Research Article

Assessment of the Adequacy of Gauge Repeatability and Reproducibility Study Using a Monte Carlo Simulation

Chunghun Ha,¹ David S. Kim,² and SeJoon Park³

¹School of Information & Computer Engineering, Hongik University, 94 Wausan-ro, Mapo-gu, Seoul 04066, Republic of Korea

²School of Mechanical, Industrial and Manufacturing Engineering, Oregon State University, Corvallis, OR 97330, USA

³Department of Industrial and Management Engineering, Myongji University, 116 Myonggi-Ro, Cheoin-Gu, Yongin-Si, Gyeonggi-Do 449-728, Republic of Korea

Correspondence should be addressed to SeJoon Park; sonmupsj@hanmail.net

Received 7 March 2017; Accepted 20 June 2017; Published 15 August 2017

Academic Editor: J.-C. Cortés

Copyright © 2017 Chunghun Ha et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ANOVA gauge repeatability and reproducibility study is the most popular tool for measurement system analysis. Two experimental designs can be applied depending on the durability of the objects. If repeated measurements are possible or sufficient homogeneous nonrepeatable samples are available, crossed design is appropriate; otherwise, nested design should be used. In this paper, we investigated the adequacy of ANOVA gauge repeatability and reproducibility study from the perspective of practitioners. We proposed a Monte Carlo simulation that is close to the realistic procedure to evaluate the adequacy of both structures. During the evaluation, we considered the average performance metrics, percentage of correct decision, histogram shape, and symmetric mean absolute percentage error for the four popular performance metrics, namely, % Study Variation, % Contribution, % Tolerance, and the number of distinct categories. The experimental results show that the nested design fails to judge the precision of the gauge while the crossed design succeeds.

1. Introduction

Gauge repeatability and reproducibility (GRR) study is a representative measurement system analysis (MSA) tool [1]. Two factors determine the adequacy of a measurement system: accuracy, such as bias, linearity, stability, and correlation, and precision such as repeatability and reproducibility. The main concern of the GRR is that a measurement system has sufficient precision to measure the variation of the manufactured products or the manufacturing process under consideration. There are three conventional GRR methods; the range method, the average and range method using control chart, and the analysis of variance (ANOVA) GRR (AGRR) [1]. After the AGRR was introduced by Montgomery and Runger [2, 3], it became the most popular tool for MSA as it considers the interaction effects and provides interval estimates for the variance components and the performance metrics [4]. The ANOVA in AGRR measures the variability of observations and estimates variance components. The performance metrics, which are composed of sums or ratios of the estimated

variance components, provide the criteria used to analyze the precision for the measurement system. Crossed designs are standard experimental layouts for AGRR. Nested designs, or hierarchical designs, are used for nonrepeatable measurements such as a destructive test. Though the measured object is nonrepeatable, if sufficient homogeneous samples are captured, the crossed design will be appropriate [5].

For the last two decades, numerous studies have been conducted on AGRR. Previous studies mainly concentrated on providing theoretical backgrounds, introductions of the AGRR, efficient approximations for narrower confidence intervals of variance components and performance metrics, and variations of AGRR for special experimental structures. In fact, AGRR is a popular tool used in the industrial field. QS-9000, a quality standard of the American automotive industry, even provides a guideline for AGRR [6]. From the practitioners' perspective, previous theoretical studies are not as such valuable. The main reason for this is the wide confidence intervals. Theoretical studies have mainly focused

on developing efficient approximations on the wide confidence intervals. Since the AGRR is based on sampling, it is reasonable that the confidence intervals give clearer evidence than the point estimates to arrive at the correct conclusions. However, as Burdick and Larsen [4] noted, the estimated confidence intervals are often too wide to be used. In many cases, the confidence intervals overlap with the decision criteria of AGRR in Table 3, making them unsuitable for assessing the adequacy of the gauge. Therefore, practitioners choose the point estimates of the performance metrics.

In this situation, it is imperative to verify the adequacy of AGRR with the point estimates. Especially, as Bergeret et al. [7] mentioned, the adequacy of the nested design is quite doubtful. Though the theoretical basis of AGRR is very firm, there are several possible sources to harm the adequacy of the AGRR. The fundamental assumption of AGRR is that all effects including interaction follow a normal distribution and those are independent. If we filter outliers during inspection or select samples arbitrarily to secure a sufficient range of variability, the normality or the independent assumption breaks. The nested design itself is another source of decreasing adequacy. The nested effects interferes to separate the variance components clearly, and, consequently, the accuracy of the performance metrics decreases.

Practitioners are not concerned with theoretical derivations or proofs. Their primary concern with the AGRR is that the tool works properly to determine the precision of the gauge and, if possible, to find ways to improve the adequacy of the AGRR within budget constraints. Theoretical, and even practical, studies have not dealt with these issues. Existing practical studies only focused on offering user guidelines or providing case studies on various applications. The purpose of this paper is to evaluate the adequacy of the AGRR for both the crossed and nested designs and to investigate the causes of any inadequacies. To accomplish this, we constructed a series of Monte Carlo simulations and verified the adequacy via four popular performance metrics, % Study Variation, % Contribution, % Tolerance, and the number of distinct categories (NDC) [8].

Section 2 introduces the conventional AGRR process for both crossed and nested designs and compares the differences in formulas for the performance metrics. In Section 3, we briefly review existing references on the AGRR. The proposed Monte Carlo simulation method and experimental environments are described in detail in Section 4. We summarize the simulation results for various perspectives, show why the nested design AGRR is unsuitable for MSA, and reveal the cause of inadequacy in Section 5. Finally, in Section 6, the conclusions and further discussion of this paper are provided.

