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

Quality Quandaries: Precision and Accuracy of Ear Thermometry

Thomas S. Akkerhuis[a], Gerard C. H. Niemeijer[b], Albert Trip[b], Reinoud J. B. J. Gemke[c] & Ronald J. M. M. Does[a]

[a] Institute for Business and Industrial Statistics (IBIS UvA), Department of Operations Management, University of Amsterdam, Amsterdam, The Netherlands

[b] University Medical Center Groningen, Groningen, The Netherlands

[c] VU Medical Center, Amsterdam, The Netherlands

Published online: 24 Aug 2015.

To cite this article: Thomas S. Akkerhuis, Gerard C. H. Niemeijer, Albert Trip, Reinoud J. B. J. Gemke & Ronald J. M. M. Does (2015): Quality Quandaries: Precision and Accuracy of Ear Thermometry, Quality Engineering, DOI: 10.1080/08982112.2015.1065324

To link to this article: http://dx.doi.org/10.1080/08982112.2015.1065324

Quality Engineering, 0:1–10, 2015

ISSN: 0898-2112 print / 1532-4222 online
DOI: 10.1080/08982112.2015.1065324

Quality Quandaries: Precision and Accuracy of Ear Thermometry

Thomas S. Akkerhuis,[1]
Gerard C. H. Niemeijer,[2]
Albert Trip,[2]
Reinoud J. B. J. Gemke,[3]
Ronald J. M. M. Does[1]
[1]Institute for Business and Industrial Statistics (IBIS UvA), Department of Operations Management, University of Amsterdam, Amsterdam, The Netherlands
[2]University Medical Center Groningen, Groningen, The Netherlands
[3]VU Medical Center, Amsterdam, The Netherlands

Edited by Ronald J. M. M. Does

Address correspondence to Ronald J. M. M. Does, IBIS UvA, Department of Operations Management, University of Amsterdam Business School, Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands. E-mail: r.j.m.m.does@uva.nl

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lqen.

INTRODUCTION

The subject of this quandary is the analysis of measurement systems. Measurement systems are everywhere—we find them at home (bathroom scales, thermostats), in our cars (speedometers, check engine lights), in hospitals (heart rate monitors, blood pressure monitors), during sports (stopwatches), and around our wrists (watches). Measurement system analysis (MSA) is the part of applied statistics that attempts to describe, categorize, and evaluate measurement error, improve the usefulness, accuracy, and precision of measurements, and propose methods for developing new and better measurement instruments (Allen and Yen 1979).

MSA is an important field of research in industrial statistics, a branch of statistics that is often applied in industry. For example, for the automotive industry, the Automotive Industry Action Group (AIAG) prescribes how measurement error should be quantified, and dictates upper bounds (AIAG 2003). This is essential: for example, we dislike break systems that fail to sufficiently decelerate a car, because there was an error in quality measurement before it was built into the car. Measurement errors can thus endanger one's wellbeing. We believe that measurements in medicine are of comparable importance. It is interesting to find out how well methods used in industry perform in a medical context. For that reason, we apply a technique that is popular in industry, to a measurement device that is very relevant in medicine.

In particular, we will report on a study of infrared ear thermometers (ETs). In contrast to the introduction of new drugs, which is bound to strict regulations and procedures, most "over the counter" diagnostic tests are not scrutinized in a similar fashion before they become available for consumers.

ETs reflect body temperature by measuring radiation from the tympanic membrane. The average body temperature for a healthy person is 36.8°C, but values between 36.0°C and 37.6°C are considered normal (Mackowiak, Wasserman, and Levine 1992). Variations within this normal range result from differences in gender, menstrual cycle, race, time of the day, and age, but temperatures outside this range may well be due to a medical condition (Mackowiak, Wasserman, and Levine 1992; Sund-Levander, Forsberg, and

Wahren 2002). There are different ways to measure body temperature. Reliable and accurate measurement of rectal temperature is the gold standard of the body's core temperature. Therefore, a rectal thermometer (RT) is frequently used to measure body temperature (Craig et al. 2002; Smitz, Van de Winckel, and Smitz 2003). ETs are more convenient for obvious reasons, although they are generally believed to be less reliable due to influences from ambient temperature, conditions of the ear that is measured (e.g., local inflammation), training of the nurse handling the thermometer, and local hyperemia (Smitz, Van de Winckel, and Smitz 2003; Amoateng-Adjepong, Del Mundo, and Manthous 1999; Doyle, Zehner, and Terndrup 1992; Heusch and McCarthy 2005; Korshid et al. 2004; Petersen and Hauge 1997; Stavem, Saxholm, and Smith-Erichsen 1997; Weiss Pue, and Smith 1991).

The technique used here is an expanded gage repeatability and reproducibility (R&R) study (Montgomery and Runger 1993a, 1993b). This study is based on the fundamentals of experimental design (Box, Hunter, and Hunter 1978) and enables subdividing measurement error into various sources. Gage R&R studies are common practice in industry but can be applied to any numerical measurement system. For categorical measurements, one can, for example, consider kappa (κ) statistics (Cohen 1960; Erdmann, De Mast, and Warrens 2015) or sensitivity/specificity probabilities (Pepe 2003).

In current medical literature, measurement error is often quantified using correlation, regression, confidence intervals, Blant-Altman plots, and *t*-tests (Craig et al. 2002; Smitz, Van de Winckel, and Smitz 2003; Korshid et al. 2004; Stavem, Saxholm, and Smith-Erichsen 1997). A gage R&R study is more complete in the sense that it provides a statistical model for measurement error. This study has been performed in limited form in a previous Quality Quandary by Erdmann, Does, and Bisgaard (2010). We revisit this subject because in this study (i) we have applied a gold standard (the RT measurements), and (ii) we have used an experimental design that allows us to estimate the effect of various factors on measurement error.

