Available online at: https://ijact.in

Date of Submission	29/02/2020
Date of Acceptance	23/03/2020
Date of Publication	30/04/2020
Page numbers	3629-3632 (4 Pages)

Cite This Paper: Wei C Ng, Sin Yin Teh, Heng C Low, Ping C Teoh. An evaluation of inter-rater reliability of the technicians in a manufacturing environment, 9(4), COMPUSOFT, An International Journal of Advanced Computer Technology. PP. 3629-3632.

ISSN:2320-0790

COMPUSOFT

An International Journal of Advanced Computer Technology

AN EVALUATION OF INTER-RATER RELIABILITY OF THE TECHNICIANS IN A MANUFACTURING ENVIRONMENT

Wei Chien Ng[1], Sin Yin Teh[1*], Heng Chin Low[2] and Ping Chow Teoh[3]

[1]School of Management, Universiti Sains Malaysia, 11800 Minden, Penang,
[2]School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Minden, Penang,
[3]School of Science and Technology, Wawasan Open University, 10050 Penang,
*Corresponding address
ngweichien@hotmail.com, *tehsyin@usm.my, hclow@usm.my, pcteoh@wou.edu.my

Abstract: In the era of digital economy, Industry Revolution 4.0 has become the aim for manufacturing organisations in order to transform into a smart factory. With the advancement of technology, company engages in continuous improvement projects to ensure high quality products being manufactured. Assessing the strength of agreement between technicians' ratings of quality problem identification results is of primary interest because an effective diagnostic procedure is dependent upon high levels of consistency between technicians. However, in practice, discrepancies are often observed between technicians' ratings and it is considered as a major quality issue in monitoring the troubleshooting and repairs of the equipment. This has motivated us to evaluate the accuracy and agreement between technicians' ratings. The primary objective of this study is to evaluate the inter-rater reliability of the technicians on the continuous improvement projects before actual implementation. A case study is conducted in one of the smart manufacturing companies in the Penang Free Trade Zone. This study utilised Fleiss's Kappa analysis because it is suitable in situations where there are more than two raters, i.e. six technicians who are responsible to identify six problems simulated for a continuous improvement project. The findings of the study show good to excellent agreement and high accuracy in problem identification. Overall, the technicians are capable in understanding the newly developed troubleshooting and repairs database and able to carry out the continuous improvement project effectively. This outcome provides top management an insight for evidence-based decision making to thoroughly execute the newly developed digital database in smart manufacturing.

Keywords:Fleiss' Kappa analysis, Smart Manufacturing, Continuous Improvement, Technicians, Inter-rater reliability

I. INTRODUCTION

Since the introduction of Industry 4.0, companies from various sectors especially manufacturing is moving towards a new level of manufacturing processes equipped with customized and flexible mass production technologies. Industry 4.0 which also known as "smart factory" is the fourth industrial revolution which focuses on Cyber-Physical system-enabled manufacturing and service innovation [1]. In order to survive in a competitive world, companies especially the manufacturing companies must engage in continuous improvement to ensure the high quality of products and services produced. Continuous improvement is defined as an improvement initiative that increases success and reduces failures [2]. Continuous improvement is also viewed as an offshoot of existing quality initiatives like total quality management or as a completely new approach of enhancing creativity and achieving competitive excellence in the market [3-5]

Often, the continuous projects introduced in the manufacturing companies are often not communicated well to the technicians. As a matter of fact, technicians such as engineers, technicians and operators came from different background, education levels and nationalities. Therefore, the technicians have a lack of agreement in the continuous improvement projects as their understanding are dissimilar. This often leads to unsuccessful project implement and ultimately causes massive costs incurred and loss of reputation by the manufacturing company.

Therefore, the main objective of this paper is to evaluate the inter-rater reliability of the technicians using Fleiss' Kappa analysis on a continuous improvement project before actual implementation. This case study is conducted in one of the smart manufacturing companies in Penang Free Trade Zone. One of the continuous improvement projects is selected in this study to investigate the inter-rater reliability of the technicians.

The rest of this paper is organized as follows: In Section 2, details of Fleiss' Kappa analysis are explained. In Section 3, the design of the experiment is presented. The results of the experiment are discussed in Section 4. Finally, a conclusion is provided in the last section.

