sciendo CzOTO 2020, volume 2, issue 1, pp. 275-284

REPEATABILITY AND REPRODUCIBILITY STUDIES FOR NON-REPLICABLE TESTS

doi: 10.2478/czoto-2020-0034
Date of submission of the article to the Editor: 20/11/2019
Date of acceptance of the article by the Editor: 18/12/2019

Pavlína Mikulová[1] – *orcid id: 0000-0002-2035-029X*
Jiří Plura[1] – *orcid id: 0000-0002-4739-0208*
Krzysztof Knop[2] – *orcid id: 0000-0003-0842-9584*
[1]VSB – Technical University of Ostrava, **Czech Republic**
[2]Czestochowa University of Technology, **Poland**

Abstract: The paper presents several approaches to gauge repeatability and reproducibility (GRR) analysis regarding non-replicable measurements. Measurement systems have to deal with processes in which, by the nature of the measured object or by the type of measurement itself, measurements are not repeatable. In these cases, each sample unit can be measured only once. Such situations are referred as non-replicable measurement systems. The aim of the paper is to map out the current approaches being used in GRR analysis in various cases of non-replicable tests and compare each other in order to find out the suitable use of analysis application. Approaches used are subject to critical analysis so that its review can serve a useful base for analysis of different non-replicable tests. At present, it is desirable to bring the improving actions in order to obtain the results of high quality from such kind of measurements. Since different non-replicable tests can measure a different quality characteristic, it is valuable to bring the appropriate designs for various tests. Subsequently, this review will serve an outline how to proceed in analyzing the results obtained by non-replicable tests. Specifically, GRR analysis works with two known designs named as “Crossed” and “Nested” design, which statistical software normally use. Doubtfully, crossed design is suggested to use at certain cases and nested at other specific cases. This is assessed and improving actions designed.
Keywords: non-replicable measurement, repeatability, reproducibility, analysis

1. INTRODUCTION

Measurement system analysis (MSA) serves an assessment of a measurement system ability to detect meaningful differences in process variables. Measurement systems refer to the collection of operators, gauges, procedures, software and operations, which are necessary for collecting the data of high quality and not to provide distorted information leading to wrong decisions. Thus, to establish an effective measurement system is an essential step for quality improvement (He et al., 2003). Moreover, MSA is even strictly required in automotive industry (IATF 16949, 2016).

The most used analysis of measurement system, which makes possible to evaluate repeatability and reproducibility of measurement system is GRR analysis. Generally, %GRR evaluates the suitability of measurement system for process control, capability studies, or statistical studies (Douglas and Keith, 2002).
Standardly in GRR analysis, the readings of measurement systems can be replicated for each part. In statistics, a replicable system is one where the same or different appraisers may measure any given part multiple times and the result obtained falls within a predictable range of values (Miner, 2016). However, not all measurement systems have this characteristic. This is the case of non-replicable tests whose general categories are:

- destructive measurement systems (impact tests, welding tests, tensile tests etc.),
- systems where the part changes during measurement process (leak tests with qualitative data, torque tests etc.).

Herewith, the complex problem occurs in that once the measurement is obtained for a particular part, this part is no longer available for additional measurements with the same or different appraiser. Hence, it is difficult to measure the repeatability and reproducibility. In order to assess the non-replicable measurement systems, few studies have been done and are introduced in chapter 4.

2. GAGE REPEATABILITY AND REPRODUCIBILITY ANALYSIS

Since measurement systems are used for making decisions about products, processes, or services, an analytic conclusion about the measurement system is necessary. The transition from enumerative to analytic results requires subject matter knowledge and expertise to assure that all expected measurement sources of variation are considered in the design.
In general cases, two major sources of measurement system variability are:

- Repeatability – occurs during the repeated measurements under the same conditions (same sample, gage, method, operator, etc.).
- Reproducibility – the variation in the averages of the measurements made by different operators using the same gage when measuring the same characteristic.

The GRR study determines how much variation is due to the measurement method and how much is due to the appraisers. Various approaches can be used to evaluate repeatability and reproducibility; the most used are Average and Range method (A&R) and ANOVA. A&R method allows the measurement system's variation to be decomposed into two separate components, repeatability and reproducibility. However, variation due to the interaction between the appraiser and the part/gage is not accounted for in the analysis. There is another method of analysis, which is able to identify the operator-part interactions in addition. It is the analysis of variance, ANOVA (MSA Work Group, 2010).

