Gauge R&R Studies for Destructive Measurements

JEROEN DE MAST and ALBERT TRIP

Institute for Business and Industrial Statistics, University of Amsterdam (IBIS UvA)
Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands

The standard method for assessing a measurement system's precision is a gauge R&R study. Such a study involves an experiment in which each of a number of objects is measured multiple times. From the results, the spread of multiple measurements on a single object (the *measurement spread*) can be estimated. A serious complication is encountered when objects are affected by the measurement or when the true value of objects is variable. Such measurements are called *destructive*. Applying the standard gauge R&R set-up to destructive measurements is either impossible or results in an overestimation of measurement spread. This article studies alternative set-ups, which can be applied to destructive measurements if certain conditions hold. Even if required conditions are not completely met, the proposed approaches will at least lead to a smaller overestimation of measurement spread.

KEY WORDS: Gauge Capability; Measurement System Analysis; Measurement Theory.

Introduction

Statistical methodologies for quality improvement, such as statistical process control, the Shainin System, and the Six Sigma program, heavily depend on data in order to identify opportunities for improvement. As a consequence, the reliability of the data is considered an important issue. The three aforementioned improvement strategies explicitly require the experimenter to verify the precision of his measurement procedures.

The standard method for assessing the precision of a measurement system is a so-called *gauge repeatability and reproducibility study* (gauge R&R study). Such a study can be applied when the measurement system in question results in measurements on a continuous scale. Measurements are collected from a crossed design: each of a certain number of products (typically 10) is measured several times (typically twice) by a number of operators (typically 3). The results are analyzed using analysis of variance models, in which certain variance components are associated with different sources of measurement spread (between operators, operator-product interaction, and within-operators error).

It is crucial for a standard gauge R&R study that objects can be measured more than once. If this is not the case, the measurement is called *destructive*. If repeated measurements on a single object are not possible, the measurement spread is necessarily confounded with the object-to-object variation. This article explores the possibilities of gauge R&R studies in the situation of destructive measurement systems. The paper is limited to the case of continuous measurements.

Some Examples

In the course of this paper, we shall consider various examples.

I Weight of Biscuits The weight of biscuits is measured using a scale. The weights can be considered constant in time, and weighing a biscuit has no effect on its weight. Therefore, the measurement is nondestructive.

II Strength of Biscuits In order to measure the strength of biscuits, pressure is exerted onto them. The pressure is slowly increased until the biscuit breaks. The pressure at which the biscuit breaks is

Dr. de Mast is Senior Consultant at IBIS UvA. His e-mail address is jdemast@science.uva.nl

Dr. Trip is Senior Consultant at IBIS UvA. His e-mail address is atrip@science.uva.nl

the measured strength. The measurement is destructive because a biscuit is lost after its strength is determined and cannot be measured a second time.

III Pressure in Pipelines The pressure in water pipelines can be measured using ordinary pressure meters. To determine the precision of such a meter, though, one comes across the complication that the water pressure continuously fluctuates. As a consequence, measurement variation is confounded with variation in the water pressure.

IV Flight Times of Paper Helicopters A popular case to illustrate experimental design and other statistical methods in the classroom uses helicopters made of paper (see Box and Liu (1999)). An important characteristic of these helicopters is their flight time. A flight time is measured by releasing a helicopter from a predesignated height and measuring with a stopwatch the lapsed time before it hits the floor. It is not possible to obtain multiple measurements from a single person from a single flight. Therefore, the experimenter should measure different flights. However, because the flights themselves will be different, part of the measured variation in flight times is not due to the measurement procedure but consists of variation between different flights.

Gauge R&R Studies and Destructive Measurements

Measurement

Each statistical study concerns experimental *units*, the objects of which properties are studied. The collection of all units is called the *population*. Units in the population can be classified according to a certain *property* (such a classification is sometimes called a *natural variable*). Measurement maps this classification onto a numerical system. Taking biscuits as units (example I), we could consider the property weight. By measuring the biscuits, we assign a number to each biscuit (its *weight*). On the one hand, there exists an empirical relation among the biscuits (some biscuits are heavier than others). On the other hand, we have mathematical relations among the weight values (such as the ordering relation and distance metric that are defined on $\mathfrak{R}$), and these mathematical relations are intended to reflect the empirical relation among the biscuits.

