Evaluation of Measurement Systems with a Small Number of Observers

Edwin R. van den Heuvel* and Albert Trip

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

ABSTRACT

The standard R&R study for evaluation of measurement systems assumes that participating observers constitute a random sample. Often there are only a few observers, all of them included in the study. An alternative measure for the gauge R&R is proposed for this situation, and it is shown that this may improve the perception of the quality of the measurement system markedly, especially with only a few observers. Finally it is shown that a simple estimator can be used, with a bias limited to just a few percent.

Key Words: Analysis of variance; Gauge R&R study; Mixed effects models; Prediction intervals; Random effects models

INTRODUCTION

Within quality programs, such as statistical process control (SPC) and Six-Sigma, it is important to evaluate the measurement system or method. When a good measurement system is in place, the measurements of a quality characteristic are precise, and therefore the characteristic may be controlled and the variation may be reduced. The usual way to investigate the variability or precision of the measurement system is to conduct a well-described experiment. The variability is often divided into two components; the first caused by observers (or operators), and the other by the measurement device (the “gauge”) itself. Since these components are called “Reproducibility” and “Repeatability,” respectively, the experiment is also known as an R&R study.

The variances of the different components are estimated with linear combinations of the mean squares...
from the analysis of variance (ANOVA) table. The combined standard deviation of all components is the variability of the measurement system, which is compared to the width of the lower and the upper specification limit for the quality characteristic (the tolerance width). Criteria of the Automotive Industry Action Group (AIAG\(^{[1]}\)) use the ratio of the two, to indicate whether the measurement system is suitable for its task. Improvements of the measurement system should be made if the measurement variation is too high; the individual standard deviations of the different components indicate which sources of variation should be improved.

Evaluation of measurement systems is usually based upon the following model:

\[
y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk}, \quad i = 1, 2, \ldots, n;
\]

\[
j = 1, 2, \ldots, m; \quad k = 1, 2, \ldots, r,
\]

where \( n \) is the number of products, \( m \) the number of observers, and \( r \) the number of replications. The value \( y_{ijk} \) is replication \( k \) of the measurement on product \( i \), done by observer \( j \). The symbols \( \mu \), \( \alpha_i \), \( \beta_j \), \( \gamma_k \), and \( \epsilon_{ijk} \) represent the true value of the process, the effect of the products, the effect of the observers, the interaction between observers and products, and the repeatability, respectively. In the remainder of this paper we will concentrate on continuous, nondestructive measurements.

If the effects of products and observers, the interaction of the two, and the repeatability are modeled with a normal distribution with variances \( \sigma^2_{\alpha} \), \( \sigma^2_{\beta} \), \( \sigma^2_{\gamma} \), and \( \sigma^2_{\epsilon} \), respectively, and zero means, then model (1) is called the random effects model (REM). In fact, the REM is the basis on which standard R&R studies are built. However, this model makes sense only if products and observers are drawn from a large population, and if the underlying distributions of the populations are approximately normal.

In real life the assumption of random observers is often not true. In smaller industries it is not unusual that all possible observers—a limited number—are contributing in the R&R study. Then every organization will have measurements requiring expertise that just a few observers possess. A common example refers to laboratory measurements for off-line quality control purposes or inspection, generally done by skilled analysts. Our case illustration—later on in this article—is an example of this type of measurement. In such situations, it is obviously wrong to model the observer effect as a random component representing a larger population. Instead, the assumption that each observer measures a fixed amount from the true value makes more sense. The modification of model (1) in this way is called the mixed effects model (MEM), since it involves both random and fixed effects. The relation between both models is that the MEM converges to the REM (in probability) if the number of observers is large, and the fixed observer deviations stem from a normal distribution.

The MEM was studied by Dolezal et al.\(^{[2]}\) and they proved that confidence intervals for the MEM are narrower than those for the REM. However, the usual mean square estimates are exactly the same for both models. In this article, we show that the method of estimating measurement variability may be improved for the MEM. This is based on the observation that a 99% prediction interval for a measurement of a randomly selected product assigned randomly to an observer is different for both models. Note that it is the 99% prediction interval for the measurement error in AIAG\(^{[1]}\) that is compared with the tolerance width to evaluate the capability of the measurement system for the REM. Unfortunately, a 99% prediction interval for the MEM cannot be established exactly, but a simple larger interval serves the purpose well. This larger interval is compared with the 99% prediction interval of the REM. Furthermore, an estimator of the larger prediction interval is proposed and evaluated through its bias. A case from Philips Semiconductors illustrates the differences between the two models. Finally, guidelines for use of our estimation method conclude the paper.

