RESEARCH ARTICLE

WILEY

A bootstrap method for the measurement error estimation in Gauge *R*&*R* Studies

Giovanni Celano 🅾

Università di Catania, Catania, Italy

Correspondence
Giovanni Celano, Università di Catania, Catania, Italy.
Email: giovanni.celano@unict.it

Abstract

A measurement system fit for use is an essential resource to perform quality control activities: its assessment must be carried out periodically to quantify its bias(location) and precision(width) error to qualify it for the purpose it is used. In particular, the measurement system capability must be determined to evaluate how much of the observed variability originates from the gauge's precision error. Gauge Repeatability and Reproducibility (*R&R*) studies are aimed at getting a reliable estimate of the precision error σ_M of the measurement system. The outcome of a Gauge *R&R* study are point estimates and confidence intervals of the precision error components and related measurement capability metrics. Here, a general bootstrap-based procedure is proposed to get confidence interval estimations of the measurement error and its components from a Gauge (*R&R*) study for continuous observations carried out with either the Average and Range (ARGG) control chart or the experimental design method. An application of the proposed bootstrap-method is presented for a dataset of observations.

KEYWORDS

ANOVA, measurement system assessment, statistical process monitoring, $\bar{X}$ and R method, uncertainty

1 | INTRODUCTION

A capable measurement system (MS) is a fundamental resource in any production environment to get reliable measures of the process parameters. For example, when performing on-line quality control, decisions about process variability can be significantly affected by the MS error. In fact, a poorly performing MS can lead a quality practitioner to make erroneous decisions about the conformity of a part or the distribution of a quality characteristic (QC). To tackle the problem, quality management standards set specific requirements about the MS. In particular, point 7.1.5 "Monitoring and Measuring Resources" of the ISO 9001-2015 Quality Management International Standard[7] states that:

> "The organization shall determine and provide the resources needed to ensure valid and reliable results when monitoring or measuring is used to verify the conformity of products and services to requirements."

Furthermore, grounding on the ISO 9001-2015 document, the standard IATF 16949-2016 "Quality management system requirements for automotive production and relevant service parts organizations" issued by the International Automotive Task Force (IATF)[6] sets more restrictive requirements to the MS qualification process in the automotive industry by specifying at point 7.1.5.1.1 that:

"Statistical studies shall be conducted to analyse the variation present in the results of each type of inspection, measurement, and test equipment system identified in the control plan. The analytical methods and acceptance criteria used shall conform to those in reference manuals on measurement system analysis."

The reference manual on Measurement Systems Analysis (MSA) has been published by the Automotive Industry Action Group (AIAG).[1] This manual provides practitioners with definitions, statistical techniques and threshold criteria to qualify a MS. The performance of a MS should be established in terms of its *resolution*, *bias* and *precision* errors. Here, we are interested to the estimation of the precision error through a capability study of the MS: Gauge *R&R* studies are aimed at getting an estimation of the precision error and its components. For *normally* distributed observations, the AIAG Measurement Systems Analysis Manual describes two widely adopted methods to carry out a Gauge *R&R* analysis: the Average and Range (ARGG) control chart and the experimental design methods. The former method is implemented by using $\bar{X}$ and R control charts to estimate the precision error components. Rarely, the S control chart is used if the study is designed to have a large number of repetitive measures on the same part, that is, the sample size n of the control charts is too large to use the sample range in the variability estimation. The experimental design method is based on (balanced or unbalanced) ANOVA models with Random(R) or Mixed(M) factors. Both the methods assume normality of the observations. They are widely adopted by practitioners and implemented in many commercial software for industrial statistics. From a practical point of view, the average and range method is easier to be implemented by practitioners because it only requires a basic statistical knowledge about $\bar{X}$ and $R(S)$ Shewhart control charts; conversely, the method based on random and mixed ANOVA models requires practitioners to have a good background on Design of Experiment (DoE) techniques. A comprehensive discussion about this second approach to perform a Gauge *R&R* study is provided by Burdick et al.[2] It is worth noting that the experimental method provides more information about the MS capability than the ARGG method, see the AIAG Manual[1] and Park and Ha.

The outcome of a Gauge *R&R* study is a set of point estimates of precision error components and performance metrics to qualify the measurement system, as defined in the next Section. These estimated values are characterized by an uncertainty originated from *practitioner-to-practitioner* variability in the selection and measurement of parts to carry out the study. For this reason, in the outcome of a Gauge *R&R* study confidence intervals (CI) are calculated for each estimated precision error component (source of variability). However, exact CIs cannot be defined for some components and performance metrics: for this reason, many approaches have been investigated to get reliable approximations. Approximated CIs have been derived for the Gauge *R&R* experiments based on ANOVA random(R) and mixed(M) models. The precision error components are derived from the sources of variability identified in the ANOVA statistical study: their approximated CIs are calculated starting from these sources of variability. In particular, the Modified Large Sample (MLS) and Generalized (GCI) confidence intervals are suggested and widely discussed in Burdick et al.,[2] who also conclude that they get a comparable performance in terms of coverage probability. For a Gauge *R&R* study, Burdick et al.[2] argue that these methods of calculation for CIs of each variance component should be preferred to other asymptotic large sample approaches based on the Maximum Likelihood (ML) and Residual Maximum Likelihood (REML) estimation of the variance components. The MLS method was firstly introduced by Graybill and Wang[4]: it has the advantage of providing practitioners with closed-form formulas that can be easily implemented on a spreadsheet and is mostly used in commercial software, like Minitab. However, its application cannot be easily generalized to any investigated ANOVA model. The Generalized Confidence Intervals GCIs were originally introduced by Weerahandi.[13] They are based on the notion of generalized pivotal quantity (GPQ) for a scalar parameter γ representing a variance component. The GCIs are obtained by means of Monte Carlo simulations of linear combinations of GPQs related to the variance components of interest. Next, a quantile-based approach lets practitioners to determine the GCIs bounds. Burdick et al.[2] provide extensive discussions and formulas to get the GCIs of interest for the ANOVA models they investigate.

An alternative approach to GCIs is implementing Bootstrap techniques to calculate the confidence intervals. The bootstrap is a computer-based simulation method based on resampling for assigning measures of accuracy to statistical estimates, see Efron and Tibshirani.[4] For a *nonparametric* bootstrap, resampling is performed with replacement from the *original* sample of observations without any assumption about their distribution. The bootstrap samples allow new outcomes of the estimated parameter to be obtained and, thus, a standard error to be estimated. With *parametric* bootstrap the observations are assumed to have a *known* distribution with *unknown* parameters, which are estimated

TABLE 1 Gauge *R&R* study notation

Parameter	Symbol	Burdick et al.[2]	AIAG Manual[1]
Process variation	σ_P	$\sqrt{\gamma_P}$	PV
Repeatability variation	σ_{RP}	***	EV
Reproducibility variation	σ_{RD}	***	AV
Measurement system variation	σ_M	$\sqrt{\gamma_M}$	GRR
Total variation	$\sigma_Y = \sqrt{\sigma_P^2 + \sigma_M^2}$	***	TV
Ratio process/measurement variance	$\gamma_R = \frac{\sigma_P^2}{\sigma_M^2}$	γ_R	***
Ratio measurement/total variance	$\gamma_{MY} = \frac{\sigma_M^2}{\sigma_Y^2}$	***	$\left(\frac{GRR}{TV}\right)^2$

from the original sample of data. In this case, the bootstrap samples are generated from the *known* distribution with the *estimated* parameters.

A *nonparametric* bootstrap approach has been considered by Wang and Li[10] to get confidence intervals for a Gauge *R&R* study performed with the Average and Range (ARGG) control chart's method. The original observations from the study are resampled to get bootstrap distributions of the calculated parameters from the $\bar{X}$ and $R(S)$ control charts. A comparison is carried out with the ML, REML and MLS approaches for the interval estimation of repeatability and reproducibility errors made with a random ANOVA model. In another paper, Yeh and Sun[14] implement a Monte Carlo simulation to replicate a Gauge *R&R* study based on the Average and Range (ARGG) method and to get the empirical distribution of some performance metrics of the MS. Is it worth highlighting that both these studies only consider the generation of CIs from a Gauge *R&R* study with the Average and Range (ARGG) method. Furthermore, both of them are focused on a small set of performance measures for the measurement system.

In this paper, a normal-theory *parametric* Bootstrap approach is proposed to calculate the CIs for the precision error of the MS, its components and several performance measures. In particular, in addition to available literature procedures:

- it can be applied to *any* selected method to perform a Gauge *R&R* study. In particular, with the ARGG method it gives comparable results to the available nonparametric procedure proposed in Wang and Li[10];
- with the experimental method, it computes approximated CIs for the precision error components and MS performance measures generally characterized by narrower widths than the Generalized Confidence Intervals proposed in Burdick et al.[2] for both the Random(R) and Mixed(M) ANOVA models;
- it provides practitioners with significant insights about the outcome of a Gauge *R&R* study by looking at the bootstrap distribution of each investigated MS performance measure.

The remainder of the paper is organized as follows: in the next Section, the adopted notation and the list of most important MS performance metrics is provided. Some considerations about the linear models to be considered for the observed measures are commented, as well. Section 3 presents the proposed Bootstrap method to get the interval estimates. A performance study is shown in Section 4. An illustrative example is presented in Section 5. Conclusions and future research complete the paper. An Appendix details formulas to perform a Gauge *R&R* study based on the Average and Range (ARGG) and the ANOVA methods.

2 | MS CAPABILITY PERFORMANCE

The capability performance quantifies the fitness for use of a MS. Several acceptance criteria based on different metrics can be considered. The same notation as in Burdick et al.[2] is adopted here. However, many practitioners usually refer to formulas reported in the Automotive Industry Action Group (AIAG) Manual[1]: for this reason, we also recall the AIAG Manual notation. Table 1 shows the sources of variability calculated in a Gauge *R&R* study, together with their symbols, respective formulas and notations adopted in related literature, see Burdick et al.,[2] and in the AIAG Manual.[1]

It is worth noting that σ_P is defined either as the "Process Variation" by Burdick et al.[2] or "Part Variation" by the AIAG Manual. If the MS is used to get measures for on-line process monitoring, for example to implement a control chart, then the definition of "Process Variation" is more appropriate and the parts involved in the Gauge *R&R* study have to be collected from a stable (in-control IC) process. Otherwise, when conformity to specifications is checked with

the measurement process, then the parts involved in the Gauge *R&R* study should be collected as to be representative of the entire specification interval: in this case, "Part Variation" is an acceptable definition. The "Measurement System Variation" is also known as the "precision error" or "measurement error" of the MS. The estimation of each source of variability can be made by either the Average and Range (ARGG) or the experimental method, as briefly summarized in Appendix A.

