Quality Engineering, 22:46–53, 2010
Copyright © Taylor & Francis Group, LLC
ISSN: 0898-2112 print/1532-4222 online
DOI: 10.1080/08982110903412924

Taylor & Francis
Taylor & Francis Group

Quality Quandaries*: A Gage R&R Study in a Hospital

Tashi P. Erdmann[1], Ronald J. M. M. Does[1], Søren Bisgaard[1,2]

[1]Institute for Business and Industrial Statistics, University of Amsterdam, The Netherlands
[2]Eugene M. Isenberg School of Management, University of Massachusetts, Amherst, Massachusetts

Downloaded By: [Erdmann, Tashi P.] At: 09:12 2 December 2009

INTRODUCTION

Measurement system analysis (MSA) is indispensable to quality management. Neither quality control nor quality improvement can be done without being able to take reliable measurements. In quality improvement projects it is standard practice to assess the reliability of measurements before doing any analyses. In particular, in the second phase of a Six Sigma project, the Measure phase, the measurement procedures need to be validated.

A very important aspect of the quality of a measurement procedure is its precision, or the measurement variation: the extent to which measurements vary around their mean. The measurement variation should be small compared to the product variation or compared to the specification interval, which is the difference between upper and lower specification limits. The precision of a measurement procedure can be assessed by a gage R&R study.

This "Quality Quandaries" column gives an example of a gage R&R experiment in a hospital: a study of the precision of temperature measurements with an ear thermometer. The remarkable conclusion of the experiment is that temperature measurements may be the most precise if each ear is measured once and then the maximum of the two measurements is taken as the body temperature.

The column starts with a discussion of the various aspects of the quality of measurement based on the first chapter of the thesis of Van Wieringen (2003), followed by the ear thermometer example.

QUALITY OF MEASUREMENT

The knowledge obtained from any measurement is determined by the quality of measurement. If the quality of measurement is low, then so is the amount of information that can be gained from it.

The objective of a numerical measurement is to determine the value of a quantity (International Organization for Standardization [ISO] 1995). The measurement is supposed to reflect a certain property of the object measured and should be as close as possible to the true value or reference value of that property. The reference value, sometimes called the *gold standard*, is the approximation of the true value that would be obtained by a metrology laboratory. The difference between the measured value and the reference

*Edited by Søren Bisgaard.

Address correspondence to Tashi P. Erdmann, Institute for Business and Industrial Statistics, Plantage Muidergracht 12, University of Amsterdam, 1018 TV, The Netherlands. E-mail: t.p.erdmann@uva.nl

Downloaded By: [Erdmann, Tashi P.] At: 09:12 2 December 2009

value is called *measurement error*. It is the probability distribution of these measurement errors that determines the quality of measurement.

In the literature of industrial statistics two aspects of the quality of measurement are usually distinguished: its accuracy and its precision. Accuracy relates to the bias of the measurement and precision to the measurement spread. The *Measurement System Analysis Reference Manual* by the Automotive Industry Action Group (AIAG; 2003) has been a guide in making this distinction.

Accuracy is the degree to which the measurement system is subject to bias or systematic measurement error. Bias is the difference between the average of multiple measurements on the same object and the reference value. Bias can be corrected for by calibration of the measurement equipment. The extent to which bias is constant over time is called *stability*, and the extent to which bias is constant over the measured range is called *linearity*.

Precision is the degree to which the measurement system is subject to measurement spread, which is the standard deviation of repeated measurements on the same object. Precision is subdivided into repeatability and reproducibility. If all circumstances such as measurement instrument, person, and location are kept equal for each of those repeated measurements, the measurement variation that is left is called *repeatability*. This is the variation that could be obtained after eliminating all sources of variation that are caused by differences in circumstances. If measurements are conducted under different circumstances, the variation often increases. The additional variation due to varying circumstances is called *reproducibility*. One could be interested in various types of reproducibility, such as the reproducibility for different operators, different measurement devices, or different environmental conditions. The extent to which precision is constant over time and over the measured range is called *consistency* and *uniformity*, respectively.