2. ANOVA Gauge Repeatability and Reproducibility Study

A standard AGRR uses the crossed design. The two-way random effect model is as follows:

$$y_{ijk} = \mu + O_i + S_j + (SO)_{ij} + E_{(ij)k}, \quad (1)$$

$$i \in \{1, 2, \dots, o\}, \quad j \in \{1, 2, \dots, s\}, \quad k \in \{1, 2, \dots, r\},$$

where y_{ijk} is an observation; μ is the unknown overall mean; $S, O, (SO)$, and E are random variables that represent the effects of the sample, the operator, the interaction between the sample and the operator, and the replicate, respectively; o, s , and r are the number of operators, samples, and replicates, respectively. It is generally assumed that $O_i \sim N(0, \sigma_O^2)$, $S_j \sim N(0, \sigma_S^2)$, $(SO)_{ij} \sim N(0, \sigma_{SO}^2)$, and $E_{(ij)k} \sim N(0, \sigma_E^2)$, and those are independent of each other. Left side of Figure 1 shows an experimental structure of the crossed design with 2 operators, 4 parts, and 2 replicates. In the figure, the number in the observations indicates ijk . In the crossed design, two operators measure four distinct samples twice.

Table 1 shows the resulting ANOVA table. If the interaction effect is negligible, that is, V_{SO}/V_E is less than the significant level, it is pooled into error terms; that is, E and the table are changed. For detailed explanation, refer to Montgomery and Runger [2, 3].

We can estimate the variance components as follows:

$$\begin{aligned} \hat{\sigma}_E^2 &= V_E, \\ \hat{\sigma}_{SO}^2 &= \frac{(V_{SO} - V_E)}{r}, \\ \hat{\sigma}_S^2 &= \frac{(V_S - V_{SO})}{sr}, \\ \hat{\sigma}_O^2 &= \frac{(V_O - V_{SO})}{or}. \end{aligned} \quad (2)$$

If the samples are destructive, we must apply the nested design instead of the standard crossed design. The right side of Figure 1 shows a nested experimental structure with two operators, four parts, and four replicates. The experiment is a counterpart of the crossed design on the left side with the same number of observations. In the experiment, two operators measure sixteen distinct samples in four batches (or lots). If the samples in a batch are homogeneous and enough samples are available, the crossed design can still be effective. The two-way random effects model for the nested design is as follows:

$$y_{ijk} = \mu + O_i + S(O)_{j(i)} + E_{(ij)k}, \quad (3)$$

$$i \in \{1, 2, \dots, o\}, \quad j \in \{1, 2, \dots, s\}, \quad k \in \{1, 2, \dots, r\},$$

where y_{ijk} is an observation and μ is the unknown overall mean. $O, S(O)$, and E are random variables that represent effects of operator, sample nested operator, and replicate, respectively. o, s , and r are the number of operators, samples per operator, and replicates, respectively. It is also assumed that $O_i \sim N(0, \sigma_O^2)$, $S(O)_{ij} \sim N(0, \sigma_{S(O)}^2)$, and $E_{(ij)k} \sim N(0, \sigma_E^2)$ and that these are independent of each other. Table 2 shows the ANOVA Table of the nested design.

FIGURE 1: Examples of experimental designs for a crossed design and a nested design.

TABLE 1: ANOVA table of the crossed design under two-way random effects model.

Source of variability	Sum of square (S)	Degrees of freedom (d)	Mean of square (V)	Expected mean square	F
Operator (O)	S_O	$d_O = o - 1$	$V_O = \frac{S_O}{d_O}$	$\theta_O = \sigma_E^2 + r\sigma_{SO}^2 + or\sigma_O^2$	V_O/V_{SO}
Sample (S)	S_S	$d_S = s - 1$	$V_S = \frac{S_S}{d_S}$	$\theta_S = \sigma_E^2 + r\sigma_{SO}^2 + sr\sigma_S^2$	V_S/V_{SO}
Interaction: $S \times O$	S_{SO}	$d_{SO} = (s - 1)(o - 1)$	$V_{SO} = \frac{S_{SO}}{d_{SO}}$	$\theta_{SO} = \sigma_E^2 + r\sigma_{SO}^2$	V_{SO}/V_E
Replicate (E)	S_E	$d_E = so(r - 1)$	$V_E = \frac{S_E}{d_E}$	$\theta_E = \sigma_E^2$	
Total	S_T	$sor - 1$			

The estimates of the variance components are as follows:

$$\begin{aligned}\hat{\sigma}_E^2 &= V_E, \\ \hat{\sigma}_{S(O)}^2 &= \frac{(V_{S(O)} - V_E)}{r}, \\ \hat{\sigma}_O^2 &= \frac{(V_O - V_{S(O)})}{or}.\end{aligned}\quad (4)$$

The goal of AGRR is to determine whether the measurement system can distinguish variation of products or processes properly. To do that, the AGRR extracts the gauge error (repeatability) and the operator error (reproducibility) from the observed measurements and judges the adequacy via performance metrics. The most popular performance metrics in practice are % Study Variation, % Contribution, % Tolerance, and the number of distinct categories (NDC). Popular software in the quality field, Minitab, provides the four metrics for AGRR [8]. We summarize the calculation formula and relevant decision criteria of the metrics for the crossed design and nested design in Table 3.

According to the formulas in Table 3, under the assumption of the same estimates of the variance components, we can surmise that the values of all performance metrics of the nested design are superior to the crossed design. In the crossed design, the measurement variation $\hat{\sigma}_{R\&R}^2$ includes the

interaction effect $\hat{\sigma}_{SO}^2$, and it makes the values worse. However, practical results oppose this theoretical analysis. Bergert et al. [7] claimed that % Contribution and % Tolerance of the nested design are overestimated when compared with the crossed design. They investigated three case studies and argued that improper estimation of repeatability results in the overestimation of the performance metrics. In this situation, it is valuable to determine whether the AGRR, specially the nested design, is indeed an appropriate tool to determine precision of the gauge and to investigate how accurate the AGRR is.

3. Previous Studies

In this section, we briefly review previous studies on AGRR. The mainstream theoretical developments on AGRR include accurate and efficient approximation approaches for narrower confidence intervals of the variance components and performance metrics, methods for improving the accuracy of the AGRR, and the methods for nonrepeatable measurements.

Montgomery and Runger [2, 3] introduced AGRR as an alternative to the conventional GRRs such as the range method and the average and range method. They also suggested a proper experimental design for AGRR and the

TABLE 2: ANOVA table of the nested design under two-way random effects model.

