ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2023.1.1.727

Measurement System Analysis applied to a competence assessment methodology in engineering training programs in Mexico.

Juan Victor Bernal - Olvera, Ph. D[1], Mónica Belem Bernal - Perez, MBA[2], and Mireya Berenice Monroy - Anieva, MBA[3]

[1,2,3,]Tecnológico Nacional de México/TES Cuautitlán Izcalli, Mexico, juan.bo@cuautitlan.tecnm.mx, monica.bp@cuautitlan.tecnm.mx, mireya.ma@ cuautitlan.tecnm.mx

***Abstract*– A reliable evaluation system must be a guarantee to obtain a qualification that reflects the degree of acquisition of competence. The evaluation must be homogeneous so that teachers assign grades closer to the degree of competence formation in engineering training programs and serve as a reference for more objective feedback. For this reason, this article shows a quantitative method of Measurement System Analysis (MSA) to determine the reliability of the system evaluation and to generate a diagnosis that allows the establishment of adequate strategies to improve the evaluation practice of participating university professors within engineering training in Mexico.**

Keywords—MSA, assessment, training, repeatability, reproducibility.

I. Introduction

Evaluation and feedback are very important factors in the development of student learning, but it has been an important source of dissatisfaction among the participants [1]. The powerful influence of feedback on the learning process is widely recognized [2], [3], [4]), and its delivery to students is too important for progress in learning [5], attending conditions of interaction between practices, context, and individuals [6]

One aspect of knowing the performance and evolution of learning is learning analytics, defined as the measurement, collection, analysis, and reporting of data on the progress of students and the contexts in which learning takes place [1], and should be oriented toward the improvement of learning - teaching processes with the participation of students [7].

Reference [8] analyzed the "quick fixes" that universities introduced to enable digital assessment and the challenges and tradeoffs they faced between scale and security, trust and fairness establishing three aspects to cover: relevant, that is, would allow universities to go beyond traditional forms of assessment, dictated by the practical limitations of analog exams, and build systems that are relevant to contemporary needs and reflect the learning process, and make use of innovative assessment methods too impractical to deliver without digital tools; adaptable, to address the needs of a diverse and growing student population, a range of providers, and any number of geographies; and trustworthy, based on solid foundations of academic integrity, security, privacy, and fairness.

Having a reliable measurement process ensures reliability in the acquisition and generation of data, reducing the risk of making erroneous decisions or delivering information or products out of specification [9]. To determine the degree of reliability of the system in this study, the Gage tool, Reproducibility, and Repeatability (Gage R&R), is used, using an analysis of variance and a crossed structure in its execution.

This article focuses on the reliability of the evaluation, with a systemic perspective that includes the evaluation instrument and the teachers who use it. To do this, use is made of a methodology used in the automotive industry to evaluate their measurement systems, which is called Measurement Systems Analysis (MSA), described in the IATF 16949 standard [10], as part of a package known as Core Tools for continuous quality improvement.

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

II. Frame of reference

A. *Literature review*

Every process has indicators that allow its control and improvement. Therefore, product evaluation and process improvement require accurate measurement and precise techniques. Because all measurements contain errors, and by the next mathematical expression: every observed value is equal to the true value plus the measurement error, its understanding and management are studied through Measurement Systems Analysis (MSA), and it is an important function in the process improvement [11].

The MSA is a comprehensive set of tools for the measurement, acceptance, and analysis of data and errors, and includes topics such as statistical process control, capability analysis, and gage repeatability and reproducibility, among others [12]. MSA recognizes that measurements are made on both simple and complex products, using physical devices and visual inspection devices that rely heavily on human judgment of product attributes [13].

There are two aspects considered in this study repeatability, which measures the degree of error, generally attributed to the measuring instrument, and is best thought of as "random error" and reproducibility, which is between raters, attributed to the differences between measurements while using the same measurement instrument [14].