3. GRR STUDIES FOR NON-REPLICABLE TESTS

In standard GRR studies, the underlying assumptions are the parts as well as operators are independent, random and multiple measurements are possible. On the contrary, when the gauges or tests are of non-replicable nature, the samples can only be measured once. This imposes a problem in the assumptions of the standard GRR analysis (Wilson, 2007). Hence, for assessing the non-replicable measurement sys-

tems, few approaches will be introduced in this chapter and two types of GRR design are described in next section.

3.1. Designs of assessing non-replicable measurement systems

In terms of crossed GRR design (Fig. 1), conducting GRR analysis requires making the critical assumption in the form of identifying the batches of homogeneous parts (McNeese, 2016). This means measuring any part of the batch for the same characteristic, its results should be similar with a small sign of variation due to repeatability, unlike the crossed design of replicable measurement systems, where the same part is really being measured more than once (repeatedly). In reality, the variation is due to the measurement system repeatability in both cases. The standard crossed design model is (Meulen, 2009):

$$y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk.} \quad (1)$$

where y_{ijk} is denoted to the data, a_i is for parts, b_j is for operators, ε_{ijk} represents an error term and k indexes replications.

In nested GRR design (Fig. 2), the difference occurs at the stage when determining if there is a sufficient amount of parts from each batch for each operator. If there are not enough parts in each batch for each operator to measure multiple times, the nested (hierarchical) design is used. Thus, the samples are nested within the operators. It has the property that the levels of at least one treatment (factor) is contained or appears with only one level of another treatment (factor). To emphasize the fact that the parts are different for each operator, the parts have consecutive numbers as seen in Fig. 2. The samples are assigned at random to the operators.

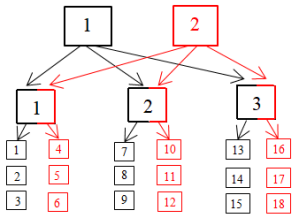

Fig. 1. Crossed GRR design

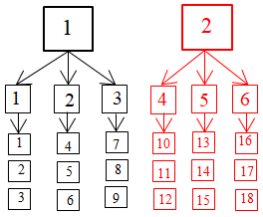

Fig. 2. Nested GRR design

The nested arrangement shows no interactions as the samples are not criss-crossed with the operators. In theory, these measurements are nested within sample and within operator, but they are confounded with the error term treated as the gauge repeatability $\varepsilon_{k(ij)}$, where the number of measurements is represented by n, k=1,…, n. The nested design model for the non-replicable GRR becomes:

$$y_{ijk} = \mu + a_i + b_{j(i)} + \varepsilon_{k(ij)} \quad (2)$$

The procedure how to properly select the right GRR design is depicted in Fig. 3.

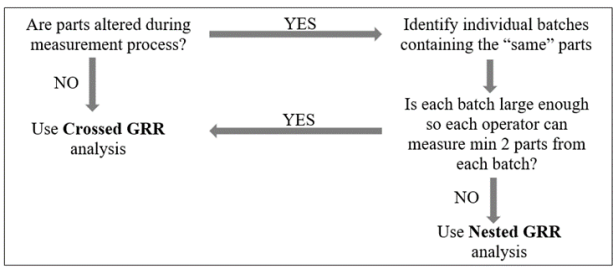

Fig. 3. Selection of GRR design (McNeese, 2016)

The detailed application of nested design procedure is described by M. P. Wilson (Wilson, 2007).

3.2. Approaches to non-replicable measurement systems analysis

This section introduces several approaches to GRR analysis of measurement systems which cannot be repeated. These approaches have been applied to particular experiments described in chapter 4.

Patterning of variation: Generally, non-replicable measurements are measurements, for which either of the following two conditions does not hold:

- × Temporal stability: the real value of the sample does not change in time.
- × Robustness against measurement: objects are not affected when measured.

In performing GRR analysis for non-replicable measurements there is no structural solution to assess it, a number of approaches works in some cases. According to (Meulen et al., 2009), either of the following conditions must hold:

- ✓ Patterned temporal variation (PTV): the variation over time of each sample follows a certain pattern.
- ✓ Patterned object variation (POV): the variation across samples follows a certain pattern.

