**International Refereed Journal of Engineering and Science (IRJES)**
*ISSN (Online) 2319-183X, (Print) 2319-1821*
*Volume 6, Issue 1 (January 2017), PP.08-16*

# On Confidence Intervals Construction for Measurement System Capability Indicators.

Daniela F. Dianda[1],Jose A. Pagura[1],Nicolás M. Ballarini[1]
[1]*Instituto De Investigaciones Teóricas Y Aplicadas De La Escuela De Estadística, Fac. De Ciencias Económicas Y Estadística. Universidad Nacional De Rosario. Argentina.*

**Abstract:** There are many criteria that have been proposed to determine the capability of a measurement system, all based on estimates of variance components. Some of them are the Precision to Tolerance Ratio, the Signal to Noise Ratio and the probabilities of misclassification.
For most of these indicators, there are no exact confidence intervals, since the exact distributions of the point estimators are not known. In such situations, two approaches are widely used to obtain approximate confidence intervals: the Modified Large Samples (MLS) methods initially proposed by Graybill and Wang, and the construction of Generalized Confidence Intervals (GCI) introduced by Weerahandi.
In this work we focus on the construction of the confidence intervals by the generalized approach in the context of Gauge repeatability and reproducibility studies. Since GCI are obtained by simulation procedures, we analyze the effect of the number of simulations on the variability of the confidence limits as well as the effect of the size of the experiment designed to collect data on the precision of the estimates. Both studies allowed deriving some practical implementation guidelinesin the use of the GCI approach.
We finally present a real case study in which this technique was applied to evaluate the capability of a destructive measurement system.
**Keywords:**Gauge R&R studies, Generalized confidence intervals, Destructive measurements.

## I. INTRODUCTION

Repeatability and reproducibility (R&R) studies are widely used in current activities of quality improvement, since they allow deciding whether the system is able to produce measurements that reflect the true behavior of the process, by quantifying its variability and comparing it with the total process variation. Many criteria have been proposed to make that comparison, all of them involving certain indicators which must be estimated from the data. Some of these indicators are the Precision to Tolerance Ratio (PTR), the %R&R indicator, the Signal to Noise Ratio (SNR) and the probabilities of misclassification. In general, estimations of these indicators are based on estimations of the variance components of the model proposed by design.

The importance of studying the capability of measurement systems relies on the fact that any quality improvement activity involves analyzing information about the process, and that information comes, in general, from measuring the variables of interest. This activity introduces an additional source of variation to data that needs to be considered. Inadequate measurement system could add enough variability so that the measurements do not represent the reality of the process. It was proved that the presence of measurement errors in data affects the performance of several of the commonly used techniques in the field of quality improvement, such as control charts, univariate process capability analysis, multivariate process capability analysis [1], [2], [3], [4], [5], [6]. So, measurement system analysis should be considered as a fundamental step in any quality improvement strategy. The possibility of having confidence intervals for the indicators from which it is decided whether the measurement system is adequate or not, provides greater support to make such decision.

Although point estimation is not a problem for most of these indicators, there are not exact confidence intervals, since the combinations of random variables from which point estimators are constructed do not have a known probability distribution. In such situations, two approaches have been proposed to construct approximate confidence intervals: one is the modified large samples method ($MLS$), first proposed by Graybill and Wang [7], and the second approach is based on the construction of generalized confidence intervals ($GCI$), which were developed by Weerahandi[8].

In the decade of 1930the possibility of constructing confidence intervals for linear combinations of variances began to be discussed. Since then several authors proposed alternatives for their construction. In 1936 Smithdefined an estimate for linear functions of variances and proposed its approximate distribution from the Chi-squared distribution with a specific determination of degrees of freedom [9]. In the 1940s, Satterthwaite studied this approach giving rise to what is known as the Satterthwaite procedure [10], [11]. In 1978, Burdick and Sielken proposed a new method that did not thrive as it led to intervals with extremely large widths compared to those obtained with the Satterthwaite procedure [12]. In 1980, a new method was published by the

authors Graybill and Wang [7], which is currently one of the approaches used to construct approximate confidence intervals. The method was called *Modified Large Sample*.

Another approach used today arose from the concept of generalized inference introduced by Tsui and Weerahandi to construct hypothesis tests when no exact methods exist [13]. Some years later Weerahandiextended this concept to construct what is known as *generalized confidence interval* [8].

Empirical comparisons suggest that both approaches produce similar results [14]. The generalized method has the advantage of offering a general procedure that can be used under complex designs that include crossed or nested factors, and both fixed and random effects. The MLS method is not so flexible in that sense, but it does offer closed expressions that are relatively easy to implement in any computational software.