The different aspects of the quality of measurements that have been discussed can be assessed by experiments. A well-known experiment is a gage R&R study, which is an experiment that assesses precision. The experiment has a factorial design, allowing for the determination of the effect of the different factors on the measurement variability (see Box et al. 1978). The variation caused by factors related to the measurement system is reproducibility, whereas the variation that remains when all factors are kept constant is repeatability.

AN INTRODUCTORY EXAMPLE

In this section we introduce an example of a gage R&R study in health care. In a quality improvement project in a hospital, it was necessary to measure the body temperature of patients. The temperature measurements were taken with an ear thermometer. A man's body temperature is under normal circumstances in the range of a lower specification limit (LSL) of 35°C and an upper specification limit (USL) of 40°C.

To make sure no false conclusions would be drawn during the investigation of the body

TABLE 1 **Data of the Gage R&R Experiment**

		Operator 1 Jolan				Operator 2 Mariska				Operator 3 Paula			
		1		2		1		2		1		2	
		r	l	r	l	r	l	r	l	r	l	r	l
Louis	1	37.3	37.5	37.3	37.5	37.5	37.7	37.3	37.6	37.5	37.6	37.4	37.5
Rene	2	37	37.3	36.7	36.8	37.5	37.3	37.4	37.2	37.4	37.4	37.3	37.1
Ben	3	36.4	37	37.3	37	37.5	37.3	37.4	37.1	37.6	37.4	37.2	37
Robert	4	37.6	37.5	37.6	37.4	37.5	37.5	37.5	37.7	37.7	37.6	37.6	37.5
Renee	5	36.7	37.6	37.8	37.5	37.9	37.5	37.6	37.6	37.9	37.6	37.9	37.8
Sandra	6	37.5	37.7	37.6	37.3	38.4	38	37.8	37.8	37.6	37.9	37.8	37.8
Ton	7	37	36.9	37.1	37.3	37.1	37.3	37.4	37.5	37.2	37.4	37.1	37.2
Annemarie	8	37.7	37.4	37.6	37.4	37.6	37.5	37.5	37.1	37.5	37.4	37.2	36.9
Lieke	9	36.4	36.5	36.6	36.1	37.1	36.9	36.7	36.8	37	36.4	36.9	36.8
Ronald	10	37.2	37.4	37	37.3	37.1	37.2	37.2	37.2	37.1	37.2	37	37.3

Downloaded By: [Erdmann, Tashi P.] At: 09:12 2 December 2009

temperature data, the quality of the temperature measurement was assessed first by means of a gage R&R experiment. The nurses handling the ear thermometer may cause extra variability in the measurements. They were taken along (as a factor) in the experiment. Different healthy persons were involved in the experiment. They contribute to the observed variation, which is object (read: person) variation, not part of the measurement variation. The experiment was therefore designed such that it allowed for separation of object variability from measurement variability. Thus, object was taken as a factor. A single ear thermometer was used by all nurses, which was also the case during the experiment. Hence, it was not a factor during the experiment. Each person was measured in the right and left ears. This procedure was done twice.

Taking all this into account, it was decided to conduct an experiment involving three nurses and 10 healthy persons. Each person was measured twice in both right and left ears by each nurse. The results are presented in Table 1.

The persons were measured in random order to eliminate disturbing effects that may occur over time and to ensure that the nurses did not remember which temperature they had measured before.

MATHEMATICAL MODEL

Traditionally, experiments for the evaluation of measurement systems involve two factors (Montgomery and Runger 1993a, 1993b), which correspond to the factors persons and nurses in our example. In such an experiment, n objects (persons) are measured l times by m operators (nurses). When dealing with continuous measurements it is assumed that the outcome of the experiment can be modeled by an (additive) two-way random-effects model. Let X_{ijk} be the kth judgment of nurse j on person i, then the random effects model is given by:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where μ is the overall mean, $\alpha_i \sim N(0, \sigma_\alpha^2), \beta_j \sim N(0, \sigma_\beta^2), \gamma_{ij} \sim N(0, \sigma_\gamma^2)$, and $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$ are random variables representing the effects of persons, nurses, person–nurse interaction, and error variance, respectively, for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, and $k = 1, 2, \ldots, l$. All these effects are assumed to be stochastically independent random variables. The expected values of the random effects associated with nurses, persons, person–nurse interaction, and error are zero. The measurement error due to person–nurse interaction should be regarded as resulting from nurses approaching persons differently; for example, having difficulty with a person's length.