We start with a definition and decomposition of measurement error, and then detail the experimental design and statistical model. Then we interpret the results and close with a discussion.

MEASUREMENT ERROR

Measurement error is defined as the discrepancy between the (hypothesized) reference value of the property of the subject and the measured value. The reference value is defined as the mean value that would be assigned to the subject's property by a standard measurement system (i.e., taken by general consent as a basis for comparison, set up, and established by an authority). This is a conceptual value. In this section, we detail the components (and subcomponents) of measurement error.

Measurement error is dissected first into accuracy and precision. Other categorizations are documented as well. For example, psychometrics uses a categorization into validity and reliability (Kerlinger and Lee 2000).

- Accuracy: The degree to which the measurement system is free of bias. *Bias* is the difference between the overall average of repetitive measurements of the property of the subject and the reference value of the subject's property. Systematic measurement error and location variation are equivalent concepts.
- Precision: The extent to which one obtains similar results if one measures the same subject multiple times. It is further split up in two components.
    - If the repeated measurements are conducted under identical circumstances (involving the same subject, the same measurement instrument, the same person, the same location, one directly after the other), the observed variation represents the best attainable precision with this measurement system. This variation is referred to as *repeatability*.
    - If a subset of the measurement is conducted under different circumstances, the observed variation will increase. The additional variation due to varying circumstances is called *reproducibility*. A valid statement of reproducibility requires specification of the conditions changed, e.g., other raters handling the measurement system, alternative measuring equipment used, changed environmental conditions.

Random measurement error and width variation are equivalent concepts.

Accuracy is inversely related to bias, in the sense that low bias means high accuracy. Bias is often quantified

as a mean value (μ). Equivalently, high precision means that repeatability and reproducibility are high, which happens when variation in measurements of the same subject is low. Precision is often quantified in terms of variance (σ^2), but strictly speaking, precision is inversely related to variance.

CASE STUDY: EXPERIMENTAL DESIGN AND STATISTICAL MODELING

This gage R&R study establishes the measurement error of ETs. Both reproducibility and repeatability will be quantified. To express accuracy, RT measurements are performed to obtain reference values. This section discusses the experiment and statistical model.

The first step in a gage R&R study is to plan an experiment. Experimental data has many advantages over observational data. The fundamental difference is the degree of control (Box, Hunter, and Hunter 1978). We start with making an inventory of the possible factors (we use the word "factor" instead of "circumstance" as the former is the common term used in the context of design of experiments) that may influence measurement. Some factors are of interest, and are manipulated in the experiment according to an experimental design. Manipulation allows for the estimation of their effects. Other factors are not of interest, and it is made sure that they remain constant during the experiment, as not to confound effect estimates. Then there are factors that are either not in the inventory, or not controllable. These are averaged out by randomization, i.e., performing the measurements in random order.

After thorough brainstorming sessions, the following factors were chosen to be manipulated in the experiment:

- Subject (p): 10 healthy volunteers participated in the experiment. These were not actual patients, but employees of the hospital. As not all 10 volunteers have the same body temperature, it is important to take subject into account as a factor.
- Nurse (j): Ear temperatures have been measured by 5 different nurses (4 registered nurses, 1 student nurse). This is relevant because the technique that is used to insert the ET in an ear can influence outcomes.
- Thermometer (k): Two different ETs are used. Although both are validated, it is interesting to see whether they influence measurement error.
- Ear (ℓ): There may be a difference between temperature measurements in the left ear and the right ear. As opposed to the previous three factors, ear is taken as a fixed effect.

In order to be able to calculate repeatability, we perform 2 measurements for each combination of settings for these factors, as is convention in gage R&R studies. A factorial design is chosen, requiring $10 \times 5 \times 2 \times 2 \times 2 = 400$ measurements. Additionally, RT measurements are performed to serve as reference values. Although RT measurement is assumed unbiased, it is not perfectly precise. For that reason, two RT measurements are performed per subject, and the average is used as reference value. The total number of measurements is 420. The complete dataset can be found in the Appendix.

The 10 subjects were gathered in a room with 10 chairs in a circle. The nurses walked around the inside of the circle to collect measurements. The subjects performed RT measurements themselves in a room within 3 meters walking distance, directly before and after the measurements with the ETs. The following factors were kept constant:

- Ambient temperature. The room was climate controlled so ear temperature measurements cannot be influenced by changes in climate.
- Body temperature. As subjects could remain seated, changes in body temperature over the course of the experiment were minimal. Subjects could perform rectal measurements very close to their chair, as not to influence body temperature.

As body temperature has been shown to fluctuate over time, it was essential to minimize the duration of the experiment. Therefore, the experimenters have chosen not to randomize the order of measurements. It is believed that the effect of other variables is negligible. Temperatures were measured and denoted with one decimal place.

Using the data, we can estimate accuracy and precision. For accuracy, we calculate an average ear temperature and an average rectal temperature for each subject, and compare them using a paired-samples t-test.