## GAUGE R&R WITH FIXED OBSERVER EFFECTS

Model (1) with all effects independent and normally distributed (the REM) is the basis for investigating and judging measurement systems, see AIAG\(^{[1]}\). The variance of the measurement system is defined as:

\[
\sigma^2_m = \sigma^2_{\alpha} + \sigma^2_{\gamma} + \sigma^2_{\epsilon}.
\]

The measurement variation consists of the repeatability of the system (\( \sigma^2_{\epsilon} \)) and the reproducibility (\( \sigma^2_{\alpha} + \sigma^2_{\gamma} \)). According to AIAG\(^{[1]}\) a measurement system is only acceptable if the so-called “gauge R&R” (defined as 5.15\( \sigma_m \)) does not cover more than 30% of the tolerance width. This criterion is used irrespective of the model. Engel and De Vries\(^{[3]}\) related the AIAG criterion to probabilities of incorrect decisions, and concluded that
the process capability should be used as well when deciding about the suitability of the measurement system. Burdick and Larsen\textsuperscript{[1]} presented confidence intervals for a.o. $\sigma^2_b$, $\sigma^2_b + \sigma^2_g$, and $\sigma^2_b + \sigma^2_g + \sigma^2_l$. They showed that the number of observers has a major impact on the length of the confidence intervals for any variance measure that includes $\sigma_b$, and concluded that an increase in the number of observers is preferred over increased replication. Vardeman and Van Valkenburg\textsuperscript{[5]} also discussed the REM in the context of R&R studies; among other things they investigated were optimal sizes among other things they investigated were optimal sizes for the numbers $n$, $m$, and $r$, as well as estimates of the uncertainty of the interesting statistics.

In the REM, where all effects are normal, the gauge R&R is equivalent to the width of a 99% prediction interval of a measurement. When some assumptions in model (1) are not true, however, the meaning of $5.15\sigma_m$ may be completely different and its relation to the gauge R&R is not clear anymore. But a 99% prediction interval is clear, whatever the assumptions in model (1). In fact, the method of AIAG\textsuperscript{[1]} to evaluate the measurement precision for attribute data is also based on the concept of a 99% prediction interval.

Assume in model (1) that the effects $b_j$ of observers $(j = 1, \ldots, m)$ are fixed (the MEM). For reasons of identifiability, one must further assume that $b_1 + b_2 + \cdots + b_m = 0$. Now given the value $\alpha$ of a randomly selected product, the cumulative distribution function of a measurement of that product (assuming that all observers measure with equal frequency) is equal to

\[ F_\alpha(y) = \frac{1}{m} \sum_{j=1}^{m} \Phi\left(\frac{y - \mu - \alpha - b_j}{\sigma}\right) \]  

(3)

with $\sigma^2 = \sigma^2_b + \sigma^2_e$ and $\Phi$ the standard normal distribution function. In addition, if the observer $j$ is given, the distribution function of a measurement of that same product is equal to $\Phi(y - \mu - \alpha - b_j)/\sigma)$. However, assigning products to observers for quality control or inspection is often done randomly and $F_{\alpha}$ in Eq. (3) represents the distribution function of the total measurement error. Indeed, the distribution function in Eq. (3) reduces to a degenerate distribution at $\mu + \alpha$, the true product value without measurement error, if and only if $b_1 = b_2 = \cdots = b_m = 0$, $\sigma_r = 0$, and $\sigma_e = 0$.

A 99% prediction interval $\text{IF} = [\text{LF}, \text{UF}]$ for the MEM is now determined by

\[ F_\alpha(\text{LF}) = 0.005 \]

\[ F_\alpha(\text{UF}) = 0.995. \]  

(4)

Solving $L_F$ and $U_F$ does not give closed expressions, but the width of this interval can be approximated by

\[ \Omega_F = \beta_m - \beta_1 + 5.15\sigma, \]  

(5)

with $\beta_1, \beta_2, \ldots, \beta_m$ the ordered observer effects of $b_1, b_2, \ldots, b_m$. In fact, $\Omega_F$ is an upper bound of the width of IF, with equality if all observer effects are equal (as shown in the Appendix). A lower bound of the width of IF is

\[ \beta_m - \beta_1 + 2\Phi^{-1}\left(1 - m \frac{m}{200}\right)\sigma, \quad m \leq 100, \]  

(6)

with $\Phi^{-1}$ the inverse of the normal distribution function. Equality holds if $\beta_1$ is much smaller and $\beta_m$ is much larger (measured in units of $\sigma$) than all other observer effects. Table 1 shows the lower bound for several numbers of observers.

Note that the lower bound may be substantially smaller than the upper bound $\Omega_F$, especially when $\sigma$ is relatively large. But this is exactly the situation that the width of IF is better approximated by the upper limit, because relatively large $\sigma$ implies that all observer effects are more or less "equal." The upper limit is therefore a realistic approximation of the width of IF.

### Table 1

<table>
<thead>
<tr>
<th>Observers (m)</th>
<th>Lower Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>$\beta_m - \beta_1 + 4.65\sigma$</td>
</tr>
<tr>
<td>3</td>
<td>$\beta_m - \beta_1 + 4.34\sigma$</td>
</tr>
<tr>
<td>4</td>
<td>$\beta_m - \beta_1 + 4.11\sigma$</td>
</tr>
<tr>
<td>5</td>
<td>$\beta_m - \beta_1 + 3.92\sigma$</td>
</tr>
<tr>
<td>10</td>
<td>$\beta_m - \beta_1 + 3.29\sigma$</td>
</tr>
<tr>
<td>20</td>
<td>$\beta_m - \beta_1 + 2.56\sigma$</td>
</tr>
</tbody>
</table>

### COMPARISON WITH RANDOM EFFECTS MODEL

Given the value $\alpha$ of a randomly selected product, the cumulative distribution function of a measurement of that product for the REM is

\[ F_\alpha(y) = \Phi\left(\frac{y - \mu - \alpha}{\sigma_m}\right), \]  

(7)

with $\sigma_m$ defined in Eq. (2). In addition, we may see that the distribution degenerates at the true value $\mu + \alpha$ if and only if $\sigma_p = 0$, $\sigma_r = 0$, and $\sigma_e = 0$. Again, the $F_R$ in Eq. (7) represents the distribution function of the total
measurement error when products are randomly assigned to observers. Further situations are examined.

