the final model. To implement the procedure, start with the full model and test hypotheses about the variance components with the analysis of variance (Table 3). If all variance components are statistically significant at some chosen α (say, $\alpha < 0.3$), then estimate the variance components, which should all be non-negative. If some of the variance components do not differ significantly from zero at the chosen α level, then eliminate the least important model factor and refit the reduced model. This procedure would be continued until all variance components are significantly different from zero. By using this model-building approach, the analysis-of-variance method estimates of the variance components will be non-negative.

Table 3 shows the P-values for the F-tests on the variance components. Because the P-value for the Operator by Part interaction variance component is large, we should drop this term from the model, resulting in a restricted model that is similar to Eq. (2) but without the interaction term. When we fit this model to the data, Table 4 results. The P-value for Operators is not small, but because the F-ratio exceeds one, a non-negative estimate of σ_O^2 will be obtained, so we will retain this term in the final model. The bottom panel of Table 4 shows the estimates of the variance components. Note that now $\hat{\sigma}_{\text{repeatability}} = \sqrt{0.88} = 0.94$ and $\hat{\sigma}_{\text{reproducibility}} = \sqrt{0.011} = 0.11$. Although these results do not differ dramatically from those obtained in the classical gauge study in Table 2, we feel that they are preferable because their statistical properties are better understood. Furthermore, as we will show in Part II of this paper, it is relatively easy to construct confidence intervals for these variance components.

Table 4. Modified Analysis of Variance for the Gauge Capability Study in Table 2

SOURCE OF VARIABILITY	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE	F_0	P-value
Operators	2.62	2	1.31	1.49	0.2304
Parts	1185.43	19	62.39	70.90	0.0001
Repeatability	86.55	98	0.88		
Total	1274.60	119			

Expected Mean Squares:

$E(MS_O) = \sigma_R^2 + pn\sigma_O^2$

$E(MS_P) = \sigma_R^2 + on\sigma_P^2$

$E(MS_R) = \sigma_R^2$

Estimates of Variance Components:

$\hat{\sigma}_O^2 = (MS_O - MS_R)/pn$

$\hat{\sigma}_P^2 = (MS_P - MS_R)/on$

$\hat{\sigma}_R^2 = MS_R$

Estimates of Variance Components:

$$\hat{\sigma}^2_{\text{repeatability}} = \sigma_R^2 = 0.88$$

$$\hat{\sigma}_P^2 = (62.39 - 0.88)/6 = 10.25$$

$$\sigma_O^2 = (1.31 - 0.88)/40 = 0.011$$

$$\hat{\sigma}^2_{\text{reproducibility}} = \hat{\sigma}_O^2 = 0.011$$

Although we concentrate here on the analysis-of-variance method of estimating variance components, other estimation methods could be used. The analysis-of-variance method produces method of moments estimators. Minimum variance quadratic unbiased estimators [see Rao (10)], maximum likelihood (4,11) and constrained maximum likelihood methods are other possibilities. In particular, maximum likelihood estimators could be attractive alternatives to the method of moments approach because non-negativity constraints could be imposed when the latter method produces large negative variance component estimates. We will illustrate this in Part II of the paper. For another example, see the work of Inman, Ledolter, Lenth, and Niemi (6).

Bias of Classical Reproducibility Estimates

We now investigate the difference between the ANOVA and classical analyses by considering the expected values of the two different estimators. From the model in Eq. (2), the mean of the readings for operator i is

$$\bar{X}_{i.} = O_i + \frac{\sum_{j=1}^{p} P_j}{p} + \frac{\sum_{j=1}^{p} OP_{ij}}{p} + \frac{\sum_{j=1}^{p} \sum_{k=1}^{n} R_{k(ij)}}{pn} \tag{4}$$

Classical gauge capability analysis estimates reproducibility as $\hat{\sigma}_{\text{reproducibility}} = R_{\bar{X}}/d_2$. The d_2 is chosen to convert a range into an unbiased estimator of standard deviation. By using the model for the data and the fact that the d_2 factors generate unbiased estimates of standard deviation, it can be shown that

$$\mathrm{E}(\hat{\sigma}_{\text{reproducibility}}) = \mathrm{E}\left[\frac{R_{\bar{X}_{i.}}}{d_2}\right] = \left[\sigma_O^2 + \frac{\sigma_{OP}^2}{p} + \frac{\sigma_R^2}{pn}\right]^{1/2} \tag{5}$$

Because the same parts are measured by each operator in the study, the derivation of the expected value depends on the result that the means of the measurement taken by each operator are not independent. Comparing Eq. (5) with Eq. (3), $\hat{\sigma}_{\text{reproducibility}}$ is seen to be a biased estimator of $\sigma_{\text{reproducibility}}$. Typically, the contribution to the bias from the term σ_R^2 is small because the magnitude of the variance component for Replications is not much larger than the variance component for Operator and the divisor of pn reduces the contribution to the bias from this term quickly as the number of parts and replications used in the gauge capability analysis increase. However, σ_{OP}^2 can be an important component of reproducibility and the divisor of p (the number of parts in the study) is often large; the result is that classical R&R gauge capability analysis can substantially underestimate reproducibility of the measurement process.