2.1 | Models for measurement error

We do reference to additive linear models to account for the measurement error effect when collecting a measure of a parameter. In particular, let us consider a quality characteristic of a part code to be produced and monitored in a manufacturing process. The *true*, but *unobservable*, value of the quality characteristic is denoted by X and is considered as a normal random variable, i.e. $X \sim N(\mu_P, \sigma_P^2)$; here, the subscript $'P'$ denotes "Process". At every time $t = 1, 2, ...$, the quality characteristic is measured by a MS to be tested for capability. The measurement process is performed by an *appraiser j*, for $j = 1, ..., A$, who is an operator or a gauge or a combination of both. The additive linear model from Linna and Woodall[10] has been widely adopted in Statistical Process Monitoring literature to account for the measurement error. It assumes that the measurement process is randomly performed by *any* available appraiser and pools all *combined* observations from the MS to quantify the measurement error, which is independent of X. Therefore, the additive linear model from Linna and Woodall[8] links X_t to Y_t as follows:

$$Y_t = BX_t + \epsilon_t \tag{1}$$

where $t = 1, 2 ...$, and $\epsilon_t \sim (\mu_M, \sigma_M^2)$ is the measurement error due to the MS. Its mean μ_M and its standard deviation σ_M represent the *bias* and the *precision error* of the MS, respectively. The constant B is related to the linearity error of the MS, see the AIAG Manual.[1] Assuming a perfect calibration, no bias-linearity errors affect the MS and we get $(\mu_M, B) = (0, 1)$. As a consequence, $\mu_Y = \mu_P$ and the observed values are $Y \sim N(\mu_P, \sigma_Y^2)$, where $\sigma_Y^2 = \sigma_P^2 + \sigma_M^2$. The precision error σ_M is the sum of the repeatability σ_{RP} and reproducibility σ_{RD} variability calculated by:

$$\sigma_M = \sqrt{\sigma_{RP}^2 + \sigma_{RD}^2} \tag{2}$$

According to the AIAG Manual:

- the *repeatability* error is referred to as the "within-appraisers" variability: it is defined as the variation in measurements obtained with one measurement tool when used several times by the same appraiser while measuring the same quality characteristic on the same part: it is considered as the inherent variation of the MS;
- the *reproducibility* error is referred to as the "between-appraisers" variability: it is defined as the variation in average of measurements collected by different appraisers while measuring the same quality characteristic on the same part; therefore, it can be considered as the between-appraisers bias.

When the number A of available appraisers is too large to be included in the Gauge *R&R* study, it provides information about the distribution of *pooled* observations from all appraisers by investigating on a random subset of $a \leq A$ appraisers. The Gauge *R&R* study can be performed with either the ARGG method or the experimental method. The latter is performed by considering an ANOVA Random(R) model with both parts and appraisers assumed as random factors.

Conversely, if all A available appraisers can be included in the Gauge *R&R* study, then the MS capability study can provide information about the *within-appraiser* distributions of observations. Denoting as $Y_{t,j}$ the *measured* value of the quality characteristic X_t at time t *by* appraiser j, the additive linear model linking X_t to $Y_{t,j}$ can be adapted as follows:

$$Y_{t,j} = B_j X_t + \epsilon_j \tag{3}$$

where $t = 1, 2 ...$, $j = 1, ..., A$ and $\epsilon_j \sim (\mu_j, \sigma_{\epsilon,j}^2)$ is the random error *for each* appraiser j involved in the measurement process. B_j is a constant accounting for the linearity error of each appraiser: for the Gauge *R&R* study, we assume $B_j =$

$1, \forall j$; that is, no linearity error is present. Therefore, the *within-appraiser* observed values of the quality characteristic are random variables $Y_{t,j} \sim N(\mu_{Y,j}, \sigma_{Y,j}^2)$, where $\mu_{Y,j} = \mu_P + \mu_j$ and $\sigma_{Y,j} = \sqrt{\sigma_P^2 + \sigma_{\epsilon,j}^2}$.

The j-th *appraiser's bias* is $\mu_j = \mu_{Y,j} - \mu_P$, for $j = 1, \dots, A$. In a Gauge *R&R* study no overall bias error is assumed for the MS, therefore $\sum_{j=1}^{A} \mu_j = 0$. The variability originated by each appraisers' bias is calculated as:

$$\gamma_A = \frac{\sum_{j=1}^{A} (\mu_{Y,j} - \mu_Y)^2}{A} \tag{4}$$

and contributes to the reproducibility error σ_{RD}. $\sigma_{\epsilon,j}$ is the j-th *appraiser's precision error*, for $j = 1, \dots, A$: it includes the repeatability error σ_{RP} and the interaction variability between parts and appraisers, if it is statistically significant.

In this case, the Gauge *R&R* study may only be performed with the experimental method by considering an ANOVA Mixed(M) model with parts(appraiser) as the random(fixed) factor, which can also detect an interaction between appraisers and parts, (unrestricted model). In Gauge *R&R* studies with the Mixed(M) model, a common precision error is calculated for each appraiser, i.e. $\sigma_{\epsilon,j} = \sigma_\epsilon$ and $\sigma_{Y,j} = \sigma_Y$, for $j = 1, \dots, A$. As a consequence, for the *within-appraiser* observed values the estimated distributions only differ with respect to location, i.e. $Y_{t,j} \sim N(\mu_{Y,j}, \sigma_Y^2)$.

2.2 | MS capability metrics

The following capability metrics can be calculated after a Gauge *R&R* study and are intended for qualifying a MS used to discriminate between *good* and *bad* parts:

- *Precision-to-tolerance ratio PTR*: this metric is useful when the MS is used for checking the conformance of a part to specifications. It is defined by doing reference to the specification interval (LSL, USL) of the quality characteristic to be measured:

$$PTR = \frac{k \times \sigma_M}{USL - LSL} \tag{5}$$

where k is either 5.15 or 6. The AIAG Manual considers a capable MS when $PTR \leq 0.1$. If $0.1 \leq PTR \leq 0.3$, then the MS can be considered acceptable, when process capability is high and the misclassification costs are low. When $PTR \geq 0.3$ the MS is not capable. If the MS is intended for use in process monitoring, then the PTR is not a good indicator of its intrinsic capability because it does not compare the MS variability to the natural process variation.

- *Misclassification rates*: they quantify how well the MS discriminates between good and bad parts. A *false failure* occurs when $LSL \leq X \leq USL$ and the MS classifies the part as nonconforming. The joint probability δ of false failure (producer's risk) is equal to:

$$\delta = \Pr[(LSL \leq X \leq USL) \text{ and } (Y \leq LSL \text{ or } Y \geq USL)]$$

$$= \int_{LSL}^{USL} \int_{-\infty}^{LSL} f(y,x) dy dx + \int_{LSL}^{USL} \int_{USL}^{\infty} f(y,x) dy dx \tag{6}$$

where $f(y,x)$ is the joint probability distribution function of Y and X, which is bivariate normal with mean $[\mu_Y, \mu_P]^\intercal$ and covariance matrix:

$$\begin{bmatrix} \sigma_P^2 + \sigma_M^2 & \sigma_P^2 \\ \sigma_P^2 & \sigma_P^2 \end{bmatrix} \tag{7}$$

Similarly, a *missed fault* occurs for $X \leq LSL$ or $X \geq USL$ and $LSL \leq Y \leq USL$, that is when the MS wrongly classifies the part as conforming to specifications. In this case, the joint probability β of missed fault (consumer's risk) is equal

to:

$$\beta = \Pr[(X \le LSL \text{ or } X \ge USL) \text{and} (LSL \le Y \le USL)]$$

$$= \int_{-\infty}^{LSL} \int_{LSL}^{USL} f(y,x)dydx + \int_{USL}^{\infty} \int_{LSL}^{USL} f(y,x)dydx \tag{8}$$

For a Gauge $R\&R$ study with the ANOVA Mixed(M) model, it is possible to define the j-th appraiser's misclassification rates $[\delta_j, \beta_j]$, for $j = 1, \dots, A$, by assuming the appraiser's population mean $\mu_{Y,j}$ instead of the *pooled* mean μ_Y.

Other MS capability metrics are suitable to qualify a MS used to run a Statistical Process Monitoring tool, for example a control chart. In this case, the measurement precision error should be small, if compared to the process variation σ_P. Therefore, these metrics depend on $\frac{\sigma_P}{\sigma_M} = \sqrt{\gamma_R}$:

- *Number of Distinct Categories NDC*: it represent the number of categories that can be reliably distinguished by the MS. It is also known as the Signal-to-Noise Ratio (*SNR*). It is calculated as:

$$NDC = \sqrt{2} \times \frac{\sigma_P}{\sigma_M} = \sqrt{2\gamma_R} \tag{9}$$

see Woodall and Borror[13] for an explanation about the formula. A value of five or greater is recommended. Another metric related to distinct categories is the discrimination ratio DR, non discussed here, see Woodall and Borror[13] for further details.

- *Proportion of total variation PTV*: it represents the proportion of total variation in the measurements due to the MS:

$$PTV = \frac{\sigma_M}{\sigma_Y} = \sqrt{\gamma_{MY}} = \frac{\sigma_M}{\sigma_P + \sigma_M} = \frac{1}{\sqrt{\gamma_R} + 1} \tag{10}$$

For convenience of notation, the symbol PTV is introduced here. In many statistical packages it is usually known as the Total gage $R\&R$ %Study Var, even if it is a ratio between two standard deviations. In the AIAG Manual it is defined as the %GRR metric. Similarly to the PTR, the threshold limits 0.1(0.3) have been settled by AIAG to decide if the MS is capable(uncapable) for quality monitoring. The PTV has the advantage of directly weighing the measurement error to the total observed variability, thus being a good criterion to decide if the MS is fit for use in process monitoring or not. It can be easily shown that there is a close relationship between PTV and NDC:

$$NDC = \sqrt{2} \times \left(\frac{1}{PTV} - 1 \right) \tag{11}$$

thus a $PTV \le 0.1$ leads to $NDC \approx 13$.