In this work we deepen the study of certain aspects related to the construction of the confidence intervals by the generalized approach.Onethoseaspectsis related to the fact that this method does not produce closed-form intervals. Instead, they have to be approximated by simulation procedures and hence different confidence intervals could be obtained when applying this method over the same data set.This inconvenient can be solved by using an adequate number of simulations in the procedure. However it does not exist in the literature rules establishing how many simulations are enough to ensure that the variability in the confidence limits is negligible. Krishnamoorthy and Mathew use, without justification, 30,000 simulations in the determination of tolerance limits [15]. Later, Romero, Zúnica and Pagura proposed that the number of simulations is a key factor to explain the uncertainty in repeated simulations [16]. In the context of the estimation of variance components in $R\&R$ studies, Burdick et al. use 100,000 simulations in the numerical examples of the methods, without justification either [14].We therefore conducted a comparative study designed so as to be able to identify the effect that the number of simulations produces on the confidence limits variability, when the method is applied repeatedly over the same data.

The other aspect analyzed is related to the effect that the *size* of the experiment can cause on the precision of the estimates of the capability indices of the measurement systems. In most practical situations, experiments that are designed to analyze the performance of measurement systems involve a small number of trials, which can cause estimates to lose precision. In this sense, we present the results obtained on a data set varying the size of the experiment.

Finally, the approach of GCIs is applied to a real data set coming from a metallurgical industry. The measurement process to be analyzed in this case is classified as *destructive*, so it requires the use of nested models which are not very common in the context of gauge R&R studies.

## II. METHODOLOGY

### 2.1Gauge R&R studies

To performan$R\&R$ study, it is necessary to assume that the obtained measurements $y_i$ can be described by a model of the form $y_i = x_i + \varepsilon_i$. Let $X \sim N(\mu_X;\ \sigma_P)$and$\varepsilon \sim N(0;\ \sigma_M)$beindependent random variables that represent the true value of the measured characteristic and the component of error introduced by the measurement process, respectively. Then the total variability of the observed measurements is:

$$\sigma_T^2 = \sigma_P^2 + \sigma_M^2 \tag{1}$$

Furthermore, the variability related to the measurement process $\sigma_M^2$ can be subdivided into two components: the repeatability -variability due to the measuring device-, and the reproducibility –variability arising from different operators-. Thus:

$$\sigma_M^2 = \sigma_{Repet}^2 + \sigma_{Reprod}^2 \tag{2}$$

Once each of these components is identified, there are several criteria to decide whether the measurement system is *capable* or not, where capable means that the measurement system has the ability to generate precise information. One of the most commonly used is the Precision to Tolerance Ratio, $PTR$, defined as:

$$PTR = \frac{k\ \sigma_M}{(USL - LSL)} \tag{3}$$

where$USL$ y $LSL$ are the specification limits of the process, and $k$ is a constant that correspond to the number of standard deviations between the natural tolerance limits that contain the middle$(1 - \alpha/2)100\%$ of a normal process. Common used values are $k = 6$, corresponding to $\alpha = 0.0027$, and $k = 5.15$ corresponding to $\alpha = 0.01$.

A rule of thumb to determine the capability of the measurement system, proposed by the Automotive Industry Action Group, is as follows: The system is considered capable if $PTR$ is less than 0.10 and not capable if it is greater than 0.30. If the percentage is between 0.10 and 0.30, the rule does not provide a decision and

other factors should be considered, such as the global behaviorof the process or the cost for misclassification of units [17].

Montgomery and Runger, among other authors, pointed out that the $PTR$ indicator does not necessarily bring a quantification of the behavior of the measurement process, since a highly capable process with respect to the specifications can tolerate measurement systems that produce higher variability than those that are less capable [18], [19]. An alternative indicator -$\%R\&R$- has been proposed, which is built comparing the variability of the measurement system with the total variability of the process:

$$\%R\&R = 100\frac{\sigma_M}{\sigma_T} \tag{4}$$

The decision about the capability of the measurement system from this indicator is made following the criteria suggested for $PTR$ (in percent terms).

In the context of these studies, the variability associated to the measurement system comprises the variability due to the repeatability and due to the reproducibility, which makes reasonable to evaluate which of the two factors $\sigma^2_{Repeat}\ /\sigma^2_M$ and $\sigma^2_{Reprod}\ /\sigma^2_M$has higher contribution.

Other widelyusedindicators are the Signal to Noise Ratio -$SNR$-, that reflects capability of discrimination of the measurement system, and those indicatorsbased on probabilities of misclassification, whose application is meaningful only on measurements systems designed to discriminate between good and bad parts or units.