To compare the widths of the prediction intervals IR and IF, assume that the observer effects \( \beta_1, \beta_2, \ldots, \beta_m \) constitute a random sample from a normal distribution with zero mean and variance \( \sigma^2 \). First note that the weak law of large numbers implies that \( F_R(y) \to F_R(y) \) (in probability) for \( m \to \infty \). Hence, the model with fixed observers looks like the REM whenever the number of observers is large and the observer effects are normally distributed. Only the situation with a few observers will be explored in more detail further.

The width of the REM prediction interval (denoted as \( \Omega_k \)) is compared with the upper bound \( \Omega_F \) of the MEM prediction interval:

\[
\Omega_R = 5.15 \sqrt{\sigma^2} \]

\[
\Omega_F = \beta(m) - \beta(1) + 5.15 \sigma
\]

First, two extreme situations are examined:

1. Observer variation is negligible \( (\sigma_{\beta} \ll \sigma) \): \( \Omega_R \) and \( \Omega_F \) both converge towards \( 5.15 \sigma \), so there is not much difference between the two widths.

2. Observer variation is dominating \( (\sigma_{\beta} \gg \sigma) \): \( \Omega_R \) converges towards \( 5.15 \sigma_{\beta} \) and \( \Omega_F \) converges towards \( \beta(m) - \beta(1) \). Since we assume (for the comparison of \( \Omega_F \) and \( \Omega_R \)) that the \( \beta \)s are samples from a normal distribution, the expectation of \( \beta(m) - \beta(1) \) will be equal to \( d_m \sigma_{\beta} \), with \( d_m \) the expected value of the range of \( m \) independent standard normal variables. For up to 20 observers \( d_m \) ranges from 1.13 to 3.73, and the probability that \( \beta(m) - \beta(1) \) exceeds 5.15 is limited to 0.036. So \( \Omega_R \) exceeds \( \Omega_F \) with high probability.

Now more generally, the expected width of \( \Omega_F \) is

\[
E\Omega_F = d_m \sigma_{\beta} + 5.15 \sigma = (d_m + 5.15 \rho) \sigma_{\beta}
\]

(8)

with \( \rho = \sigma/\sigma_{\beta} \). In Table 2, the ratio of \( E(\Omega_F) \) to \( \Omega_R \) is given for several combinations of \( m \) and \( \rho \).

Table 2 shows that \( \Omega_F \) as a measure for gauge R&R in the model with fixed observers:

<table>
<thead>
<tr>
<th>Values of ( \rho = \sigma/\sigma_{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m )</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>20</td>
</tr>
</tbody>
</table>

1. Will be much smaller (on average) than \( \Omega_R \) if \( \rho \) is small;
2. Will be a fraction higher (on average) than \( \Omega_R \) for large \( \rho \) and/or high \( m \);
3. Will be equal to \( \Omega_R \) if \( \rho \) is very large.

Our conclusion is that serious errors can be made in the determination of the gauge R&R in the model with fixed observers, when the traditional REM is assumed. The proposed alternative, which is a larger value for the width of a 99% prediction interval (or real gauge R&R), is nonetheless in many cases (much) smaller when the observer variation is in fact dominating. The fewer the observers, the stronger this effect. Additionally, the proposed alternative is never much worse than the traditional method.

**ESTIMATION**

In the REM the standard deviation of the measurement system is estimated with the square root of certain linear combinations of mean squares from the ANOVA table. Dolezal et al.\(^3\) and others use this estimate also for models with fixed observer effects, even though it estimates a parameter that may well be utterly wrong. In this section we study an estimator for the most appropriate parameter \( \Omega_F = \beta(m) - \beta(1) + 5.15 \sigma \) of the MEM that quantifies the measurement variability.

An obvious estimator of \( \Omega_F \) is \( W_F \), defined by

\[
W_F = B(m) - B(1) + 5.15 \sigma,
\]

(9)

with

\[
B_j = \frac{1}{ni} \sum_{i=1}^n \sum_{k=1}^r y_{ijk},
\]

\( B_1, B_2, \ldots, B_m \), the ordered sample of \( B_1, B_2, \ldots, B_m \), and \( s \) the usual estimate for \( \sigma \) based on the mean squares.
To study the moments of $\Omega_F$, one needs knowledge of the moments of $B_{(m)} - B_{(1)}$. This leads to a study of the range of independent normally distributed random variables $X_i$ with mean value $\beta_i$ and variance $\tau^2 = \sigma^2/n + \sigma_e^2/nr$. For three situations the expectation can be evaluated in a closed form.