Representativeness of alternative objects: The appraiser must know a historical representativeness of alternative objects in order to use pattern. This means he knows the measurement error homogeneity (identical for all units that are measured) is extended to the alternative objects, i.e. such pattern can be used if the differences in the measured characteristic are known and fixed during measurement (De Mast and Trip, 2005).

Assessing the repeatability: William D. Kappele (Kappele, 2010), an expert in analyzing the designs of experiments, claims there are two options for assessing repeatability of non-replicable tests:

- ✓ To find a replicable test correlating with the results of the non-replicable test and use it instead.
- ✓ To collect parts which are so similar in the property to be measured that it is assumed they have identical measurements.

Methodologies characterized in this chapter have been applied in various non-replicable tests which are the subject of following chapter.

4. METHODOLOGIES USED IN VARIOUS NON-REPLICABLE MEASUREMENT SYSTEMS ANALYSIS

Following sections introduce the specific cases of non-replicable measurements and approaches of GRR analysis applied to them. In chapter 5, these approaches will be compared and general methodology designed.

4.1. Adjustment of GRR estimation

Company producing linear and switch-mode chargers for mobile terminals measured the pulling force of chargers' cover after ultrasonic welding (Han and He, 2007). The machine recorded maximal force required to pull cover of chargers. The measurement system was verified in terms of the chargers were good or bad due to pulling force. Since the homogenous batch size was large enough to assign at least two parts from each batch to the operator, the crossed design was adopted. Because the same operator measured the same response more than once, and nothing else in changing,

this variation in measurements was assumed that it belonged to the gauge. The P/T ratio was used to measure the suitability of gauge in order to make pass/fail decision to a specification. Generally, the %P/T with two-sided specification limits is defined. For such destructive test of pulling force (the larger the better characteristic), only one specification limit (LSL) was considered. Hence, the P/T ratio was modified (Han and He, 2007):

$$\%P/T = \frac{3 * \sigma_{MSE}}{\mu - LSL}\ x\ 100\% \tag{3}$$

where μ means process mean. The variation range of measurement system becomes $6\sigma_{MSE}$, in this case $3\sigma_{MSE}$, with the probability of 99,73% when the process follows a normal distribution. There is a small %P/T when process mean is far enough from LSL, which means a small probability to judge a good product as a bad one or vice versa. Generally, the %GRR is defined:

$$\%GRR = \frac{\sigma_{MSE}}{\sqrt{\sigma_{MSE}^2 + \sigma_P^2}}\ x\ 100\% \tag{4}$$

where σ_p is the variation due to parts. Through designing and analyzing the experiment of pulling force testing by crossed GRR design, the conclusion is the test was appropriate for process control.

4.2. Application of nested design

The impact strength of steel was measured. The impact test was used to measure the energy required to break a notched metal bar of specified geometry (Douglas and Keith, 2002). The test samples were prepared from ingots randomly selected from a wider range of ingots. It was assumed the samples created from the same part of ingot were more homogeneous than samples created from different parts of ingot. In this case, only three samples could be prepared from each ingot because of its size. Due to the destructive nature of the test and small batch size, it was not possible for each operator to measure each ingot multiple times. Therefore, the nested GRR study was used.

A measurement system found inadequate mostly due to repeatability error and this should have raised the question about the homogenous batch assumption. The graphical results indicated the operator 1 reached higher values compared with operator 2 and 3. However, it could not be automatically judged it had been the operator 1's blame. Instead, the appropriate option was to design an additional study to assess procedurally what the operators should have done differently.

4.3. Patterning of variation by using PTV and POV

Two cases of non-replicable measurement systems were evaluated by F. Meulen (Meulen et al., 2009). In section 3.2 Patterning of variation, the conditions of PTV and POV have been introduced. The idea of the following practical experiments was to fit the model for the PTV or POV, and correct the data for it that leads to a smaller estimation of measurement spread. Measurement spread represents the variation within rows of standard GRR study layout, see (Montgomery, 2005), and exploits replications for estimating the variance components. The following two cases are introduced:

A. Temporal stability is violated, PTV holds: samples vary over time, but the variation over time follows a certain pattern that can be modeled.

B. Robustness against measurement is violated, POV holds: samples change during measurement, but their variation follows a pattern that can be modeled.