The underestimate of gauge reproducibility did not occur in the previous example because Operator by Part interaction was small; it was not significantly different from zero. In this example, either analysis would indicate that improvements in the measurement process need to focus on the repeatability of the system because the majority of the variation is due to that component.

Although the classical gauge capability analysis results are not dramatically different from the ANOVA results in the example from Montgomery (8), another example is shown in Table 5. The data are adopted from an example in a manual produced by IBM (5). Where needed, estimates from the classical and ANOVA analysis are denoted with a C and an A, respectively. The estimate $\hat{\sigma}^C_{\text{repeatability}} = 0.93$ and the estimate $\hat{\sigma}^C_{\text{reproducibility}} = 0.59$ are from the classical analysis. From the ANOVA analysis shown in Table 6, $\hat{\sigma}^2_{\text{repeatability}} = 0.81$, $\hat{\sigma}^2_{OP} = 1.94$, and $\hat{\sigma}^2_O = 0.013$. Therefore, the ANOVA analysis estimates

$$\hat{\sigma}^2_{\text{reproducibility}} = 1.94 + 0.013 = 1.95,$$

and by taking square roots, $\hat{\sigma}^A_{\text{reproducibility}} = 1.40$ and $\hat{\sigma}^A_{\text{repeatability}} = 0.90$. The two analyses provide similar estimates of repeatability. However, the ANOVA

Table 5. Classical Gauge Repeatability and Reproducibility Study from IBM (5)

OPERATOR 1			$\bar{x}$	OPERATOR 2			$\bar{x}$	OPERATOR 3			$\bar{x}$
56	55	57	56.0	57	58	56	57.0	56	57	56	56.3
63	62	62	62.3	58	59	57	58.0	62	64	64	63.3
56	54	55	55.0	57	55	56	56.0	55	55	55	55.0
57	55	56	56.0	56	57	55	56.0	56	57	55	56.0
58	58	57	57.7	59	60	60	59.7	57	60	60	59.0
56	55	54	55.0	60	59	57	58.7	55	57	56	56.0
56	55	56	55.7	58	56	56	56.7	59	60	59	59.3
57	57	56	56.7	54	55	53	54.0	57	58	57	57.3
65	65	64	64.7	64	64	65	64.3	65	64	65	64.7
58	57	57	57.3	61	60	60	60.3	58	59	60	59.0
$\bar{\bar{x}}_1 = 57.6$				$\bar{\bar{x}}_2 = 58.1$							$\bar{\bar{x}}_3 = 58.6$
$\bar{R}_1 = 1.4$				$\bar{R}_2 = 1.8$							$R_3 = 1.5$

Repeatability

$$\bar{\bar{R}} = \tfrac{1}{3}(\bar{R}_1 + \bar{R}_2 + \bar{R}_3) = 1.57$$

$$\hat{\sigma}_{\text{repeatability}} = \frac{\bar{\bar{R}}}{d_2} = \frac{1.57}{1.693} = 0.93$$

Reproducibility

$$\bar{\bar{x}}_{\max} = \max(\bar{\bar{x}}_1, \bar{\bar{x}}_2, \bar{\bar{x}}_3) = 58.6$$

$$\bar{\bar{x}}_{\min} = \min(\bar{\bar{x}}_1, \bar{\bar{x}}_2, \bar{\bar{x}}_3) = 57.6$$

$$R_{\bar{\bar{x}}} = \bar{\bar{x}}_{\max} - \bar{\bar{x}}_{\min} = 1.0$$

$$\hat{\sigma}_{\text{reproducibility}} = \frac{1.0}{1.693} = 0.59$$

$$\hat{\sigma}^2_{\text{gauge}} = \hat{\sigma}^2_{\text{repeatability}} + \hat{\sigma}^2_{\text{reproducibility}} = (0.93)^2 + (0.59)^2 = 1.21$$

Table 6. Analysis of Variance for the Gauge Capability Study in Table 5

SOURCE OF VARIABILITY	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE	F_0	P-value
Operators	2	14.07	7.03	1.06	0.3671
Parts	9	673.88	74.88	11.28	0.0001
Operator by Part	18	119.49	6.64	8.18	0.0001
Repeatability	60	48.67	0.81		
Total	89	856.10			

$$\hat{\sigma}_R^2 = 0.81$$

$$\hat{\sigma}_{OP}^2 = \frac{6.64 - 0.81}{3} = 1.94$$

$$\hat{\sigma}_O^2 = \frac{7.03 - 6.64}{30} = 0.013$$

$$\hat{\sigma}_{\text{reproducibility}} = \sqrt{\hat{\sigma}_O^2 + \hat{\sigma}_{OP}^2} = \sqrt{0.013 + 1.94} = 1.40$$

reproducibility estimate is substantially higher than the classical estimate and indicates that reproducibility problems should be addressed first in this measurement system. The difference between the classical and ANOVA analyses in this example is due to the large Operator by Part interaction.