We arrive at the following definition. The *measurement* of a property of a unit is the assignment of a numeral to that unit, which reflects a classification of the units according to the property under study (cf., definitions by Lord and Novick (1968), Allen and Yen (1979), Wallsten (1988)). In this paper, we consider only measurement procedures that use a continuous scale, in which case the numerals are an element of $\mathfrak{R}$ or a subset thereof. (In fact, measurements are on a discrete scale at best; we take the term 'continuous measurement' as a *façon de parler* for a measurement of sufficiently high resolution.) A measurement could then be described as a map $Y : U \to \mathfrak{R}$, where U is the collection of units to which the measurement procedure applies.

To understand the problem of gauge R&R studies for destructive measurements (and—as we shall see—for nondestructive measurements as well), it is important to realize that objects evolve in time. Therefore, we modify our definition of units slightly and regard a single object that is considered on two moments in time as two different experimental units. For this reason, we denote the elements of U as $u_{i,t}$, i referring to a particular object and t indexing time. The measured value of a unit $u_{i,t} \in U$ is denoted $Y(u_{i,t})$.

In examples I and II, a unit is a certain biscuit i considered at time t. In example III, a unit $u_{i,t}$ is a pipeline i and the water that is in it at time t.

Gauge R&R Studies

Because many measurement procedures suffer from a random measurement error, part of the map, Y, is stochastic. We define

$$F_{Y|u}(y) = P(Y(u) \leq y \,|\, u) \quad \text{for fixed } u \in U,$$

which defines the probability distribution of the random measurement error. The usefulness of a measurement system is, to a large extent, determined by properties of $F_{Y|u}$. Measurement system analysis typically distinguishes the usefulness of a measurement system into its *accuracy* and its *precision* (AIAG (2002)).

Accuracy relates to the extent to which the measurement suffers from a bias. For any $u \in U$, let $\mu_u = \mathrm{E}_{Y|u}(Y(u)) = \int y \, dF_{Y|u}(y)$, the expected value of a measurement of a unit u. Bias is the difference between μ_u and the object's true value, $T(u)$ (i.e., a reference value that would be assigned to the object by a standard and authoritative measurement system). In the remainder of this article, the following assumption is made:

(a) Constancy of bias: $\mu_u - T(u)$ is constant over u (linearity) and constant in time (stability).

In this paper, we shall not elaborate on accuracy.

Precision relates to the variability of a measurement system. We define the measurement spread of measurement Y for unit $u \in U$ as the square root of

$$\sigma_u^2 = \mathrm{E}_{Y|u}\left(Y(u) - \mu_u\right)^2. \qquad (1)$$

Usually, it is assumed that

(b) Homogeneity of measurement error:

$$F_{Y|u_{i,t}}(y - \mu_{u_{i,t}}) = F_{Y|u_{j,s}}(y - \mu_{u_{j,s}}), \quad \text{for all } u_{i,t},\ u_{j,s} \in U,$$

that is, the distribution of the measurement error is identical for all units that are measured.

If (b) holds, the measurement spread is independent of the unit that is measured, and we have $\sigma_u^2 = \sigma^2$ for all $u \in U$. We shall assume throughout this paper that (b) holds.

If measurements on an object u are conducted under identical circumstances (the same person, immediately after each other), the precision of a measurement is at its best. The corresponding precision is called *repeatability*. If measurements are conducted under varying circumstances, precision will usually be worse. *Reproducibility* is the precision of measurements during which certain specified conditions vary in a realistic manner. A gauge R&R study (Burdick et al. (2003)) is a statistical study aimed at quantifying a measurement system's repeatability and reproducibility. In its typical setup, it involves an experiment in which a number of objects are measured multiple times by each of a number of operators. Measurement spread is decomposed into three variance components: $\sigma^2 = \sigma_O^2 + \sigma_{PO}^2 + \sigma_e^2$, where σ_O^2 is the variance that can be attributed to systematic differences among operators; σ_{PO}^2 is the variance resulting from additional differences among operators per object; and σ_e^2 the remaining error variance. Repeatability relates to σ_e^2, reproducibility to σ_O^2 and σ_{PO}^2.

The Fundamental Problem of Gauge R&R Studies

The fundamental problem of gauge R&R studies is that, in general, it is impossible for fixed $u_{i,t} \in U$ to observe more than one realization of $Y(u_{i,t})$ and, therefore, that it is impossible to estimate σ^2 without making some assumptions.

The fundamental problem of gauge R&R studies is solved in standard (nondestructive) gauge R&R studies by making two homogeneity assumptions:

(c) Temporal stability of objects: for an object i, $T(u_{i,t}) = T(u_{i,s})$ for any two moments t, s in a relevant time interval (and therefore, given (a), $\mu_{u_{i,t}} = \mu_{u_{i,s}}$). In words: it does not matter at which time objects are measured.