The case A is explained by measuring the core temperature of a food product. The food product was baked until its core reached the temperature of 80°C. It was measured by inserting a digital thermometer into the product. Since the product was cooling down quite rapidly (about 1°C per minute), the repeated measurements would have confounded measurement spread with variation in the product's true core temperature (temporal stability was violated here). It was assumed the measurements of part were done at *n* times instances (each operator measured each part at each time instant). However, at any given time instant, one operator could only measure the one food specimen. Nonetheless, an experimental design, in which each operator measured (though not always the same part) at each time instant, was possible to create. This was able to easily accomplish by a Latin-square design. The examples of a design satisfying this requirements is the Latin-square design given by

	t_1	t_2	t_3	t_4	t_5	t_6
sample 1	A	B	C	A	C	B
sample 2	B	C	A	C	B	A
sample 3	C	A	B	B	A	C

Any permutation of the columns of this design can yield a design that suits the specific purpose. In the shown design, each operator measured each part twice (n=2) and at all-time instants. ANOVA test showed the interaction term could be dropped from the model. The estimated measurement spread was decreased by refining the standard model (1) that led to decreased % of measurement variance due to repeatability. It was proven the temperature decreased linearly in time for each sample. Therefore, the times of measurements $t_1,\ldots t_k$ must have been denoted into experiment. Then, y_{ijk} denoted the temperature of the sample *i* at time t_k and fitted mixed analysis of covariance model has been proposed (Meulen et al., 2009):

$$y_{ik} = \mu + \mathrm{a_i} + \gamma(t_k - 150) + \varepsilon_{ik} \qquad (5)$$

where a_i is a random sample effect and γ is the fixed position effect. This approach well-worked since the analysis clearly showed the estimated measurement spread was in fact much lower compared with the use of standard model.

The case B is explained by measuring the shrinkage of carpet tiles. In order to stress test of the carpet tile and its performance with respect to shrinkage, it was exposed to extreme temperatures and moisture conditions during the measurement. The GRR study of such stress test refers to the consistency of the test results that would be obtained, if the tests were performed multiple times on the same tile. However, these test were irreversible, therefore the measurement was nonreplicable (violated condition of robustness against measurement). The analysis of this example was similar to the case A. Also, the proportion of measurement variance due to repeatability decreased here.

4.4. Patterning by historical estimate of representativeness of alternative objects

In order to measure the strength of biscuits, pressure was exerted onto them. The pressure had slowly increased until the biscuit broke. The pressure at which the biscuit broke was the measured strength (De Mast and Trip, 2005). Here, the condition of POV held. The operator used a historical estimate of representativeness of alternative objects for the pattern. The differences in the strengths of biscuits which were

taken from different positions of the oven belt were fixed. In *I* time instants, the operator can select *JK* biscuits from different positions. Each of *J* operators measured *K* of these biscuits. The measurements were modeled (De Mast and Trip, 2005):

$$y_{ijk} - \gamma_{jk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk} \quad (6)$$

where γ_{jk} is assumed known difference in strength (compared with the overall mean) of biscuits from position *j*, *k*. Within-part variation among biscuits, that was not accounted by the γ_{jk}, was confounded with error variance σ_e^2. Given homogeneity assumption was only met to a certain extent and confounding problem has not been solved entirely.

4.5. Approaches to repeatability expression

In this section, the following two cases are introduced (Kappele, 2010):

A. Use of replicable test correlating with the results of non-replicable test instead.
B. Parts are so similar in the property to be measured that it is assumed they are identically measured.

The case A is represented by measuring the thickness of copper plate in electronic device. Printed circuit boards used for electronic devices had copper traces plated onto them to act as conductors. The thickness of this plating could be measured by cutting a piece from the board (through the middle of a hole). The copper thickness could then be measured directly, however, this would have left the board destroyed and unusable. There was the correlated measuring way existed for printed circuit board manufacturers – a digital meter was used to measure the copper thickness in the hole. A small probe was inserted into the hole and the copper thickness dimension was displayed. This measurement system was non-destructive, the board was not affected by the measurement and could still be sold to a customer. Although, this was a better option compared with the original destructive test, some precautions were necessary to hold:

- to compare the readings from the meter with the cross section of circuit board to make sure they agree,
- to inspect the methods agree, not only assume that they correlate and result in the same measurement (i.e. to determine the best technique for using the meter to obtain agreement).

An additional uncertainty occurred in the correlation with the original method. At this stage, it would have been desirable a deeper statistical analysis.