1. Negligible $\sigma_e$ and $\sigma_e/\tau = 0$. Then $B_{(m)} - B_{(1)}$ is an unbiased estimator of $\beta_{(m)} - \beta_{(1)}$, because $B_j = \mu + \beta_j$, with $\beta$ the average of the $\alpha$s.

2. No observer effect ($\beta_{(m)} - \beta_{(1)} = 0$). In this case $B_{(m)} - B_{(1)}$ has a positive bias:

$$E(B_{(m)} - B_{(1)}) = d_m\tau,$$

(10)

with $d_m$ as before the expected value of the range of $m$ independent standard normal variables.

Equation (10) indicates the particular sensitivity of the estimator $W_F$ through the interaction component $\sigma_e^2$, since this component is leading in $\tau^2$ (for $\tau > 1$).

3. Two observers ($m = 2$). The random variable $B_{(2)} - B_{(1)}$ equals $|B_1 - B_2|$ and $B_1 - B_2$ is normally distributed with mean $\nu = \beta_1 - \beta_2$ and variance $2\tau^2$. The bias is

$$E(B_{(2)} - B_{(1)}) - |\nu| = -2|\nu|\Phi\left(-\frac{|\nu|}{\tau\sqrt{2}}\right) + 2\tau\sqrt{2}\phi\left(-\frac{|\nu|}{\tau\sqrt{2}}\right);$$

(11)

with $\phi$ the standard normal density function. This equation is determined in the Appendix. It is a positive, monotone decreasing function of $|\nu|$ with a maximum of $2\tau/\sqrt{\pi}$ at $\nu = 0$. This follows from the observation that the derivative of the bias with respect to $|\nu|$ is negative for $|\nu| > 0$.

The three cases above illustrate that if the relevant observer effects (the extremes $\beta_{(1)}$ and $\beta_{(m)}$) cannot be distinguished properly from the other effects—in terms of $\tau$—there will be a positive bias. The less clear the distinction, the larger the bias.

Another issue in the evaluation of $W_F$ is the estimation of $\sigma$. The standard AIAG approach uses ranges—in an incorrect way, as Vardeman and VanValkenburg point out. In this paper we prefer the use of the appropriate linear combinations of the mean squares from the ANOVA table, based on model (1). First of all for its transparency, but also for its efficiency.

Assume that $s$ is an estimator of $\sigma$, and $fs^2/\sigma^2$ has a $\chi^2$-distribution with $f$ degrees of freedom. Then the expectation of $s$ is $E_s = c_4\sigma$, with

$$c_4 = \sqrt{\frac{2\Gamma(f + 1/2)}{\Gamma(f/2)}}. \quad (12)$$

Note that $c_4$ is smaller than 1, in fact $c_4 \approx 1 - 1/(4f)$, which implies that $s$ estimates $\sigma$ with a negative bias.

For model (1), an unbiased estimate of $\sigma^2 = \sigma_g^2 + \sigma_e^2$ is based on the usual linear combination of two mean squares, but this estimator has not a simple $\chi^2$-distribution. An explicit formula for the correction factor $c_4$ is therefore unknown in this particular case, but it remains smaller than 1 which could be observed by applying Jensen’s inequality.

A combination of the two observations that $B_{(m)} - B_{(1)}$ has a positive bias in estimating $\beta_{(m)} - \beta_{(1)}$ and $s$, based on the usual linear combinations of the mean squares, has a negative bias in estimating $\sigma$ made us curious about the combination of the two in estimating $\Omega_F$. The remainder of this section is therefore committed to simulation results. The following parameter settings were examined:

$$\sigma_g = 0.5, 1.0, 2.0; \sigma_e = 0.1, 0.5, 1.0, 2.0; \sigma_e = 0.1, 0.5, 1.0, 2.0, \text{ and } m = 3, 5, 10.$$ 

The parameters $n = 10$ and $r = 3$ were fixed. For each combination of the parameter settings 10 random configurations of observer effects ($\beta_s$) were examined with 5000 “experiments.” Not every possible combination of the settings was examined, however, as may be seen in Table 3.

The results of the simulations indicate that $W_F$ is on average a little larger than $\Omega_F$. Within the examined region the average bias ranges from $-0.5$ to $7.4\%$ (the average taken over the 10 separate $\beta$-configurations). The average bias belonging to a specific $\beta$-configuration ranges from $-1.4$ to $10.2\%$. The bias is large when $m$ is large, and at the same time $\sigma_\beta$ is small in comparison to (especially) $\sigma_\gamma$. A negative bias may be expected when $m$ is small, and $\sigma_\beta$ is large compared to $\sigma_e$. These results can be understood when the estimator $W_F$ is examined more closely.

1. $B_{(m)} - B_{(1)}$ converges to $d_m\tau$ [Eq. (10)] when $\sigma_\beta << \sigma_\gamma$; this quantity increases with $m$. But the degrees of freedom for estimating $\sigma_\gamma$ and $\sigma_e$ increase when $m$ increases, so the bias in $s$ tends to disappear.