(d) Robustness against measurement: $T(u_{i,t})$ is equal before and after object i is measured. In words: objects are not affected when they are measured.

If (c) and (d) hold, the distribution of the measurement error can be estimated by collecting measurements $Y(u_{i,t_j})$, i fixed, $j = 1, \ldots, k$. Then σ^2 can be estimated, for instance, by

$$\hat{\mu} = \frac{1}{k}\sum_{j=1}^{k} Y(u_{i,t_j}),$$

and

$$\hat{\sigma}^2 = \frac{1}{k-1}\sum_{j=1}^{k}\left(Y(u_{i,t_j}) - \hat{\mu}\right)^2.$$

In practice, one will involve multiple objects, $i = 1, \ldots, n$, and compute a pooled estimate for σ. In example I, we can assume both (c) and (d) and perform a standard gauge R&R study.

Destructive measurements are those measurement procedures for which either (c) or (d) does not hold. The measurement procedure for determining the strength of biscuits is an example of a measurement for which (d) does not hold. Measuring pressure in pipelines is an example for which (c) does not hold.

In the preceding setup, we have limited experimental units to objects. If the units under study are other phenomena, U should be indexed differently. For example, the units of interest in example IV are *flights*. We could denote flight j of helicopter i as $u_{i,j}$ and the corresponding measured flight time as $Y(u_{i,j})$. In the remainder of this paper, we assume that units are objects considered at moments in time. The theory could be extended in a straightforward manner to populations with a different structure.

Solutions for the Fundamental Problem

For destructive measurements, the fundamental problem can be solved if the homogeneity conditions (c) and (d) can be replaced with alternative homogeneity conditions that do hold and that amount to the same effect. This section proposes a number of homogeneity conditions that could replace (c) and

(d) and thus enable estimation of σ. In practice, homogeneity conditions will not hold perfectly. The consequences of this fact will be studied in the next section.

If (c) holds (temporal stability), there is no need to differentiate between different time instants, and we can take $U = \{u_i\}_{i=1,\ldots,n}$, dropping the time index.

Homogeneity of Objects

In the situation where we have temporal stability but not robustness against measurement (assumption (c) holds, but (d) does not), we could exploit a potential (near) homogeneity across objects.

(e) Homogeneity of objects: there is a subset $H \subset U$ for which $T(u_i) = T(u_j)$ for all u_i and u_j in H. In words: some objects can be considered identical with respect to the measurement under study. Moreover, (c) is assumed to hold.

If (e) can be assumed, σ^2 can be estimated from a sample $u_1, \ldots, u_k$ from H by

$$\hat{\mu} = \frac{1}{k} \sum_{i=1}^{k} Y(u_i),$$

and

$$\hat{\sigma}^2 = \frac{1}{k-1} \sum_{i=1}^{k} \left(Y(u_i) - \hat{\mu}\right)^2 .$$

If (e) does not hold for the objects of interest, we might—with some ingenuity—be able to find some alternative objects, $A = \{v_i\}_{i=1,\ldots,k}$, beyond the units under study ($A \not\subset U$) to which the measurement procedure can be applied, for which assumption (e) does hold (H replaced with A) and for which the following homogeneity assumption can be made:

(f) Representativeness of alternative objects:

$$F_{Y|v_i}(y - \mu_{v_i}) = F_{Y|u}(y - \mu_u) \quad \text{for each } v_i \in A$$

and for arbitrary $u \in U$. In words: assumption (b) extends to the alternative objects $v_i \in A$.

Instead of measuring the strength of biscuits (example II), we could measure the strength of certain plastic bars, which are known to be very homogeneous (assumption (e), H replaced with A), whence they have approximately identical strengths. Choosing bars that break at more or less the same pressures as the biscuits, we may as well assume that the distribution of the measurement error while measuring these bars is indicative for the measurement error while measuring biscuits (assumption (f)). In the case of the paper helicopters (example IV), a similar trick is possible: one could record on video a number of helicopter flights (example IV). Measurement spread could be determined by playing back each recording several times and measuring the duration of the flight. Phillips et al. (1997) use a similar strategy to determine the measurement variability of a material strength test.

Correcting Heterogeneity Across Objects

Still assuming that we have temporal stability (assumption (c)) but not robustness against measurement (d), we study what we can do if the trick above (replacing (d) with (e)) does not work. The idea is to model the heterogeneity across objects and correct for it, so that the objects are made homogeneous in an artificial manner.