The$SNR$ is defined as $\sqrt{2\sigma^2_P/\sigma^2_M}$and its value indicates the number of distinct categories the measurement system can reliably distinguish. It is recommended to obtain values of $SNR$ of at least five.

Regarding misclassification, it can arise in two different ways: false failure, when a good unit is classified as a bad one or failure; or missed fault, when a failure is misclassified as a good unit. In any case, desirable values for the false failure rate -$\delta$- and for the missed fault rate -$\beta$-are established and compared with the sample results to decide about the capability of the system.

Regardless the selected criterion or indicatorto evaluate the measurement system capability, the parameters involved in its expression must be estimated from sample data. Such data is usually obtained from designed experiments and analyzed by an analysis of variance. This technique allows to evaluate the significance of the selected factors and to obtain point estimates of the parameters needed to construct the measurement system capability indicator chosen.

In the particular case of the balanced two-factor crossed random model with interaction, where the two factors are referred to as "parts" (factor "$P$") and "operators" (factor "$O$"), measurements can be represented by the model:

$$Y_{ijk} = \mu_Y + P_i + O_j + (PO)_{ij} + \varepsilon_{ijk};\ \ i = 1,\dots,p;\ \ j = 1,\dots,o;\ \ k = 1,\dots,r \tag{5}$$

The random effects of the model are assumed to be jointly independent and normally distributed with zero meansand the following variances:

$$V(P_i) = \sigma^2_P;\quad V(O_j) = \sigma^2_O;\quad V((PO)_{ij}) = \sigma^2_{PO};\quad V(\varepsilon_{ijk}) = \sigma^2_E \tag{6}$$

The variance components of interest in the context of $R\&R$ studies (those involved in (2)), are obtained as a combination of the variance components of the factors in the model, keeping in mind that the variation of the measurement system is attributed to all sources of variation except parts. Additionally, residual variance represents the repeatability of the system; and variance due to operators and interaction between operators and parts represent its reproducibility:

$$\sigma^2_{Repet} = \sigma^2_E;\quad \sigma^2_{Reprod} = \sigma^2_{PO} + \sigma^2_O;\quad \sigma^2_M = \sigma^2_{Repet} + \sigma^2_{Reprod} \tag{7}$$

Expressions for point estimates of variance components in (6) are easily obtained from the expressions of the expected mean squares($E(S_i^2)$) of the model, which in this case are:

$$\begin{gathered} E(S^2_O) = \theta_O = \sigma^2_E + r\sigma^2_{PO} + pr\sigma^2_O \\ E(S^2_P) = \theta_P = \sigma^2_E + r\sigma^2_{PO} + or\sigma^2_P \\ E(S^2_{PO}) = \theta_{PO} = \sigma^2_E + r\sigma^2_{PO} \\ E(S^2_E) = \theta_E = \sigma^2_E \end{gathered} \tag{8}$$

Then:

$$\hat{\sigma}^2_E = MS_E;\quad \hat{\sigma}^2_{PO} = \frac{MS_{PO} - MS_E}{r};\quad \hat{\sigma}^2_O = \frac{MS_O - MS_{PO}}{pr};\quad \hat{\sigma}^2_P = \frac{MS_P - MS_{PO}}{or} \tag{9}$$

where$MS_i$is the sample value of the mean squares $S_i^2$.

The variance components in $R\&R$ studies can be rewritten as combinations of the expected mean squares in (8), and their point estimates can be obtained in terms of the point estimates in (9). For example, the variability of the measurement system can be estimated as:

$$\hat{\sigma}_M^2 = \hat{\sigma}_{Repet}^2 + \hat{\sigma}_{Reprod}^2 = \hat{\sigma}_E^2 + \hat{\sigma}_{PO}^2 + \hat{\sigma}_O^2 = \frac{p(r-1)MS_E + (p-1)MS_{OP} + MS_O}{pr} \quad (10)$$

The purpose of the point estimations is to make a decision about the measurement system capability. Therefore, it would be of great benefit to have a confidence interval to quantify the uncertainty associated with the estimation process. However, for most of the mentioned indicators, there are not exact confidence intervals due to the fact that the exact distributions of the point estimators are unknown.

**2.2Generalized confidence intervals**

The traditional approach for constructing confidence intervals is based on a pivotal quantity, i.e., a function of the unknown parameter of interest whose distribution does not depend on that parameter, from which the desired interval is obtained. Using a similar approach, $GCI$ construction is based on what it is called a *generalized pivotal quantity* ($GPQ$), defined by Weerahandias follows [8].