The example in Tables 5 and 6 demonstrates that the classical reproducibility estimates should be replaced with ANOVA estimates when there is substantial Operator by Part interaction in the study. When there is substantial Operator by Part interaction in the study, just increasing the sample size will not improve the classical analysis; instead, the underestimate of reproducibility is magnified. From Eq. (5), although the bias of the classical estimator of $\sigma_{\text{reproducibility}}$ due to the $\sigma_{\text{repeatability}}$ term decreases as the number of parts in the study increases, the underestimate of reproducibility by the classical analysis increases as the divisor of σ_{OP}^2 in Eq. (5) increases.

Alternative, Simple Reproducibility Estimator

The classical R&R analysis is valuable because of its simplicity. A simple modification of the classical analysis can be used in cases with large Operator by Part interactions to approximate the ANOVA analysis.

The first step is to average the replicate measurements from each operator on each part to obtain $\bar{X}_{ij.}$. Next, compute the range over operators for each part and then average these ranges to obtain $\bar{R}$. Dividing $\bar{R}$ by the d_2 factor corresponding to the number of operators in the study provides an estimate of $\sigma_{\text{reproducibility}}$ denoted as $\hat{\sigma}^*_{\text{reproducibility}}$. It is easily shown that

$$E(\hat{\sigma}^*_{\text{reproducibility}}) = \left[\sigma_O^2 + \sigma_{OP}^2 + \frac{\sigma_R^2}{n}\right]^{1/2} \tag{6}$$

The bias in this estimator is due to the term $\sigma^2_{\text{repeatability}}/n$. If $\sigma^2_{\text{repeatability}}$ is small relative to the variance components σ_O^2 and σ_{OP}^2, or if the number of replicates n is large, the estimator in Eq. (6) should be a reasonable estimate for $\sigma_{\text{reproducibility}}$. In the example from Montgomery (8), the estimate of $\sigma_{\text{reproducibility}}$ calculated by this approach is 0.58. This estimate is higher than the classical R&R estimate because $\sigma^2_{\text{repeatability}}$ is relatively large in this example and n is small. In the second example, the modification to the classical estimator yields $\hat{\sigma}^*_{\text{reproducibility}} = 1.38$. This is very close to the ANOVA estimate of reproducibility and correctly identifies reproducibility as being the major component of variability in the measurement process. The classical analysis compares $\hat{\sigma}_{\text{reproducibility}} = 0.59$ to $\hat{\sigma}_{\text{repeatability}} = 0.93$ and incorrectly concludes that repeatability is the major component of measurement error.

Furthermore, a simple approach that maintains the simplicity of the classical R&R study is to compute reproducibility estimates by the classical method and by our proposed method. A large difference between the two estimates suggests that Operator by Part interaction should be investigated as one of the contributors to measurement error.

Although classical gauge capability analysis is an established methodology, in many processes technological improvements for measurement have increased the ratio of σ_{OP}^2 to σ_O^2 compared to historical standards. This does not diminish the need for a complete understanding of the measurement process because the requirements for defect prevention in manufacturing have simultaneously increased. Knowledge of measurement precision is an essential component of Total Quality. However, the presentation above shows that a change in this ratio degrades the estimates used in classical gauge capability analysis. Although the classical analysis has been useful in the past, our methods should be upgraded for the present manufacturing environment.

Overcorrection from Standards

It is common practice in many industries to use standards or controls on a daily (or shift) basis to adjust readings from the measurement process. The adjustment is often applied as follows: if a standard or control reads 10% higher than its known value, all subsequent readings for that day are adjusted downward by 10%. This practice is useful for eliminating systematic bias from the measurement process. A constant bias that will persist throughout the course of the day can sometimes be compensated for by the adjustment above. However, the approach is clearly ineffective against measurement precision. Random measurement variability in the daily measurement of the standard will be propagated as a bias throughout all subsequent measurements made on that day. In fact, this can be considered an insidious example of a form of overcontrol of the measurement process. A similar problem can arise if measurement instruments are not correctly calibrated. If calibrations are difficult, the error in the calibrations will be propagated as a bias in all subsequent measurements. With difficult calibrations, it might not be appropriate to recalibrate on a routine basis.