(g) Patterned object variation: $T(u_i) = f(i)$ for all u_i in a subset H of U, with f a function that has a limited number of parameters. In words: the variation across the objects follows a certain pattern. Moreover, (c) is assumed to hold.

The pattern could be a polynomial function, such as $f(i) = \beta_0 + \beta_1 i$ for successive products $i = 1, 2, \ldots, k$, or positional differences among objects. Also, moving average or autoregressive models can be considered. Given that the variation of the units is completely determined by this pattern (as assumption (g) states) and given a model $\hat{f}$ for this pattern, the measurement error σ can be estimated by

$$\hat{\sigma}^2 = \frac{1}{k-p} \sum_{i=1}^{k} \left(Y(u_i) - \hat{f}(i)\right)^2 ,$$

with p the number of parameters of function f that must be estimated and $u_i, i = 1, \ldots, k$ a sample from H.

Suppose that, for a certain sequence of biscuits, the strength increases following a linear trend:

$$T(u_i) = f(i) = \beta_0 + \beta_1 i, \quad i = 1, 2, \ldots, k, \qquad (2)$$

with u_i the ith biscuit in the sequence. We can estimate the parameters of the pattern f by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{k} Y(u_i)(i - \bar{i})}{\sum_{i=1}^{k} (i - \bar{i})^2},$$

and

$$\hat{\beta}_0 = \overline{Y(u_i)} - \hat{\beta}_1 \bar{i}.$$

Thereupon, σ^2 can be estimated by

$$\hat{\sigma}^2 = \frac{1}{k-2}\sum_{j=1}^{k}\left(Y(u_j) - \hat{f}(j)\right)^2, \qquad (3)$$

with $\hat{f}(j) = \hat{\beta}_0 + \hat{\beta}_1 j$. Of course, in practice, the strength of a sequence of biscuits does not precisely follow a model such as Equation (2); there will also be some random variation. The implications of this fact will be studied in the next section.

Comparison to a Nondestructive Measurement Procedure

In the situation that (d) does not hold, it might be possible to find an alternative measurement procedure that is not destructive. If an experimenter is willing to determine the precision of the destructive measurement system anyway, he can do so with the help of this nondestructive measurement system.

(h) There is an alternative measurement procedure for which (d) holds. Moreover, (c) is assumed to hold.

We select a sample of k objects and do a standard gauge R&R study on them using the alternative measurement procedure. This allows us to estimate the variance σ_P^2 of the objects' true values. Next, all objects are measured once using the measurement procedure of interest. The variance of these measurements, after subtracting $\hat{\sigma}_P^2$, estimates σ^2.

As an example, consider the determination of the thickness of phosphor layers on displays. This thickness can be determined with the help of a device that sends a ray of light through the display and measures how much light is blocked. Due to its immobility, this nondestructive measurement system cannot be used, however, on the site where it is needed. For this reason, the operators use an alternative measurement procedure during manufacturing, in which the phosphor in a specified area is scraped off and weighed. Because this (destructive) measurement procedure was to be used for manufacturing, its precision needed to be assessed. The operators used the procedure described above.

Comparison to a Perfect Destructive Measurement Procedure

Suppose we have an alternative, destructive measurement procedure for which the measurement spread is practically zero (expensive laboratory equipment).

(i) There is an alternative measurement procedure X for which $X(u_i) = T(u_i)$ for all $u_i \in U$. Moreover, (c) is assumed to hold.

We select a sample of k objects and randomly split it in two subsamples. One subsample is measured using the alternative procedure. The variance of these measurements gives us an estimate of the variance σ_P^2 of the objects in the sample. The other subsample is measured using the procedure of interest. The variance of these measurements, minus $\hat{\sigma}_P^2$, estimates σ^2.

Correcting for Temporal Instability

If temporal stability (assumption (c)) does not hold, one could model the fluctuation over time and correct for it:

(j) Patterned temporal variation: $T(u_{i,t}) = f(i,t)$ for all $u_{i,t}$ in a subset H of U, with f a function that has a limited number of parameters. In words: the variation over time of objects follows a certain pattern.

Given multiple measurements at moments $t_1, \ldots, t_k$ of a single object i, the measurement error can be estimated by

$$\hat{\sigma}^2 = \frac{1}{k-p}\sum_{j=1}^{k}\left(Y(u_{i,t_j}) - \hat{f}(i,t_j)\right)^2.$$

In practice, one can involve multiple objects and pool the estimated variances. Following this approach, we can estimate an ARIMA model $T(u_t) = f(t)$ that describes the pressure $T(u_t)$ in a pipeline at time t (example III) and take the variance of the residuals of this model as an estimate for σ^2.