2. $B_{(m)} - B_{(1)}$ will have a small bias in estimating $\beta_{(m)} - \beta_{(1)}$ when $\sigma_\beta >> \sigma_e$ and $m$ is small (and $\sigma_\gamma$ not larger than $\sigma_\beta$). But for estimating $\sigma = \sigma_\gamma$ very few degrees of freedom are available, so $s$ will have a negative bias.
parameter settings. Results of additional simulation of the extremes are that: (1) with $s_b = 0$, $s_g = s_1 = 1$, and $m = 10$ the bias amounts to 15%; and (2) the bias is about $-2\%$ when $m = 2$, $s_b = 2$, $s_g = 2$, and $s_1 = 0$.

From this section we conclude that $W_F$ is a good estimator of $V_F$, because the bias is in general limited to (say) 2–3%. Many observers with small effects lead to the largest bias, especially when observer–product interactions are large. With only a few observers and large observer effects, the estimator $W_F$ will slightly underestimate the parameter $V_F$. The reader may recall from the earlier discussion about the width of the MEM prediction interval IF, however, that in this very situation (observer effects clearly separated) the lower bound $(6)$ is a better approximation than the upper bound $V_F$. The damage of a negative bias of $W_F$ in estimating $V_F$ is therefore even more limited as far as gauge R&R is concerned.

### CASE ILLUSTRATION

Philips Semiconductors Stadskanaal (The Netherlands) is a QS 9000 certified manufacturer of diodes. For investigations of physical and chemical nature beyond the possibilities of operators on the shop floor, it has an analytical laboratory. To characterize the measurement variability of an advanced optical microscope, the following R&R study was done by the two analysts. Both measured a single product three times in a row at five different spots, seven days long. The location of the spots was roughly defined, as good as could be, nonetheless subject to variation. The quality characteristic is the distance from the edge of a semiconductor crystal to the epitaxial layer (in mm). The measurements are given in Table 4.

The variation in the data can be analyzed with a slight modification of model (1).

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \delta_{ijk} + \epsilon_{ijkl}$$

spot $i = 1, 2, 3, 4, 5$

analyst $j = 1, 2$

day $k = 1, 2, 3, 4, 5, 6, 7$

repetition $l = 1, 2, 3$

### Table 3

Simulation Results of Bias of $W_F$

<table>
<thead>
<tr>
<th>$\sigma_\beta$</th>
<th>$\sigma_\gamma$</th>
<th>$\sigma_s$</th>
<th>$m$</th>
<th>% Bias</th>
<th>$\sigma_\beta$</th>
<th>$\sigma_\gamma$</th>
<th>$\sigma_s$</th>
<th>$m$</th>
<th>% Bias</th>
<th>$\sigma_\beta$</th>
<th>$\sigma_\gamma$</th>
<th>$\sigma_s$</th>
<th>$m$</th>
<th>% Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
<td>3</td>
<td>1.06</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>3</td>
<td>-0.11</td>
<td>2</td>
<td>0.5</td>
<td>0.1</td>
<td>3</td>
<td>-0.52</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
<td>5</td>
<td>1.61</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>5</td>
<td>0.56</td>
<td>2</td>
<td>0.5</td>
<td>0.1</td>
<td>5</td>
<td>0.01</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
<td>10</td>
<td>1.77</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>10</td>
<td>1.15</td>
<td>2</td>
<td>0.5</td>
<td>0.1</td>
<td>10</td>
<td>0.33</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>2</td>
<td>3</td>
<td>1.52</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>3</td>
<td>0.06</td>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>3</td>
<td>0.02</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>2</td>
<td>5</td>
<td>2.77</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>5</td>
<td>0.54</td>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>5</td>
<td>0.05</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>2</td>
<td>10</td>
<td>3.86</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>10</td>
<td>1.09</td>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>10</td>
<td>0.09</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>3</td>
<td>1</td>
<td>0.77</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>3</td>
<td>0.09</td>
<td>2</td>
<td>0.5</td>
<td>1</td>
<td>3</td>
<td>-0.01</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>3</td>
<td>5</td>
<td>2.89</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>5</td>
<td>1.83</td>
<td>2</td>
<td>0.5</td>
<td>1</td>
<td>5</td>
<td>0.08</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>3</td>
<td>10</td>
<td>4.37</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>10</td>
<td>1.73</td>
<td>2</td>
<td>0.5</td>
<td>1</td>
<td>10</td>
<td>0.07</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
<td>3</td>
<td>1.58</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0.61</td>
<td>2</td>
<td>0.5</td>
<td>2</td>
<td>3</td>
<td>0.20</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2.68</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1.72</td>
<td>2</td>
<td>0.5</td>
<td>2</td>
<td>5</td>
<td>0.81</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>4.68</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>1.36</td>
<td>2</td>
<td>0.5</td>
<td>2</td>
<td>10</td>
<td>0.43</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2.23</td>
<td>2</td>
<td>0.1</td>
<td>0.1</td>
<td>3</td>
<td>-0.17</td>
<td>2</td>
<td>1</td>
<td>0.1</td>
<td>3</td>
<td>-0.53</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>3.89</td>
<td>2</td>
<td>0.1</td>
<td>0.1</td>
<td>5</td>
<td>0.02</td>
<td>2</td>
<td>1</td>
<td>0.1</td>
<td>5</td>
<td>0.50</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>3.55</td>
<td>2</td>
<td>0.1</td>
<td>0.1</td>
<td>10</td>
<td>-0.01</td>
<td>2</td>
<td>1</td>
<td>0.1</td>
<td>10</td>
<td>0.92</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
<td>0.5</td>
<td>3</td>
<td>1.99</td>
<td>2</td>
<td>0.1</td>
<td>0.5</td>
<td>3</td>
<td>0.09</td>
<td>2</td>
<td>1</td>
<td>0.5</td>
<td>3</td>
<td>0.00</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
<td>0.5</td>
<td>5</td>
<td>4.51</td>
<td>2</td>
<td>0.1</td>
<td>0.5</td>
<td>5</td>
<td>0.02</td>
<td>2</td>
<td>1</td>
<td>0.5</td>
<td>5</td>
<td>0.39</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
<td>0.5</td>
<td>10</td>
<td>7.40</td>
<td>2</td>
<td>0.1</td>
<td>0.5</td>
<td>10</td>
<td>0.06</td>
<td>2</td>
<td>1</td>
<td>0.5</td>
<td>10</td>
<td>0.60</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2.00</td>
<td>2</td>
<td>0.1</td>
<td>1</td>
<td>3</td>
<td>0.28</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0.15</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>5.78</td>
<td>2</td>
<td>0.1</td>
<td>1</td>
<td>5</td>
<td>0.35</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>0.66</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>6.96</td>
<td>2</td>
<td>0.1</td>
<td>1</td>
<td>10</td>
<td>0.32</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>0.77</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3.20</td>
<td>2</td>
<td>0.1</td>
<td>2</td>
<td>3</td>
<td>0.85</td>
<td>2</td>
<td>2</td>
<td>0.1</td>
<td>3</td>
<td>-0.25</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>4.72</td>
<td>2</td>
<td>0.1</td>
<td>2</td>
<td>5</td>
<td>0.46</td>
<td>2</td>
<td>2</td>
<td>0.1</td>
<td>5</td>
<td>1.17</td>
</tr>
<tr>
<td>0.5</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>6.78</td>
<td>2</td>
<td>0.1</td>
<td>2</td>
<td>10</td>
<td>0.93</td>
<td>2</td>
<td>2</td>
<td>0.1</td>
<td>10</td>
<td>1.54</td>
</tr>
</tbody>
</table>
with \(\alpha, \beta, \gamma, \delta,\) and \(\varepsilon\) representing the effects of spot, analyst, spot–analyst interaction, day, and repetition, respectively. The ANOVA is shown in Table 5.