To establish precision, we start out with a linear model for the m th measurement of subject p by nurse j

TABLE 1 Descriptive Statistics of Temperature Measurements

		1	2	3	4	5	6	7	8	9	10
ET	Mean	35.9	36.3	37.1	36.7	36.4	36.9	36.4	37.2	36.4	37.0
	St. Dev.	0.26	0.19	0.24	0.25	0.15	0.24	0.20	0.22	0.29	0.22
	Skewness	−0.20	−0.03	−0.07	−0.54	−0.65	0.29	−1.01	−0.63	−0.36	−0.79
	Min	35.3	35.8	36.7	36.1	36.0	36.5	35.7	36.7	35.8	36.4
	Max	36.3	36.8	37.5	37.1	36.6	37.5	36.7	37.6	37.0	37.3
RT	Mean	36.5	36.9	37.5	36.9	37.2	37.1	36.9	37.7	37.1	37.4

using ET k in ear ℓ, T_{pjklm}:

$$T_{pjklm} = \mu + \beta_p + \beta_j + \beta_k + \beta_\ell + \beta_{pj} + \beta_{pk} + \beta_{p\ell} + \beta_{jk} + \beta_{j\ell} + \beta_{k\ell} + \varepsilon_{pjk\ell m}, \quad [1]$$

for $p = 1,\ldots,10$, $j = 1,\ldots,5$, $k = 1,2$, and $\ell = 1,2$. μ is the grand average, the β's are main and interaction effects, and ϵ is error. Equation [1] can be estimated using least squares, under the restriction that the average effects are 0. For the subject effect, for example, this restriction is implemented by setting β_{p5}, the effect of subject 5, equal to $-(\beta_{p1} + \beta_{p2} + \beta_{p3} + \beta_{p4})$. The implementation of the restriction for other effects is equivalent. Writing the model in terms of the associated variances, we obtain the gage R&R model:

$$\sigma^2_{\text{total}} = \sigma^2_{\text{subject}} + \sigma^2_{\text{measurement}} = \sigma^2_{\text{subject}} + \sigma^2_{\text{reproducibility}} + \sigma^2_{\text{repeatability}} = [\sigma^2_p] + \left[\sigma^2_j + \sigma^2_k + \sigma^2_\ell + \sigma^2_{pj} + \sigma^2_{pk} + \sigma^2_{p\ell} + \sigma^2_{jk} + \sigma^2_{j\ell} + \sigma^2_{k\ell}\right] + [\sigma^2_\epsilon]. \quad [2]$$

Equation [2] shows how variation in all measurement outcomes, σ^2_{total}, is decomposed. σ^2_p is subject variation, also called part-to-part spread: variation due to different subjects. It is not part of measurement error. $\sigma^2_j, \sigma^2_k, \sigma^2_\ell$ represent random measurement error due to differences in nurses, ETs, and ears, respectively. The interactions are included as well, as is customary in analysis of experiments in industry, because some sources of variation may be influenced by other factors. The main and interaction effects together represent reproducibility. σ^2_ε is repeatability: the measurement error that remains if the factors are held constant.

RESULTS

In Table 1, descriptive statistics of the measurements are given per subject. Note that, per subject, $5 \times 2 \times 2 \times 2 + 2 = 42$ measurements are performed.

All subjects seem healthy based on the average temperatures (being between 36.0°C and 37.6°C). Looking, however, at the extreme values, we see that subjects 1, 2, 7, and 9 were at least once measured to have a temperature below 36.0°C, indicative of a potential medical condition. This illustrates how measurement error can influence a diagnosis.

Interestingly, the distribution of ear measurements is negatively skewed for nine out of ten subjects. This can be explained as follows. If the ET is not inserted in the outer ear canal deep enough, which may be due to insertion under a wrong angle, measurements will be lower due to the ambient temperature. In contrast, there is a limit to how far a thermometer can be inserted, and thus there is an upper boundary to the measurement outcomes.

Equation [1] was estimated: see Table 2, which tabulates the main effect estimates. For simplicity, interaction effect estimates have not been included in the table.

TABLE 2 Main Effects Estimated for Eq. [1]. Interaction Effects Omitted. *Means Significant at 5 Percent Level

Baseline: $\mu = 36.3$°C	Subject	Nurse	Side	Thermometer
1	−0.78*	−0.10*	0.01	−0.04*
2	−0.32*	0.04*	−0.01	0.04*
3	0.49*	0.03		
4	0.06*	−0.14*		
5	−0.24*	0.17*		
6	0.31*			
7	−0.26*			
8	0.60*			
9	−0.19*			
10	0.34*			

TABLE 3 Outcome of Gage R&R Analysis of Ear Thermometry Measurements

Baseline: $\mu = 36.6^{\circ}C$ Source	ET measurements Variance	Variance (percent)
Total measurement spread	0.0611	25.3%
Repeatability (σ_{ε}^2)	0.0289	12.0%
Reproducibility	0.0322	13.4%
Nurse (σ_j^2)	0.0130	5.4%
Thermometer (σ_k^2)	0.0029	1.2%
Subject–Nurse interaction (σ_{pj}^2)	0.0084	3.5%
Subject–Ear interaction (σ_{pl}^2)	0.0066	2.7%
Nurse–Ear interaction (σ_{jl}^2)	0.0013	0.5%
Subject variation (σ_p^2)	0.1802	74.7%
Total variation	0.2413	100%

Effects with $p > 0.05$ have been removed, but as the experimental design is orthogonal due to its factorial structure, the estimates in Table 2 remain unchanged. The variance components of the remaining factors are given in Table 3.

Table 3 shows the total variation in the measurements by the ETs, decomposed into repeatability, reproducibility, and subject variation. The dominant source of variation for the ear measurements is the subject variation (74.7%), and the measurement spread is roughly equally distributed over repeatability (12.0%) and reproducibility (13.4%). The total measurement spread is $\pm\Phi^{-1}(0.995)\ \sqrt{0.0611} = \pm 0.637^{\circ}C$ (99% confidence).

The major component of reproducibility is the nurse effect (0.0130). This suggests that nurses have different ways of measuring. The effect of which ET is used is significant, but relatively small (0.0029). The Subject-Nurse interaction is also of considerable influence (0.0084). A reason could be body length: the difference in body length of the nurse and of the subject determines the angle under which an ET is inserted. The interaction between Nurse and Ear (0.0013) may point to left- or right-handedness, but is very small. For left-handed nurses, left ears may be slightly easier to measure, and vice versa. Finally, the interaction between Subject and Ear (0.0066) shows that, for some subjects, there are differences between the two ears. This asymmetry may be a result of one ear being cleaner than the other.