3 | THE BOOTSTRAP CIs PROCEDURE

In this Section, we discuss the integration of the proposed bootstrap procedure to a Gauge $R\&R$ study based either on the ARGG or the experimental method with ANOVA Random(R) or Mixed(M) model. We remember from the Appendix that p denotes the number of parts, a is the number of appraisers and r is the number of measurement readings for every part from each appraiser. Let us consider the generic observation Y_{ijk}, for $i = 1, \dots, p$, $j = 1, \dots, a$ and $k = 1, \dots, r$ from the Gauge $R\&R$ study. For each appraiser j involved in the Gauge $R\&R$ study, the observations are normal random variables $Y_{ijk} \sim N(\mu_{Y,j}, \sigma_Y^2)$. We denote as $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_a]^T = [Y_{111}, Y_{112}, \dots, Y_{par}]^T$ the vector of Gauge $R\&R$ study observations stacked by appraiser. Whichever is the selected method, the vector $\hat{\mathbf{\Gamma}} = [\hat{\sigma}_P, \hat{\sigma}_{RD}, \dots, \hat{\sigma}_M]^T$ of estimated sources of variability and the vector of observation means $\hat{\boldsymbol{\mu}} = [\hat{\mu}_Y, \hat{\mu}_{Y,1}, \dots, \hat{\mu}_{Y,a}]^T$ are obtained from the Gauge $R\&R$ study: given these two vectors, any MS capability performance measure can be derived, as explained in Section 2. We denote as $\hat{\theta}$ the estimate of a generic parameter $\theta = s(\mathbf{Y})$, (source of variability, mean or a function of them) obtained through the Gauge $R\&R$ study.

TABLE 2 Comparison with Wang and Li[11] bootstrap method. Shaft diameter dataset adapted from Hoguet[5]

	Wang and Li[13]	Bootstrap ARGG	Bootstrap ANOVA (R)	(M)	Generalized CI (R)	(M)
σ_{RP}	0.89 (0.41, 1.06)	0.89 (0.72, 1.05)	0.96 (0.80, 1.11)	0.96 (0.80, 1.11)	0.96 (0.83, 1.14)	0.96 (0.83, 1.15)
σ_{RD}	0.81 (0.75, 0.89)	0.70 (0.46, 0.94)	0.66 (0.42, 1.06)	0.54 (0.34, 0.77)	0.66 (0.15, 4.43)	0.54 (0.18, 0.77)
σ_{M}	1.20 (0.87, 1.41)	1.13 (0.95, 1.35)	1.16 (0.98, 1.36)	1.10 (0.94, 1.27)	1.16 (1.15, 4.56)	1.10 (0.93, 1.30)

TABLE 3 Comparison with Wang and Li[11] bootstrap method. Automobile mini-motor dataset

	Wang and Li[11]	Bootstrap ARGG	Bootstrap CI (R)	(M)	Generalized CI (R)	(M)
σ_{RP}	1.34 (1.00, 1.54)	1.31 (1.10, 1.54)	1.18 (1.00, 1.31)	1.18 (1.00, 1.31)	1.18 (1.05, 1.35)	1.18 (1.05, 1.35)
σ_{RD}	0.30 (0.00, 0.49)	0.20 (0.00, 0.52)	0.20 (0.00, 1.23)	0.17 (0.00, 0.64)	0.20 (0.03, 1.52)	0.17 (0.00, 0.52)
σ_{M}	1.37 (1.00, 1.61)	1.33 (1.12, 1.57)	1.20 (1.06, 1.35)	1.19 (1.06, 1.35)	1.20 (1.23, 2.10)	1.19 (1.08, 1.39)

The CIs are obtained by running a normal-theory *parametric* bootstrap procedure to generate new samples of observations from the original vector **Y** collected during the Gauge R&R study. The bootstrap procedure works as follows:

1. Design the Gauge R&R study by selecting the number of parts p, appraisers $a \leq A$ and trials r. Collect the original vector of observations **Y** and perform the Gauge R&R study with a proper method;
2. Calculate the vectors of estimated sources of variability $\widehat{\mathbf{\Gamma}}$ and means $\widehat{\boldsymbol{\mu}}$. Fix the number B of bootstrap samples to be generated;
3. For each appraiser j, for $j = 1, \dots, a$, generate the B bootstrap sample vectors $\mathbf{Y}_j^{*(b)} = (Y_{1j1}^{*(b)}, Y_{1j2}^{*(b)}, \dots, Y_{ijk}^{*(b)}, \dots, Y_{pjr}^{*(b)})^T$ of $p \times r$ Gauge R&R observations and stack them into the vectors:

$$\mathbf{Y}^{*(1)} = [\mathbf{Y}_1^{*(1)}, \mathbf{Y}_2^{*(1)}, \dots, \mathbf{Y}_a^{*(1)}]^T$$

$$\mathbf{Y}^{*(2)} = [\mathbf{Y}_1^{*(2)}, \mathbf{Y}_2^{*(2)}, \dots, \mathbf{Y}_a^{*(2)}]^T$$

$$\dots \qquad \dots$$

$$\mathbf{Y}^{*(B)} = [\mathbf{Y}_1^{*(B)}, \mathbf{Y}_2^{*(B)}, \dots, \mathbf{Y}_a^{*(B)}]^T \qquad (12)$$

where $Y_{ijk}^{*(b)} = X_i^{*(b)} + \epsilon_{ijk}^{*(b)}$, $X_i^{*(b)} \sim N(\widehat{\mu}_P, \widehat{\sigma}_P^2)$, $\epsilon_{ijk}^{*(b)} \sim N(\widehat{\mu}_j, \widehat{\sigma}_\epsilon^2)$, for $i = 1, \dots, p$, $j = 1, \dots, a$, $k = 1, \dots, r$ and $b = 1, \dots, B$.

4. Calculate the b-th bootstrap estimate $\widehat{\theta}^{*(b)}$, for $b = 1, \dots, B$, of each parameter of interest θ by using the selected Gauge R&R method, see Appendix A;
5. Store the bootstrap distribution of $\widehat{\theta}^*$ and calculate the α-th percentile confidence interval $(\widehat{\theta}^{*(\frac{\alpha}{2})}; \widehat{\theta}^{*(1-\frac{\alpha}{2})})$, where α is the confidence level.

4 | PERFORMANCE STUDY

The performance study starts by comparing the proposed bootstrap approach to the nonparametric bootstrap procedure developed for the Gauge R&R study with ARGG method in Wang and Li.[11] The Gauge R&R study datasets shown in Table 1 (adapted from Hoguet[6]), and Table 4 of Wang and Li[13] are considered. For the sake of brevity, in Tables 2 and 3 we only consider the CIs obtained with the two Bootstrap procedures for the measurement error σ_M and its reproducibility/repeatability components (σ_{RP}, σ_{RD}). To give a comprehensive overview of the proposed Bootstrap procedure, the CIs are shown for the ARGG method and the experimental method with both the Random(R) and Mixed(M) ANOVA models. The generalized CIs are shown, as well. We note that the point estimates of σ_{RD} are larger in Wang and Li[10] than those found in this paper with the ARGG method because they did not account for the repeatability variability contamination suggested in the AIAG Manual, as shown in Appendix, Equation (A.2). Furthermore, in Table 3 the estimation values of

measurement error components are different because Wang and Li[10] proposed a $\bar{X}$ and S control chart to implement the ARGG method although setting $r = 2$ would suggest using the range R control chart. From the two tables, it is evident that the proposed parametric bootstrap CIs are the narrowest for the repeatability error σ_{RP} and the measurement error σ_M. Only for the reproducibility component σ_{RD}, the Wang and Li[10] bootstrap CIs with the ARGG method are narrower than the other two procedures.

For the proposed bootstrap CIs, it is important to verify the *coverage probability* versus the nominal $(1 - \alpha)\%$ confidence level. For $B = \{50, 100, 150, 300, 500\}$, we have simulated the coverage probability for the CIs of the measurement error σ_M, its variability components σ_{RP}, σ_{RD} and the process error σ_P. We have assumed $(p, a, r) = (10, 3, 3), (20, 6, 6), \alpha = 0.05$ and for the pooled observations' distributions $\mu_X = \mu_Y = 0, \sigma_Y = 1$; furthermore, $\sigma_M = 0.2$ and $\sigma_{RP} = 0.2 \times \sigma_M$ have been assumed as input values. With the experimental method, both the Random(R) and Mixed(M) models have been considered. For $B \geq 100$, the calculated coverage probability always meets the nominal 95% confidence level, whichever is the selected method to perform the Gauge *R&R* study.

The performance study has been continued by investigating the effect of Gauge *R&R* study design parameters and measurement error components on the calculation of bootstrap CIs. In particular, we have considered the design parameters $p = (10, 20)$, $a = (3, 6)$, $r = (3, 6)$ and $\sigma_{RP}^2 = 0.2(0.8) \times \sigma_M^2$, where $\sigma_M = 1$. The reproducibility error $\sigma_{RD}^2 = 0.8(0.2) \times \sigma_M^2$ has been simulated by considering no interaction part-appraiser, $\sigma_{RD}^2 = \frac{\sum_{j=1}^{a}(\mu_{Y,j} - \mu_Y)^2}{2k}$ and assuming the coefficient $k = \frac{a-1}{2}(\frac{a}{2})$ for the Random(Mixed) model. Finally, we have generated within-appraiser observations with $\mu_{Y,1} = -\mu_{Y,a} = \sqrt{k/2} * \sigma_{RD}$ and $\mu_{Y,j} = \mu_Y$ for $j = 2, \ldots, (a-1)$; the simulated variability is $\sigma_Y = 1$. The study has been restricted to the experimental method, which should be generally preferred to the ARGG method, because it can account for the interaction parts-appraisers, whenever needed. To have a reference index for the direct comparison among CI widths of different parameters θ, the ratio $R_{CI} = \frac{UL - LL}{\theta}$ is calculated; $LL(UL)$ denote the lower(upper) confidence limits, respectively. Tables 4–7 show the CIs obtained with the bootstrap and the generalized methods, respectively, and the corresponding values of R_{CI}. With respect to the measurement error σ_M and its components $(\sigma_{RP}, \sigma_{RD})$, the obtained results can be summarized as follows:

- the bootstrap CIs are significantly narrower than the Generalized CIs for every investigated combination of design parameters (p, a, r) and ANOVA model; in fact, the R_{CI} values are always smaller for every estimated parameter. The difference is stronger for the reproducibility error σ_{RD} and, consequently, σ_M;
- a narrower width of CIs is achieved when implementing the Mixed(M) model with the experimental method, as explained by the lower R_{CI} values. Thus, the Mixed(M) model should always be preferred by practitioners when all available appraisers can be included in the study, i.e. assuming $(a = A)$, without increasing too much the number $N = p \times a \times r$ of observations. For the proposed Bootstrap CIs, the width difference between the two ANOVA models is generally smaller than the Generalized CIs;
- for a given value of σ_M, the width of CIs calculated for σ_{RD} and σ_M increases with σ_{RP}. This result depends on the highest variability in the bootstrap distribution of σ_{RD}, as the reproducibility error contribution to σ_M decreases. For example, with the Random(R) model, when $(p, a, r) = (10, 3, 3)$ and $\sigma_{RP} = 0.2(0.8) \times \sigma_M$, values of Bootstrap $R_{CI} = 0.206(1.638)$ and Generalized $R_{CI} = 5.836(7.985)$ are obtained for σ_{RD}. The same trend occurs with the Mixed(M) model, but with smaller absolute values of R_{CI};

With respect to the process variability σ_P, the *pooled* measurements variability σ_Y and related ratios, Tables 4–7 show that:

- the uncertainty associated to σ_P is higher than σ_M. Depending on the design parameters (p, a, r), the R_{CI} values for σ_P are between three and seven times R_{CI} values for σ_M. This depends on the higher simulated value of $\sigma_P = 0.995$ than $\sigma_M = 0.1$. Conversely, R_{CI} values are approximately the same for σ_P and $\sigma_Y = 1$;
- the uncertainty associated to $\sqrt{\gamma_{MY}} = \frac{\sigma_P}{\sigma_Y}$ is always larger than $\sqrt{\gamma_R} = \frac{\sigma_P}{\sigma_M}$. This finding depends on the fact that $\sigma_Y = \sqrt{\sigma_P^2 + \sigma_M^2}$ and, therefore, it accumulates the uncertainty in the estimation of both process and MS variability. As a consequence, the estimation of metric *NDC* is more uncertain than *PTV*.

Finally, Generalized CIs highlighted in bold characters in Tables 4–7 fail to contain the true value of the estimated parameter. This occurs for many parameters when $p = 20$ and $\sigma_{RP}^2 = 0.2 \times \sigma_M^2$ for the Mixed(M) model. A similar problem

TABLE 4 Sensitivity analysis on Gauge R&R study design parameters, $\mu_X = \mu_Y = 0$, $(\sigma_M, \sigma_Y) = (0.1, 1)$, $\sigma_{RP}^2 = 0.2 \times \sigma_M^2$, $p = 10$

| | Bootstrap CI | | | | Generalized CI | | | |
| | Model (R) | | Model (M) | | Model (R) | | Model (M) | |
$(p, a, r) = (10, 3, 3)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.4348, 1.1248)	0.693	(0.4345, 1.1263)	0.695	(0.5494, 1.4528)	0.908	(0.5363, 1.4052)	0.873
σ_{RD}	(0.0816, 0.1070)	0.284	(0.0829, 0.1039)	0.234	(0.0477, 0.6210)	6.410	(0.1002, 0.1356)	0.396
σ_{RP}	(0.0409, 0.0559)	0.335	(0.0403, 0.0555)	0.340	(0.0416, 0.0576)	0.357	(0.0421, 0.0576)	0.346
σ_M	(0.0949, 0.1164)	0.215	(0.0952, 0.1147)	0.196	(0.0776, 0.6240)	5.464	(0.1116, 0.1443)	0.327
σ_Y	(0.4467, 1.1300)	0.683	(0.4469, 1.1310)	0.684	(0.5709, 1.5349)	0.964	(0.5383, 1.4063)	0.868
$\sqrt{\gamma_{MY}}$	(0.0920, 0.2378)	1.458	(0.0921, 0.2358)	1.437	(0.0719, 0.6356)	5.636	(0.0891, 0.2424)	1.533
$\sqrt{\gamma_R}$	(4.0852, 10.8255)	0.677	(4.1211, 10.8063)	0.672	(1.2148, 13.8665)	1.272	(4.1086, 11.2214)	0.715
$(p, a, r) = (10, 3, 6)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.4257, 1.1470)	0.725	(0.4155, 1.1456)	0.734	(0.5524, 1.4532)	0.905	(0.5389, 1.4124)	0.878
σ_{RD}	(0.0771, 0.0941)	0.190	(0.0794, 0.0928)	0.150	(0.0446, 0.5741)	5.921	(0.1020, 0.1249)	0.255
σ_{RP}	(0.0414, 0.0514)	0.224	(0.0410, 0.0514)	0.232	(0.0417, 0.0521)	0.233	(0.0420, 0.0521)	0.227
σ_M	(0.0895, 0.1051)	0.156	(0.0912, 0.1043)	0.131	(0.0763, 0.5773)	5.010	(0.1120, 0.1334)	0.214
σ_Y	(0.4365, 1.1512)	0.715	(0.4277, 1.1496)	0.722	(0.5703, 1.5532)	0.983	(0.5407, 1.4132)	0.873
$\sqrt{\gamma_{MY}}$	(0.0840, 0.2247)	1.407	(0.0847, 0.2300)	1.454	(0.0692, 0.6044)	5.351	(0.0870, 0.2302)	1.432
$\sqrt{\gamma_R}$	(4.3363, 11.8578)	0.756	(4.2307, 11.7682)	0.758	(1.3183, 14.4112)	1.316	(4.3296, 11.4908)	0.720
$(p, a, r) = (10, 6, 3)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.4349, 1.1092)	0.678	(0.4416, 1.1568)	0.719	(0.5463, 1.4531)	0.911	(0.5413, 1.4054)	0.868
σ_{RD}	(0.0848, 0.1027)	0.201	(0.0854, 0.0998)	0.161	(0.0574, 0.2291)	1.920	(0.1462, 0.1912)	0.503
σ_{RP}	(0.0426, 0.0527)	0.228	(0.0426, 0.0531)	0.234	(0.0433, 0.0535)	0.229	(0.0432, 0.0537)	0.235
σ_M	(0.0968, 0.1120)	0.151	(0.0973, 0.1108)	0.135	(0.0826, 0.2371)	1.545	(0.1540, 0.1974)	0.434
σ_Y	(0.4475, 1.1143)	0.667	(0.4535, 1.1615)	0.708	(0.5560, 1.4597)	0.904	(0.5432, 1.4063)	0.863
$\sqrt{\gamma_{MY}}$	(0.0928, 0.2354)	1.426	(0.0891, 0.2312)	1.421	(0.0703, 0.3028)	2.324	(0.1231, 0.3294)	2.063
$\sqrt{\gamma_R}$	(4.1290, 10.7302)	0.663	(4.2073, 11.1752)	0.700	(3.1478, 14.1825)	1.109	(**3.0242, 8.1180**)	0.512
$(p, a, r) = (10, 6, 6)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.4341, 1.1386)	0.708	(0.4356, 1.1363)	0.704	(0.5518, 1.4238)	0.876	(0.5240, 1.4229)	0.903
σ_{RD}	(0.0814, 0.0917)	0.115	(0.0823, 0.0909)	0.095	(0.0549, 0.2271)	1.925	(0.1483, 0.1847)	0.407
σ_{RP}	(0.0392, 0.0455)	0.142	(0.0391, 0.0457)	0.147	(0.0393, 0.0461)	0.152	(0.0391, 0.0459)	0.152
σ_M	(0.0916, 0.1008)	0.092	(0.0922, 0.1006)	0.084	(0.0695, 0.2315)	1.620	(0.1509, 0.1867)	0.358
σ_Y	(0.4443, 1.1428)	0.698	(0.4462, 1.1404)	0.694	(0.5658, 1.4277)	0.862	(0.5258, 1.4236)	0.898
$\sqrt{\gamma_{MY}}$	(0.0840, 0.2157)	1.317	(0.0839, 0.2181)	1.343	(0.0625, 0.3006)	2.381	(0.1151, 0.3256)	2.105
$\sqrt{\gamma_R}$	(4.5272, 11.8642)	0.737	(4.4739, 11.8831)	0.745	(3.1727, 15.9646)	1.286	(**3.0601, 8.6829**)	0.565

occurs with the Random R model for σ_M, when $p = 10$ and $\sigma_{RP}^2 = 0.8 \times \sigma_M^2$. The proposed bootstrap CIs never show this problem.

5 | ILLUSTRATIVE EXAMPLE

We show the implementation of the bootstrap procedure with the Gauge R&R study dataset presented in the AIAG Manual,[1] see Table 8. The Gauge R&R study is arranged by considering $(p, a, r) = (10, 3, 3)$. It is performed with the experimental method: for the sake of comparison, it is implemented with both the ANOVA Random(R) and Mixed(M) models. To provide more details, we assume a specification interval $(LSL, USL) = (-4.5, 4.5)$. Table 8 shows the AIAG MSA Manual dataset.[1]

A summary of obtained results from the Gauge R&R study is shown in Table 9. This Table also shows the bootstrap and generalized CIs for the sources of variation and the related MS performance measures. Finally, it also presents the