The case B is represented by measuring the penetration of bullet-proof vests. The bullet-proof vest was shot and observed if the bullet penetrated it. Whether or not it did, the vest was irreparably damaged. For this kind of destructive test, there was no nondestructive substitute existed. Test vests were made from the same roll of material using such material that was as close on the roll as possible to avoid within lot variation. Additionally, it was verified the variation in the material met requirements and the vests were sewn together as identically as possible including stitch placement and spacing. This required care above normal production procedures. The more homogenous preparation, the more accurate the results were.

5. RESULTS OF APPROACHES APPLIED IN NON-REPLICABLE TESTS

In this chapter, the measurement systems are summarized and applied approaches compared. Based on the critical analysis of the approaches, its applications will be

generalized in order to gain the uniform methodology how to proceed in cases of non-replicable measurement systems analysis.

5.1. Comparison of approaches used in specific non-replicable measurement systems analyzes

The particular measurements and its analyzes characterized in chapter 4 are compared in Table 1.

Table 1
Comparison of study cases analyzes

	Measurement	Conditions of non-replicable measurement systems	GRR design	Additional analysis
1	Pulling force of chargers	✓ Sufficient number of homogenous parts for multiple measurements	Crossed	not required
2	Impact strength of steel	× Insufficient number of homogeneous parts for multiple measurements	Nested	required
3	Core temperature of food product	× Temporal stability is violated ✓ PTV holds	Crossed (Latin-square)	not required
	Shrinkage of carpet tile	× Robustness against measurement is violated ✓ POV holds	Crossed (Latin-square)	not required
4	Strength of biscuits	× Insufficient homogeneity of parts ✓ Representativeness of alternative objects holds ✓ Temporal stability of objects holds ✓ POV holds	Crossed	required
5	Thickness of copper plate	× Insufficient number of homogeneous parts for multiple measurements, ✓ Substitution of non-replicable test by replicable test correlating with the non-replicable test results	Crossed	required
	Penetration of bullet-proof vests	✓ Sufficient number of homogenous parts for multiple measurements	Crossed	not required

Source: own research.

Table 1 summarizes the approaches to analysis of the range of introduced non-replicable tests. As seen in three cases, there is an additional analysis required what is the sign of insufficient or inappropriate use of GRR analysis approach. The particular approaches have been compared and will be generalized in next section.

5.2. Generalization of applied approaches

The above approaches show that there are several aspects need to emphasize in non-replicable measurements:

- The variation range represents the combined variability from measurement system repeatability and the within batch variability (Han and He, 2007). If a non-replicable measurement system is found inadequate due to mostly the repeatability error, it cannot be directly concluded the system is not valid.
- Nested design is used when homogenous batch sizes are limited and each batch can only be tested multiple times by one operator. Anyway, repeatability variation

is indistinguishable from the within-batch variation. Hence, the design is considered to be merely instrumental in results analysis.

- The R-chart identifies whether certain operators had difficulty to test samples constantly as well as identifies specific part that is not homogenous. The $\bar{X}$-chart shows if the average of one operator appear higher/lower than those of others (Mikulová and Plura, 2019). If necessary, an additional study should be done to assess what the operators might do differently (to inspect how the samples were prepared or fixtured into device).
- Use of Latin square designs can be obtained from standard Latin squares via permutation of rows, columns and labels, see (Neter et al., 1985).
- To replace replications with measurements of different samples leads to confounding measurement variation with between-samples within part variation. Methodology for reducing the resulting estimation of measurement spread lies in exploiting patterns Patterned Temporal Variation and Patterned Object Variation. To estimate measurement spread, an appraiser needs multiple measurements of a single part. This can be done by meeting the assumptions of homogeneity not to confound the measurement spread with other sources of variation. The homogenization of parts can be reached by several approaches characterized in (Miner, 2016).
- If the nondestructive (replicable) test corresponding to destructive (non-replicable) is determined, the standard GRR study can be performed using this one instead. If using the nondestructive substitute, there can an additional uncertainty occur in the correlation with the original method. Then, the deeper statistics is helpful.
- Use of patterns is appropriate when some of the assumptions do not hold.

Meeting the assumptions (consistency, temporal stability of objects, robustness against measurement, representativeness of alternative objects) is the essential step to ensure a well-working non-replicable measurement system.