Let $R = r(\boldsymbol{X}; \boldsymbol{x}, \boldsymbol{\upsilon})$ be a function of $\boldsymbol{X}, \boldsymbol{x}$ and $\upsilon$ (but not necessarily a function of all), where $\boldsymbol{\upsilon} = (\theta, \boldsymbol{\delta})$ is a vector of unknown parameters, $\theta$ being the parameter of interest and $\boldsymbol{\delta}$ a vector of nuisance parameters. Then, $R$ is said to be a generalized pivotal quantity if it has the following two properties:

***Property A:***R has a probability distribution free of unknown parameters
***Property B:*** The observed pivotal, defined as $r_{obs} = r(x; x, \upsilon)$ does not depend on the nuisance parameters in $\delta$.

Once theGPQ is defined, let$C_{1-\alpha}$ a region in the sample space of Rsatisfying$P(R \in C_{1-\alpha}) = 1 - \alpha$. Then a $100(1-\alpha)\%$ generalized confidence interval for the parameter$\theta$ is the subset of the parametric space $\Theta$defined as:

$$\Theta_C(r) = \{\theta \in \Theta \mid r_{obs} \in C_{1-\alpha}\} \quad (11)$$

Applying this method, the confidence intervals depend on the distribution of the $R$ but there are only few instances in which this method leadsto closed-form intervals, being the general rule to approximate the limits of the region by simulation procedures [20], [15], [14].

It is important to realize that deducing an appropriate $GPQ$for each particular problem is not a trivial task. Although several authors have developed expressions for $GPQ$s in many particular situations, it does not exist yet a general method for constructing these functions.

Burdick et al. summarizes the $GPQ$s already developed for estimating most of the capability indicators used in measurement capability studies, in several different scenarios [14].

Consider the case of the indicator$\%R\&R$, under the model specified in (4). It is possible to obtain a $GPQ$ from results by Hamada y Weerahandi [21], who proposed $GPQ$ for both parameters$\sigma_M^2$ and$\sigma_P^2$, and then applying the proposition due to Iyer and Patterson [22], who proposed a general procedure to construct $GPQ$s for functions that dependon various parameters.

The identity in (1) makes it possible to re-write the $\%R\&R$ indicator as a function of $\sigma_M^2$and$\sigma_P^2$. Then, the $GPQ$is obtained by applying that function to the $GPQ$s ofeach individual parameters involved in the expression.

Thus, being:

$$\%R\&R = 100\frac{\sigma_M}{\sigma_T} = 100\frac{\sigma_M}{\sqrt{\sigma_M^2 + \sigma_P^2}}$$

$$GPQ_{\sigma_P^2} = max\left[0, \frac{(p-1)CM_P}{or\ W_1} - \frac{(p-1)(o-1)CM_{PO}}{or\ W_3}\right]$$

$$GPQ_{\sigma_M^2} = \frac{(o-1)CM_O}{pr\ W_2} + \frac{(p-1)^2(o-1)CM_{PO}}{pr\ W_3} + \frac{po(r-1)^2CM_E}{r\ W_4}$$

We have that:

$$GPQ_{\%R\&R} = 100\sqrt{\frac{GPQ_{\sigma_M^2}}{GPQ_{\sigma_P^2} + GPQ_{\sigma_M^2}}} \quad (12)$$

where $W_1$, $W_2$, $W_3$and$W_4$ are jointly independent chi-squared random variables with $(p-1)$, $(o-1)$, $(p-1o-1$and$po(r-1)$degrees of freedom, respectively.

Since we already have the expression of the $GPQ$ forthe$\%R\&R$ indicator, its distribution needs to be estimated by means of a simulation procedure. Then, lower and upper confidence bounds are defined as those values corresponding to the $100(\alpha/2)$thand$100(1-\alpha/2)$th percentiles in the simulated distribution of the $GPQ$, respectively.

## III. Results

### 3.1The Effect Of The Number Of Simulations On The Variability Of Confidence Limits

As it was mentioned, obtaining confidence intervals using the generalized approach involves simulation procedures. One of the disadvantages of this is that, for the same data set, the replication of the simulation procedure could result in different confidence intervals each time; which is undesirable in practice. In order to evaluate how the number of simulations affectsthe variability of the resulting intervals, we performed three simulation studies varying the number of simulations (*N*) among $10{,}000;\ 50{,}000$ and $100{,}000$.Each study was repeated 5,000 times, so we finally have 5,000 GCIs computed from each value of *N*.