TABLE 5 Sensitivity analysis on Gauge R&R study design parameters, $\mu_X = \mu_Y = 0$, $(\sigma_M, \sigma_Y) = (0.1, 1)$, $\sigma_{RP}^2 = 0.2 \times \sigma_M^2$, $p = 20$

| | Bootstrap CI | | | | Generalized CI | | | |
| | Model (R) | | Model (M) | | Model (R) | | Model (M) | |
$(p,a,r) = (20,3,3)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.6248, 1.2073)	0.585	(0.6367, 1.1914)	0.557	(0.7116, 1.3443)	0.636	(0.7073, 1.3546)	0.651
σ_{RD}	(0.0767, 0.0951)	0.206	(0.0790, 0.0923)	0.149	(0.0421, 0.5641)	5.836	**(0.1011, 0.1243)**	0.259
σ_{RP}	(0.0406, 0.0509)	0.232	(0.0403, 0.0507)	0.232	(0.0411, 0.0515)	0.234	(0.0413, 0.0515)	0.230
σ_M	(0.0886, 0.1042)	0.156	(0.0907, 0.1033)	0.125	(0.0712, 0.5671)	4.959	**(0.1110, 0.1326)**	0.216
σ_Y	(0.6334, 1.2112)	0.578	(0.6444, 1.1953)	0.551	(0.7225, 1.4245)	0.702	(0.7087, 1.3555)	0.647
$\sqrt{\gamma_{MY}}$	(0.0786, 0.1537)	0.750	(0.0807, 0.1507)	0.701	(0.0658, 0.5229)	4.571	(0.0892, 0.1748)	0.856
$\sqrt{\gamma_R}$	(6.4308, 12.6807)	0.628	(6.5586, 12.3563)	0.583	(1.6301, 15.1746)	1.361	(5.7084, 11.2066)	0.553
$(p,a,r) = (20,3,6)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.6231, 1.1938)	0.574	(0.6196, 1.1870)	0.570	(0.7039, 1.3244)	0.624	(0.7056, 1.3497)	0.647
σ_{RD}	(0.0829, 0.0950)	0.134	(0.0841, 0.0934)	0.104	(0.0457, 0.5739)	5.906	**(0.1097, 0.1251)**	0.172
σ_{RP}	(0.0412, 0.0479)	0.149	(0.0410, 0.0476)	0.148	(0.0413, 0.0483)	0.158	(0.0413, 0.0483)	0.155
σ_M	(0.0939, 0.1044)	0.104	(0.0947, 0.1037)	0.090	(0.0744, 0.5770)	5.026	**(0.1181, 0.1328)**	0.147
σ_Y	(0.6308, 1.1978)	0.567	(0.6274, 1.1910)	0.564	(0.7155, 1.3920)	0.676	(0.7070, 1.3504)	0.643
$\sqrt{\gamma_{MY}}$	(0.0820, 0.1580)	0.759	(0.0827, 0.1580)	0.753	(0.0698, 0.5200)	4.502	(0.0939, 0.1791)	0.852
$\sqrt{\gamma_R}$	(6.2499, 12.1475)	0.593	(6.2491, 12.0457)	0.583	(1.6425, 14.2905)	1.271	(5.5738, 10.6465)	0.510
$(p,a,r) = (20,6,3)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.6251, 1.2088)	0.587	(0.6389, 1.1941)	0.558	(0.7013, 1.3409)	0.643	(0.6999, 1.3165)	0.620
σ_{RD}	(0.0814, 0.0944)	0.145	(0.0822, 0.0912)	0.101	(0.0543, 0.2170)	1.819	**(0.1512, 0.1811)**	0.334
σ_{RP}	(0.0409, 0.0480)	0.158	(0.0409, 0.0479)	0.157	(0.0415, 0.0483)	0.152	(0.0415, 0.0484)	0.155
σ_M	(0.0926, 0.1024)	0.097	(0.0931, 0.1018)	0.087	(0.0773, 0.2239)	1.465	**(0.1577, 0.1867)**	0.290
σ_Y	(0.6322, 1.2129)	0.581	(0.6461, 1.1983)	0.552	(0.7096, 1.3498)	0.640	(0.7014, 1.3173)	0.616
$\sqrt{\gamma_{MY}}$	(0.0801, 0.1542)	0.740	(0.0805, 0.1516)	0.711	(0.0677, 0.2451)	1.774	**(0.1280, 0.2525)**	1.245
$\sqrt{\gamma_R}$	(6.4091, 12.4412)	0.606	(6.5201, 12.3812)	0.589	(3.9556, 14.7298)	1.083	**(3.9527, 7.8083)**	0.388
$(p,a,r) = (20,6,6)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.6410, 1.2085)	0.570	(0.6227, 1.2056)	0.586	(0.6983, 1.3333)	0.638	(0.6983, 1.3175)	0.622
σ_{RD}	(0.0847, 0.0929)	0.092	(0.0850, 0.0916)	0.074	(0.0564, 0.2179)	1.806	**(0.1588, 0.1793)**	0.229
σ_{RP}	(0.0424, 0.0471)	0.106	(0.0424, 0.0472)	0.108	(0.0427, 0.0474)	0.106	(0.0426, 0.0474)	0.107
σ_M	(0.0957, 0.1023)	0.066	(0.0960, 0.1023)	0.063	(0.0815, 0.2258)	1.443	**(0.1649, 0.1850)**	0.201
σ_Y	(0.6483, 1.2125)	0.564	(0.6304, 1.2097)	0.579	(0.7058, 1.3407)	0.635	(0.6998, 1.3183)	0.618
$\sqrt{\gamma_{MY}}$	(0.0820, 0.1529)	0.709	(0.0820, 0.1561)	0.742	(0.0711, 0.2503)	1.792	**(0.1308, 0.2532)**	1.224
$\sqrt{\gamma_R}$	(6.4648, 12.1519)	0.572	(6.3267, 12.1587)	0.586	(3.8679, 14.0211)	1.020	**(3.9427, 7.6422)**	0.372

Modified Large Sample (MLS) CIs that Minitab commercial software generates only for the Random(R) model. The not shown bootstrap CIs for each appraiser μ_j, and the misclassification rates (δ_j, β_j), for $j = 1, 2, 3$, are available upon request. The CIs have been obtained from $B = 10,000$ sampling replications. In Table 9 the estimates and CIs are provided for both ANOVA models: column "*estimate*" shows the results obtained with the Mixed(M) model and the Random(R) model between round parentheses.

No significant interaction between appraisers and parts is present for the MSA Manual dataset, that is, $\hat{\sigma}_{PA} = 0$: therefore, the restricted version of the ANOVA mixed(M) is used in the study. As expected, the Random(R) model gets a larger estimate for the reproducibility error σ_{RD} than the Mixed(M) model because the appraisers are considered as a random factor. This influences the estimates of the derived MS performance parameters from σ_{RD}. Furthermore, with the Random(R) model the width of CIs is wider than the Mixed(M) model: the CIs width increase is larger for the generalized CIs than the bootstrap CIs, being the latter more robust to the ANOVA model selection. In any case, the bootstrap CIs are narrower in width than the generalized CIs, whichever is the selected ANOVA model. For the Random(R) model, they are also narrower than the MLS CIs generated by Minitab. As an outcome from the Gauge R&R study, the estimated

TABLE 6 Sensitivity analysis on Gauge R&R study design parameters, $\mu_X = \mu_Y = 0$, $(\sigma_M, \sigma_Y) = (0.1, 1)$, $\sigma_{RP}^2 = 0.8 \times \sigma_M^2$, $p = 10$

	Bootstrap CI				Generalized CI			
	Model (R)		Model (M)		Model (R)		Model (M)	
$(p, a, r) = (10, 3, 3)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.4281, 1.1238)	0.699	(0.4317, 1.1264)	0.698	(0.5485, 1.4518)	0.908	(0.5354, 1.4042)	0.873
σ_{RD}	(0.0309, 0.1041)	1.638	(0.0336, 0.0808)	1.057	(0.0316, 0.3886)	7.985	(0.0000, 0.0954)	2.134
σ_{RP}	(0.0818, 0.1118)	0.335	(0.0806, 0.1110)	0.340	(0.0833, 0.1152)	0.357	(0.0842, 0.1152)	0.346
σ_M	(0.0966, 0.1315)	0.349	(0.0957, 0.1284)	0.327	(**0.1142**, **0.4064**)	2.922	(0.0943, 0.1396)	0.453
σ_Y	(0.4424, 1.1295)	0.687	(0.4478, 1.1313)	0.683	(0.5705, 1.4715)	0.901	(0.5438, 1.4084)	0.865
$\sqrt{\gamma_{MY}}$	(0.0979, 0.2570)	1.592	(0.0957, 0.2583)	1.626	(0.0930, 0.4776)	3.845	(0.0783, 0.2251)	1.469
$\sqrt{\gamma_R}$	(3.7598, 10.1689)	0.644	(3.7406, 10.4024)	0.670	(1.8398, 10.7020)	0.891	(4.3682, 12.7443)	0.842
$(p, a, r) = (10, 3, 6)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.4243, 1.1461)	0.725	(0.4198, 1.1419)	0.726	(0.5512, 1.4507)	0.904	(0.5377, 1.4100)	0.877
σ_{RD}	(0.0181, 0.0701)	1.164	(0.0232, 0.0529)	0.664	(0.0316, 0.2455)	4.785	(0.0063, 0.0703)	1.432
σ_{RP}	(0.0829, 0.1029)	0.224	(0.0820, 0.1028)	0.232	(0.0833, 0.1041)	0.233	(0.0840, 0.1043)	0.227
σ_M	(0.0890, 0.1107)	0.217	(0.0897, 0.1110)	0.213	(**0.1156**, **0.2734**)	1.578	(0.0917, 0.1197)	0.279
σ_Y	(0.4361, 1.1502)	0.714	(0.4327, 1.1464)	0.714	(0.5687, 1.4668)	0.898	(0.5451, 1.4132)	0.868
$\sqrt{\gamma_{MY}}$	(0.0864, 0.2339)	1.476	(0.0868, 0.2358)	1.491	(0.0885, 0.3459)	2.574	(0.0738, 0.1981)	1.243
$\sqrt{\gamma_R}$	(4.1560, 11.5349)	0.742	(4.1209, 11.4800)	0.740	(2.7125, 11.2499)	0.858	(4.9795, 13.5173)	0.858
$(p, a, r) = (10, 6, 3)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.4334, 1.1092)	0.679	(0.4449, 1.1559)	0.715	(0.5450, 1.4504)	0.910	(0.5401, 1.4028)	0.867
σ_{RD}	(0.0370, 0.0985)	1.376	(0.0374, 0.0699)	0.726	(0.0316, 0.1319)	2.244	(0.0000, 0.1264)	2.828
σ_{RP}	(0.0851, 0.1055)	0.228	(0.0852, 0.1061)	0.234	(0.0865, 0.1071)	0.229	(0.0865, 0.1075)	0.235
σ_M	(0.0987, 0.1216)	0.229	(0.0985, 0.1207)	0.222	(**0.1150**, **0.1781**)	0.631	(0.0958, 0.1594)	0.636
σ_Y	(0.4469, 1.1144)	0.668	(0.4573, 1.1610)	0.704	(0.5596, 1.4564)	0.897	(0.5482, 1.4064)	0.858
$\sqrt{\gamma_{MY}}$	(0.0969, 0.2503)	1.533	(0.0933, 0.2415)	1.482	(0.0882, 0.2581)	1.700	(0.0841, 0.2473)	1.632
$\sqrt{\gamma_R}$	(3.8686, 10.2688)	0.643	(4.0185, 10.6760)	0.669	(3.7425, 11.2948)	0.759	(3.9787, 11.8711)	0.793
$(p, a, r) = (10, 6, 6)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.4338, 1.1342)	0.704	(0.4334, 1.1327)	0.703	(0.5492, 1.4177)	0.873	(0.5215, 1.4168)	0.900
σ_{RD}	(0.0286, 0.0656)	0.826	(0.0303, 0.0487)	0.411	(0.0286, 0.1028)	1.660	(**0.0564**, **0.1218**)	1.464
σ_{RP}	(0.0783, 0.0910)	0.142	(0.0783, 0.0914)	0.147	(0.0786, 0.0921)	0.152	(0.0782, 0.0918)	0.152
σ_M	(0.0864, 0.1006)	0.142	(0.0867, 0.1005)	0.137	(0.0881, 0.1340)	0.459	(0.0793, 0.1332)	0.540
σ_Y	(0.4433, 1.1383)	0.695	(0.4431, 1.1367)	0.694	(0.5601, 1.4208)	0.861	(0.5286, 1.4194)	0.891
$\sqrt{\gamma_{MY}}$	(0.0815, 0.2106)	1.291	(0.0823, 0.2138)	1.315	(0.0685, 0.2008)	1.324	(0.0694, 0.2140)	1.446
$\sqrt{\gamma_R}$	(4.6427, 12.2330)	0.763	(4.5690, 12.1147)	0.758	(4.8777, 14.5692)	0.974	(4.6202, 14.3862)	0.982

$\widehat{PTV} = 0.28$ is barely acceptable by customer approval because the measurement error σ_M is around 30% of total variability σ_Y.