Neither the spot effect nor the spot–analyst interaction is significant. Both \(s_a\) and \(s_g\) can therefore be estimated with 0. From Table 5 the REM estimates can be computed using the EMS:

\[
\sigma^2_s = 0.266, \quad \sigma^2_b = 0.338, \quad \sigma^2_\beta = 1.047.
\]

The estimate of the measurement variation \(\sigma^2_m = \sigma^2_s + \sigma^2_b + \sigma^2_\varepsilon\) is now 1.651. Hence, the estimated gauge R&R for the REM is 6.617 (5.15 \(s\)).

However, what we have here is a typical example of a situation with fixed observers. The two analysts of the laboratory are the only ones who use the microscope. The observer effects \(\beta_s\) in model (13) are therefore fixed, and \(\Omega_F\) in Eq. (9) is a better estimate of the gauge R&R. In this case the averages of both analysts are \(\hat{\beta}_1 = 57.467\) and \(\hat{\beta}_2 = 56.011\), respectively. The estimate for the random variation is \(\sigma^2 = \sigma^2_s + \sigma^2_b + \sigma^2_\varepsilon = 0.604\) using

Table 5. The estimated gauge R&R for the MEM is now

\[
\hat{\beta}_1 - \hat{\beta}_2 + 5.15\sqrt{\bar{\sigma}} = 57.467 - 56.011 + 5.15\sqrt{0.604} = 5.458.
\]

This is 18% smaller than the “standard” (REM) gauge R&R. Note that the lower bound in Eq. (6) for the gauge R&R with two analysts is estimated by \(\hat{\beta}_1 - \hat{\beta}_2 + 4.65\sqrt{\bar{\sigma}} = 5.070\). The measurement system (equipment and analysts) is better than the standard method would indicate.