A recommended alternative is routinely to measure a standard (sometimes called a *control*) part or sample. These measurements can be graphed on a control chart and a calibration performed only if there is evidence that the measurements are out-of-control. This can be an effective approach when calibrations are difficult or time-consuming. Resources can be saved, and without frequent, suspect calibrations, the measurement system can be better controlled. The center line for the control chart of standards would most likely be chosen to equal the known reading of the standard. Because the action required from false alarms from this control chart would only be a calibration, control limits might be placed as close to the center line as two standard deviations to increase the sensitivity of the chart to measurement problems. In operations where the measurement process is a substantial component of total variability and calibrations are difficult or time-consuming, a chart for standards is a recommended starting point for gaining control of the measurement process.

Other Aspects of Gauge Capability Analysis

We have concentrated on treating the classical gauge R&R study as a designed experiment. The most obvious analysis format is that of a two-factor factorial, random effects model. This will produce estimates of variance components that are meaningful measures of gauge capability. If there is significant Operator by Part interaction, the classical analysis will provide biased estimates of the variance components. The analysis of variance is a natural format for testing hypotheses about variability due to Parts, Operators, and Operator by Part interaction.

There are other aspects of this problem that are important. For example, it might be very helpful to construct confidence intervals on the variance components. These will give an idea about the relative uncertainty associated with the estimates and provide some basis for sample size decisions. Furthermore, the random model might not always be appropriate; if operators in the examples above were a fixed factor, then the analysis should be based on a mixed-effects model. All of our analyses assumed that the factorial design analysis is appropriate. This is only true if all parts are measured by all operators. If each operator uses different parts, then a *nested* design model should be used. Finally, other sources of variation may need to be included. For example, a gauge capability analysis could investigate the effects of different tools on variability, and then an experiment involving tools, operators, and parts must be designed. Part II of this paper will deal with some of these problems.

References

1. Box, G. E. P., Hunter, W. G., and Hunter, J. S., *Statistics for Experimenters*, John Wiley & Sons, New York, 1978.
2. Deming, W. E., *Quality, Productivity and Competitive Position*, MIT Center for Advanced Engineering Study, Cambridge, MA, 1982.
3. Duetler, T., Grubbs-Type Estimators for Reproducibility Variances in an Interlaboratory Test Study, *J. Qual. Technol.* 23(4), 324–335 (1991).
4. Hemmerle, W. J. and Hartley, H. O., Computing Maximum Likelihood Estimates for the Mixed A.O.V Model Using the W Transformation, *Technometrics*, 15, 819–831 (1973).
5. IBM Corporation, *Process Control Capability and Improvement*, Southbury, CT, 1984
6. Inman, J., Ledolter, J., Lenth, R. V., and Niemi, L., Two Case Studies Involving an Optical Emission Spectrometer, *J. Qual. Technol.* 24(1), 27–36 (1992).
7. Milliken, G. A. and Johnson, D. E., *Analysis of Messy Data, Volume I: Designed Experiments*, Van Nostrand Reinhold, New York, 1984.
8. Montgomery, D. C., *Introduction to Statistical Quality Control*, 2nd ed., John Wiley & Sons, New York, 1991.
9. Montgomery, D. C., *Design and Analysis of Experiments*, 3rd ed., John Wiley & Sons, New York, 1991.
10. Rao, C. R., Minimum Variance Quadratic Unbiased Estimation of Variance Components, *J. Multivariate Anal.*, 1, 445–456 (1971).
11. Searle, S. R., *Linear Models*, John Wiley & Sons, Inc. New York, 1971.
12. Tsai, P., Variable Gage Repeatability and Reproducibility Study Using the Analysis of Variance Method. *Qual. Eng.* 1(1), 107–115 (1988).

About the Authors: Douglas C. Montgomery is Professor of Industrial & Management Systems Engineering at Arizona State University. He is author of seven books, including widely-used texts on statistical quality control and experimental design, and over 100 papers and articles. He is a Fellow of the American Society for Quality Control, and a Senior Member of the Institute of Industrial Engineers, and has held several national offices in these organizations, including Director of the Statistics Division of ASQC. Dr. Montgomery is also a member of the Society of Manufacturing Engineers, the Operations Research Society of America, and the American Statistical Association. He is on the editorial board of the *Journal of Quality Technology*.

George C. Runger is Assistant Professor in the Department of Decision Sciences and Engineering Systems at Rensselaer Polytechnic Institute. Formerly, Dr. Runger was a senior engineer at IBM where he implemented statistical process control in chemical and discrete manufacturing operations. His research interests are in quality and productivity improvement, integrating manufacturing data for management decisions, and experimental design strategies.