Repeatability is the random measurement error that remains when all factors are unchanged, and equals 0.0289. Nurses never exactly insert the ET at constant depths, under constant angles, leading to errors. This source is only marginally smaller than measurement error that is attributable to the factors considered (reproducibility).

In order to determine whether the precision is sufficient, we use the precision-to-tolerance ratio (P/T%). It compares the 99% confidence interval of the measurement spread $\sigma_{\text{measurement}}$ to the tolerance interval. In this case, the tolerance interval is the range of temperatures that are considered normal. If the ratio is large, much of the tolerance interval is "consumed" by measurement spread, and the measurement system is considered imprecise. In symbols, we have

$$P/T = \frac{5.15\sigma_{\text{measurement}}}{37.6 - 36.0} \times 100\%.$$

To guarantee the quality of measurement, AIAG (2003) has proposed criteria for the P/T-ratio. Although the boundaries are debatable (Engel and De Vries 1997), they are commonly used in industry.

- $P/T > 30\%$: Quality of measurements is inadequate.
- $10\% < P/T < 30\%$: Quality of measurements is moderate.
- $P/T < 10\%$: Quality of measurements is adequate.

In this case, we have that $P/T = 79.6\%$, which is considered too large.

In cases where there is no clear tolerance interval to compare the measurement spread to, the gage R&R percentage (GRR%) is used, which is defined by $\sigma_{\text{measurement}}/\sigma_p$: The ratio between measurement variation and subject variation. AIAG (2003) prescribes that

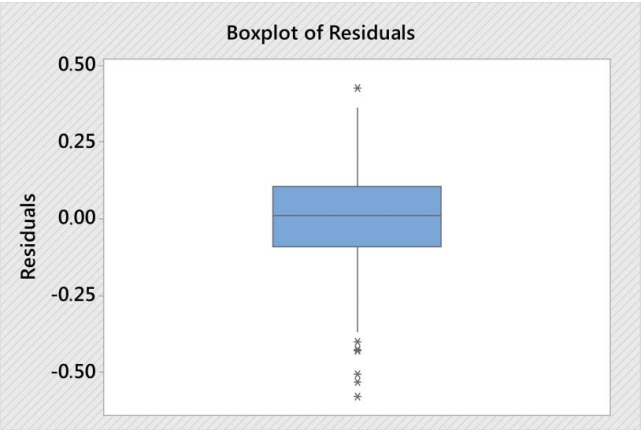

FIGURE 1 Boxplot of residuals.

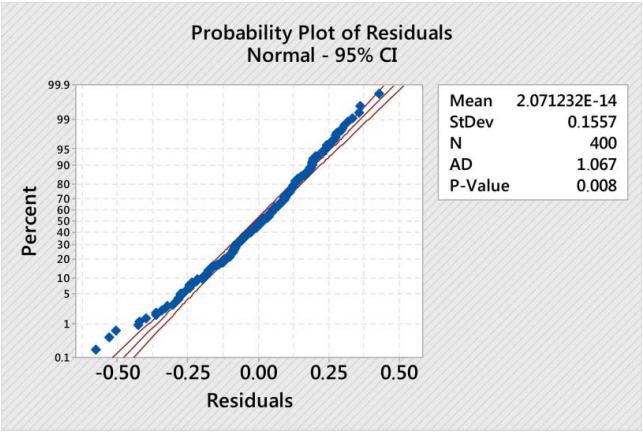

FIGURE 2 Probability plot of the residuals.

this ratio should be below 10%. In this case, we have that $GRR\% = 58.3\%$, which confirms that the precision of the ET is insufficient.

Regarding accuracy, we compared 10 pairs of average ET measurements, with an average difference of -0.466 ± 0.140 (with 95 percent confidence, the RT measurements being higher). The associated p-value for the paired sample t-test is 0.000.

Finally, we study the residuals of the linear model. Figure 1 shows a boxplot of the residuals and Figure 2 the probability plot. The boxplot contains asterisks for measurements outside the interval $[Q_1 - \frac{3}{2}(Q_3 - Q_1), Q_3 + \frac{3}{2}(Q_3 - Q_1)]$, with Q_u the uth quartile, $u = 1, 3$. Outliers are indicated mostly on the lower side. We do not believe these are actual outliers, but rather a result of the negative skewed distribution of the measurements. The Anderson–Darling test rejects normality with $p = 0.008$, but the deviation from normality seems to be negligible.

Importantly, there are no ET measurements that really stand out, ruling out disturbing influences during the experiment. As there are no irregularities, we view the data as reliable.

DISCUSSION

A gage R&R study is used to decompose random measurement error into repeatability and reproducibility, and further attributes reproducibility to various factors. Reference measurements are used to determine and test accuracy.

The main conclusions are that ETs give downward biased estimates of body temperature by roughly 0.5°C (using rectal temperatures as gold standard), and have a P/T-ratio of roughly 80%. The bias is in the same order of magnitude as the 0.29°C found in (Craig et al. 2002). The P/T-ratio is many times larger than the industrial standard of $10 - 30\%$. These numbers indicate that the ETs are insufficiently precise and accurate in the way that they are commonly used in practice.

Some modern ETs apply bias corrections automatically, and some are able to perform multiple measurements in a small timeframe and automatically report the maximum. The first approach may solve issues with accuracy, but does not improve precision. The latter can work well for both. First, the selection of a maximum may partially offset the bias of 0.5 degrees (note that the measurement spread is roughly ± 0.6°C). Furthermore, as we have shown that the distribution of measured temperatures is negatively skewed in general, this will improve precision. The improved measurement quality of selecting the highest temperature out of multiple measurements is proven in research (Smitz et al. 2009; Haugan et al. 2012). It has been suggested to perform the repeated measurements over both ears (Erdmann, Does, and Bisgaard 2010).