This model is appropriate if persons and nurses are drawn from large populations, and the underlying distributions are approximately normal. There are circumstances in which the operators involved are the only available; for example, with laboratory measurements. In such a case, the operator's effects should be treated as fixed (Van den Heuvel and Trip 2003).

In this model the variance component σ_ε^2 is the repeatability, because it represents the variation observed among repeated measurements with unchanged conditions. Reproducibility is defined as

$$\sigma_\beta^2 + \sigma_\gamma^2$$

The variance component related to the factor person has no relation with the measurement process. The total measurement variance σ_m^2 is defined as:

$$\sigma_m^2 = \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$$

The measurement spread is the total measurement standard deviation σ_m, which is the square root of the total measurement variance.

STATISTICAL ANALYSIS

The model described above is analyzed as an analysis of variance (ANOVA). For the temperature measurement data in Table 1, the ANOVA table is as shown in Table 2. The various variance components can be estimated by taking linear combinations of the mean sums of squares, following Vardeman

TABLE 2 **ANOVA Results**

ANOVA table	df	SS	MS	F	p-value
Persons	9	10.125	1.125	15.659	0.000
Nurses	2	1.109	0.554	7.715	0.004
Persons × Nurses	18	1.293	0.072	1.742	0.046
Repeatability	90	3.712	0.041		
Total	119	16.239			

Downloaded By: [Erdmann, Tashi P.] At: 09:12 2 December 2009

and Van Valkenburg (1999):

$$\sigma_\alpha^2 = 0.088$$
$$\sigma_\beta^2 = 0.012$$
$$\sigma_\gamma^2 = 0.008$$
$$\sigma_\varepsilon^2 = 0.041$$
$$\sigma_m^2 = 0.012 + 0.008 + 0.041 = 0.061$$

The standard deviations representing reproducibility, repeatability, and measurement spread of the temperature measurements are easily estimated from these results:

$$\sigma_{Reproducibility} = 0.140$$
$$\sigma_{Repeatability} = 0.203$$
$$\sigma_m = 0.247$$

The measurement spread can be used to construct a confidence interval for a person's body temperature. In order to obtain such an interval, the measurement spread is multiplied by a coverage factor c:

$$X \pm c\sigma_m$$

where X is a measurement and c is a suitable constant, such that the specified interval can be regarded as a confidence interval for the person's true body temperature. In industry the constant c in the equation is taken to be 2.575, corresponding to a 99% confidence interval. This results in $X \pm 0.64$ degrees, the 99% confidence interval for the body temperature.

CRITERIA FOR MEASUREMENT ERROR

Criteria are necessary in order to decide whether a measurement system is useful for a certain goal. These criteria should prescribe the amount of disparity of measurements on the same person that is acceptable. In industry the measurement spread is compared to the difference in specification limits. If the measurement spread is too large compared to the width of the tolerance interval, then the measurement system is considered inadequate. To decide whether this is the case, industry uses the precision-to-tolerance ratio (P/T ratio), which compares the width of the 99% confidence interval for the reference value to the width of the tolerance interval.

$$\mathrm{P/TRatio} = [5.15\sigma_m/(\mathrm{USL} - \mathrm{LSL})] \times 100\%$$

TABLE 3 AIAG Criteria for P/T Ratio

Criterion	Quality of Measurements
P/T-ratio > 30%	Inadequate
10% < P/T-ratio <30%	Moderate
P/T-ratio < 10%	Adequate

The precision-to-tolerance ratio gives the percentage of the tolerance interval that is "consumed" by the measurement spread. If the precision-to-tolerance ratio is large, this indicates poor measurement capability; that is, poor capability to determine whether the reference value falls inside the tolerance interval.

To decide upon the adequacy of the measurement system based on the precision-to-tolerance ratio, AIAG (2003) provided the standards shown in Table 3. The criteria of the AIAG relating the P/T ratio and the quality of measurement are debatable (cf. Engel and De Vries 1997).