B. *Background*

The Technological National of Mexico (TecNM) is a decentralized administrative institution of the Ministry of Public Education, with technical autonomy, academic and management; it has as its mission comprehensively train competitive professionals in science, technology, and other areas of knowledge. This has 254 campuses in the country. One of these, Campus Cuautitlan Izcalli has done this project, and the results are shown.

Located 30 km northeast of Mexico City this campus is a decentralized public institution of the Government of the State of Mexico; it has 8 engineering careers, a degree in Public Accounting, and 2 postgraduate programs with a focus on technological research. Seven of the 8 engineering careers have international accreditations, leaving only one for being recently created. It was opened in 1998, and currently, it has an enrollment of more than 5,000 students, and infrastructure made up of classroom buildings, laboratories, green yards, sports, and recreation areas. The Industrial Engineering career is the Division with the largest number of students, about 20%, who are divided into two shifts, morning, and evening. Its teaching-learning processes have a competency-based training approach, it is having constant training for its teaching staff in pedagogical training. The courses are for six months, the first period from February to July, and the second from August to January of the following year.

For the training of engineers, a competence development scheme is applied, which is planned with teams of professors who teach the same subject, is executed in class, and evaluated in a formative and summative way to give a final grade. In this evaluation stage, learning products are generated that are evaluated based on a rubric, which is a whole coherent set of criteria for students' work that includes descriptions of levels of performance quality on the criteria [15].

III. Methodology

A. *General procedure*

The methodology used is based on the scientific method, is of an experimental systemic nature, and consists of the following steps.

- Description of the problem.
- Hypothesis formulation.
- Selection of the subject to take the sample.
- Insulation of the sample.
- Statistical analysis of the information.
- Discussion of the results.
- Conclusions and recommendations

B. *Statement of the Hypothesis*

The hypothesis to be tested is stated as follows: The competency assessment system is adequate because the dispersion of the system is less than the value indicated in the AIAG manual.

IV. Results and discussion

A. *Device: the rubric*

The evaluation instrument that is applied is a rubric; references [16], and [17] say that it articulates expectations for student work by listing rules for the work and performance level descriptions across a continuum of quality. It is suited for formative and summative use by containing descriptions of quality work and not evaluation [18]. It is a good tool that helps the teacher in the assessment process of scholarly performance.

The rubric for this paper process has 8 items, each of these measures 3 aspects by the Likert scale and the criteria reviewed by the rubric in this article are:

- Cover and delivery format.
- Orthography.
- Delivery on time.
- Use of bibliography.
- Objective of learning.
- Introduction to the topic.
- Development of the report.
- Conclusion.

It has been evaluated with Cronbach's Alpha with a value of 0.8016. Reference [19] says its threshold is 0.80, while other authors suggested alpha values of 0.70 can be accepted for the early stages of research [20]. Its use considers the dispersion of all data in a whole, it is the mean of all possible split-half coefficients, is a lower bound for the coefficient of precision and estimates and, also, it is a lower bound to the proportion of variance test attributable to common factors among the items [21].

B. *Measurement Development*

For this research work, the information obtained from the Economics subject in the Industrial Engineering program, carried out in person during the first semester of 2022, is used. The first batch of products received was 26; it is considered an adequate sample of 5 of these reports to submit to the analysis process in the two aspects the methodology establishes, and it can get a group of 30 results. Through a process of insulation supported by a simple sampling, each specimen of the sample is obtained.

Three professors were asked to evaluate, each of them, twice the same product, which, in this case, is the report of the referred subject. The results were controlled by identifying each participant as P1, P2, and P3 to denote Professor 1, Professor 2, and Professor 3, respectively. Each of the teachers considered meets three requirements to carry out this experiment: an experience of more than three years qualifying by competencies, a master's degree as the minimum degree of study, and knowledge of the subject of study.

The results of these evaluations (measure) were concentrated in a matrix, shown in Table 1. It is convenient to clarify that the time space between the first and second review of the same report was one week.

TABLE I
MATRIX WITH THE RESULTS OF THE EVALUATION TO THE REPORTS.