Based on the results, we offer three recommendations, as follows:

- This particular group of nurses may benefit from training. First, learning a standardized technique will help in reducing variation due to different nurses (decreasing σ_j^2). Second, such a technique is developed to give more consistent outcomes (decreasing σ_ϵ^2). Thirdly, it may give suggestions regarding how to compensate for the difference in height among subjects (decreasing σ_{pj}^2), and which ear to measure depending on the handedness of the nurse (decreasing σ_{jl}^2 and σ_{pl}^2). In summary, the goal would be making sure that the thermometer is inserted deep enough.
- We have seen that the differences between the two thermometers are a significant source of variation, even though the same brand and type of thermometer was used. This may be due to wear or small differences in calibration. Although this source of variation is relatively small, routine checks for wear and judicious calibration are expected to enhance precision.
- Finally, we found that the interaction between "ear" (left or right) and "subject" significantly induced variation. It has been suggested that cerumen may reduce measurement outcomes. This can be solved

by either cleaning an ear before measurement, or by measuring both ears and taking the maximum (cleanest ear).

The result of the gage R&R study is encouraging because it confirms and combines the findings of many of the articles that are cited in the introduction. That is, ear thermometers are downward biased and can be relatively imprecise. Moreover, we see that the recommendation of performing multiple measurements and taking the maximum is already implemented in modern ear thermometers. Further research could focus on a comparable gage R&R study on the rectal thermometers, which is expected to yield a precision that is substantially higher. In the meanwhile, it is recommended to make decisions that influence a patient's health based on rectal thermometry, and to use ear thermometry only for screening purposes.

The results are thus only valid for the type of thermometer used in this research (Genius II), and it is not clear whether they hold for other ETs as well. Other thermometers can be evaluated using the techniques outlined here. It also must be noted that only healthy subjects have been used in this experiment, and that it is not clear how well ETs perform in terms of measurement error, when true temperatures are outside the normal range.

ABOUT THE AUTHORS

Thomas S. Akkerhuis is a consultant and Ph.D. Student at the Institute for Business and Industrial statistics at the University of Amsterdam (IBIS UvA). The institute operates as an independent consultancy firm within the University of Amsterdam. His consulting activities are focused on Lean Six Sigma, and the topic of his research is measurement system analysis.

Gerard C. Niemeijer is a Certified Management Consultant (CMC) and certified Lean Six Sigma Master Black Belt at the University Medical Center Groningen.

Albert Trip was Lean Six Sigma Master Black Belt at the University Medical Centre Groningen. Now he is an independent consultant for Lean and Six Sigma projects, as well as alderman in the municipality of Borger-Odoorn.

Reinoud J. B. J. Gemke is consultant pediatrician and professor in pediatrics and child health at VUmc University Medical Center, Amsterdam. He director of the pediatric residents training program and member of the Society of Pediatric Program Directors. His research focusses on the epidemiology and outcome of congenital and acquired diseases in early life.

Ronald J. M. M. Does is Professor of Industrial Statistics at the University of Amsterdam; Director of the Institute for Business and Industrial Statistics; Head of the Department of Operations Management; and Director of the Institute of Executive Programmes at the Amsterdam Business School. He is a Fellow of the ASQ and ASA, and Academician of the International Academy for Quality. His current research activities include the design of control charts for nonstandard situations, healthcare engineering and operations management methods.

REFERENCES

Allen, M. J., Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole.

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

Amoateng-Adjepong, Y., Del Mundo, J., Manthous, C. A. (1999). Accuracy of an infrared tympanic thermometer. *Chest*, 155(4):1002–1005.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20:37–46.

Craig, J. V., Lancaster, G. A., Taylor, S., Williamson, P. R., Smyth, R. L.. (2002). Infrared ear thermometry compared with rectal thermometry in children: A systematic review. *The Lancet* 360:603–609.

Doyle, F., Zehner, W. J., Terndrup, T. E. (1992). The effect of ambient temperature extremes on tympanic and oral temperatures. *American Journal of Emergency Medicine* 10 (4):285–289.

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

Erdmann, T. P., De Mast, J., Warrens, M. (2015). Some common errors of experimental design, interpretation and inference in agreement. *Statistical Methods in Medical Research* (in press).

Erdmann, T. P., Does, R. J. M. M., Bisgaard, S. (2010). Quality quandaries: A gage R&R study in a hospital. *Quality Engineering* 22 (1):46–53.

Haugan, B., Langerud, A. K., Kalvøy, H., Frøslie, K. F., Riise, E., Kapstad, H. (2012). Can we trust the new generation of infrared tympanic thermometers in clinical practice? *Journal of Clinical Nursing* 22:698–709.

Heusch, A. I., McCarthy, P. W. (2005). The patient: A novel source of error in clinical temperature measurement using infrared aural thermometry. *Journal of Alternative and Complementary Medicine* 11 (3):473–476.

Kerlinger, F. N., Lee, H. B. (2000). *Foundations of Behavioral Research*. 4th ed. New York, NY: Harcourt.

Korshid, L., Eser, I., Zaybak, A., Yapucu, U. (2004). Comparing mercury-in-glass, tympanic and disposable thermometers in measuring body temperature in healthy young people. *Journal of Clinical Nursing* 14:496–500.

Mackowiak, P. A., Wasserman, S. S., Levine, M. M. (1992). A critical appraisal of 98.6 F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich. *Journal of the American Medical Association* 268 (12):1578–1580.