The study was made on the data set based on the experiment described by Houf and Berman [23]. The response variable is the thermal performance of a module measured in Celsius per watt. Each response has been multiplied by 100 for convenience of scale. The data represent measurements of ten power modules recorded by three operators. Each part was measured three times by each operator, thus generating a total of 90 randomized trials.ANOVA technique was used to analyze this data set, assuming a two-factor crossed random model with interaction. Estimations of the variance components were:

$$\hat{\sigma}_E^2 = 0.5111;\quad \hat{\sigma}_{PO}^2 = 0.7280;\quad \hat{\sigma}_O^2 = 0.5646;\quad \hat{\sigma}_P^2 = 48.2926$$

thus:

$$\hat{\sigma}_{Repet}^2 = 0.5111;\quad \hat{\sigma}_{Reprod}^2 = 1.2926;\quad \hat{\sigma}_M^2 = 1.8037;\quad \widehat{\%R\&R} = 18.9749\%$$

Generalized confidence intervals for $\%R\&R$were computed using MonteCarlo simulation. The simulation procedure was defined to simulate $N$ values of the $GPQ$ in (12), by first generating random values of each of the four chi-squared variables$W_1$, $W_2$, $W_3$and$W_4$ previously mentioned, and later combining them with the observed mean squares in the ANOVA.

Table 1 shows descriptive statistics for the limits of the 95% GCIs across the 5,000 determinations in each simulation study.

**Table 1.**Simulation results: Descriptive measures for GCIs limits.

| | | Number of simulations | | | |
|---|---|---|---|---|---|
| | | **10,000** | **50,000** | **100,000** | **200,000** |
| **Lower bound** | **Mean** | 10.7923 | 10.7917 | 10.7906 | 10.7915 |
| | **Std. Dev** | 0.1076 | 0.0466 | 0.0336 | 0.0233 |
| | **Minimum** | 10.4460 | 10.6436 | 10.6990 | 10.7217 |
| | **Maximum** | 11.1363 | 10.9224 | 10.9047 | 10.8663 |
| **Upper bound** | **Mean** | 60.2151 | 60.1766 | 60.1698 | 60.1670 |
| | **Std. Dev.** | 1.1338 | 0.5188 | 0.3760 | 0.2559 |
| | **Minimum** | 56.7991 | 58.5664 | 59.0575 | 59.3237 |
| | **Maximum** | 64.6730 | 61.7135 | 61.4816 | 60.9538 |

As it was expected, the variability on the confidence limits decreases as the number of simulations increases. However, it should be noted that the variability on upper limits is fairly greater than the variabilityof lower limits, regardless of the number of simulations considered. The shape of the $GPQ$ distributions could be a reason for this behavior. In fact, empirical distributions of the $GPQ$ were found very asymmetric, with a huge positive skew. Table 2 shows descriptive statistics of skewness of the $GPQ$'s empirical distributions and Fig.1 shows one of the 5,000 $GPQ$'s empirical distributions obtained in each study. Furthermore, Fig. 2 shows de empirical distributions of the generalized confidence limits by number of simulations, from which is evident the reduction in variability accounted by increasing the value of *N*.

**Table 2.Simulations results:** Descriptive statistics for skewnesscoefficientof GPQ's empirical distributions.

| | | Number of simulations | | | |
|---|---|---|---|---|---|
| | | **10,000** | **50,000** | **100,000** | **500,000** |
| **Skewness coefficient** | **Mean** | 2.4355 | 2.4382 | 2,4389 | 2.4389 |
| | **Minimum** | 2.2750 | 2.3615 | 2.3893 | 2.3999 |
| | **Maximum** | 2.6249 | 2.5282 | 2.4892 | 2.4760 |

**Figure 1.**Empirical distributions of *GPQ*.

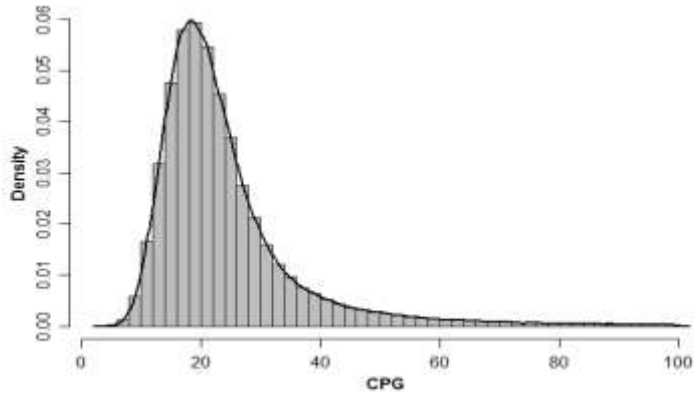
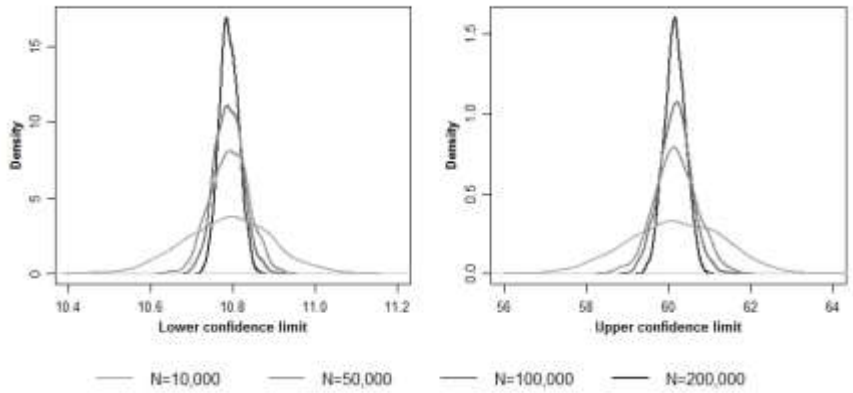