Source of variability	Sum of square (S)	Degrees of freedom (d)	Mean of square (V)	Expected mean square	F
Operator (O)	S_O	$d_O = o - 1$	$V_O = \frac{S_O}{d_O}$	$\theta_O = \sigma_E^2 + r\sigma_{S(O)}^2 + or\sigma_O^2$	$V_O/V_{S(O)}$
Sample (Operator) (S(O))	$S_{S(O)}$	$d_{S(O)} = o(s - 1)$	$V_{S(O)} = \frac{S_{S(O)}}{d_{S(O)}}$	$\theta_{S(O)} = \sigma_E^2 + r\sigma_{S(O)}^2$	$V_{S(O)}/V_E$
Replicates (E)	S_E	$d_E = so(r - 1)$	$V_E = \frac{S_E}{d_E}$	$\theta_E = \sigma_E^2$	
Total	S_T	$sor - 1$			

Satterthwaite approximation for confidence intervals of the variance components. Borrer et al. [9] compared two approximations, the restricted maximum likelihood estimation (REML) using SAS PROC MIXED, and a modified large sample (MLS) method to estimate confidence interval of $\sigma_{R\&R}^2$ for the two-way random effects model. They claimed that the REML is superior to the MLS due to the narrower confidence interval. Burdick and Larsen [4] compared five approximation approaches, that is, MLS [10], Satterthwaite approximation [3], AIAG [1], REML [11], and the Milliken and Johnson method [12] for five variations, namely, σ_E^2 , σ_O^2 , σ_{RPD}^2 , $\sigma_{R\&R}^2$, and $\sigma_S^2/\sigma_{R\&R}^2$. The simulation results showed that the MLS is superior to others since it satisfies the confidence coefficient in spite of wider confidence interval. Dolezal et al. [13] investigated the confidence interval of $\sigma_{R\&R}^2$ for a two-way mixed effects model with fixed operators. They suggested the mixed effects model for a limited number of operators because the interval length is shorter than conventional random effects model. Hamada and Weerahandi [14] proposed a modified generalized inference approximation for $\sigma_{R\&R}^2$ and argued that it provides a shorter confidence interval than the MLS [4]. Chiang [15] also proposed an approximation for $\sigma_{R\&R}^2$ using the surrogate variables. He compared the confidence interval to the MLS and insisted that it is an effective general method for the balanced random effects model. Daniels et al. [16] employed the generalized confidence interval approach using a generalized pivotal quantity. They stated that the approach is superior to the MLS if $\sigma_O^2/\sigma_{R\&R}^2$ is less than or equal to 0.2. Wang and Li [17] proposed a bootstrapping method that can estimate the confidence interval when the control chart GRR is applied.

As for the performance metrics and their confidence intervals, Burdick et al. [18] stated that the confidence interval of $\sigma_S^2/\sigma_{R\&R}^2$ is too wide to be used, hence they recommended the Cochran method based on Satterthwaite approximation [19]. Chiang [20] argued that the confidence coefficients of the MLS and Satterthwaite approximation for $\sigma_S^2/\sigma_{R\&R}^2$ become low when σ_S^2 is less than 0.5. To overcome this phenomenon, he suggested the F-screened MLS that applies the MLS only when σ_S^2 is statistically significant. Burdick et al. [21] reviewed previous research on AGRR and stated that precision-to-tolerance ratio (% Tolerance), signal-to-noise ratio (SNR), and discrimination ratio (DR) are popular performance metrics for AGRR. Adamec and Burdick [22] compared the performance of the MLS and the generalized

inference procedure for the DR in a three-way random effects model. Burdick et al. [23] proposed the generalized inference procedure for the misclassification rate. Woodall and Borrer [24] reviewed and analyzed relationships for popular performance metrics, % Study Variation [1, 8], NDC [1, 8], SNR [1, 21], DR [1], and misclassification rates.

There were a few studies on improving the accuracy of AGRR. Pan [6] calculated the optimal set of (o, s, r) to provide the shortest confidence interval of $\sigma_{R\&R}^2$ for variety combinations of σ_O^2 , $\sigma_{S(O)}^2$, and σ_E^2 under the same number of observations. Browne et al. [25] proposed two-staged AGRR to increase adequacy of AGRR. At a baseline stage, a number of operators measure a sample to obtain an appropriate range of samples. Then, at the second stage, a standard AGRR is conducted with the samples. They argued that the approach provides shorter standard deviations for $\sigma_{R\&R}/\sigma_T$ and $\sigma_O/\sigma_{R\&R}$ than the normal one-staged AGRR. Pan et al. [26] suggested a revised % Tolerance for the multivariate GRR that provides smaller mean squared error and mean absolute percentage error than the conventional % Tolerance. They also calculated the optimal (o, s, r) in terms of the new performance metric using the principal component analysis.

Research on the nested design is rare. Bergeret et al. [7] applied the nested design for three case studies with destructive samples and argued that the nested design overestimates % Study Variation and % Tolerance. Mast and Trip [27] introduced four assumptions for AGRR that are the consistency of bias, homogeneity of measurement errors, temporal stability of objects, and robustness against measurement. They defined a nonrepeatable measurement as the measurement that the last two assumptions are not satisfied. Furthermore, they proposed several alternative AGRRs that are suitable for various homogeneity assumptions. Van Der Meulen et al. [28] developed a compensation method to improve overestimation of the nested design for nonrepeatable measurements.

There are many practical studies on AGRR, but those mainly focused on introducing basic theoretical knowledge, suggesting systematic user guidelines, and providing case studies for various applications. We will skip the review for the references.

4. Experimental Setup

4.1. Procedure for Simulation Experiment. Adequacy implies the ability to perform the desired goal. The purpose of AGRR is to determine the sufficiency of precision of a measurement

TABLE 3: Variance components, performance metrics, and their decision criteria of the crossed design and the nested design.