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.
Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Berlin, Germany: Springer-Verlag.
Petersen, M. H., Hauge, H. N.. (1997). Can training improve the results with infrared tympanic thermometers? *Acta Anaesthesiologica Scandinavica* 40:1066–1070.
Smitz, S., Van de Winckel, A., Smitz, M. (2009). Reliability of infrared ear thermometry in the prediction of rectal temperature in older inpatients. *Journal of Clinical Nursing* 18:451–456.
Stavem, K., Saxholm, H., Smith-Erichsen, N. (1997). Accuracy of infrared ear thermometry in adult patients. *Intensive Care Med* 23:100–105.
Sund-Levander, M., Forsberg,C., Wahren, L. K. (2002). Normal oral, rectal, tympanic and axillary body temperature in adult men and women: A systematic literature review. *Scandinavian Journal of Caring Sciences* 16 (2):122–128.
Weiss, M. E., Pue, A. F., Smith, J.. (1991). Laboratory and hospital testing of new infrared tympanic thermometers. *Journal of Clinical Engineering* 16:137–144.

APPENDIX: DATA

The first dataset is of the gage R&R experiment with the ear thermometers and can be analyzed with the "Gage R&R Study (Expanded)" function in Minitab.

The rightmost column "Temp" contains measurement outcomes. The other columns contain the settings for the factors that have been manipulated.

- The first column ("Ne") shows which nurse did the measurement (1 through 5).
- The second column ("Te") shows which of the two measurement devices has been used.
- The third column ("Pe") shows which of the 10 patients were measured.
- The fourth column ("Se") shows whether the right or left ear has been measured.
- The fifth column ("Re") indicates whether it is the first measurement (1) or a repeat (2).

The dataset below gives the gold standard measurements in column "Tempr." The patient is indicated by "Pr," and the replication by "Rr." The first replication took place before the gage R&R experiment above, and the second afterwards.

TABLE A1 Data of the Ear Thermometers

Ne	Te	Pe	Se	Re	Temp.	Ne	Te	Pe	Se	Re	Temp.	Ne	Te	Pe	Se	Re	Temp.	Ne	Te	Pe	Se	Re	Temp.	Ne	Te	Pe	Se	Re	Temp.
1	1	1	1	1	35.4	3	1	1	1	1	35.8	5	1	1	1	1	35.9	2	2	1	1	1	35.6	4	2	1	1	1	35.6
1	1	2	1	1	36.2	3	1	2	1	1	36.3	5	1	2	1	1	36.5	2	2	2	1	1	36.3	4	2	2	1	1	36.3
1	1	3	1	1	37	3	1	3	1	1	37.2	5	1	3	1	1	37.3	2	2	3	1	1	37.2	4	2	3	1	1	36.9
1	1	4	1	1	36.6	3	1	4	1	1	36.7	5	1	4	1	1	37	2	2	4	1	1	36.8	4	2	4	1	1	36.4
1	1	5	1	1	36.4	3	1	5	1	1	36.4	5	1	5	1	1	36.4	2	2	5	1	1	36.2	4	2	5	1	1	36.3
1	1	6	1	1	36.6	3	1	6	1	1	36.9	5	1	6	1	1	37.1	2	2	6	1	1	36.8	4	2	6	1	1	36.7
1	1	7	1	1	36.5	3	1	7	1	1	36.3	5	1	7	1	1	36.6	2	2	7	1	1	36.2	4	2	7	1	1	36.3
1	1	8	1	1	37.1	3	1	8	1	1	37.3	5	1	8	1	1	37.2	2	2	8	1	1	37.3	4	2	8	1	1	37.1
1	1	9	1	1	36.4	3	1	9	1	1	36.6	5	1	9	1	1	36.7	2	2	9	1	1	36.3	4	2	9	1	1	36
1	1	10	1	1	36.4	3	1	10	1	1	36.5	5	1	10	1	1	36.6	2	2	10	1	1	36.6	4	2	10	1	1	36.7
1	1	1	2	1	35.3	3	1	1	2	1	35.4	5	1	1	2	1	35.9	2	2	1	2	1	35.9	4	2	1	2	1	35.7
1	1	2	2	1	35.8	3	1	2	2	1	36.1	5	1	2	2	1	36.3	2	2	2	2	1	36.1	4	2	2	2	1	36.5
1	1	3	2	1	36.7	3	1	3	2	1	36.9	5	1	3	2	1	37.1	2	2	3	2	1	37.3	4	2	3	2	1	36.8
1	1	4	2	1	36.4	3	1	4	2	1	36.5	5	1	4	2	1	37.1	2	2	4	2	1	36.7	4	2	4	2	1	36.4
1	1	5	2	1	36.2	3	1	5	2	1	36.3	5	1	5	2	1	36.6	2	2	5	2	1	36	4	2	5	2	1	36.2
1	1	6	2	1	36.5	3	1	6	2	1	36.8	5	1	6	2	1	37.3	2	2	6	2	1	37	4	2	6	2	1	36.7
1	1	7	2	1	35.7	3	1	7	2	1	36.2	5	1	7	2	1	36.3	2	2	7	2	1	36.2	4	2	7	2	1	36
1	1	8	2	1	37.2	3	1	8	2	1	37.1	5	1	8	2	1	37.5	2	2	8	2	1	37.6	4	2	8	2	1	36.8
1	1	9	2	1	35.9	3	1	9	2	1	36.1	5	1	9	2	1	36.6	2	2	9	2	1	36.6	4	2	9	2	1	36
1	1	10	2	1	36.8	3	1	10	2	1	36.8	5	1	10	2	1	36.8	2	2	10	2	1	37.1	4	2	10	2	1	36.9
1	1	1	1	2	35.7	3	1	1	1	2	36.2	5	1	1	1	2	35.8	2	2	1	1	2	35.8	4	2	1	1	2	35.8

(*Continued on next page*)