The P/T ratio of the temperature measurements is as follows:

$$\mathrm{P/T - Ratio} = \frac{5.15 \cdot 0.247}{40 - 35} \times 100\% = 25.4\%$$

According to AIAG the quality of the measurements is moderate. About a quarter of the tolerance interval is consumed by the measurement spread.

GRAPHICAL ANALYSIS

There are several graphs that can be of help in validating and interpreting the results of a gage R&R experiment. First of all, the assumptions of the model should be verified by a residual analysis. Secondly, the results of the experiment should be visualized to correctly interpret the results and draw appropriate conclusions.

For the residual analysis we use a normal probability plot and a scatterplot of the residuals against fitted values to check for heteroscedasticity. These plots are shown in Figure 1.

In the normal probability plot two negative outliers are visible. Further examination of the outliers tells us that they were measurements by the first nurse of the third and fifth persons. In both cases, it was the first measurement of the four measurements that were taken. Because of these outliers, the null hypothesis of a normal distribution of residuals is rejected at the 5% significance level

Downloaded By: [Erdmann, Tashi P.] At: 09:12 2 December 2009

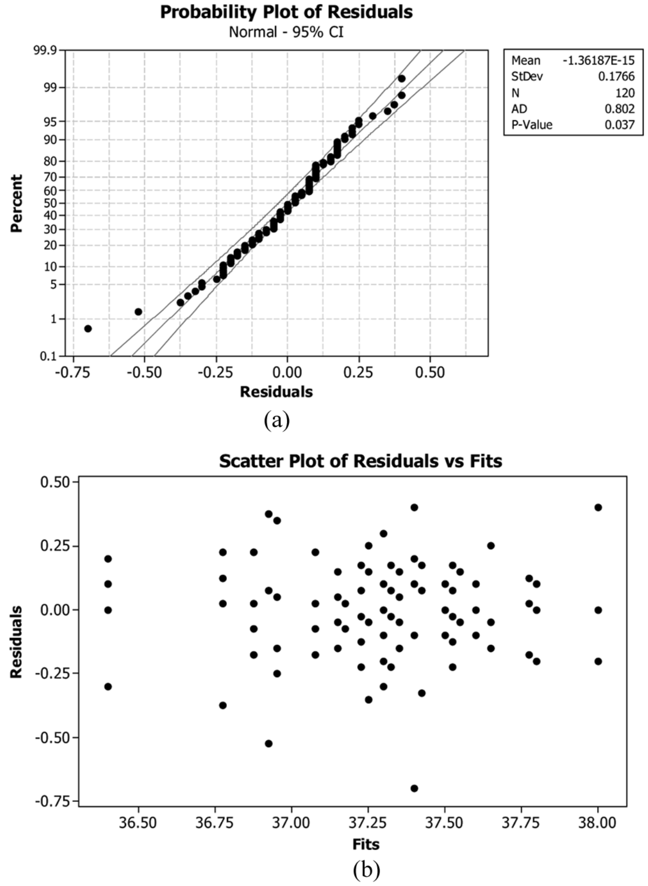

FIGURE 1 Probability plot of residuals and scatterplot of residuals against fitted values.

(p value = 0.037). Therefore, the model assumption of normally distributed errors may not be satisfied. However, this seems to be caused only by the two outliers. The plot of residuals against fits does not show signs of heteroscedasticity.

The gage R&R command of the statistical software package Minitab 15 produced six useful graphs for interpreting the results of the experiment. They are displayed in Figure 2.

In the first graph it can be seen that most of the measurement variation is caused by repeatability, but a considerable part is caused by differences among operators (reproducibility). The variation can probably be reduced quite a bit by giving the nurses clearer instructions on exactly how to put the thermometer into the ear and how to measure the temperature.