The bootstrap distributions of performance measures for the MS can also add significant information to the Gauge R&R study outcome. For example, the bootstrap distributions are particularly useful to interpret the R&R plot of *NDC* versus *PTR*. To show it, we assume $k = 6$ for the *PTR* calculation in this example. Figure 1 shows the R&R plots for a subset of $N = 1500$ bootstrap samples of *NDC* and *PTR* obtained by implementing both the ANOVA Random(R) (left panel) and Mixed(M) (right panel) models. The horizontal and vertical dotted lines divide the panel in four regions. The upper left part is the acceptance region, which is delimited by the threshold limits $NDC_L = 5$, $PTR_U = 0.3$. The red circle corresponds to the estimated values $(\widehat{PTR}, \widehat{NDC})$ with the two models, which barely meet the *NDC* lower threshold. An investigation of the joint bootstrap distribution for the couple (PTR, NDC) reveals that the probability $P(NDC \geq 5, PTR \leq 0.3) = 0.539(0.402)$ with the Mixed(Random) model. This outcome depends on a poor performance versus *NDC* due to the excessively large value of σ_M. Therefore, although the MS has a good capability versus the specification interval, (good

TABLE 7 Sensitivity analysis on Gauge R&R study design parameters, $\mu_X = \mu_Y = 0$, $(\sigma_M, \sigma_Y) = (0.1, 1)$, $\sigma_{RP}^2 = 0.8 \times \sigma_M^2$, $p = 20$

	Bootstrap CI				Generalized CI			
	Model (R)		Model (M)		Model (R)		Model (M)	
$(p, a, r) = (20, 3, 3)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.6331, 1.2155)	0.585	(0.6416, 1.2026)	0.564	(0.7176, 1.3563)	0.642	(0.7133, 1.3667)	0.657
σ_{RD}	(0.0174, 0.0868)	1.551	(0.0225, 0.0544)	0.712	(0.0316, 0.2328)	4.501	(0.0000, 0.0696)	1.556
σ_{RP}	(0.0811, 0.1018)	0.232	(0.0807, 0.1014)	0.232	(0.0821, 0.1031)	0.234	(0.0826, 0.1031)	0.230
σ_M	(0.0873, 0.1093)	0.220	(0.0890, 0.1096)	0.206	(0.1060, 0.2591)	1.532	(0.0903, 0.1186)	0.283
σ_Y	(0.6405, 1.2198)	0.579	(0.6495, 1.2067)	0.557	(0.7295, 1.3725)	0.643	(0.7189, 1.3704)	0.651
$\sqrt{\gamma_{MY}}$	(0.0786, 0.1559)	0.773	(0.0799, 0.1539)	0.740	(0.0869, 0.2695)	1.826	(0.0747, 0.1483)	0.736
$\sqrt{\gamma_R}$	(6.3345, 12.6788)	0.638	(6.4202, 12.4687)	0.608	(3.5734, 11.4610)	0.793	(6.6871, 13.3560)	0.670
$(p, a, r) = (20, 3, 6)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.6246, 1.1930)	0.571	(0.6227, 1.1878)	0.568	(0.7038, 1.3246)	0.624	(0.7056, 1.3499)	0.648
σ_{RD}	(0.0311, 0.0712)	0.898	(0.0341, 0.0535)	0.434	(0.0316, 0.2786)	5.525	(0.0390, 0.0718)	0.734
σ_{RP}	(0.0824, 0.0957)	0.149	(0.0820, 0.0953)	0.148	(0.0826, 0.0967)	0.158	(0.0827, 0.0965)	0.155
σ_M	(0.0912, 0.1067)	0.155	(0.0917, 0.1058)	0.142	(0.1152, 0.3025)	1.873	(0.0955, 0.1166)	0.211
σ_Y	(0.6326, 1.1969)	0.564	(0.6305, 1.1915)	0.561	(0.7178, 1.3460)	0.628	(0.7110, 1.3527)	0.642
$\sqrt{\gamma_{MY}}$	(0.0814, 0.1600)	0.786	(0.0821, 0.1594)	0.772	(0.0951, 0.3075)	2.125	(0.0778, 0.1518)	0.741
$\sqrt{\gamma_R}$	(6.1711, 12.2432)	0.610	(6.1948, 12.1368)	0.597	(3.0943, 10.4715)	0.741	(6.5311, 12.8315)	0.633
$(p, a, r) = (20, 6, 3)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.6274, 1.2153)	0.591	(0.6402, 1.1992)	0.562	(0.7044, 1.3471)	0.646	(0.7029, 1.3225)	0.623
σ_{RD}	(0.0287, 0.0845)	1.249	(0.0300, 0.0509)	0.467	(0.0316, 0.0961)	1.442	(0.0350, 0.1061)	1.590
σ_{RP}	(0.0818, 0.0959)	0.158	(0.0817, 0.0957)	0.157	(0.0830, 0.0966)	0.152	(0.0829, 0.0968)	0.155
σ_M	(0.0907, 0.1051)	0.145	(0.0905, 0.1046)	0.140	(0.1065, 0.1469)	0.405	(0.0948, 0.1390)	0.441
σ_Y	(0.6344, 1.2193)	0.585	(0.6479, 1.2033)	0.555	(0.7143, 1.3529)	0.639	(0.7090, 1.3256)	0.617
$\sqrt{\gamma_{MY}}$	(0.0794, 0.1543)	0.749	(0.0792, 0.1506)	0.714	(0.0850, 0.1789)	0.939	(0.0841, 0.1788)	0.947
$\sqrt{\gamma_R}$	(6.4044, 12.5525)	0.618	(6.5640, 12.5815)	0.605	(5.4995, 11.7188)	0.625	(5.5499, 11.8626)	0.634
$(p, a, r) = (20, 6, 6)$	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}	(LL, UL)	R_{CI}
σ_P	(0.6419, 1.2111)	0.572	(0.6251, 1.2084)	0.586	(0.6991, 1.3351)	0.639	(0.6991, 1.3193)	0.623
σ_{RD}	(0.0351, 0.0690)	0.758	(0.0359, 0.0503)	0.322	(0.0316, 0.1042)	1.623	(0.0587, 0.1024)	0.978
σ_{RP}	(0.0847, 0.0942)	0.106	(0.0847, 0.0944)	0.108	(0.0853, 0.0948)	0.106	(0.0853, 0.0948)	0.107
σ_M	(0.0942, 0.1045)	0.103	(0.0944, 0.1044)	0.100	(0.1176, 0.1585)	0.409	(0.1061, 0.1363)	0.302
σ_Y	(0.6492, 1.2152)	0.566	(0.6328, 1.2127)	0.580	(0.7101, 1.3436)	0.634	(0.7052, 1.3223)	0.617
$\sqrt{\gamma_{MY}}$	(0.0823, 0.1541)	0.718	(0.0823, 0.1569)	0.747	(0.0933, 0.1920)	0.987	(0.0893, 0.1781)	0.888
$\sqrt{\gamma_R}$	(6.4139, 12.1155)	0.573	(6.2928, 12.1168)	0.585	(5.1106, 10.6695)	0.559	(5.5715, 11.1728)	0.563

TABLE 8 Gauge R&R study. Dataset of observations reprinted from the AIAG MSA Manual,[1] 4th ed. p. 118

Part #	Appraiser A			Appraiser B			Appraiser C		
1	0.29	0.41	0.64	0.08	0.25	0.07	0.04	-0.11	-0.15
2	-0.56	-0.68	-0.58	-0.47	-1.22	-0.68	-1.38	-1.13	-0.96
3	1.34	1.17	1.27	1.19	0.94	1.34	0.88	1.09	0.67
4	0.47	0.50	0.64	0.01	1.03	0.20	0.14	0.20	0.11
5	-0.80	-0.92	-0.84	-0.56	-1.20	-1.28	-1.46	-1.07	-1.45
6	0.02	-0.11	-0.21	-0.20	0.22	0.06	-0.29	-0.67	-0.49
7	0.59	0.75	0.66	0.47	0.55	-0.83	0.02	0.01	0.21
8	-0.31	-0.20	-0.17	-0.63	0.08	-0.34	-0.46	-0.56	-0.49
9	2.26	1.99	2.01	1.80	2.12	2.19	1.77	1.45	1.87
10	-1.36	-1.25	-1.31	-1.68	-1.62	-1.50	-1.49	-1.77	-2.16