Components	Equation	Crossed design	Nested design	Criteria
Repeatability	$\hat{\sigma}_{RPT}^2$	$\hat{\sigma}_E^2$	$\hat{\sigma}_E^2$	
Reproducibility	$\hat{\sigma}_{RPD}^2$	$\hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2$	$\hat{\sigma}_O^2$	
R&R	$\hat{\sigma}_{RRR}^2 = \hat{\sigma}_{RPT}^2 + \hat{\sigma}_{RPD}^2$	$\hat{\sigma}_E^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2$	$\hat{\sigma}_E^2 + \hat{\sigma}_O^2$	
Sample	$\hat{\sigma}_S^2$ or $\hat{\sigma}_{S(O)}^2$	$\hat{\sigma}_S^2$	$\hat{\sigma}_{S(O)}^2 = \hat{\sigma}_S^2 + \hat{\sigma}_{SO}^2$	
Total	$\hat{\sigma}_T^2$	$\hat{\sigma}_S^2 + \hat{\sigma}_E^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2$	$\hat{\sigma}_S^2 + \hat{\sigma}_E^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2$	
% Study Variation	$\hat{\sigma}_{RRR}/\hat{\sigma}_T \times 100$	$\sqrt{\frac{\hat{\sigma}_E^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2}{\hat{\sigma}_S^2 + \hat{\sigma}_E^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2}} \times 100$	$\sqrt{\frac{\hat{\sigma}_E^2 + \hat{\sigma}_O^2}{\hat{\sigma}_S^2 + \hat{\sigma}_E^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2}} \times 100$	≤10% acceptable 10~30% pending >30% unacceptable
% Contribution	$\hat{\sigma}_{RRR}^2/\hat{\sigma}_T^2 \times 100$	$(\hat{\sigma}_E^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2) / (\hat{\sigma}_S^2 + \hat{\sigma}_E^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2) \times 100$	$(\hat{\sigma}_E^2 + \hat{\sigma}_O^2) / (\hat{\sigma}_S^2 + \hat{\sigma}_E^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2) \times 100$	≤1% acceptable 1~9% pending >9% unacceptable
% Tolerance	$5.15 \times \hat{\sigma}_{RRR} / \text{Tolerance} \times 100$	$5.15 \times (\hat{\sigma}_E^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2) / \text{Tolerance}$	$5.15 \times (\hat{\sigma}_E^2 + \hat{\sigma}_O^2) / \text{Tolerance}$	≤10% acceptable 10~30% pending >30% unacceptable
Number of distinct category (NDC)	$1.41 \times \hat{\sigma}_S / \hat{\sigma}_{RRR}$	$1.41 \times \sqrt{\hat{\sigma}_S^2 / (\hat{\sigma}_E^2 + \hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2)}$	$1.41 \times \sqrt{\hat{\sigma}_S^2 / (\hat{\sigma}_E^2 + \hat{\sigma}_O^2)}$	≥5 acceptable <5 unacceptable

Performance metric

system by statistics. To verify the adequacy of AGRR, we require information on the population, in other words, the true values must be known. However, it is impossible to obtain complete data for a population. One alternative, for overcoming this problem, is to use simulation. Most existing studies applied a Monte Carlo simulation for verification. If an effect $Q \sim N(0, \sigma_Q^2)$, then $d_Q V_Q / \theta_Q \sim \chi^2(d_Q)$ or $V_Q \sim \theta_Q \chi^2(d_Q) / d_Q$, where d_Q is the degree of freedom of Q and V_Q is the mean of squares of Q . From the relationship, the population of V_Q can be generated using Chi-squared distribution, and subsequently, the estimates of variance components and performance metrics can be calculated. This simulation approach, however, has two weaknesses. First, the normality assumption of effects must be satisfied. If we limit the random sampling by inspection processes or select samples or operators arbitrarily to obtain better results, we cannot generate the population due to the broken normality condition. Second, the true values of variance components and performance metrics are still unknown, therefore the adequacy of AGRR cannot be judged. Therefore, we propose a new Monte Carlo simulation for verification of the adequacy of AGRR. This approach generates populations of all effects instead of populations of the mean of the squares. An observation will then consist of the sampled effects from the populations. Since the true variance components can be calculated from the populations, intensive evaluation is possible. In addition, during the procedure, it is possible to employ various realistic constraints such as inspection, without loss of generality. The detailed procedure of the proposed Monte Carlo simulation is in Algorithm 1.

At the beginning of each scenario, the levels of $\tilde{\sigma}_S^2$, $\tilde{\sigma}_O^2$, $\tilde{\sigma}_{SO}^2$, and $\tilde{\sigma}_E^2$ are assigned. The total number of scenarios N_{scenario} is 140 since the number of levels of $\tilde{\sigma}_S^2$ is 14 in Table 4 and there are 10 combinations of $\tilde{\sigma}_O^2$, $\tilde{\sigma}_{SO}^2$, and $\tilde{\sigma}_E^2$ at each level of $\tilde{\sigma}_S^2$ in Table 5. The population of each effect is generated by the truncated normal distribution bounded by $\pm 3\tilde{\sigma}_Q$. In practice, an inspection process may filter outliers of the products, so the truncated normal assumption is reasonable for the actual samples. Normal distribution can be used if this is insignificant. The size of each population S_{pop} is 10,000 except for the interaction, which is 10,000 by 10,000. The true values of the performance metrics are computed by the crossed design formulas in Table 3 with population parameters, σ_S^2 , σ_O^2 , σ_{SO}^2 , and σ_E^2 . A set of observations is generated by Equations (1) and (3) with the sampled effects from the populations. At this time, the structure of observations follows the experimental design in Table 6. Section 4.3 explains the experimental design in more detail. At next step, AGRRs using the crossed and nested designs estimate the variance components, where the significance level for determining pooling of the interaction is 0.05. For every AGRR, the performance metrics, % Study Variation, % Contribution, % Tolerance, and NDC are calculated by the formulas in Table 3 where the Tolerance is $6\tilde{\sigma}_S$. These steps are repeated N_{repeat} (100) times.

4.2. Simulation Parameters. The sample variance σ_S^2 affects the % Study Variation, % Contribution, and NDC among the

TABLE 4: Levels of $\tilde{\sigma}_S^2$ and $\tilde{\sigma}_O^2 + \tilde{\sigma}_{SO}^2 + \tilde{\sigma}_E^2$ for population generation.