II. Fleiss' Kappa Analysis

Several methods are available for measuring agreement between two raters. This study utilised Fleiss's Kappa analysis because it is suitable in situations where there are more than two raters. The Kappa statistic was first proposed by Cohen [6] after Scott [7] proposed the π statistic as a measurement of the inter-rater reliability of two raters. Since then some extension works on evaluation of agreement between raters can be found in Cohen [8], Everitt [9], Maxwell [10], Fleiss [11], Fleiss [12], Bangdiwala [13], Barlow [14], among others.

Inter-rater agreement analysis originated from the Measurement System Analysis. According to McHugh [6], there are two types of inter-rater agreement analysis such as Cohen's Kappa [6] and Fleiss' Kappa [11]. Cohen's Kappa analysis is a measure of interrater agreement for qualitative items between two raters [15]. Meanwhile, Fleiss' Kappa analysis is a measure of interrater agreement for qualitative items for three or more raters [16]. It can be interpreted as expressing the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters made their ratings completely random.

This study utilised Fleiss's Kappa analysis because it is suitable in situations where there are more than two rates, i.e. six technicians who are responsible to identify six problems simulated for a continuous improvement project in the manufacturing company. The Kappa statistic that is implemented in the Minitab version 17 software for multiple raters is originally proposed by Fleiss [6]. A review of the statistical calculation of Fleiss Kappa statistic is warranted as follows.

The Kappa (K) score represents the possibility of the agreement. The higher the K, the higher the agreement between the raters. The degree of agreement between the multiple raters provides some indication as to the consistency of the values. In other words, high agreement between the rates indicates the ratings reflect the actual circumstance. While, low agreement between the raters indicates less confidence in the results. Table 1 shows the interpretation of K in Fleiss' Kappa analysis.

TABLE 1**.** Summary of Fleiss' Kappa analysis results

K	Interpretation
< 0	No agreement
0.0 – 0.40	Poor agreement
0.41 – 0.75	Moderate agreement
0.75 – 1.00	Good to excellent agreement

The general form of *K* score is defined as [11],

$$K = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{1}$$

Kappa statistic is a measure of agreement which naturally controls for chance. The factor $1 - \bar{P}_e$ provides the degree of agreement that is attainable above chance while $\bar{P} - \bar{P}_e$ provides the degree of agreement achieved above chance. If the ratersare able to achieve a complete agreement of the matter, the *K* score equals to 1. On the contrary, if there is no agreement between the raters, then the *K* score should be less than 0.

The detailed steps of calculating the *K* score are as follows.

First, $\bar{P}_e$ is defined as:

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2 \tag{2}$$

where

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \quad , \quad 1 = \sum_{j=1}^{k} p_j \tag{3}$$

Let *N* be the total number of subjects, and *n* be the number of ratings per subject. Thereafter, let *k* denote the number of categories into the assignments are made. The subjects are indexed by *i* = 1, …, *N* and the categories are indexed by *j* = 1, …, *k*. The symbol n_{ij} represents the number of raters who assigned the *i*-th target subject independently into the *j*-th category. It is worth noting that the categories are mutually exclusive and exhaustive.

Then $\bar{P}$ is defined as:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{4}$$

where,

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij} \left(n_{ij} - 1\right)$$

$$= \frac{1}{n(n-1)} \sum_{j=1}^{k} \left(n_{ij}^2 - n_{ij}\right)$$

$$= \frac{1}{n(n-1)} \left[\left(\sum_{j=1}^{k} n_{ij}^2\right) - (n)\right] \quad (5)$$

Next, insert Equation (5) into Equation (4),

$$\bar{P} = \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn\right) \quad (6)$$

III. Design of the Experiment

In the realm of continuous quality improvement, a project on a newly developed troubleshooting and repairs database that used to capture the troubleshooting input of the technicians on the equipment is focused. As a matter of fact, in the era of the industrial revolution, fast-paced smart manufacturing process is running in a high speed. Hence equipment breakdowns are very common. In order to monitor the troubleshooting and repairs of the equipment systematically and effectively via knowledge management system, a predetermined drop-down selection of troubleshooting and repairs database has been developed.

Since this is a predetermined drop-down selection for the technicians, assessing the strength of agreement between technicians' ratings of quality problem identification results is of primary interest. This is because an effective diagnostic procedure is dependent upon high levels of consistency between technicians. Note that the discrepancies observed between technicians' ratings are considered as major quality issues in monitoring the troubleshooting and repairs of the equipment.