Measuring Reference Material

If either (c) or (d) does not hold, it might be possible to obtain units for which there is a perfect measurement available, i.e., their true value is known. Typically, this is the case when one has at one's disposal calibration material with known true value.

(k) Known true value: $T(v_{i,t_i})$ is known for units $v_{i,t_i} \in A, i = 1, \ldots, k$. In words: we can find alternative units the true value of which is known. Moreover, (f) is assumed to hold for A, and the bias of the measurement system is assumed to be zero (or known, so that it can be corrected for).

Measurement error can be estimated by

$$\hat{\sigma}^2 = \frac{1}{k}\sum_{i=1}^{k}\left(Y(v_{i,t_i}) - T(v_{i,t_i})\right)^2. \qquad (4)$$

It is possible to buy plastic bars that break at a specified pressure (example II). These bars could be measured instead of biscuits. The measured spread, estimated from Equation (4), could be assumed to represent the measurement spread when measuring the biscuits' strength.

Note: if the bias of the measurement system is unknown, it can be estimated from

$$\hat{b} = \frac{1}{k}\sum_{i=1}^{k}\left(Y(v_{i,t_i}) - T(v_{i,t_i})\right).$$

The measurements should be corrected for the estimated bias, and the scalar $1/k$ in Equation (4) should be replaced with $1/(k-1)$.

What Happens if Homogeneity Conditions Do Not Hold Perfectly?

In the preceding section, we have seen that—in order to overcome the fundamental problem of gauge R&R studies—the experimenter must make one or more of the assumptions (c) through (k). Often, however, these assumptions hold only to a certain extent. In general, the consequence of this is that the obtained estimate for measurement spread overestimates the true measurement spread.

Consider, for instance, the biscuits example (example II). Suppose that we have assumed that the strength in a certain sequence of biscuits follows the model in Equation (2). Acknowledging that the variation in the strength of the biscuits also has a random component, we describe the true strength of biscuit i by

$$T(u_i) = \beta_0 + \beta_1 i + \epsilon_i,$$

with ϵ_i independent drawings from the $\mathcal{N}(0, \sigma_P^2)$ distribution. If this model gives an accurate representation of reality, then the estimator defined in Equation (3) has as its expected value not σ^2 but

$$\mathrm{E}(\hat{\sigma}^2) = \sigma^2 + \sigma_P^2,$$

i.e., random object-to-object variation and measurement spread are confounded. The good news is that the estimated measurement spread is on the safe side: if the estimated measurement spread is satisfactory, then so will be the true measurement spread. The negative effect of the overestimation of the measurement spread is that the chance increases that the measurement system is falsely judged inadequate. To reduce this chance, efforts should be made to obtain units that meet the appropriate ones of the assumptions (c) through (k) to a larger extent. This is the subject of the next section.

To be clear: it is theoretically impossible to distinguish object variation from measurement spread if the homogeneity assumptions do not hold perfectly, simply because this information is not contained in the data that can be collected. Bergeret et al. (2001) claim to have found a method to accomplish this distinction anyway. Their method involves a two-stage experiment that is modeled using nested analysis of variance models. The flaw in their method is that their $MS_{p\&e}$ should be written $MS_{p\&e\&loc}$ because it not only confounds parts variation and repeatability but location-to-location variation as well. The variance of the means of the units might be σ_p^2, but the variance of measurements on a single location of each unit is $\sigma_p^2 + \sigma_{\mathrm{loc}}^2$. Consequently, their estimator S_{repeat}^2 has expected value $\sigma_{\mathrm{loc}}^2 + \sigma^2$ instead of σ^2.

Experimental Designs for Destructive Gauge R&R Studies

Gauge R&R Study Under Assumptions (c) and (d)

The standard gauge R&R study has the experimenter collect measurements following a crossed design: each of I objects is measured K times by each of J operators. The kth measurement by operator j on object i is denoted y_{ijk}, $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$. Assuming (c) and (d), these measurements are typically modeled as

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

with $\alpha_i \sim \mathcal{N}(0, \sigma_P^2)$ the random object effects, $\beta_j \sim \mathcal{N}(0, \sigma_O^2)$ the random operator effects, $(\alpha\beta)_{ij} \sim \mathcal{N}(0, \sigma_{PO}^2)$ the object-operator interaction, and $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma_e^2)$ the error. The data can be analyzed using the ANOVA method (see Table 1, which uses standard ANOVA notation; see Montgomery (1997)). Taking appropriate linear combinations of mean squares, the experimenter finds estimates of σ_O^2 , σ_{PO}^2 and σ_e^2. Total measurement spread is estimated as $\sqrt{\hat{\sigma}_O^2 + \hat{\sigma}_{PO}^2 + \hat{\sigma}_e^2}$.