$\tilde{\sigma}_S^2$	$\tilde{\sigma}_O^2 + \tilde{\sigma}_{SO}^2 + \tilde{\sigma}_E^2$
8192	2^{13}
4096	2^{12}
2048	2^{11}
1024	2^{10}
512	2^9
256	2^8
128	2^7
64	2^6
32	2^5
16	2^4
8	2^3
4	2^2
2	2^1
1	2^0

TABLE 5: Levels of $\tilde{\sigma}_O^2$, $\tilde{\sigma}_{SO}^2$, and $\tilde{\sigma}_E^2$ for population generation.

$\tilde{\sigma}_O^2$	$\tilde{\sigma}_{SO}^2$	$\tilde{\sigma}_E^2$
1	1	1
0.8	1.4	0.8
1.4	0.8	0.8
0.8	0.8	1.4
1	1.4	0.6
0.6	1.4	1
1	0.6	1.4
1.4	0.6	1
1.4	1	0.6
0.6	1	1.4

TABLE 6: Levels of factors of experimental design.

Factors	Level (low, high)	
	Crossed design	Nested design
Number of operators	$o_C \in \{3, 5\}$	$o_N = o_C$
Number of samples per operator	$s_C = o_C s_N$	$s_N \in \{3, 5\}$
Number of replicates	$r_C \in \{2, 4\}$	$r_N = o_N r_C$
Number of observations	$o_C s_C r_C$	$o_N s_N r_N$

four performance metrics. To investigate the effect of σ_S^2 , we set $\tilde{\sigma}_S^2$ from 2^0 (1) to 2^{13} (8192) in Table 4, while $\tilde{\sigma}_O^2 + \tilde{\sigma}_{SO}^2 + \tilde{\sigma}_E^2$ is three. The variations, σ_O^2 , σ_{SO}^2 , and σ_E^2 affect all the performance metrics. To analyze the effects of the variance components, ten orthogonal sets of $\tilde{\sigma}_O^2$, $\tilde{\sigma}_{SO}^2$, and $\tilde{\sigma}_E^2$ are designed as shown in Table 5. Since $\tilde{\sigma}_O^2 + \tilde{\sigma}_{SO}^2 + \tilde{\sigma}_E^2$ is fixed at three, the experimental design seems to be limited. However, except for % Tolerance, the performance metrics do not depend on the magnitude but the ratio among the variance components. In this perspective, our experimental design can cover a total of 140 scenarios and the number is sufficient.

```

For 1 to  $N_{\text{scenario}}$  Do:
  Step 1. Assign a set of variance components,  $(\tilde{\sigma}_O^2, \tilde{\sigma}_S^2, \tilde{\sigma}_{SO}^2, \tilde{\sigma}_E^2)$  according to the assigned experimental design in Tables 4 and 5.
  Step 2. Generate  $S_{\text{pop}}$  size of populations of effects with  $\text{TN}(0, \tilde{\sigma}_O^2)$ ,  $\text{TN}(0, \tilde{\sigma}_S^2)$ ,  $\text{TN}(0, \tilde{\sigma}_{SO}^2)$ , and  $\text{TN}(0, \tilde{\sigma}_E^2)$ , where TN denotes a truncated Normal distribution bounded by  $[-3\tilde{\sigma}_Q, +3\tilde{\sigma}_Q]$  for  $Q \in \{O, S, SO, E\}$ .
  Step 3. Calculate true variance components  $(\sigma_O^2, \sigma_S^2, \sigma_{SO}^2, \sigma_E^2)$  and the true values of performance metrics from the populations of effects.
  For 1 to  $N_{\text{repeat}}$  Do:
    For 1 to  $N_{\text{experimental design}}$  Do:
      Step 4. Generate observations.
      Step 4-1. Select samples from each population of effects by the experimental design in Table 6.
      Step 4-2. Generate observations by the structural models of Equation (1) and (3), respectively.
      Step 5.
      Step 5-1. Perform the AGRR with the observations.
      Step 5-2. Calculate variance components  $(\hat{\sigma}_O^2, \hat{\sigma}_S^2, \hat{\sigma}_{SO}^2, \hat{\sigma}_E^2)$ 
      Step 5-3. Calculate the performance metrics for the crossed design and the nested design, respectively.
    End
  Step 6. Calculate mean of performance metrics for the crossed design and the nested design.
End
End

```

ALGORITHM 1

4.3. Experimental Design for the Crossed and Nested Designs. Direct comparison of performance metrics between the crossed design and the nested design is meaningless because they are used in different environments. However, it is valuable to evaluate and compare the adequacy levels. To do that, we designed two structures that use the same observations, for the experiment. For example, in Figure 1, we use the same observations in both the designs. The experimental design to maintain the same number of observations for the crossed and the nested design is as follows.

The number of operators, samples, and replicates also affects the performance metrics. To investigate their effects, we set up a 2^3 factorial design for the number of operators using the crossed design (o_C). The number of samples per operator in the nested design (s_N) and the number of replicates in the crossed design (r_C) are shown in Table 6. In order to match the number of observations, the number of operators in the nested design (o_N), the number of samples in the crossed design (s_C), and the number of replicates in the nested design (r_N) are assigned o_C , $o_C s_N$, and $o_N r_C$, respectively.