TABLE A1 Data of the Ear Thermometers *(Continued)*

Ne	Te	Pe	Se	Re	Temp.	Ne	Te	Pe	Se	Re	Temp.	Ne	Te	Pe	Se	Re	Temp.	Ne	Te	Pe	Se	Re	Temp.	Ne	Te	Pe	Se	Re	Temp.
1	1	2	1	2	36.3	3	1	2	1	2	36.4	5	1	2	1	2	36.4	2	2	2	1	2	36.3	4	2	2	1	2	36.5
1	1	3	1	2	36.8	3	1	3	1	2	37.2	5	1	3	1	2	36.7	2	2	3	1	2	37.4	4	2	3	1	2	37
1	1	4	1	2	36.1	3	1	4	1	2	36.6	5	1	4	1	2	36.9	2	2	4	1	2	36.8	4	2	4	1	2	36.6
1	1	5	1	2	36.6	3	1	5	1	2	36.4	5	1	5	1	2	36.3	2	2	5	1	2	36.5	4	2	5	1	2	36.5
1	1	6	1	2	36.9	3	1	6	1	2	37	5	1	6	1	2	37.1	2	2	6	1	2	37	4	2	6	1	2	37.1
1	1	7	1	2	36.5	3	1	7	1	2	36.6	5	1	7	1	2	36.7	2	2	7	1	2	36.6	4	2	7	1	2	36.7
1	1	8	1	2	37.3	3	1	8	1	2	37.5	5	1	8	1	2	37	2	2	8	1	2	37.5	4	2	8	1	2	37.3
1	1	9	1	2	36.6	3	1	9	1	2	36.7	5	1	9	1	2	36.8	2	2	9	1	2	36.5	4	2	9	1	2	36.4
1	1	10	1	2	37.2	3	1	10	1	2	37.1	5	1	10	1	2	37	2	2	10	1	2	37.2	4	2	10	1	2	37.1
1	1	1	2	2	35.9	3	1	1	2	2	36.1	5	1	1	2	2	36.1	2	2	1	2	2	36.1	4	2	1	2	2	35.9
1	1	2	2	2	36.2	3	1	2	2	2	36.1	5	1	2	2	2	36.3	2	2	2	2	2	36.2	4	2	2	2	2	36.4
1	1	3	2	2	36.7	3	1	3	2	2	37.3	5	1	3	2	2	37.4	2	2	3	2	2	37.3	4	2	3	2	2	36.9
1	1	4	2	2	36.7	3	1	4	2	2	36.7	5	1	4	2	2	37	2	2	4	2	2	36.9	4	2	4	2	2	36.7
1	1	5	2	2	36.5	3	1	5	2	2	36.6	5	1	5	2	2	36.2	2	2	5	2	2	36.4	4	2	5	2	2	36.4
1	1	6	2	2	36.9	3	1	6	2	2	37.2	5	1	6	2	2	37.2	2	2	6	2	2	37.5	4	2	6	2	2	36.9
1	1	7	2	2	36.4	3	1	7	2	2	36.5	5	1	7	2	2	36.4	2	2	7	2	2	36.3	4	2	7	2	2	36.4
1	1	8	2	2	37.1	3	1	8	2	2	37.3	5	1	8	2	2	37.6	2	2	8	2	2	37.4	4	2	8	2	2	37.1
1	1	9	2	2	36.4	3	1	9	2	2	36.5	5	1	9	2	2	36.7	2	2	9	2	2	36.4	4	2	9	2	2	36
1	1	10	2	2	37.1	3	1	10	2	2	36.9	5	1	10	2	2	37.1	2	2	10	2	2	37.3	4	2	10	2	2	37.1
2	1	1	1	1	36.1	4	1	1	1	1	35.5	1	2	1	1	1	35.4	3	2	1	1	1	36.1	5	2	1	1	1	36
2	1	2	1	1	36.5	4	1	2	1	1	36.5	1	2	2	1	1	36.5	3	2	2	1	1	36.4	5	2	2	1	1	36.8
2	1	3	1	1	37.4	4	1	3	1	1	36.8	1	2	3	1	1	37.1	3	2	3	1	1	37.1	5	2	3	1	1	37.5
2	1	4	1	1	36.6	4	1	4	1	1	36.3	1	2	4	1	1	36.7	3	2	4	1	1	36.8	5	2	4	1	1	37
2	1	5	1	1	36.3	4	1	5	1	1	36.1	1	2	5	1	1	36.5	3	2	5	1	1	36.5	5	2	5	1	1	36.5
2	1	6	1	1	36.9	4	1	6	1	1	36.6	1	2	6	1	1	36.7	3	2	6	1	1	36.8	5	2	6	1	1	37.1
2	1	7	1	1	36.3	4	1	7	1	1	36.5	1	2	7	1	1	36.2	3	2	7	1	1	36.5	5	2	7	1	1	36.6
2	1	8	1	1	37.3	4	1	8	1	1	37.1	1	2	8	1	1	37.2	3	2	8	1	1	37.5	5	2	8	1	1	37.3
2	1	9	1	1	36.7	4	1	9	1	1	36.1	1	2	9	1	1	36.2	3	2	9	1	1	36.8	5	2	9	1	1	36.8
2	1	10	1	1	36.7	4	1	10	1	1	36.8	1	2	10	1	1	36.8	3	2	10	1	1	36.8	5	2	10	1	1	37.1
2	1	1	2	1	35.9	4	1	1	2	1	35.7	1	2	1	2	1	35.9	3	2	1	2	1	36	5	2	1	2	1	36.3
2	1	2	2	1	36.4	4	1	2	2	1	36.4	1	2	2	2	1	36.3	3	2	2	2	1	36.3	5	2	2	2	1	36.7
2	1	3	2	1	37.1	4	1	3	2	1	36.9	1	2	3	2	1	37.1	3	2	3	2	1	37.1	5	2	3	2	1	37.5
2	1	4	2	1	36.7	4	1	4	2	1	36.3	1	2	4	2	1	36.5	3	2	4	2	1	36.6	5	2	4	2	1	37
2	1	5	2	1	36.2	4	1	5	2	1	36.3	1	2	5	2	1	36.5	3	2	5	2	1	36.4	5	2	5	2	1	36.5
2	1	6	2	1	37	4	1	6	2	1	36.6	1	2	6	2	1	37	3	2	6	2	1	37.1	5	2	6	2	1	37.4
2	1	7	2	1	36.3	4	1	7	2	1	36.1	1	2	7	2	1	36.2	3	2	7	2	1	36.4	5	2	7	2	1	36.4
2	1	8	2	1	37.4	4	1	8	2	1	36.7	1	2	8	2	1	37.4	3	2	8	2	1	37.4	5	2	8	2	1	37.5
2	1	9	2	1	36.1	4	1	9	2	1	35.8	1	2	9	2	1	36.2	3	2	9	2	1	36.5	5	2	9	2	1	36.8
2	1	10	2	1	36.9	4	1	10	2	1	37	1	2	10	2	1	37.2	3	2	10	2	1	37	5	2	10	2	1	37.1
2	1	1	1	2	35.6	4	1	1	1	2	35.7	1	2	1	1	2	35.7	3	2	1	1	2	36.3	5	2	1	1	2	36
2	1	2	1	2	36.5	4	1	2	1	2	36	1	2	2	1	2	36.3	3	2	2	1	2	36.3	5	2	2	1	2	36.3
2	1	3	1	2	37.3	4	1	3	1	2	37	1	2	3	1	2	36.9	3	2	3	1	2	37.4	5	2	3	1	2	37.4
2	1	4	1	2	37	4	1	4	1	2	36.1	1	2	4	1	2	36.8	3	2	4	1	2	36.8	5	2	4	1	2	36.9
2	1	5	1	2	36.4	4	1	5	1	2	36.4	1	2	5	1	2	36.6	3	2	5	1	2	36.5	5	2	5	1	2	36.5
2	1	6	1	2	36.9	4	1	6	1	2	36.8	1	2	6	1	2	36.6	3	2	6	1	2	36.8	5	2	6	1	2	37.1
2	1	7	1	2	36.5	4	1	7	1	2	36.6	1	2	7	1	2	36.5	3	2	7	1	2	36.4	5	2	7	1	2	36.5
2	1	8	1	2	37.2	4	1	8	1	2	37.2	1	2	8	1	2	37.3	3	2	8	1	2	37	5	2	8	1	2	37.1
2	1	9	1	2	36.7	4	1	9	1	2	36.3	1	2	9	1	2	36.6	3	2	9	1	2	36.8	5	2	9	1	2	37
2	1	10	1	2	37	4	1	10	1	2	37.1	1	2	10	1	2	37.3	3	2	10	1	2	37.1	5	2	10	1	2	37.1
2	1	1	2	2	36	4	1	1	2	2	35.7	1	2	1	2	2	35.9	3	2	1	2	2	36.1	5	2	1	2	2	36.3
2	1	2	2	2	36.1	4	1	2	2	2	36.3	1	2	2	2	2	36.1	3	2	2	2	2	36.2	5	2	2	2	2	36.2