Report	Review	Professor	Measure	Professor	Measure	Professor	Measure
1	1	P1	63	P2	75	P3	85
	2	P1	85	P2	87	P3	80
2	1	P1	82	P2	67	P3	75
	2	P1	75	P2	80	P3	85
3	1	P1	68	P2	72	P3	98
	2	P1	68	P2	82	P3	93
4	1	P1	57	P2	67	P3	90
	2	P1	72	P2	77	P3	90
5	1	P1	75	P2	67	P3	72
	2	P1	70	P2	67	P3	70

A first analysis of the information using a box plot indicates different levels of appreciation for the evaluation of the reports, even between the rating assigned by each teacher to the same work, as seen in Figure 1.

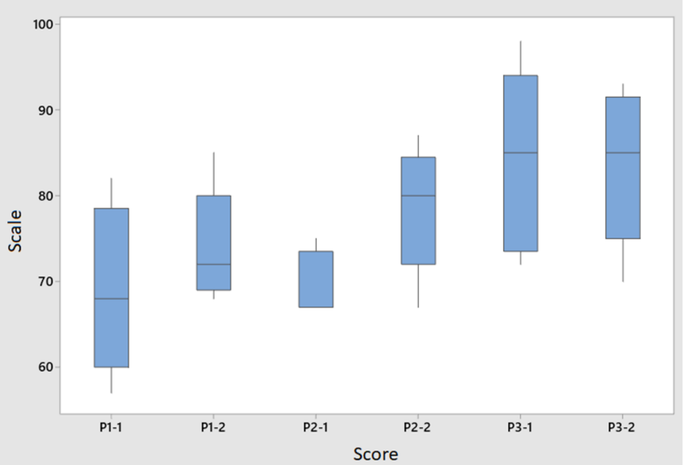
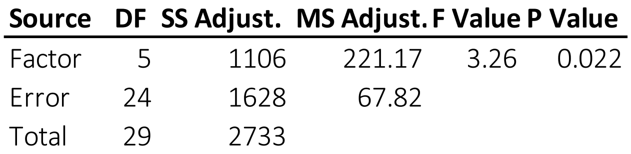


Fig.1 Box – plot showing the results of each round of evaluation to the report by each participating professor.

Applying a one-way ANOVA to check if there is a significant difference in the means of each measurement, the Minitab program shows that there is no such difference, considering a significance level α = 0.05 since a p-value of 0.022 is obtained so that the null hypothesis about the difference in means is equal to zero can be rejected, as shown in Table 2, and can be concluded that at least one means is different.

TABLE II
ANOVA VALUES FOR THE MEAN TEST.



Source	DF	SS Adjust.	MS Adjust.	F Value	P Value
Factor	5	1106	221.17	3.26	0.022
Error	24	1628	67.82		
Total	29	2733			

Making comparisons of means, Tukey does not indicate that the differences are significant, considering a significance level of 5%, as shown in Figure 2 because all the intervals contain a value of zero, which means that there is no variation in the compared means.

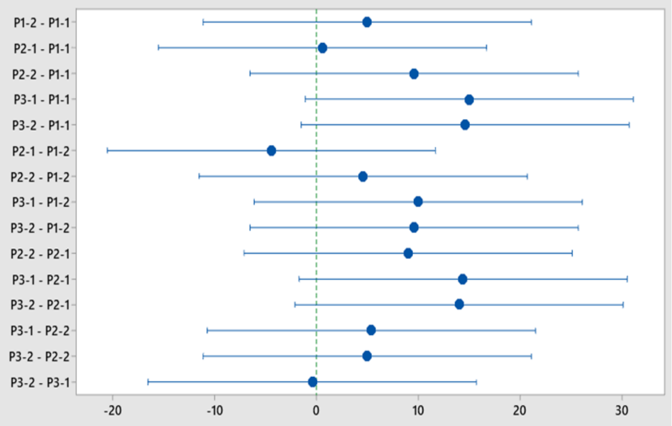


Fig. 2 Tukey comparisons at 5% significance.