**Figure 2.**Empirical distributions of generalized confidence limits, by number of simulations.




Regarding the issue of interest, the effect of the number of simulations on confidence limits variability, it is recommended to use at least 100,000 simulations when computing a GCI. Clearly, if a larger number of simulations is chosen, higher precision is achieved, especially for the upper confidence limit which is the most variable. However, the objective of the study was to identify a value of *N* that achieves a compromise between precision of the results and computational cost required. We consider that acceptable results are obtained with 100,000 simulations, in the sense that discrepancies between minimum and maximum possible value for each confidence limit are very small (0.21% and 2.42% in lower and upper limits, respectively).

These results correspond to the study of the behavior of GPQs on a particular case and they should be confirmed with more general studies. However, given the nature of the property evaluated, it is reasonable to assume that similar results can be expected in different situations.

**3.2Effect of the size of the experiment in the precision of estimates**

The data set analyzed in previous section corresponds to an experiment involving three operators who measured three times each of ten parts, thus generating a total of 90 randomized trials.
The point estimate of the measurement system capability indicator was:

$$\%R\widehat{\&}R = 18.9749\%$$

and its generalized 95% confidence interval, computed from 100,000 simulations as it was recommended, is:

$$IC_{95\%}(\%R\&R) = (10.7673; 60.1473)$$

The point estimate of the indicator leads to a situation of indecision, in which the analyst should decide about the capability of the process taking into account other factors of the process. The confidence interval can be helpful in this decision, offering in this particular case evidence against the adequacy of the measurement system, since it indicates that it can be expected that the index be as large as 60%.

Nevertheless, it can be noticed that the interval width is very large. One possible reason for this is the *size* of the experiment, i.e., the number of levels of each factor in the design of the experiment, since those numbers of levels determine the degrees of freedom of the chi-squared variables involved in the generalized interval computations.

In order to evaluate this aspect, we modified the experiment to increase the number of operators used by Houf and Berman. The modified data set has now six operators who measured three times each of the ten modules, leading to 180 randomized trials.

Over this data set, the point estimate of the measurement system capability indicator is similar the estimate obtained under the original data:

$$\%R\widehat{\&}R = 17.8670\%$$

but its generalized 95% confidence interval results:

$$IC_{95\%}(\%R\&R) = (10.0060;\ 32.0643)$$

which width is reduced by 55%. This result is a bit more informative about the performance of the measurement system, since it indicates that its capability index is not so great to conclude that the measurement system is not adequate.

In the same way, modifying the data set so that the experiment has six operators, ten modules and six replications instead of three, leads to thesimilar results:

$$\%R\widehat{\&}R = 17.3874\%;\ \ IC_{95\%}(\%R\&R) = (9.7324;\ \ 31.0761)$$

If modification only implies increasing the number of replications or parts, no differences are found in the width of confidence interval compared with that obtained under the original data.

Moreover, by doubling the number of operators and reducing a half the number of parts the reduction in interval width is achieved as well:

$$\%R\widehat{\&}R = 17.2896\%;\ \ IC_{95\%}(\%R\&R) = (6.3073;\ \ 31.63.97)$$

We can deduce from these results that which makes the difference is increasing the number of operators considered in the experiment. This is reasonable with the fact that having $o = 3$ leads to a chi-squared variable with only two degrees of freedom which random values will be very small.

From these results, we consider reasonable to recommend that the experiments proposed to analyze the capability of a measurement system, are designed so as to ensure at least four degrees of freedom for each source of variation in the ANOVA.