In case (c) or (d) do not hold perfectly (objects are either not perfectly stable or slightly affected by the measurements), some of the variance components are overestimated. Assuming that the measurements

TABLE 1. ANOVA Table

Source	df	Sum of Squares	Expected mean square
Objects	$I-1$	$JK\sum_i(\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdots})^2$	$\sigma_e^2+K\sigma_{PO}^2+JK\sigma_P^2$
Operator	$J-1$	$IK\sum_j(\bar{y}_{\cdot j\cdot}-\bar{y}_{\cdots})^2$	$\sigma_e^2+K\sigma_{PO}^2+IK\sigma_O^2$
Object × 0perator	$(I-1)(J-1)$	$K\sum_{i,j}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y}_{\cdots})^2$	$\sigma_e^2+K\sigma_{PO}^2$
Error	$IJ(K-1)$	$\sum_{i,j,k}(y_{ijk}-\bar{y}_{ij\cdot})^2$	σ_e^2

$\{y_{ijk}\}_{jk}$ are done in random order, only $\hat{\sigma}_e^2$ is affected.

Gauge R&R Study Under Assumptions (e) or (f)

A similar design could be used, but with repetitive measurements on single objects replaced with measurements on different objects. The experimenter selects I samples of JK objects each, which are assumed to be homogeneous with respect to the measurement under study. Each of J operators measures K of these objects once. The data are modeled as

$$y_{ijk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+\epsilon_{ijk},$$

with $\alpha_i \sim \mathcal{N}(0,\sigma_P^2)$ the random sample effects, $\beta_j \sim \mathcal{N}(0,\sigma_O^2)$ the random operator effects, $(\alpha\beta)_{ij} \sim \mathcal{N}(0,\sigma_{PO}^2)$ the sample-operator interaction, and $\epsilon_{ijk} \sim \mathcal{N}(0,\sigma_e^2)$ the error. The effect of the JK objects of a sample is nested within the operators factor, which might explain why many Six Sigma courses refer to this design as being nested. The name is somewhat deceptive because the samples factor is still crossed with the operators factor, whereas the objects factor—though it is indeed nested—does not show up in the model.

The ANOVA analysis is similar to the one in Table 1, but Objects should be replaced with Samples. Again, the experimenter can estimate σ_O^2, σ_{PO}^2, and σ_e^2. If (e) or (f) holds only by approximation, σ_e^2 will be overestimated (assuming that the experimenter has appropriately randomized the order of the measurements).

Gauge R&R Study Under Assumption (g)

The experimenter could use historical estimates for the pattern f. This option applies when the pattern is constant in time. In example II, for example, it could be the case that there are known and fixed differences in the strengths of biscuits that are taken from different positions of the oven belt. In I time instants, the experimenter can select JK biscuits from different positions. Each of J operators measures K of these biscuits. The measurements are modeled as

$$y_{ijk}-\gamma_{jk}=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+\epsilon_{ijk},$$

where γ_{jk} is the assumedly known difference in strength (compared with the overall mean) of biscuits from position j,k. As before, $\beta_j \sim \mathcal{N}(0,\sigma_O^2)$ and $\epsilon_{ijk} \sim \mathcal{N}(0,\sigma_e^2)$. Instead of object effects, we now have the sample effects, $\alpha_i \sim \mathcal{N}(0,\sigma_P^2)$, and the operator-sample interaction effects, $(\alpha\beta)_{ij} \sim \mathcal{N}(0,\sigma_{PO}^2)$. Using the mean squares from the analysis in Table 1, the experimenter can estimate σ_O^2, σ_{PO}^2 and σ_e^2, and thus the total measurement spread $\sqrt{\sigma_O^2+\sigma_{PO}^2+\sigma_e^2}$. Within-sample variation among biscuits that is not accounted for by the γ_{jk} is confounded with σ_e^2.

The other option is to estimate the pattern f from the same data that are used to estimate measurement variability. Depending on the pattern that is considered, the design should be tailored. We provide two examples.

Suppose that we can take objects from different positions and that we assume that there are fixed differences among these positions. A possible experimental design for $I=6$ samples and for $J=6$ operators is found in Table 2. The design is based on the 6×6 Latin-square design. Each sample consists of objects taken from positions $k=1,2,\ldots 6$.