### CONCLUSIONS AND RECOMMENDATIONS

This paper deals with analysis of measurement systems when only few observers perform the measurements. We showed that the standard model with random observer effects is not correct and may give misleading

### Table 4

**Distances of the Optical Microscope for the R&R Study**

<table>
<thead>
<tr>
<th></th>
<th>Analyst 1</th>
<th></th>
<th>Analyst 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spot 1</td>
<td>Spot 2</td>
<td>Spot 3</td>
<td>Spot 4</td>
</tr>
<tr>
<td>Day 1</td>
<td>60.0</td>
<td>57.5</td>
<td>58.7</td>
<td>59.0</td>
</tr>
<tr>
<td></td>
<td>59.5</td>
<td>57.7</td>
<td>59.5</td>
<td>59.0</td>
</tr>
<tr>
<td>Day 2</td>
<td>57.2</td>
<td>57.0</td>
<td>58.2</td>
<td>58.2</td>
</tr>
<tr>
<td></td>
<td>57.2</td>
<td>57.5</td>
<td>58.0</td>
<td>56.5</td>
</tr>
<tr>
<td></td>
<td>56.0</td>
<td>56.7</td>
<td>58.2</td>
<td>57.5</td>
</tr>
<tr>
<td>Day 3</td>
<td>58.7</td>
<td>57.7</td>
<td>57.2</td>
<td>57.7</td>
</tr>
<tr>
<td></td>
<td>58.0</td>
<td>57.5</td>
<td>57.0</td>
<td>57.0</td>
</tr>
<tr>
<td>Day 4</td>
<td>57.7</td>
<td>57.2</td>
<td>58.0</td>
<td>57.2</td>
</tr>
<tr>
<td></td>
<td>57.5</td>
<td>57.2</td>
<td>57.5</td>
<td>57.5</td>
</tr>
<tr>
<td></td>
<td>57.5</td>
<td>56.5</td>
<td>56.7</td>
<td>57.2</td>
</tr>
<tr>
<td>Day 5</td>
<td>58.2</td>
<td>58.2</td>
<td>57.5</td>
<td>57.2</td>
</tr>
<tr>
<td></td>
<td>56.7</td>
<td>57.2</td>
<td>56.2</td>
<td>57.2</td>
</tr>
<tr>
<td>Day 6</td>
<td>56.7</td>
<td>57.7</td>
<td>56.7</td>
<td>57.2</td>
</tr>
<tr>
<td></td>
<td>56.5</td>
<td>57.7</td>
<td>56.5</td>
<td>56.0</td>
</tr>
<tr>
<td></td>
<td>56.2</td>
<td>58.0</td>
<td>57.7</td>
<td>56.5</td>
</tr>
<tr>
<td>Day 7</td>
<td>58.5</td>
<td>59.0</td>
<td>58.7</td>
<td>57.7</td>
</tr>
<tr>
<td></td>
<td>57.7</td>
<td>58.0</td>
<td>58.0</td>
<td>55.7</td>
</tr>
<tr>
<td></td>
<td>57.5</td>
<td>57.5</td>
<td>56.0</td>
<td>56.5</td>
</tr>
</tbody>
</table>
Table 5
The ANOVA Table of the R&R Study

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spot</td>
<td>4.377</td>
<td>4</td>
<td>1.094</td>
<td>42\sigma_s^2 + 21\sigma_g^2 + 3\sigma_d^2 + \sigma_e</td>
</tr>
<tr>
<td>Analyst</td>
<td>111.180</td>
<td>1</td>
<td>111.180</td>
<td>105\sigma_s^2 + 21\sigma_g^2 + 3\sigma_d^2 + \sigma_e</td>
</tr>
<tr>
<td>Spot x Analyst</td>
<td>5.145</td>
<td>4</td>
<td>1.286</td>
<td>21\sigma_g^2 + 3\sigma_d^2 + \sigma_e</td>
</tr>
<tr>
<td>Day</td>
<td>76.851</td>
<td>60</td>
<td>1.281</td>
<td>3\sigma_d^2 + \sigma_e</td>
</tr>
<tr>
<td>Repetition</td>
<td>37.187</td>
<td>140</td>
<td>0.266</td>
<td>\sigma_e</td>
</tr>
<tr>
<td>Total</td>
<td>234.740</td>
<td>209</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

results, especially when the number of observers is small. We proposed to use an alternative measure for the gauge R&R, based on the width of a 99% prediction interval. This distance was approximated by the parameter $R_R$, based on the width of a 99% prediction interval. The estimator $W_F$ of $\Omega_F$, which turned out to be an overestimation of the newly defined gauge R&R. Nonetheless, $\Omega_F$ is dramatically smaller than the random model gauge R&R ($\Omega_R$), when the number of observers is small, and their effects are large (Table 2). For 20 observers (or more) $\Omega_F$ will generally be larger than $\Omega_R$. If the number of observers is limited to 10, however, then $\Omega_F$ is at most 16% larger than $\Omega_R$, while it is much smaller when the observer effects are dominating. Then we examined the estimator $W_F$ of $\Omega_F$, defined in Eq. (9). This appeared to be a good estimator of $\Omega_F$, because the bias is limited to a few percent. The combination of many observers with small effects leads to the largest bias, especially when the observer–product interaction is large. With just a few observers and large observer effects, $W_F$ will slightly underestimate the parameter $\Omega_F$, but the harm is restricted to about 2% (with only 2 observers).

Our conclusion is that $\Omega_F$ should be used to measure the gauge R&R, when only a few (at most 5) observers are present. Note that this parameter describes the variation of the measurement system if products are randomly assigned to observers. It is a better reflection of the total amount of variation due to measurements than the standard method based on the REM, especially when observer effects are relatively large. In addition, not much harm is done if the other sources of variation are more important. $W_F$ is a good but conservative estimator of $\Omega_F$ unless observer effects are really dominating, then $W_F$ might underestimate $\Omega_F$ by a few percent (though this will not be harmful as far as the real gauge R&R is concerned).