However, Hsu's multiple comparisons with the best (MCB) taking the group of measurements with the highest mean as a reference, does detect a significant difference at this same level of significance of 5% because an interval has zero as a bound, which indicates that there is a significant difference between the comparisons, shown in Figure 3, so it is concluded that there is insufficient statistical evidence to show an equality of the mean of the measurements in the measurement system.

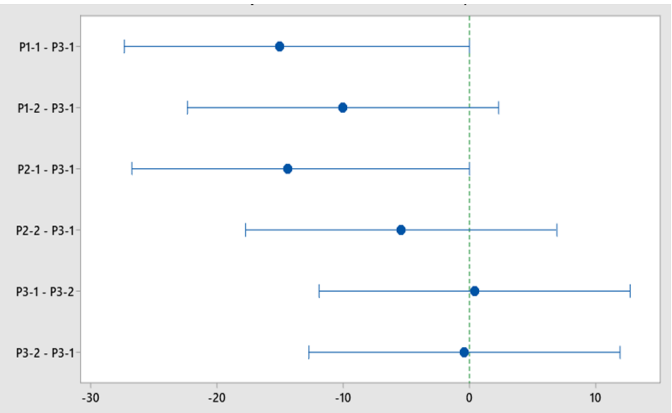


Fig. 3 Results of Hsu's MCB method.

By performing the crossover Gage R&R study using the one-way ANOVA method, Minitab outputs a series of graphs that provide more information about the performance of the measurements in the system. Thus, Figure 4 provides the box – plot on the mean of the grades assigned by each teacher; in this way, Professor 1 is the strictest, Professor 2 the most focused, and Professor 3 the softest.

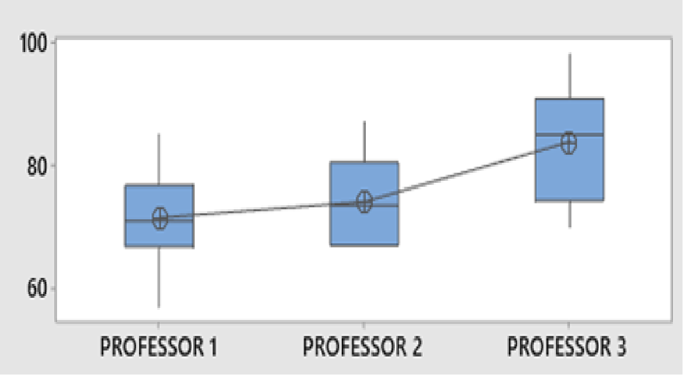


Fig. 4 Box - plot showing the average grade assigned by the Professor.

Figure 5 shows the interaction of each Professor rating the same product, having similar results in products 1, 2, and 5 but greater dispersion in 3 and 4.

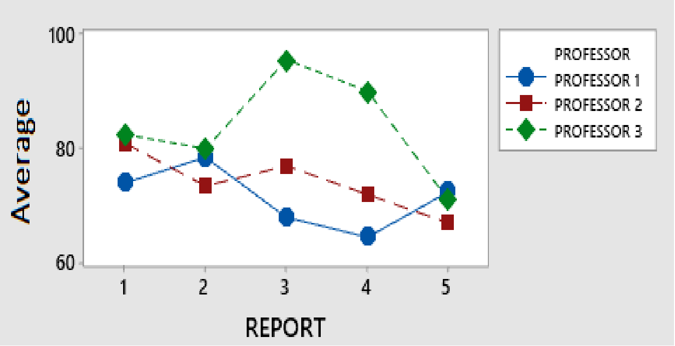


Fig. 5 Graph that shows the interaction between the grades assigned by the Professor to each of the Reports.

Figure 6 shows the averages to the grades assigned by the professors to the reviews of the reports, in which it is observed that the smallest dispersion in report 5, and in the one with the greatest amplitude were those numbered 3 and 4, which that reaffirms non-standard evaluation criteria in the evaluation process.