The second and third graphs on the left show control charts of range and mean of the measurements of each operator (nurse). Each point in the graph is the range or mean, respectively, of the four measurements that one nurse did on one person. In the range chart it is visible that the first nurse had trouble measuring the third and the fifth people. This corresponds to the two outliers that we saw earlier in the residual analysis. The second and third nurses

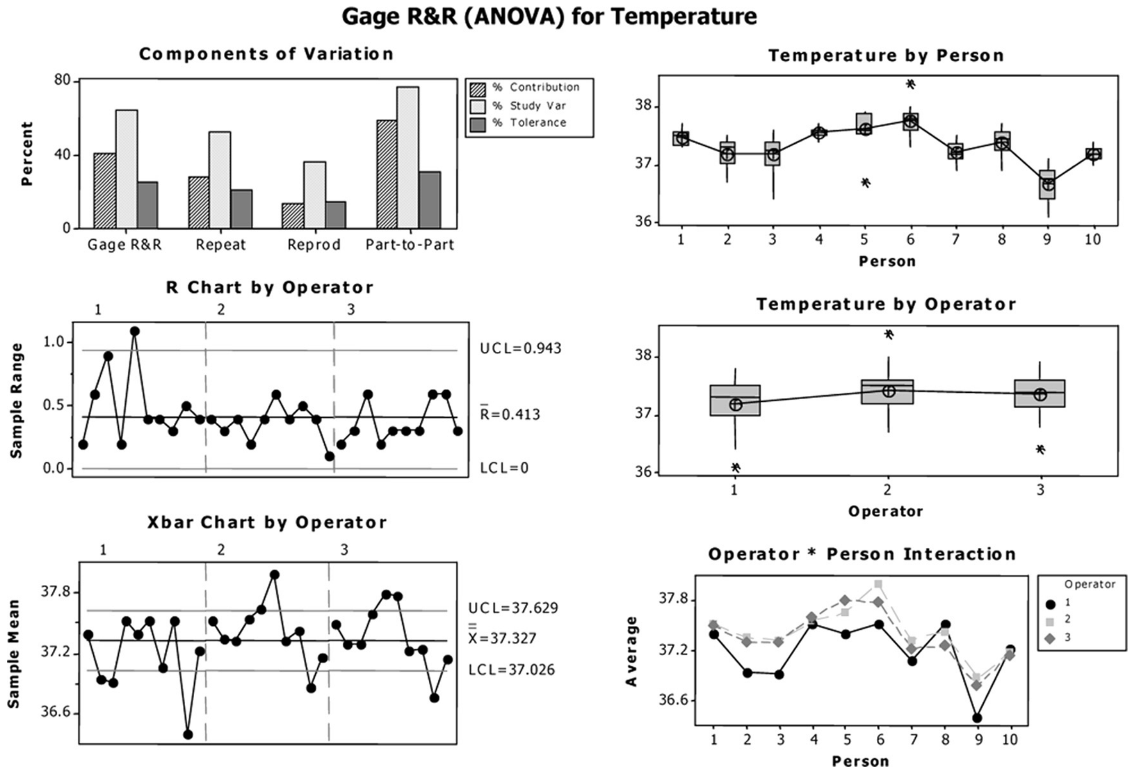

FIGURE 2 Six graphs produced by Minitab's gage R&R command.

measured more consistently than the first nurse, as can be seen from the range chart.

On the right side two graphs are displayed giving box plots. The upper graph shows for each person a box plot of the 12 measurements done on him or her. It can be seen that the fifth person's temperature was once measured a lot lower than the other 11 times, and the sixth person's temperature was once measured much higher than the other 11 times. The middle graph on the right shows for each nurse a box plot of the 40 measurements she took. Nurse one measured one very low temperature (person 5), and nurse two measured one very high temperature (person 6). Another remarkable result is that the median and mean of the measurements by the second and third nurse are higher than those of the first nurse.

Possibly the most interesting of the six plots is the interaction plot on the bottom right in Figure 2. There is clearly an interaction effect between the nurse and the person who is measured. This can be seen because the lines between the dots are not parallel. For all persons except for the first and the tenth person, nurse one recorded a lower temperature than nurses 2 and 3.

IMPROVEMENTS IN THE MEASUREMENT PROCEDURE

Now that we have an idea of the quality of the current measurement procedure, we would like to make improvements to it based on the gage R&R experiment. A first improvement is to give clearer instructions about the way the thermometer should be inserted into the ear and how the measurement should be performed, in order to decrease the differences between nurses. Secondly, we would like to do something about the outliers that were found. The largest measurement errors will occur if the ear thermometer is not inserted deep enough into the ear, causing the temperature measurement to be too low. This happened twice in the experiment, leading to the two negative outliers.