The proper determination of variance components is the further important point. In general cases, a complete factorial design with replications includes *I* parts and *J* operators, each operator measures each of *I* parts *n* times. Considering GRR analysis for non-replicable measurement systems, where parts with repeated measurements are replaced by batches of similar samples, slight changes are made in replacing the part parameter "*a*" by batch parameter. Thus, the study consists of *J* operators and *I* batches of homogenous parts (each batch formed by *J* x *n* parts). The task is to measure the variability between different parts and operators by assuming that a_i, b_j, $(ab)_{ij}$, e_{ijk} are mutually independent and normally distributed random variables with zero means and particular variances σ^2 (Jarošová, 2018).

6. DISCUSSION AND CONCLUSION

The usefulness of Latin-square design in cases of non-replicable measurement systems analysis has been proven. From Table 1, it is obvious there are three cases which require an additional analysis. In all, the problem of homogeneity violation occurs. It is concluded the estimate of measurement spread can be done by meeting the assumption that parts are invariable in time and the measuring do not affect them. The homogeneity assumptions are based on the idea, that either the effects of disturbing sources of variation are negligible, or that the results can be corrected for their influence. Naturally, the measurement system can be thought valid if the assumptions

are met. Though the option of crossed design is preferred by the majority, because nested is somehow deceptive due to samples factor is still crossed with the operators factor, the use of nested design has its justification. Nested design is recommended to use in cases that, firstly, the number of measured parts is not sufficient, secondly, some of measurement conditions stated above hold and, thirdly, following the patterns is possible. By holding these conditions, the credible results about measurement system suitability can be reached.

ACKNOWLEDGEMENTS

The work has been supported by the specific university research of the Ministry of Education, Youth and Sports of the Czech Republic No. SP2019/62, from ERDF „A Research Platform focused on Industry 4.0 and Robotics in Ostrava“, No. CZ.02.1.01/0.0/0.0/17_049/0008425 and elaborated in the framework of the grant programme „Support for Science and Research in the Moravia-Silesia Region 2018" (RRC/10/2018), financed from the budget of the Moravian-Silesian Region.

REFERENCES

De Mast, J. and Trip, A., 2005. *Gauge R&R Studies for Destructive Measurements,* Journal of Quality Technology*,* 37(1), 40–49.

Douglas, G., Keith, M. B., 2002. *Measurement System Analysis and Destructive Testing.* ASQ Six Sigma Forum Magazine, 1, 16-19.

Han, Y., He, Z., 2007. *An Applied Study of Destructive*, 2nd IEEE Conference on Industrial Electronics and Applications, Tianjin, China, 30-34.

He, Z., Sheng, J., Shi, L., 2003. *Application of Measurement System R&R Analysis in Quality Improvement,* Journal of Industrial Engineering, 6, 62-66.

IATF 16949, 2016. *Norma pro systém managementu kvality v automobilovém průmyslu*, Česká společnost pro jakost, z.s. Praha, Česká republika.

Jarošová, E., 2018. *Destructive R&R Study – Evaluation Problems,* 17th Conference on Applied Mathematics APLIMAT 2018, Bratislava, 535-544.

Kappele, W.D., Raffaldi, J., 2010. *Gage R&R for Destructive Measurement Systems,* Quality Magazine, 5, 32-34.

Meulen, F., Koning, H., De Mast, J., 2009. *Nonrepetable GRR Studies Assuming Temporal or Patterned Object Variation,* Journal of Quality Technology, 4, 41.

Mikulová, P., Plura, J., 2019. *Detailed analysis of GRR study results and their visualization,* 13th International Conference Quality Production Improvement QPI, Zaborze, Poland.

Miner, G., 2016. *Intro to MSA of Continuous Data – Part 6: R&R for Non-Replicable Measurements,* In Quality Forum, Miner’s blogs.

Montgomery, D. C., 2005. *Design and Analysis of Experiments*, 6th edition, New York, NY: John Wiley and Sons, Inc.

MSA Group, 2010. *Measurement System Analysis, Reference Manual,* 4th edition, Chrysler Group LLC, Ford Motors Corporation.

Neter, J., Wasserman, W., Kutner, M. H., 1985. *Applied Linear Statistical Models*, 2nd edition, Richard D. Irwin, Inc.

Wilson, M. P., 2007. *Gauge R&R Studies for Destructive and Non-destructive Testing*, Advanced Systems Consultants, Scottsdale-Arizona,USA.