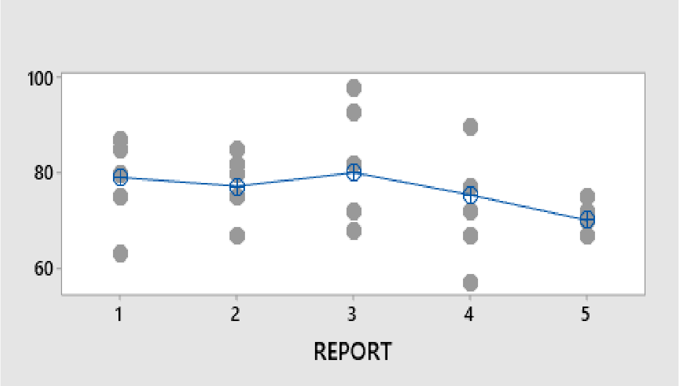


Fig. 6 Graph that shows the grades assigned by the Professor and its mean to Reports.

The results generated by Minitab include a graph of means and ranges ($\bar{\bar{X}} - \bar{R}$), in which it can be seen that the process is not in statistical control due Professor 3 has a point outside the control limits, there is no a significant dispersion in the ranges, and this may be an indication that the system does not have a standard of appreciation in the application of the rubric by the teachers who rated the products (see Figure 7).

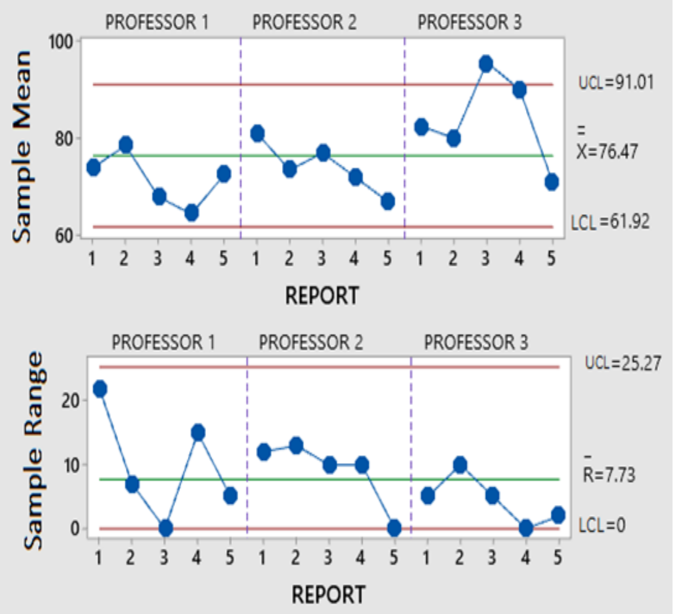


Fig. 7 Control chart $\bar{\bar{X}} - \bar{R}$ with the data obtained in the Gage R&R study applied by Professors.

Regarding the components of the variation, most of the dispersion is explained in the measurement instrument and in the way in which the measurement is carried out by the teachers, as can be seen in Figure 8.