An interesting question in this context is whether it matters whether the right or the left ear is measured. Figure 3 is an individual value plot of the residuals per ear. It can be seen that the two outliers that we saw earlier occurred in measurements of the right ear. Looking at Figure 3, the right side seems more difficult to measure than the left side. This might give us the idea to do gage R&R studies for the left ear and the right ear separately. We can use the data from the same experiment and analyze them per ear. Doing so leads to the results given in Table 4.

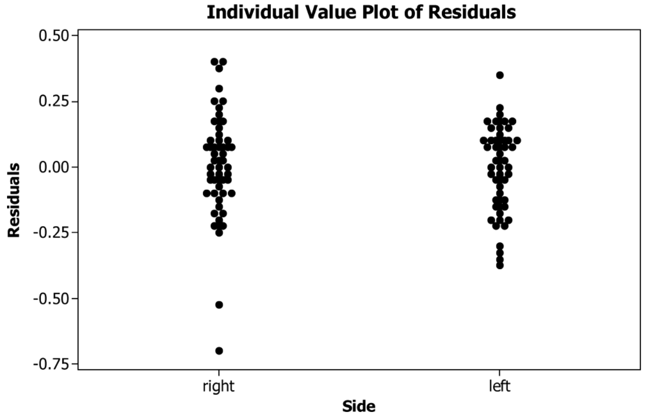

FIGURE 3 Individual value plots of residuals versus ear side.

This confirms the observation that the right ear is harder to measure than the left ear. The total measurement standard deviation is 0.28 for measurements of the right ear and 0.19 for the left. It is an interesting question what the reason for this might be. One possible reason is that most nurses are right-handed, and this makes it difficult to put the thermometer into the right ear when they are standing in front of the person measured. If this is true, then of course for left-handed people the left ear will be harder to measure.

We saw that sometimes measurement errors occur if the "difficult" side is measured and the thermometer is not inserted deep enough into the ear, causing the temperature measurement to be too low. One way to prevent this from happening would

TABLE 4 Gage R&R Standard Deviations and Precision-To-Tolerance Ratio for All Measurements, Measurements on the Left Ear, and Measurements on the Right Ear

	All measurements	Left ear	Right ear
$\sigma_{Measurement}$	0.247	0.193	0.283
σ_{Nurses}	0.110	0.078	0.137
$\sigma_{Persons \times Nurses}$	0.203	0.177	0.236
$\sigma_{Repeatability}$	0.087	0.000	0.073
5.15 ×	1.272	0.995	1.455
$\sigma_{Measurement}$	(1.112, 3.972)	(0.853, 2.967)	(1.229, 4.961)
P/T ratio	25.4%	19.9%	29.1%

Downloaded By: [Erdmann, Tashi P.] At: 09:12 2 December 2009

Downloaded By: [Erdmann, Tashi P.] At: 09:12 2 December 2009

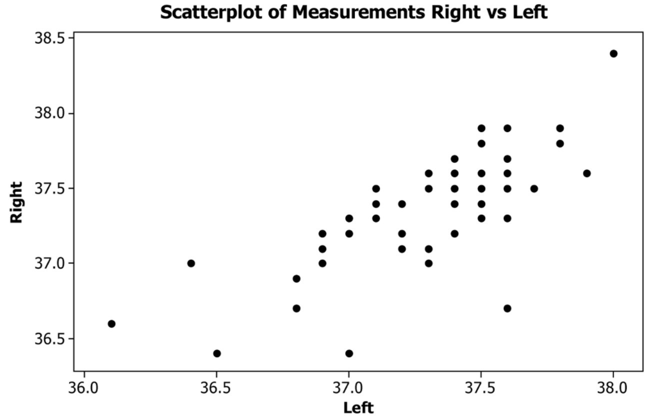

FIGURE 4 Scatterplot of the measurements on the right ear against the measurements on the left ear.

be to let nurses always take two measurements, one right and one left, and take the maximum of the two. With the same gage R&R experiment we can also analyze this new measurement procedure. Of each two subsequent measurements of different ears the maximum is taken, and these maxima are summarized in a new data set. Then the analysis of variance of the new data set is done. The total measurement standard deviation of the maxima is only 0.19. We conclude that taking the maximum of two measurements, one on the right ear and one on the left ear, may lead to a better precision.