5. Experimental Results

5.1. Performance Metrics. Figure 2 shows the trajectories of the averages of the four performance metrics for the crossed design, the nested design, and the population over $\tilde{\sigma}_S^2$. The two dashed horizontal lines indicate the rule-of-thumb decision criterion of each performance metric. The regions I, II, and III represent the acceptable, the pending, and the unacceptable regions, respectively, based on the performance metrics of population. Since $\tilde{\sigma}_O^2 + \tilde{\sigma}_{SO}^2 + \tilde{\sigma}_E^2$ and tolerance are constants, the performance metrics of the crossed design are functions of $\tilde{\sigma}_S^2$. As for the nested design, the performance metrics are also functions of $\tilde{\sigma}_S^2$ because averaging

and orthogonal designs compensate the individual effects of $\tilde{\sigma}_O^2$, $\tilde{\sigma}_{SO}^2$, and $\tilde{\sigma}_E^2$. In Figure 2, all average performance metrics of the crossed design are very close to the values of the population that are true values. It implies that the adequacy level of the crossed design is very high. However, the metrics of the nested design differ from the values of the population. In particular, the trajectory of nested design for the % Tolerance differs from the trajectory of the population. It is caused by the overestimation of $\tilde{\sigma}_S^2$ (Section 5.4. will elaborate on this). Moreover, the population shares the formulas for performance metrics with the crossed design. The σ_{SO}^2 is nested into σ_S^2 in the nested design but it is added to $\sigma_{R\&R}^2$ in the crossed design and the population. In theory, the gap between the nested design and the population should decrease in region III since $\tilde{\sigma}_S^2$ is much bigger than $\tilde{\sigma}_{SO}^2$. Region III is important because the metrics of a good measurement system are positioned at the region. Therefore, we can conclude that nested design does not provide the correct result on gauge precision.

5.2. Percentage of Correct Decision. To investigate the nested design further, we employed a new metric, the percentage of correct decision (PCD). The PCD is the percentage ratio of the same decision of the crossed or nested designs, to the population. Figure 3 shows the PCD over $\tilde{\sigma}_S^2$ for each performance metric and shows that the vertical dashed lines and the regions are equivalent to Figure 2. The PCD of the crossed design decreases to about 50% around the borderlines of decision criteria and close to one at the other range of $\tilde{\sigma}_S^2$. It is reasonable because even a small variation of the performance metrics at the border lines results in a different decision. If the AGRR works correctly, the PCD must be close to 100% except for the borderlines. However, the trajectories of the PCD of the nested design decrease rapidly up to almost zero and do

FIGURE 2: Average performance metrics over the sample variance.

not recover to 100% until a range of $\tilde{\sigma}_S^2$. In region III, those are about 60% (in case of NDC, 70%). It implies that the decision of the nested design is almost random.

5.3. Effects of the Allocation of the Variance Components and the Experimental Design. In this subsection, we investigate the causes of the poor quality of the nested design in the perspective of allocation of variance components and the experimental design. Figure 4 shows the % Study Variation for the allocations of variance components ($\tilde{\sigma}_O^2, \tilde{\sigma}_{SO}^2, \tilde{\sigma}_E^2$) in Table 5, and Figure 5 shows % Study Variation for the

experimental designs, (o_C, s_C, r_C) and (o_N, s_N, r_N) , in Table 6. As mentioned in Section 5.1, the metric should be close to the population as $\tilde{\sigma}_S^2$ increases. However, the gap is still large in the region III irrespective of the allocation of $\tilde{\sigma}_O^2$, $\tilde{\sigma}_{SO}^2$, and $\tilde{\sigma}_E^2$, while the gap of the crossed design is close to zero. This phenomenon is very similar to that in the experimental design. In general, increasing degree of freedom improves the estimation quality of variance components in AGRR; thus it decreases the gap to the true value. However, in Figure 6, changing the experimental design does not improve the gap of the nested design significantly. The adequacy of the

FIGURE 3: Percentage of correct decision by performance index for all populations.

nested design is still very low at all regions. From the above results, we can conclude that the allocation of the variance components and the experimental design are not critical to the poor adequacy of the nested design.

5.4. Robustness of AGRR. Next, we draw histograms of the repeatability and reproducibility at four distinct $\tilde{\sigma}_S^2$, as shown in Figures 6 and 7, respectively. We fixed the other parameters to eliminate the side effects: $(\tilde{\sigma}_O^2, \tilde{\sigma}_{SO}^2, \tilde{\sigma}_E^2)$ as (1, 1, 1), (o_C, s_C, r_C) as (3, 9, 2), and (o_N, s_N, r_N) as (3, 3, 6). As for repeatability, as shown in Figure 6, both designs have histograms similar to the Chi-squared distribution at every $\tilde{\sigma}_S^2$.

It coincides with the theoretical result. However, this is not so with reproducibility. In the case of the crossed design, the shape of the histogram still seems to be Chi-squared distribution. However, as shown in Figure 7, the histograms of the nested design differ, which have many zeroes and spread widely over all ranges. The zeroes of reproducibility make the effect of the operator statistically insignificant; hence, the results of AGRR are unstable. This implies that the estimation of reproducibility of the nested design is inadequate and it could be the main reason for the inadequacy of the design.

Table 7 shows the estimates of the variance components at fourteen $\tilde{\sigma}_S^2$ s for the population, the crossed design, and

FIGURE 4: Average % Study Variation for the initial variance components.

FIGURE 5: Average % Study Variation for the experimental designs.

the nested design. All estimates of the crossed design are very close to the estimates of the population. On the other hand, the AGRR of the nested design overestimates the reproducibility, $\hat{\sigma}_O^2$, while it properly estimates the repeatability and $\hat{\sigma}_O^2 + \hat{\sigma}_{SO}^2$. The overestimation of $\hat{\sigma}_O^2$ increases according

to σ_S^2 . Since several estimates of $\hat{\sigma}_O^2$ are zero in Figure 7, its overestimation implies that the nonzero estimates are very large. In the nested design, an operator does not share samples with other operators; therefore, it is hard to separate the variability of the operator and the sample in ANOVA effectively.

FIGURE 6: Histogram of sample repeatability from four different populations.

That is, the variability of σ_s^2 leaks to estimates of $\hat{\sigma}_O^2$, and, consequently, it increases $\hat{\sigma}_O^2$ and makes the adequacy low.