The measurement of an object from position k and in sample i by operator j is denoted y_{ijk} (note that, given i and j, k is completely determined by Table 2). We consider the following model:

$$y_{ijk}=\mu+\alpha_i+\beta_j+\gamma_k+\epsilon_{ijk}. \qquad (5)$$

The sample and operator effects α_i and β_j are considered random. The fixed differences among positions

TABLE 2. Latin-Square Design
(entries indicate positions $k \in \{1, \ldots, 6\}$)

	Operator					
Sample	1	2	3	4	5	6
1	1	4	3	5	2	6
2	2	1	5	3	6	4
3	3	5	4	6	1	2
4	4	3	6	2	5	1
5	6	2	1	4	3	5
6	5	6	2	1	4	3

$k = 1, 2, \ldots, 6$ are modeled in the γ_k. The ANOVA analysis follows the template of Table 3. Models for this type of design are always additive (i.e., do not contain interaction terms; see Montgomery (1997, Section 5-2)). As a consequence, this approach is only suitable if it can be assumed that there is no interaction effect for samples and operators. More advanced designs allow inclusion of an interaction term, but the complexity of the corresponding analyses brings these beyond the scope of this article. Taking linear combinations of mean squares, Table 3 allows estimation of σ_O^2 and σ_e^2.

As a second example, we study example II (strength of biscuits). We assume that the strength of consecutive biscuits increases or decreases linearly, at least when we consider just a brief period. We hope that this pattern explains a major part of the variation in strength between successive biscuits. For simplicity of the example, we only estimate the total measurement spread, σ, here, but we shall indicate how to extend the approach to enable estimation of the measurement spread components, σ_O^2, σ_{OP}^2, and σ_e^2.

The experimenter selects $I = 6$ samples of $J = 6$ consecutive biscuits each (see Table 4). The data are modeled using the ANCOVA model:

$$y_{ij} = \mu + \alpha_i + \beta_i(j - (J+1)/2) + \epsilon_{ij}.$$

TABLE 4. Strength of Biscuits

	Serial Number					
Sample	1	2	3	4	5	6
1	11.0	11.0	11.1	11.2	11.1	11.4
2	9.4	9.5	9.6	9.6	9.9	10.4
3	8.5	8.8	9.1	9.3	9.9	9.4
4	10.3	10.1	10.0	10.4	10.6	10.5
5	9.7	9.7	9.7	9.8	10.3	10.0
6	9.0	9.1	9.5	9.6	9.5	9.9

The α_i denote the sample effects. The slope of the linear trend within a sample i is represented by β_i. The variance σ^2 of the ϵ_{ij} now represents total measurement variation. Standard analysis of covariance (Montgomery (1997, Section 4-7) and Horton (1978)) results in the ANCOVA table (Table 5).

We use the error mean square to estimate measurement spread. It is calculated in ANCOVA models as

$$MS_{\text{error}} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y}_{i\cdot})^2 - \sum_{i=1}^{I}\hat{\beta}_i^2 \sum_{j=1}^{J}\left(j - \frac{J+1}{2}\right)^2}{I(n-1) - J},$$

with

$$\hat{\beta}_i = \frac{\sum_{j=1}^{J}\left(j - \frac{J+1}{2}\right)(y_{ij} - \bar{y}_{i\cdot})}{\sum_{j=1}^{J}\left(j - \frac{J+1}{2}\right)^2}.$$

Measurement error is estimated as $\sigma = \sqrt{0.0326} = 0.18$, which is a pessimistic estimate because MS_{error} confounds measurement error and biscuit-to-biscuit variation that is not explained by the linear trend. Had we not attributed a major part of the within-

TABLE 3. ANOVA Table for Latin-Square Design. Note that $I = J = K = p$

Samples	$p-1$	$p\sum_i(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2$	$\sigma_e^2 + p\sigma_P^2$
Operator	$p-1$	$p\sum_j(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2$	$\sigma_e^2 + p\sigma_O^2$
Position	$p-1$	$p\sum_k(\bar{y}_{\cdot\cdot k} - \bar{y}_{\cdots})^2$	$\sigma_e^2 + [p/(p-1)]\sum_k \gamma_k^2$
Error	$(p-2)(p-1)$	$\sum_{i,j,k}(y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot k} + 2\bar{y}_{\cdots})^2$	σ_e^2

TABLE 5. ANCOVA Table

Source	df	Adj SS	Adj MS	F	P
Trend	1	1.9612	1.9612	60.08	0.000
Sample	5	4.6887	0.9377	28.73	0.000
Sample × trend	5	0.3504	0.0701	2.15	0.094
Error	24	0.7834	0.0326		

sample variation to the linear trends, the model would have reduced to a one-factor ANOVA. Estimating measurement error as the square root of the error mean square, we would have obtained $\sigma = 0.32$.