With 10 or more observers there is not much gain in using $\Omega_F$. Moreover, the estimator $W_F$ may be seriously biased, especially when observer effects are relatively small. Then the assumption that observers constitute a random sample of a great many gives better results. At the same time this assumption is also more plausible, because the “uniqueness” of 10 or more observers is questionable. For 5–10 observers with fixed effects the methods based on the fixed effects model is preferable, unless the effects are small. In the latter case $\Omega_F$ (not very different from $\Omega_R$), can be better estimated with the ANOVA-estimator of $\Omega_R$.

In this article we discussed only the analysis of measurement variation. This assumes that products are measured randomly by different observers and we need to quantify the amount of variation coming from the total measurement system (reproducibility and repeatability). We showed that measurement systems may well perform better than viewed by the “traditional” method. Unnecessary investments to improve measurement systems might be avoided; and sometimes the only solutions are expensive. For the model we discussed, Wheeler gives a simple and cheap solution though: he proposes to compensate computationally for the fixed observer effects. Note that this requires careful model checking and additional effort to estimate the observer effects, because compensation has a permanent influence on the measurements.

**APPENDIX**

First we will show that the 99% prediction interval for the MEM is enclosed in two intervals and secondly we will prove Eq. (11).

**Bounds on the width of the MEM prediction interval:** Let $\beta_1, \beta_2, \ldots, \beta_m$ be the ordered observer effects and let $m \leq 199$. Then define $q_i$ by

$$q_i = \beta_{(i)} + \sigma \Phi^{-1}(i/200), \quad i = 1, \ldots, 199.$$ 

Then the MEM prediction interval $IF = [LF, UF]$ defined
Measurement Systems with a Small Number of Observers

by condition (4), is enclosed in two intervals:

\[ [q^m_1, q^{200-m}_m] \subseteq [LF, UF] \subseteq [q^1_m, q^{199}_m]. \]

Without loss of generality we assume that \( \mu = 0 \) and the product effect is zero (\( \sigma_x = 0 \)). Furthermore, symmetry allows us to show the inequalities for LF only. The proof follows from the following two inequalities:

\[
F(q^1_m) = \frac{1}{m} \sum_{j=1}^{m} \Phi \left( \frac{q^1_j - \beta_j}{\sigma} \right) \leq \frac{1}{m} \sum_{j=1}^{m} \Phi \left( \frac{q^1_j - \beta_{(1)}}{\sigma} \right) = 0.005
\]

and

\[
F(q^m_m) = \frac{1}{m} \sum_{j=1}^{m} \Phi \left( \frac{q^m_m - \beta_j}{\sigma} \right) \geq \frac{1}{m} \Phi \left( \frac{q^m_m - \beta_{(1)}}{\sigma} \right) = 0.005
\]

Now \( F(LF) = 0.005 \) and \( F \) is monotone nondecreasing function; therefore \( q^1_m \leq LF \leq q^m_m \). Equality holds in the first inequality if and only if \( \beta_j = \beta_{(1)} (\forall j) \); and equality holds in the second inequality if and only if the term with \( \beta_{(1)} \) is the only nonzero term in the summation. Finally, the widths of the enclosing intervals are the upper and lower bounds of Eqs. (5) and (6).

**Proof of Eq. (11):** Let \( X \) be normally distributed with mean \( \mu \) and variance \( \sigma^2 \), and denote \( \phi \) for the standard normal density function and \( \Phi \) for the standard normal distribution function. Then:

\[
E[X] = \int_{-\mu/\sigma}^{\infty} (\mu + \sigma t) \phi(t) dt
\]

\[
- \int_{-\infty}^{-\mu/\sigma} (\mu + \sigma t) \phi(t) dt
\]

\[
= \mu - 2\mu \Phi \left( -\frac{\mu}{\sigma} \right) + 2\sigma \phi \left( \frac{\mu}{\sigma} \right)
\]

\[
= |\mu| - 2|\mu| \Phi \left( -\frac{|\mu|}{\sigma} \right) + 2\sigma \phi \left( \frac{\mu}{\sigma} \right).
\]

The second equality follows from the relationship \( \phi'(t) = -t\phi(t) \).

**ACKNOWLEDGMENTS**

The authors would like to thank Gert ten Brink and Bert Dol from Philips Semiconductors, Stadskanaal, for their measurements.

**ABOUT THE AUTHORS**

Edwin van den Heuvel is a senior consultant at the Institute for Business and Industrial Statistics at the University of Amsterdam (IBIS UvA BV). He is also an associate professor at the Department of Mathematics at the University of Amsterdam. He received his Ph.D. in mathematical statistics at the University of Amsterdam in 1996.

Albert Trip is a senior consultant at IBIS UvA BV. He was a statistician at Philips Semiconductors, Stadskanaal, The Netherlands, from 1994 until 2001. He received his Ph.D. in industrial statistics at the University of Amsterdam in 2000.

**REFERENCES**