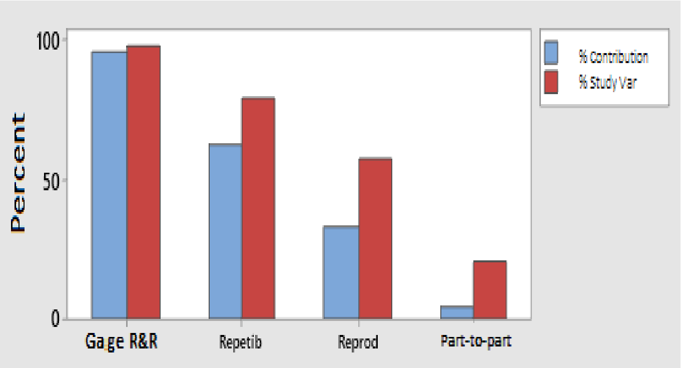
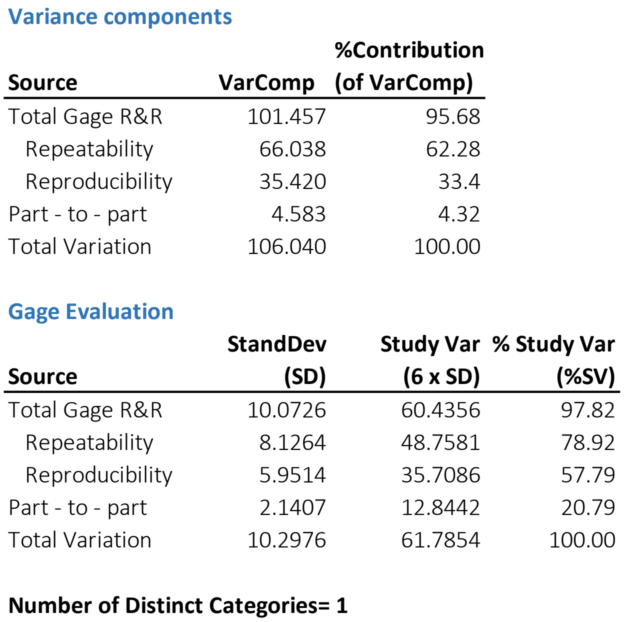


Fig. 8 Components of Variation in the Gage R&R Study.

The total R&R of the system measurement is equivalent to 97.82% of the variation of the study, so it is not adequate according to AIAG [10]. The repeatability and reproducibility values are also high, 78.92 and 57.79, respectively. Regarding the number of categories that the study can find, it is 1, very low, when the recommended standard is a minimum of 5 (see Figure 9).



Variance components

Source	VarComp	%Contribution (of VarComp)
Total Gage R&R	101.457	95.68
Repeatability	66.038	62.28
Reproducibility	35.420	33.4
Part - to - part	4.583	4.32
Total Variation	106.040	100.00

Gage Evaluation

Source	StandDev (SD)	Study Var (6 x SD)	% Study Var (%SV)
Total Gage R&R	10.0726	60.4356	97.82
Repeatability	8.1264	48.7581	78.92
Reproducibility	5.9514	35.7086	57.79
Part - to - part	2.1407	12.8442	20.79
Total Variation	10.2976	61.7854	100.00

Number of Distinct Categories= 1

Fig. 9 Results of the Gage R&R study on the competency assessment system.

V. Conclusions and Recommendations

A competency-based evaluation system must be uniform and consistent in its results so that participating teachers evaluate with fairness and adhere to criteria defined by a rubric, which must show adequate consistency. Both points constitute the two aspects to consider in the analysis of measurement systems.

This exercise, unprecedented in the Institution, has identified three types of behavior and appreciation in the granting of qualification to a product generated in the acquisition of competencies. Thus, it has found a low rating corresponding to a very demanding criterion; another that mediates the differences; and one more, assigning high marks due to very slight appreciation. Also, it can be shown that the measurement system does not satisfy a standard suggested in the AIAG manual because the total variation of the system exceeds the maximum allowed values, for which the following actions are suggested:

a) It is necessary to review and adapt the form of evaluation in the system by competencies.

b) There must be rubrics that consistently measure each product to be evaluated, being generically applicable to all engineering programs.

c) It is necessary to unify the evaluation criteria of each professor so that they are consistent in the issuance of qualifications for each product they receive from the students.

With this article, it is shown that an evaluation system can be diagnosed in the teaching – learning processes by competences, with more uniform criteria to assign a fairer qualification. The two important factors to consider enter: on the one hand, the measurement instrument, which is the rubric that can be applied to any product, and on the other hand, the teachers, who must adjust their criteria for a correct interpretation of each point at evaluate, so that the qualification is fairer and with natural variations that are controllable. In this paper, the training of engineers is considered as the context, but its application can be to any evaluation process to the acquisition of competences.