5.5. Evaluation of Robustness. The AGRR is based on sampling statistics. Even though the population is the same, each run of the AGRR can differ. If an AGRR method is reliable and robust, the variance of the results should be small. To investigate the robustness of AGRR, we employed the symmetric mean absolute percentage error (sMAPE) as follows [29]:

$$\text{sMAPE}_k = \frac{1}{n} \cdot \sum_{i=1}^n \frac{|\hat{\sigma}_{Q,i}^2 - \sigma_{Q,p}^2|}{(\hat{\sigma}_{Q,i}^2 + \sigma_{Q,p}^2)/2} \cdot 100, \quad (5)$$

where $\hat{\sigma}_{Q,i}^2$ denotes estimates of Q with the i th set of observations and $\sigma_{Q,p}^2$ is the true value of Q from the population, n is the number of repetition, and k is the index of scenario. sMAPE is a non-scale dependent variability measure; therefore the range is from -200% to 200% . However, since the values of $\hat{\sigma}_{Q,i}^2$ and $\sigma_{Q,p}^2$ are all nonnegative, in our case, sMAPE_k is nonnegative. The smaller the sMAPE_k , the lower the variability. Figure 8 shows the averaged sMAPE_k of repeatability, reproducibility, $R\&R$, and $\hat{\sigma}_s^2$ for the crossed design or $\hat{\sigma}_s^2 + \hat{\sigma}_{SO}^2$ for the nested design. As for the crossed design, all averaged sMAPE_k are very stable. On the other hand, the averaged sMAPE_k of the nested design are unstable except repeatability. The severity of the reproducibility and, consequently, the

TABLE 7: Average estimates of variance components of the population, the crossed design, and the nested design.

Population			Crossed design			Nested design		
σ_E^2	σ_O^2	σ_{SO}^2	σ_S^2	Repeatability $\frac{\sigma_E^2}{\sigma_O^2 + \sigma_{SO}^2}$	Reproducibility $\frac{\sigma_O^2 + \sigma_{SO}^2}{\sigma_S^2}$	Repeatability $\frac{\sigma_E^2}{\sigma_O^2}$	Reproducibility $\frac{\sigma_O^2}{\sigma_S^2 + \sigma_{SO}^2}$	$\sigma_S^2 + \sigma_{SO}^2$
0.978323	0.983632	0.973362	0.977343	0.991771	1.948983	0.977858	1.066857	1.971704
0.967414	0.966841	0.973394	1.946821	0.981132	1.919912	0.968178	1.109688	2.912493
0.975347	0.976984	0.973246	3.874717	0.987337	1.925811	0.97622	1.266511	4.876734
0.970432	0.970866	0.973523	7.757004	0.984564	1.944843	0.971788	1.63028	8.772708
0.97772	0.971914	0.973365	15.48564	0.995433	1.942836	0.982204	2.356923	16.44207
0.969531	0.972552	0.973272	30.93152	0.981883	1.940877	0.970869	3.81615	31.86321
0.973194	0.97107	0.973456	62.02678	0.986751	1.937034	0.975096	6.844492	62.9203
0.969944	0.973756	0.973213	124.708	0.980781	1.950806	0.968457	12.68279	126.2736
0.968533	0.972231	0.973198	249.9384	0.982749	1.956499	0.969171	25.92697	251.5965
0.969467	0.96702	0.973414	497.5466	0.983512	1.931515	0.968052	50.91509	495.8767
0.972785	0.971262	0.973209	994.6194	0.98618	1.940338	0.975168	99.09859	997.4504
0.9712	0.970518	0.973416	1999.386	0.984346	1.936489	0.973287	189.851	1992.567
0.97488	0.967591	0.97352	3993.367	0.987316	1.933699	0.974775	383.4533	4018.364
0.97684	0.975879	0.973405	7910.442	0.991661	1.943502	0.979389	781.8932	7959.333

FIGURE 7: Histogram of sample reproducibility from four different populations.

$R\&R$ are critical in regions II and III, where the value is over 150%. This result implies that the reproducibility of the nested design is not robust.

6. Conclusions

In this paper, we evaluated the adequacy of the AGRR from the perspective of practitioners. To this end, we designed a Monte Carlo simulation, which is different from conventional approaches but close to the actual AGRR process. We considered and compared two main experimental structures, crossed and nested designs, regarding four popular performance metrics for various combinations of σ_S^2 , σ_O^2 ,

σ_{SO}^2 , and σ_E^2 . The experimental results show that the adequacy of the crossed design is appropriate for all the evaluation perspectives: the average performance metrics, PCD, histogram shape, and sMAPE. However, the adequacy of the nested design is very low for all evaluation terms. We revealed that inadequacy comes from the overestimation of $\hat{\sigma}_O^2$. We tried to solve this problem by increasing the number of operators, but the problem is still unsolved. In conclusion, we highly recommend not applying the nested design as a tool of AGRR unless a solution to this problem is found. This solution could be any compensation coefficient of $\hat{\sigma}_O^2$ or other experimental design. This could be a topic of further research.

FIGURE 8: Averaged $sMAPE_k$ of the crossed design and the nested design.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science, and Technology (NRF-2013R1A1A2006947).

References

- [1] AIAG, *Measurement Systems Analysis Reference Manual*, 4th edition, 2010, <http://www.amazon.com/Measurement-Systems-Analysis-MSA-AIAG/dp/B004Z0V40G>.
- [2] D. C. Montgomery and G. C. Runger, "Gauge capability and designed experiments. Part I: basic methods," *Quality Engineering*, vol. 6, pp. 115–135, 1993, <http://www.tandfonline.com/doi/pdf/10.1080/08982119308918710>.
- [3] D. C. Montgomery and G. C. Runger, "Gauge capability analysis and designed experiments. Part II: Experimental design models and variance component estimation," *Quality Engineering*, vol. 6, no. 2, pp. 289–305, 1993.
- [4] R. K. Burdick and G. Larsen, "Confidence intervals on measures of variability in RR studies," *Journal of Quality Technology*, vol. 29, no. 3, pp. 261–273, 2015, <http://search.proquest.com/openview/6ab62569696a5aaf2e7c3d078b6e99c1/1?pq-origsite=gscholar>.
- [5] D. Gorman and K. M. Bower, "Measurement System Analysis And Destructive Testing," *Six Sigma Forum Magazine*, pp. 16–19, 2002.