Instead of taking the maximum, an alternative would be to use the mean of the left and right ear measurements. However, the mean is sensitive to negative outliers, whereas the maximum is not. The measurement standard deviation is 0.21, which is slightly more than the standard deviation if the maximum is used. Note that in general the mean of two i.i.d. normally distributed variables Y_1 and Y_2 has a smaller standard deviation than their maximum; in particular, the standard deviation of their mean is 70.7% and that of their maximum is 82.6% of the standard deviation of Y_i (see the tables in Godwin 1949, for example). However, this result does not hold here, because firstly it is questionable whether the measurements on the left and right ears are stochastically independent (in Figure 4 correlation is clearly visible) and secondly they are not normally distributed because of the occasional negative outliers we discussed earlier.

The gage R&R results of the two suggestions for an improved measurement procedure, one based on the maximum of two measurements and the other based on their mean, are shown in Table 5. It seems that taking the maximum is the best choice.

TABLE 5 Gage R&R Standard Deviations and Precision-To-Tolerance Ratio for the Mean and for the Maximum of Both Ears

	Mean of left and right	Maximum of left and right
$\sigma_{Measurement}$	0.208	0.190
σ_{Nurses}	0.110	0.079
$\sigma_{Persons \times Nurses}$	0.068	0.068
$\sigma_{Repeatability}$	0.163	0.159
$5.15 \times \sigma_{Measurement}$	1.072	0.981
	(0.895, 3.913)	(0.835, 3.026)
P/T ratio	21.4%	19.6%

CONCLUSIONS

This article discusses the principles of measurement system analysis and illustrates these with an example of a measurement system analysis in healthcare.

The quality of measurements is determined by the accuracy and the precision of the measurement. The precision of a numerical measurement can be determined by a gage R&R experiment. The purpose of a gage R&R experiment is to estimate the measurement spread and to find out what part of the measurement spread is caused by repeatability and what part by reproducibility of the measurement system.

The principles of measurement system analysis that are common in industry can very well be applied to measurement systems in health care. In this article body temperature measurements with an ear thermometer are used as an example. The gage R&R study shows that with the current measurement procedure, a measurement can differ from the real temperature by as much as 0.64 degrees (99% confidence). Sometimes measurements are far too low, because the thermometer is not inserted into the ear properly. A more precise method would be to measure both ears and then take the maximum of the two measurements. The measurement error will then be less than a half degree with 99% confidence.

REFERENCES

Automotive Industry Action Group. (2003). *Measurement System Analysis: Reference Manual,* 3rd ed. , Detroit, MI: Automotive Industry Action Group.

Box, G. E. P., Hunter, W. G., Hunter, J. S. (1978). *Statistics for Experimenters*. New York: Wiley.

Engel, J., De Vries, B. (1997). Evaluating a well-known criterion for measurement precision. *Journal of Quality Technology*, 29: 469–476.

Godwin, H. J. (1949). Some low moments of order statistics. *Annals of Mathematical Statistics*, 20(2):279–285.

International Organization for Standardization. (1995). *Guide to the Expression of Uncertainty in Measurement,* 1st ed. Geneva, Switzerland: International Organization for Standardization.

Montgomery, D. C., Runger, G. C. (1993a). Gage capability and designed experiments. Part I: Basic methods. *Quality Engineering*, 6:115–135.

Montgomery, D. C., Runger, G. C. (1993b). Gage capability and designed experiments. Part II: Experimental design methods and variance component estimation. *Quality Engineering*, 6:289–305.

Van den Heuvel, E. R., Trip, A. (2003). Evaluation of measurement systems with a small number of observers. *Quality Engineering*, 15:323–331.

Van Wieringen, W. N. (2003). *Statistical Models for the Precision of Categorical Measurement Systems*. Ph.D. thesis, University of Amsterdam, The Netherlands.

Vardeman, S. B., Van Valkenburg, E. S. (1999). Two-way random effects analyses and gage R&R studies. *Technometrics*, 41:202–211.

Downloaded By: [Erdmann, Tashi P.] At: 09:12 2 December 2009