**3.4Real case application**

This last subsection presents a case study of the $R\&R$ methods with destructive testing in a metallurgical company, where a productivity problem detected made it necessary to obtain inferential results for the measurement system capability indicator, which was achieved using the generalized approach to construct confidence intervals. The response variable of interest was time in seconds spent to carry out certain task. The nature of this variable makes the measuring process destructive, since once the time has passed it is not possible for it to be measured again.

This issue was solved using one of the alternatives suggested by De Mast y Trip [24] to perform a Gauge R&R study under destructive measurements, which led to consider a nested model with four parts($p = 4$), three operators($o = 3$)and two replications($r = 2$), to represent the time measurements obtained [25].

The estimates of variance components, according to the nested model assumed were:

$$\hat{\sigma}_E^2 = MS_E = 0{,}006557$$
$$\hat{\sigma}_O^2 = \frac{MS_O - MS_{P(O)}}{pr} = 0{,}003368$$
$$\hat{\sigma}_P^2 = \frac{MS_{P(O)} - MS_E}{r} = 0{,}012139$$

From these results the capability of the measurement system is computed, taking into account that according to the design that has been used, the repeatability is associated to the variability due to the experimental error and the reproducibility is associated to the variability due to the operators, that is:

$$\hat{\sigma}^2_{Repet} = 0{,}006557 \quad y \quad \hat{\sigma}^2_{Reprod} = 0{,}003368$$

Thus, the estimated capability indicator results:

$$\%R\widehat{\&}R = 44.9828\%$$

It is evident that the measurement system is not capable, since it is responsible for 45% of the variability observed in the measurements. The point estimate was accompanied by the generalized confidence limits, obtained from 100,000 simulations according to the recommendation derived from the previous study. Its calculation required derivation of the corresponding GPQ, which differs from that expressed in (12), since this case supposes a different ANOVA model.

The GPQs associated with $\sigma_M^2$ and $\sigma_T^2$ under this model are[6]~~(reftesis)~~:

$$GPQ_{\sigma_M^2} = \frac{po(r-1)MS_E}{W_3} - \frac{o(p-1)MS_{P(O)}}{pr\,W_2} + \frac{(o-1)MS_O}{prW_1}$$

$$GPQ_{\sigma_T^2} = \frac{po(r-1)^2MS_E}{r\,W_3} + \frac{(o-1)MS_O}{pr\,W_1} + \frac{o(p-1)^2MS_{P(O)}}{pr\,W_2}$$

$W_1, W_2$ and $W_3$ are independent chi-squared random variables with $(o-1), o(p-1)$ and $po(r-1)$ degrees of freedom, respectively.

Finally, the generalized 95% confidence interval resulted:

$$\mathrm{IC}_{95\%}(\%\mathrm{R\&}R) = (11.7297;\ 55.1981)$$

From which the conclusion about the performance of the measurement system is confirmed, since the expected range of values is above the value required for a measurement system to be considered as adequate.

## IV. Conclusion

The analysis of measurement systems capability involves quantifying the variability that they introduce into the measurements and evaluating their contribution to the total variability observed. Repeatability and reproducibility studies are designed for this purpose and they are widely known in the industrial field, although their use is generally limited to the calculation of point estimates of the measurement system capability indicator. The possibility of making inferences about the indicators, such as the calculation of confidence intervals, requires unconventional methodologies. One of the alternatives to construct confidence intervals is based in the concept of generalized inference, leading to what is known as generalized confidence intervals. Under this approach, intervals are approximated by simulation procedures. This particularity implies that the repeated application of the procedure on the same data set can lead to varying confidence limits. In this work, we investigate the effect of the number of simulations that are necessary to guarantee a minimum variability in the confidence limits, achieving a balance between computational cost and precision of results. This study provides a basis for understanding the importance of this factor, the number of simulations, on the determination of the interval, mainly because we find that the probability distribution of the pivotal quantity used to compute the intervals is highly asymmetric to the right. This implies a greater variability in the upper confidence limit.

The implementation of a Gauge R&R study requires designing an experiment to collect the information that will be analyzed. In most of the practical situations these experiments are small, in the sense that they are designed considering few levels of the factors of interest. Since the number of levels of factors in the experiment has an important role in the determination of generalized intervals, we also analyze the effect of the experiment size on the precision of the estimates. The results show that better results are obtained when experiments are designed so as to ensure at least four degrees of freedom for each source of variation.

Finally, GCIs are applied in a real case study in which the measurement to be analyzed, processing times,matches with what is called destructive measurement. The use of the generalized allowed to deriveinferential results on the capability of the measurement system, thus adding useful information for the evaluation of productivity parameters.