This approach can be extended to incorporate operator and operator × sample interaction effects in the analysis. Assuming a number of $K = 3$ operators, the experimenter randomly assigns two biscuits out of each sample to each operator. The model becomes

$$y_{ijk} = \mu + \alpha_i + \gamma_k + (\alpha\gamma)_{ik} + \beta_i(j - (J+1)/2) + \epsilon_{ijk}.$$

The γ_k and $(\alpha\gamma)_{ik}$ are the operator and operator × sample interaction effects. The terms in the model can be estimated following an ANCOVA approach; suitable combinations of mean squares allow estimation of $\sigma_O^2, \sigma_{PO}^2$, and σ_e^2. This analysis is not straightforward, however, and for that reason is beyond the scope of this article.

Conclusion

In order to estimate measurement spread, an experimenter needs multiple measurements of a single object. In many situations, this can be done by making the assumptions that objects are invariable in time and that measuring does not affect them. When these assumptions do not hold, as is the case with *destructive measurements*, measurement variation is confounded with other sources of variation. The experimenter can obtain a good estimate of measurement spread if he can exploit certain forms of homogeneity. In the article, various examples of such homogeneity assumptions were introduced (assumptions (c) through (k); see Table 6). They are based on the idea that either the effects of disturbing sources of variation are negligible or that the results can be corrected for their influence.

Given that the homogeneity assumptions that the experimenter makes will only be met to a certain extent, the confounding problem is not solved entirely.

TABLE 6. Overview of Approaches and Assumptions Developed in This Article

Standard Assumptions		
a	Constancy of bias	
b	Homogeneity of measurement error	
c	Temporal stability of objects	
d	Robustness against measurement	
Approaches for Destructive Measurements		**Additional Assumptions**
e	Homogeneity of objects Within small samples, object-to-object variation is negligible	a,b,c
f	Representativeness of alternative objects There are alternative objects for which object-to-object variation is negligible	a,b,c,e
g	Patterned object variation The object-to-object variation can be modeled and thus corrected for	a,b,c
h	Alternative, nondestructive measurement system There is an alternative measurement system that is nondestructive	a,b,c
i	Alternative perfect, destructive measurement system There is an alternative measurement system that is destructive but has virtually no measurement spread	a,b,c
j	Patterned temporal variation The temporal variation can be modeled and thus corrected for	a,b
k	Known true values There are reference objects with known true values	a,b,f

As a consequence, measurement spread will be overestimated. The better the experimenter succeeds in arranging his gauge R&R experiment such that conditions are homogeneous, the smaller this overestimation is.

Although suggested sometimes otherwise in literature (and in spite of expectations that practitioners often have), there is not a statistical trick that solves the confounding problem when no homogeneity assumptions can be justified: there is a strain between destructive measurements and the basic principles of gauge R&R studies.

References

AIAG (2002). *Measurement System Analysis; Reference Manual, 3rd ed.* Automotive Industry Action Group, Detroit, MI.

Allen, M. J. and Yen, W. M. (1979). *Introduction to Measurement Theory.* Wadsworth, Belmont, CA.

Bergeret, F.; Maubert, S.; Sourd, P.; and Puel, F. (2001). "Improving and Applying Destructive Gauge Capability". *Quality Engineering* 14(1), pp. 59–66.

Box, G. E. P. and Liu, P. Y. T. (1999). "Statistics as a Catalyst to Learning by Scientific Method; Part I—An example". *Journal of Quality Technology* 31(1), pp. 1–15.

Burdick, R. K.; Borror, C. M.; and Montgomery, D. C. (2003). "A Review of Methods for Measurement Systems Capability Analysis". *Journal of Quality Technology* 35(4), pp. 342–354.

Horton, R. L. (1978). *The General Linear Model.* McGraw-Hill, New York.

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Addison-Wesley, Reading, MA.

Montgomery, D. C. (1997). *Design and Analysis of Experiments, 4th ed.* Wiley, New York.

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(2), pp. 283–290.

Wallsten, T. S. (1988). "Measurement Theory". *Encyclopedia of Statistical Sciences, 8th ed.,* Kotz, S.; and Johnson, N. L. (eds.). Vol. 5. Wiley, New York.