(*Continued on next page*)

TABLE A1 Data of the Ear Thermometers *(Continued)*

Ne	Te	Pe	Se	Re	Temp.	Ne	Te	Pe	Se	Re	Temp.	Ne	Te	Pe	Se	Re	Temp.	Ne	Te	Pe	Se	Re	Temp.	Ne	Te	Pe	Se	Re	Temp.
2	1	3	2	2	37.3	4	1	3	2	2	36.9	1	2	3	2	2	36.9	3	2	3	2	2	37.2	5	2	3	2	2	37.5
2	1	4	2	2	37	4	1	4	2	2	36.7	1	2	4	2	2	36.7	3	2	4	2	2	36.5	5	2	4	2	2	37
2	1	5	2	2	36.3	4	1	5	2	2	36.4	1	2	5	2	2	36.4	3	2	5	2	2	36.5	5	2	5	2	2	36.6
2	1	6	2	2	37	4	1	6	2	2	36.7	1	2	6	2	2	36.9	3	2	6	2	2	37.2	5	2	6	2	2	37.4
2	1	7	2	2	36.3	4	1	7	2	2	36.3	1	2	7	2	2	36.3	3	2	7	2	2	36.3	5	2	7	2	2	36.5
2	1	8	2	2	37.3	4	1	8	2	2	36.7	1	2	8	2	2	36.9	3	2	8	2	2	37.2	5	2	8	2	2	37.4
2	1	9	2	2	36.4	4	1	9	2	2	36.4	1	2	9	2	2	36.3	3	2	9	2	2	36.4	5	2	9	2	2	36.8
2	1	10	2	2	37.2	4	1	10	2	2	37.2	1	2	10	2	2	37.1	3	2	10	2	2	37	5	2	10	2	2	37

TABLE A2 Data of the Rectal Thermometer

Pr	Rr	Temp.	Pr	Rr	Tr	Pr	Rr	Temp.	Pr	Rr	Tr	Pr	Rr	Temp.
1	1	36.6	5	1	37.3	9	1	37.2	3	2	37.6	7	2	37
2	1	36.9	6	1	37.1	10	1	37.5	4	2	36.8	8	2	37.6
3	1	37.4	7	1	36.8	1	2	36.4	5	2	37	9	2	36.9
4	1	37	8	1	37.8	2	2	36.8	6	2	37	10	2	37.3