TABLE 9 Gauge $R\&R$ parameter estimates and CIs with the ANOVA method, (Random(R) and Mixed(M) model). Dataset of observations reprinted from the AIAG MSA manual, 4th ed. p. 118

| Gauge $R\&R$ parameter	Estimate	Bootstrap CI (R)	Bootstrap CI (M)	Generalized CI (R)	Generalized CI (M)	MLS CI (R)
$\widehat{\mu}_Y$	0.001	(-0.646, 0.631)	(-0.621, 0.654)	(-0.972, 0.912)	(-0.726, 0.754)	(-)
$\widehat{\sigma}_P$	1.042	(0.533, 1.519)	(0.573, 1.516)	(0.730, 1.862)	(0.721, 1.876)	(0.715, 1.906)
$\widehat{\sigma}_{RD}$	0.227 (0.185)	(0.175, 0.294)	(0.146, 0.231)	(0.101, 1.442)	(0.178, 0.319)	(0.114, 1.443)
$\widehat{\sigma}_{RP}$	0.200	(0.164, 0.233)	(0.167, 0.231)	(0.173, 0.237)	(0.171, 0.237)	(0.173, 0.237)
$\widehat{\sigma}_M$	0.302 (0.273)	(0.259, 0.348)	(0.237, 0.310)	(0.263, 1.467)	(0.268, 0.381)	(0.227, 1.457)
$\widehat{\sigma}_Y$	1.085 (1.077)	(0.619, 1.553)	(0.636, 1.542)	(0.807, 2.197)	(0.752, 1.889)	(0.776, 2.106)
$\widehat{\gamma}_R$	11.88 (14.63)	(3.08, 26.13)	(4.25, 30.67)	(0.49, 31.03)	(4.46, 33.23)	(-)
$\widehat{\gamma}_{MY}$	0.078 (0.064)	(0.037, 0.245)	(0.032, 0.190)	(0.031, 0.673)	(0.030, 0.209)	(-)
$\widehat{PTR}$	0.302 (0.273)	(0.259, 0.348)	(0.237, 0.310)	(0.263, 1.467)	(0.267, 0.382)	(-)
$\widehat{NDC}$	4.86 (5.39)	(2.48, 7.21)	(2.91, 7.81)	(0.99, 7.88)	(3.02, 8.16)	(-)
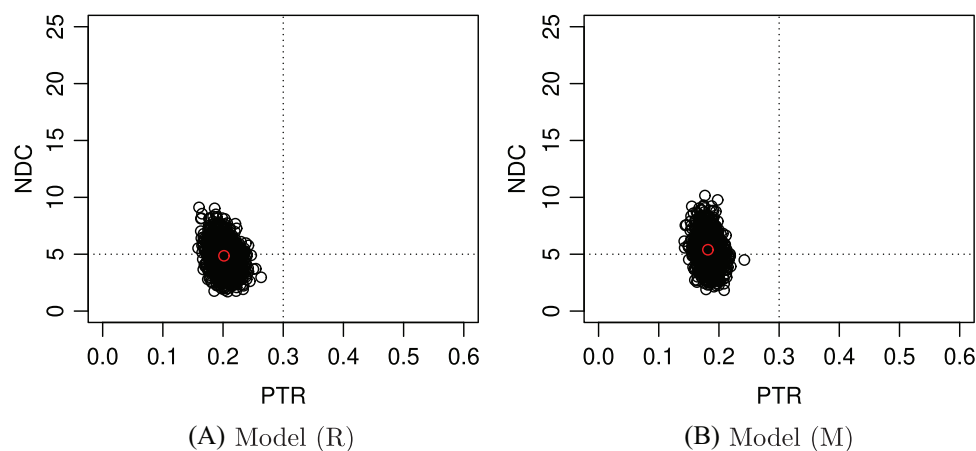
FIGURE 1 Plot of PTR versus NDC, ($R\&R$ graph). Gauge $R\&R$ study with experimental method. Dataset of observations from the AIAG MSA manual,[1] 4th ed. p. 118

performance vs. PTR), the $R\&R$ plot confirms that it cannot be used to perform on-line monitoring by means of variable control charts, (poor performance with a high probability to have an unsatisfactory value of NDC).

6 | CONCLUSIONS

The assessment of a MS is an important step to be periodically performed in order to limit the error introduced in any quality monitoring activity. The Gauge $R\&R$ study is an important statistical method to quantify the added noise from the MS to the inherent process variability associated to any quality characteristic. In fact, its impact on observed values of the quality characteristic can dramatically affect the correct estimation of process performance, like the calculation of the capability indexes. In this paper, a novel normal-theory parametric bootstrap approach has been proposed to get reliable interval estimates of sources of variability and related performance metrics for a MS. The proposed approach has been compared with other procedures available in literature and commercial software to test its effectiveness. It shows a good performance and robustness versus the design parameters and the selected method to carry out the Gauge $R\&R$ study.

The obtained results for the CIs estimation highlight the importance of correctly choosing the ANOVA model when implementing the experimental method for the Gauge $R\&R$ study. Whenever possible, it is preferable implementing the ANOVA Mixed(M) model by testing all appraisers available to the MS, as the resulting confidence intervals are narrower than other models. This is an important issue to be considered in future research on Statistical Process Monitoring in presence of measurement error: in fact, it is well known that measurement error can influence the performance of a control

chart. All contributions investigating the effect of measurement error on the performance of control charts have always implicitly assumed the *pooled* distribution of observations for the MS by overlooking the within-appraiser performance. That is, these studies implicitly assume Gauge *R&R* studies performed with the ARGG or ANOVA Random(R) model methods. However, when the estimation of the within-appraiser distributions is possible with the ANOVA Mixed(M) model, the current *pooled* measurement error model proposed in Linna and Woodall[8] for testing control charts performance is no longer effective. Distributions of observations should be considered for each appraiser and the control chart's performance should be investigated by assuming each appraiser schedule in the measurement process.

DATA AVAILABILITY STATEMENT

All datasets used in the manuscript are referenced throughout the text and available in the original papers, where they have been originally published. The dataset used in the illustrative example is fully reported in Table 8 of the manuscript.

ORCID

Giovanni Celano https://orcid.org/0000-0001-7871-7499

REFERENCES

1. Automotive Industry Action Group. *Measurement Systems Analysis*. 4th ed.; 2010.
2. Burdick RK, Borror CM, Montgomery DC. *Design and Analysis of Gauge R&R Studies*. ASA-SIAM; 2005.
3. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall; 1993.
4. Graybill FA, Wang CM. Confidence intervals for proportions of variability in two-factor nested variance component models. *J Am Statist Assoc*. 1979;74:368-374.
5. Hoguet RG. A simplified gage method for test reliability. *Ceram Eng Sci Proc*. 1994;15:59-71.
6. International Automotive Task Force. IATF 16949 - Quality management system requirements for automotive production and relevant service parts organizations. 2016.
7. International Organization for Standardization (ISO), TC 176/SC 2. ISO 9001 - Quality management systems - Requirements. 2015.
8. Linna KW, Woodall WH. Effect of measurement error on Shewhart control charts. *J Qual Technol*. 33(2):213–222.
9. Park S, Ha C. Determination of optimal experimental design for ANOVA Gauge R and R using stochastic programming. *Measurement*. 2020;156.
10. Wang FK, Li EY. Confidence intervals in repeatability and reproducibility using the bootstrap method. *Total Qual Manag*. 2003;14(3):341-354.
11. Weerahandi S. Generalized confidence intervals. *J Am Stat Assoc*. 1993;88:899-905.
12. Woodall WH, Borror CM. Some relationships between gage R and R criteria. *Qual Reliab Eng Int*. 2008;24(1):99-106.
13. Yeh TM, Sun JJ. Using the Monte Carlo simulation methods in gauge repeatability and reproducibility of measurement system analysis. *J Appl Res Technol*. 2013;11(5):780-796.

AUTHOR BIOGRAPHY

Giovanni Celano is an associate professor at the University of Catania (Italy). He holds a PhD in production engineering from the University of Palermo (Italy). His current research is focused on developing and implementing statistical process monitoring techniques for on-line quality control. He has authored/co-authored about 130 papers in international journals and in proceedings of national and international conferences. He is an advisory editor of the *Engineering Reports* journal and an associate editor of the *Quality Technology and Quantitative Management* journal.

APPENDIX A: GAUGE $R\&R$ STUDY METHODS

To design a Gauge $R\&R$ study, the number p of parts, appraisers $a \leq A$, (fe.g., crews, lab technicians, operators or instruments), and trial measurements r on each part should be fixed. In general, to conduct a Gauge $R\&R$ study $N = p \times a \times r$ observations are needed. Measurements should be collected randomly without any part identification by each appraiser. We denote as Y_{ijk} the observed value of the quality characteristic of interest on the i-th part, $i = 1, \dots, p$, measured by the j-th appraiser, $j = 1, \dots, a$, at the k-th trial, $k = 1, \dots, r$. Then, we use the following notation:

- $Y_{ij.} = \sum_{k=1}^{r} Y_{ijk}$ and $\bar{Y}_{ij.} = \frac{1}{r} \times Y_{ij.}$ are the total and average of measurements collected on the i-th part by the j-th appraiser, respectively;
- $Y_{.j.} = \sum_{i=1}^{p} \sum_{k=1}^{r} Y_{ijk}$ and $\bar{Y}_{.j.} = \frac{1}{p \times r} \times Y_{.j.}$ are the total and average of measurements collected on the p parts by the j-th appraiser, respectively;
- $Y_{i..} = \sum_{j=1}^{a} \sum_{k=1}^{r} Y_{ijk}$ and $\bar{Y}_{i..} = \frac{1}{a \times r} \times Y_{i..}$ are the total and average of measurements collected on each i-th part by the a appraisers, respectively;
- $Y_{...} = \sum_{i=1}^{p} \sum_{j=1}^{a} \sum_{k=1}^{r} Y_{ijk}$ and $\bar{Y}_{...} = \frac{Y_{...}}{p \times a \times r}$ are the total and average of collected measurements, respectively;

Now we briefly present formulas for the sources of variability calculations with the Average and Range AGRR and the experimental methods.

A.1 | Average and Range (ARGG) method

With this method the Gauge $R\&R$ study is performed by running $\bar{X}$ and R control charts for each appraiser involved in the study. A S control chart can be used when r is sufficiently large and the range R is no longer an efficient estimator of the variability. With the ARGG method, the repeatability and reproducibility errors are estimated from the $\bar{X}$ and R control charts plotted for each appraiser. This method overlooks a potential source of variability due to interaction between parts and appraisers.