## References


[1]. H-J.Mittag, Measurement error effect on control chart performance, Annual Quality Congress, 49(0), 1985, 66-73.

[2]. H-J. Mittag, Measurement error effects on the performance of process capability indices,Frontiers in Statistical Quality Control, 5,1997, 195-206.

[3]. S. Bordignon and M. Scagliarini, Statistical analysis of process capability indices with measurement errors,Quality and Reliability Engineering International, 18, 2002, 321-332.

[4]. D.Dianda, M.Quaglino, J. Pagura and M.L. De Castro, Efecto del error de medición en índices de capacidad de procesos, Revista de Ciencias Económicas y Estadística, SaberEs, 8(2), 2016, 91-110.

[5]. D. Shishebori and A.Z.Hamadami, The effect of gauge measurement capability and dependency measure of process variables on the $MC_p$,Journal of Industrial and Systems Engineering, 4(1), 2009, 59-76.

[6]. D. Dianda, Estudio estadístico de sistemas de medida e indicadores de capacidad de procesos multivariados, en contextos de mejora de la calidad y la productividad, doctoral diss., Universidad Nacional de Rosario, Argentina, Rosario, 2015.

[7]. F. Graybill and C. Wang, Confidence intervals on nonnegative linear combinations of variances. Journal of the American Statistical Association, 75, 1980, 869-873.

[8]. S. Weerahandi, Generalized confidence intervals. Journal of the American Statistical Association, 88, 1993, 899-905.

[9]. H. Smith, The problem of comparing the results of two experiments with unequal errors. Journal of the Council of Scientific and Industrial Research, 9, 1936, 211-212
[10]. F.E. Satterthwaite, Synthesis of variance.Psychometrika, 6, 1941, 309-316.
[11]. F.E. Satterthwaite, An approximate distribution of estimates of variance components. Biometrics bulletin, 2,1946, 110-114.
[12]. R.K. Burdick and R.L. Sielken, Exact confidence intervals for linear combinations of variance components in nested classifications.Journal of the American Statistical Association, 73, 1978, 632-635.
[13]. K. Tsui and S.Wheerahandi, Generalized p-values in significance testing of hypothesis in the presence of nuisance parameters. Journal of the American Statistical Association, 84, 1989, 602-607.
[14]. R.K. Burdick, C.M. BorrorandD.C. Montgomery, A review of methods for measurement systems capability analysis. Journal of Quality Technology, 35(4), 2003, 342-354.
[15]. K. Krishnamoorthy and T. Mathew, Generalized confidence limits and one-sided tolerance limits in balanced and unbalanced one-way random models. Technometrics, 46, 2004, 44-52
[16]. R. Romero, L.Zunica, R. Romero Zunica and J.Pagura, One side tolerance limits for unbalanced one-way random effects models: a generalized Mee and Owen procedure. Journal of Statistical Computation and Simulation, 78(12), 2008, 1213-1225.
[17]. AIAG-Automotive Industry Action Group. Measurement System Analysis3era Ed(Detroit, MI, 2002)
[18]. D.C. Montgomery and G.C.Runger, Gauge capability and designed experiments. Part I: Basic methods. Quality Engineering, 6(1), 1993, 115-135.
[19]. D.C. Montgomery and G.C.Runger, Gauge capability and designed experiments. Part II: Experimental design models and variance component estimation. Quality Engineering, 6(2), 1993, 289-305.
[20]. A.K. Chiang, A simple general method for constructing confidence intervals for functions of variance components. Technometrics, 43, 2001, 356-367.
[21]. M. Hamada and S. Weerahandi, Measurement system assessment via generalized inference, Journal of quality technology, 32, 2000, 241-253.
[22]. H.K. Iyer and P.L. Patterson, A recipe for constructing generalized pivotal quantities and generalized confidence intervals. Technical Report 2002/10 (Department of Statistics, Colorado State University, Fort Collins, CO, 2002), http://www.stat.colostate.edu/statresearch/stattechreports/Technical%20Reports/2002/02_10.pdf.
[23]. R.E. Houf and D.B. Berman, Statistical analysis of power modules thermal test equipment performance, IEEE Transactions on Components, Hybrids and Manufacturing Technology, 11, 1988, 516-520.
[24]. J. De Mast and A. Trip, Gauge R&R studies for destructive measurements,Journal of Quality Technology, 37(1), 2005, 40-49.
[25]. M. Quaglino, J.Pagura, D. Dianda and E.Lupachini, Estudio de sistemas de medida con ensayos destructivos. Una aplicación sobre tiempos de producción. Revista de Ciencias Económicas y EstadísticaSaberEs, 2, 2010, 59-72.