- [6] J.-N. Pan, "Determination of the optimal allocation of parameters for gauge repeatability and reproducibility study," *International Journal of Quality and Reliability Management*, vol. 21, no. 6, pp. 672–682, 2004.
- [7] F. Bergeret, S. Maubert, P. Sourd, and F. Puel, "Improving and applying destructive gauge capability," *Quality Engineering*, vol. 14, no. 1, pp. 59–66, 2001.
- [8] Minitab, *Data Analysis and Quality Tools User's Guide 2*, Minitab Inc., 2000.
- [9] C. M. Borror, D. C. Montgomery, and G. C. Runger, "Confidence intervals for variance components from gauge capability studies," *Quality and Reliability Engineering International*, vol. 13, no. 6, pp. 361–369, 1997.
- [10] F. A. Graybill and C. M. Wang, "Confidence intervals of non-negative linear combinations of variances," *Journal of the American Statistical Association*, vol. 75, no. 372, pp. 869–873, 1980.
- [11] SAS Institute, "SAS Technical Report P-229, SAS - STAT Software: Changes and Enhancements, Release 6. 07," Tech. Rep., SAS Publishing, 1992, <http://www.barnesandnoble.com/w/sas-technical-report-p-229-sas-stat-software-s-a-s-institute-incorporated/1014558798?ean=9781555444730>.
- [12] G. A. Milliken and D. E. Johnson, *Analysis of Messy Data, Volume I: Designed Experiments*, CRC Press, 2nd edition, 1984, https://scholar.google.co.kr/scholar?q=analysis+of+messy+data+milliken&btnG=&hl=ko&as_sdt=0%2C5#1.
- [13] K. K. Dolezal, R. K. Burdick, and N. J. Birch, "Analysis of a two-factor R & R study with fixed operators," *Journal of Quality Technology*, vol. 30, no. 2, pp. 163–170, 2015, <http://search.proquest.com/openview/4779c01d797cabf0e397f9628a263da7/1?pq-origsite=gscholar>.
- [14] M. Hamada and S. Weerahandi, "Measurement system assessment via generalized inference," *Journal of Quality Technology*, vol. 32, no. 3, pp. 241–253, 2015, <http://search.proquest.com/openview/0f0733ae998d9328b59078f57edf677d/1?pq-origsite=gscholar>.
- [15] A. K. Chiang, "A simple general method for constructing confidence intervals for functions of variance components," *Technometrics*, vol. 43, no. 3, pp. 356–367, 2001.
- [16] L. Daniels, R. K. Burdick, and J. Quiroz, "Confidence intervals in a gauge RR study with fixed operators," *Journal of quality technology*, vol. 37, no. 3, pp. 179–185, 2015, <http://search.proquest.com/openview/3a7cd62a2f71bc321da92f1763835158/1?pq-origsite=gscholar>.
- [17] F.-K. Wang and E. Y. Li, "Confidence intervals in repeatability and reproducibility using the Bootstrap method," *Total Quality Management and Business Excellence*, vol. 14, no. 3, pp. 341–354, 2003.
- [18] R. K. Burdick, E. A. Allen, and G. A. Larsen, "Comparing variability of two measurement processes using RR studies," *Journal of Quality Technology*, vol. 34, no. 1, pp. 97–105, 2015, <http://search.proquest.com/openview/8f3f897b4084bd2af582-e7c186737db6/1?pq-origsite=gscholar>.
- [19] W. G. Cochran, "Testing a linear relation among variances," *Biometrics. Journal of the Biometric Society*, vol. 7, pp. 17–32, 1951.
- [20] A. Chiang, "Improved confidence intervals for a ratio in an RR study," *Communications in Statistics-Simulation and Computation*, vol. 31, no. 3, pp. 329–344, 2002, <http://www.tandfonline.com/doi/abs/10.1081/SAC-120003845>.
- [21] R. K. Burdick, C. M. Borror, and D. C. Montgomery, "A review of methods for measurement systems capability analysis," *Journal of Quality Technology*, vol. 35, no. 4, pp. 342–354, 2015, <http://search.proquest.com/openview/95d1b0694be0520e1da9-f900eda08f61/1?pq-origsite=gscholar>.
- [22] E. Adamec and R. K. Burdick, "Confidence intervals for a discrimination ratio in a gauge RR study with three random factors," *Quality Engineering*, vol. 15, no. 3, pp. 383–389, 2015, <http://www.tandfonline.com/doi/abs/10.1081/QEN-120018036>.
- [23] R. K. Burdick, Y.-J. Park, and D. C. Montgomery, "Confidence intervals for misclassification rates in a gauge RR study," *Journal of Quality Technology*, vol. 37, no. 4, pp. 294–303, 2015, <http://search.proquest.com/openview/95ae579bd695c0f14bc0a7a8d-32154df/1?pq-origsite=gscholar>.
- [24] W. H. Woodall and C. M. Borror, "Some relationships between gage R & R criteria," *Quality and Reliability Engineering International*, vol. 24, pp. 99–106, 2008, <http://onlinelibrary.wiley.com/doi/10.1002/qre.870/abstract>.
- [25] R. Browne, J. MacKay, and S. Steiner, *Leveraged Gauge RR Studies. Technometrics*, vol. 52, no. 3, pp. 294–302, 2010.
- [26] J.-N. Pan, C.-I. Li, and S.-C. Ou, "Determining the optimal allocation of parameters for multivariate measurement system analysis," *Expert Systems with Applications*, vol. 42, no. 20, pp. 7036–7045, 2015.
- [27] J. De Mast and A. Trip, "Gauge R&R studies for destructive measurements," *Journal of Quality Technology*, vol. 37, no. 1, pp. 40–49, 2005.
- [28] F. Van Der Meulen, H. De Koning, and J. De Mast, "Nonrepeatable gauge R&R studies assuming temporal or patterned object variation," *Journal of Quality Technology*, vol. 41, no. 4, pp. 426–439, 2009.
- [29] P. Wallström and A. Segerstedt, "Evaluation of forecasting error measurements and techniques for intermittent demand," *International Journal of Production Economics*, vol. 128, no. 2, pp. 625–636, 2010.

Hindawi

Submit your manuscripts at
<https://www.hindawi.com>