To plot the R control chart for each appraiser j, for $j = 1, \dots, a$, the ranges $R_{ij} = max_k(Y_{ijk}) - min_k(Y_{ijk})$ are calculated for each part i, for $i = 1, \dots, p$. Then, the average range $\bar{R}_{.j}$ for each appraiser is calculated as $\bar{R}_{.j} = \frac{1}{p} \times \sum_{i=1}^{p} \bar{R}_{ij}$. The grand average $\bar{R}_{...} = \frac{1}{a} \times \sum_{j=1}^{a} \bar{R}_{.j}$ is immediately derived. The grand average range $\bar{R}_{...}$ is used to calculate the control limits of each R control chart, whose width is proportional to the estimated value of the *repeatability* variability σ_{RP} (Equipment Variation EV in the AIAG Manual):

$$\hat{\sigma}_{RP} = \frac{\bar{R}_{...}}{d_2^*(p \times a, r)} \tag{A.1}$$

where $d_2^*(p \times a, r)$ is an unbiasing constant associated to the distribution of average range. Its values can be found in the AIAG Manual, Appendix C.[1]

The average $\bar{Y}_{.j.}$ of measurements collected on the p parts for each appraiser j is used to estimate $\mu_{Y,j}$ and to plot the central line of each $\bar{X}$ control chart, for $j = 1, \dots, A$. The range $R_{\bar{Y}_j} = max_j(\bar{Y}_{.j.}) - min_j(\bar{Y}_{.j.})$ among appraisers is calculated to get the estimate $\hat{\sigma}_{RD}$ of the *reproducibility* variability (Appraiser Variation AV in the AIAG Manual), which measures the bias among a appraisers:

$$\hat{\sigma}_{RD} = \sqrt{\left(\frac{R_{\bar{Y}_j}}{d_2^*(1, r)} \right)^2 - \frac{\hat{\sigma}_{RP}^2}{p \times r}} \tag{A.2}$$

where a fraction of repeatability variability is subtracted since the $\hat{\sigma}_{RD}$ is contaminated by the $\hat{\sigma}_{RP}$, see the AIAG manual.[1] The estimated *process* variability $\hat{\sigma}_P$ is obtained by considering the averages of measurements $\bar{Y}_{i..}$ for each part i, for $i = 1, \dots, p$, and their range $R_{i..} = max_i(\bar{Y}_{i..}) - min_i(\bar{Y}_{i..})$:

$$\hat{\sigma}_P = \frac{R_{i..}}{d_2^*(1, p)} \tag{A.3}$$

TABLE A.1 ANOVA for the Random(R) model with interaction

Source of variation	Degrees of freedom	Mean square	Expected mean square
Parts (P)	$p-1$	S_P^2	$\theta_P = \sigma_E^2 + r\sigma_{PA}^2 + ar\sigma_P^2$
Appraiser (A)	$a-1$	S_A^2	$\theta_A = \sigma_E^2 + r\sigma_{PA}^2 + pr\sigma_A^2$
P x A	$(p-1)(a-1)$	S_{PA}^2	$\theta_{PA} = \sigma_E^2 + r\sigma_{PA}^2$
Replicates (Error)	$pa(r-1)$	S_E^2	$\theta_E = \sigma_E^2$

TABLE A.2 Gauge $R\&R$ parameters and point estimators for the Random(R) model with interaction

Gauge $R\&R$ parameter	Point estimator
μ_Y	$\bar{Y}_{...}$
$\mu_{Y,j}$	$\bar{Y}_{.j.}$
σ_P^2	$\frac{S_P^2 - S_{PA}^2}{ar}$
$\sigma_{RD}^2 = \sigma_A^2 + \sigma_{PA}^2$	$\frac{S_A^2 + (p-1)S_{PA}^2 - pS_E^2}{pr}$
$\sigma_{RP}^2 = \sigma_E^2$	S_E^2
$\sigma_M^2 = \sigma_{RD}^2 + \sigma_{RP}^2$	$\frac{S_A^2 + (p-1)*S_{PA}^2 + p(r-1)S_E^2}{pr}$

TABLE A.3 ANOVA for the Random(R) model without interaction

Source of variation	Degrees of freedom	Mean square	Expected mean square
Parts (P)	$p-1$	S_P^2	$\theta_P = \sigma_E^2 + ar\sigma_P^2$
Appraiser (A)	$a-1$	S_A^2	$\theta_A = \sigma_E^2 + pr\sigma_A^2$
Replicates (Error)	$par - p - a - 1$	S_E^2	$\theta_E = \sigma_E^2$

Finally, the estimated mean of combined observations from the Gauge $R\&R$ study is $\mu_Y = \bar{Y}_{...}$.

A.2 | Experimental method with ANOVA models

With this method, the Gauge $R\&R$ study is designed as an experimental plan. Here, we briefly recall formulas for the most common scenarios, that is, the balanced two factor *Random*(R) and *Mixed*(M) models with and without interaction. Tables showing the ANOVA sources of variation and formulas to get the point estimators are reported for each model with and without interaction. Details about these formulas can be found in Burdick et al.[2]

Two-factor Random(R) model

Both "parts" and "appraisers" are considered as random factors. This model should be assumed by the practitioner to carry out the Gauge $R\&R$ study when the a appraisers are randomly selected from a set of cardinality A: for example, the crew of operators who randomly perform the measurement process during quality inspections. The reference model with interaction is:

$$Y_{ijk} = \mu_Y + P_i + A_j + (PA)_{ij} + E_{ijk} \tag{A.4}$$

$$i = 1, \dots p, \ j = 1, \dots a, \ k = 1, \dots r$$

where μ_Y is a constant, $P_i, A_j, (PA)_{ij}$ and the error E_{ijk} are jointly independent normal random variables with means of zero and variances $\sigma_P^2, \sigma_A^2, \sigma_{PA}^2$ and σ_E^2. When there is no interaction, the term $(PA)_{ij}$ is noise adding error. Tables A.1–A.4 show the ANOVA sources of variation and the point estimators, respectively, for the Random(R) model with and without interaction.

Two-factor mixed (M) model

The practitioner considers this model when all A appraisers performing the measurement process are included in the Gauge $R\&R$ study: thus, "Parts" should be assumed as a random factor and "appraisers" as a fixed factor, with $a = A$. This is a very frequent situation in small companies, where a few appraisers are engaged in the measurement process activities.

TABLE A.4 Gauge R&R parameters and point estimators for the Random(R) model without interaction

Gauge R&R parameter	Point estimator
μ_Y	$\bar{Y}_{...}$
$\mu_{Y,j}$	$\bar{Y}_{.j.}$
σ_P^2	$\frac{S_P^2 - S_E^2}{ar}$
$\sigma_{RD}^2 = \sigma_A^2$	$\frac{S_A^2 - S_E^2}{pr}$
$\sigma_{RP}^2 = \sigma_E^2$	S_E^2
$\sigma_M^2 = \sigma_{RD}^2 + \sigma_{RP}^2$	$\frac{S_A^2 + (pr-1)S_E^2}{pr}$

TABLE A.5 ANOVA for the Mixed(M) model with interaction

Source of variation	Degrees of freedom	Mean square	Expected mean square
Parts (P)	$p - 1$	S_P^2	$\theta_P = \sigma_E^2 + r\sigma_{PA}^2 + ar\sigma_P^2$
Appraiser (A)	$a - 1$	S_A^2	$\theta_A = \sigma_E^2 + r\sigma_{PA}^2 + \frac{par}{a-1}\gamma_A$
P x A	$(p-1)(a-1)$	S_{PA}^2	$\theta_{PA} = \sigma_E^2 + r\sigma_{PA}^2$
Replicates (Error)	$pa(r-1)$	S_E^2	$\theta_E = \sigma_E^2$

TABLE A.6 Gauge R&R parameters and point estimators for the Mixed(M) model with interaction

Gauge R&R parameter	Point estimator
μ_Y	$\bar{Y}_{...}$
$\mu_{Y,j}$	$\bar{Y}_{.j.}$
σ_P^2	$\frac{S_P^2 - S_{PA}^2}{ar}$
γ_A	$\frac{(a-1)(S_A^2 - S_{PA}^2)}{par}$
$\sigma_{RD}^2 = \gamma_A + \sigma_{PA}^2$	$\frac{(a-1)S_A^2 + [a(p-1)+1]S_{PA}^2 - paS_E^2}{par}$
$\sigma_{RP}^2 = \sigma_E^2$	S_E^2
$\sigma_M^2 = \sigma_{RD}^2 + \sigma_{RP}^2$	$\frac{(a-1)S_A^2 + [a(p-1)+1]S_{PA}^2 + pa(r-1)S_E^2}{par}$

The reference linear model is:

$$Y_{ijk} = P_i + \mu_{Y,j} + (PA)_{ij} + E_{ijk} \tag{A.5}$$

$$i = 1, \cdots p, j = 1, \cdots a, k = 1, \cdots r \tag{A.5}$$

where $\mu_{Y,j}$ is a constant to the jth appraiser, P_i, $(PA)_{ij}$ and the error E_{ijk} are jointly independent normal random variables with means equal to zero and variances $\sigma_P^2, \sigma_{PA}^2$ and σ_E^2: therefore, we consider an *unrestricted* mixed model, as explained in Burdick et al.[2] When there is no interaction, the term $(PA)_{ij}$ is included within the error and a *restricted* model is studied. Tables A.5–A.8 show the ANOVA sources of variation and the point estimators, respectively, for the Mixed(M) model with and without interaction.

For the sake of brevity, the formulas of S_P^2, S_A^2, S_{PA}^2 and S_E^2 are not reported here: see Burdick et al.[2] or any textbook about experimental design for more details.

TABLE A.7 ANOVA for the Mixed(M) model without interaction

Source of variation	Degrees of freedom	Mean square	Expected mean square
Parts (P)	$p - 1$	S_P^2	$\theta_P = \sigma_E^2 + ar\sigma_P^2$
Appraiser (A)	$a - 1$	S_A^2	$\theta_A = \sigma_E^2 + pr(\frac{a\gamma_A}{a-1})$
Replicates (Error)	$par - p - a - 1$	S_E^2	$\theta_E = \sigma_E^2$

TABLE A.8 Gauge *R&R* parameters and point estimators for the Mixed (M) model without interaction

Gauge *R&R* parameter	Point estimator
μ_Y	$\bar{Y}_{...}$
$\mu_{Y,j}$	$\bar{Y}_{.j.}$
σ_P^2	$\frac{S_P^2 - S_E^2}{ar}$
$\sigma_{RD}^2 = \gamma_A$	$\frac{(a-1)(S_A^2 - S_E^2)}{par}$
$\sigma_{RP}^2 = \sigma_E^2$	S_E^2
$\sigma_M^2 = \sigma_{RD}^2 + \sigma_{RP}^2$	$\frac{(a-1)S_A^2 + (par-a+1)S_E^2}{par}$