Acknowledgment

Special thanks are made to Technological National of Mexico, TES Cuautitlan Izcalli for its valuable support in allowing Institution's information to be considered in this article.

References

[1] Bikanga, M., and Stansfield, M. The potential of learning analytics in understanding students' engagement with their assessment feedback. Proccedings of 2017 IEEE 17th International Conference on Advanced Learning Technologies (ICATL), 227 – 229, 2017.

[2] Quality Assurance Agency for Higer Education (QAAHE). Focus on: Feedback from Assessment. 2018. http://www.qaaa.ac.uk/scotland/focus-on/feedback-from-assessment

[3] Winstone, N., and Boud, D. Exploring cultures of feedback practice: The adption of learning – focused feedback practices in the UK and Australia. *Higer Education Research and Development*, 38(2), 411-425, 2019.

[4] Bikanga Ada, M. Evaluation of a Mobile Web Application for Assessment Feedback. *Tech Know Learn*. 2021. https://doi.org/10.1007/s10758-021-09575-6

[5] Gibbs, G. and Simpson, C. Conditions under which assessment supports students´ learning. Learning and Teaching in Higer Education, 1 (1), 3-31,2004.

[6] Henderson, M., Ryan, T., and Phillips, M. The challenges of feedback in higher education. Assessment & Evaluation in Higher Education.

[7] Sclater, N., Peasgood, A, and Mullan, J. Learning Analitics in Higher Education: A review of UK and international practice. Bristol: Jics 2016. https://www.jisc.ac.uk/reports/learning-analytics-in-higher-education,

[8] Attewell, S., Iosad, A., & Pauli, M. Assessment rebooted. Emerge Education, 2020. https://repository.jisc.ac.uk/7854/1/assessment-rebooted-report.pdf

[9] Schmitt, R., & Bauza, M. Measurement System Analysis. In: Chatti, S., Laperrière, L., Reinhart, G., Tolio, T. (eds) CIRP Encyclopedia of Production Engineering. Springer, Berlin, Heidelberg, 2019.

[10] Automotive Industry Action Group (AIAG) QS 9000: measurement systems analysis, 4th edn. Chrysler Group LLC/Ford Motor Company/General Motors Corporation, Michigan, USA, 2010. https://www.aiag.org/quality/automotive-core-tools/msa

[11] Montgomery, D. Introduction to Statistical Quality Control. New York: John Wiley and Sons, (2005).

[12] Besterfield, D. H. Quality Control. Englewood Cliffs, New Jersey: Prentice Hall, 2009.

[13] Smith, R., McCrary, S, and Callahan, N. Gauge Repeatability and Reproducibility Studies and Measurement System Analysis: A Multimethod Exploration of the State of Practice. Journal of Industrial Technology, 23(1), 1 – 12, 2007.

[14] Kappele, W., and Raffaldi, J An Introduction to Gauge R&R. Quality, 44(13), 24-25, 2005.

[15] Brookhart, S. How to Create and Use Rubrics for Formative Assessment and Grading. Alexandria, VA: ASCD, 2013.

[16] Andrade, H. G. Using rubrics to promote thinking and learning. Educational Leadership 57, 13–18, 2000.

[17] Arter, J. A., and Chappuis, J. Creating and Recognizing Quality Rubrics. Boston: Pearson, 2006.

[18] Brookhart, S., and Chen, F. The quality and effectiveness of descriptive rubrics. Educational Review, 2014, DOI: 10.1080/00131911.2014.929565

[19] Norcini, John J. "Standards and Reliability in Evaluation: When Rules of Thumb Don't Apply." Academic Medicine 74 (10): 1088–1090, 1999.

[20] Iacobucci, Dawn, and Adam Duhachek. "Advancing Alpha: Measuring Reliability with Confidence." Journal of Consumer Pscyhology 13 (4): 478–487, 2003.

[21] Cronbach, L. Coefficient alpha and the internal structure of tests. Psychometrika, 16 (3), pp 297 – 334, 1951.