As presented in Section 2, Fleiss' Kappa analysis is applied to examine the effectiveness of the troubleshooting and repairs database. There are six technicians from all three shifts in the manufacturing floor who are responsible in troubleshooting the equipment. All six technicians are involved in the Fleiss' Kappa analysis. There are six simulated equipment quality problems and the problems are purposely repeated 3 times randomly to fulfil the accuracy, repeatability and reproducibility in the analysis. The sample size of the Fleiss' Kappa analysis is consistent with the research done by [17]. With the assistance from a senior technician, the simulated problems were designed for one of the test stations in the manufacturing floor. The troubleshooting of the same set of simulated problems is conducted by each technician individually and randomly to avoid any bias during the experiment.

Technicians are required to choose the paired problem-and-solution respectively from the newly developed troubleshooting and repairs database when they were troubleshooting the simulated problems. The inputs are captured in an online form given to them. Troubleshooting inputs selected by the technicians on the simulated problems are presented in Table 2, Appendix A. Note that it is necessary to prepare a Minitab dataset as shown in Table 2 in order to compute the Fleiss' Kappa statistic and the associated statistical tests and precision measures. From Table 2, the troubleshooting input dataset have 18 problems (subjects), 6 technicians (raters) and 18×6 (categories). The variables in the input dataset could be characters or numeric. Consequently, this input dataset is used to run the Fleiss' Kappa analysis using Minitab version 17 software and the *K* score of the input data is computed.

IV. Results and Discussion

After the experiment is conducted, the results of Fleiss' Kappa analysis are compiled in Table 3. From Table 3 of the Fleiss' Kappa analysis, there are 4 sets of outputs generated such as the *K* scores within the technicians, each technician versus standard, all technicians versus standard and technicians to technicians (between technicians). For Fleiss' Kappa analysis (within technicians), the purpose is to measure the repeatability of the technicians in selecting the predetermined input given when troubleshooting the simulated problems. From Table 3, technicians 1, 2, 4 and 6 (*K*=1.00000) managed to achieve good to excellent complete agreement while technicians 3 and 5 (*K*=0.61702) managed to achieve moderate agreement. In other words, the technician well agrees with himself across three trials.

TABLE 2**.** Summary of Fleiss' Kappa analysis results

Appraiser	***K* Score**
Fleiss' Kappa Analysis – Within Technicians (Repeatability)	
Technician 1	1.00000
Technician 2	1.00000
Technician 3	0.61702
Technician 4	1.00000
Technician 5	0.61702
Technician 6	1.00000
Fleiss' Kappa Analysis – Each Technician versus Standard (Accuracy)	
Technician 1	1.00000
Technician 2	1.00000
Technician 3	0.73982
Technician 4	1.00000
Technician 5	0.80539
Technician 6	1.00000
Fleiss' Kappa Analysis – Between Technicians (Reproducibility)	
Overall	0.85315
Fleiss' Kappa Analysis – All Technicians versus Standard	
Overall	0.92420

Note: All results are significant at α = 0.05

For Fleiss' Kappa analysis (each technician versus standard), the purpose is to determine the accuracy of each of the technicians. From Table 3, technicians 1, 2, 4, 5 and 6 (*K*=0.80539 and *K*=1.00000) managed to achieve good to excellent agreement while technician 3 (*K*=0.73982) managed to achieve moderate agreement in Fleiss' Kappa analysis for each technician versus standard. The technician's assessment across trials is in accordance with the known standard.

For Fleiss' Kappa analysis between technicians, the main purpose is to measure the reproducibility of the technicians in using the new database when troubleshooting the simulated problems. From Table 3, the *K* score (between technicians) 0.85315 is good to excellent agreement. This shows that the reproducibility of the technicians has shown excellent result. All technicians' assessments agree with each other.

Lastly, for the Fleiss' Kappa analysis (all technicians versus standard), the ultimate purpose is to determine the repeatability, reproducibility and accuracy of all technicians in understanding the newly developed database when troubleshooting the simulated problems. The Fleiss Kappa analysis shows that the overall *K* score is 0.92420 which is good to excellent agreement. All technicians' assessments agree with the known standard. Since the *K* score is more than 0.75, then the newly developed troubleshooting and repairs database is proven effective to be implemented in the manufacturing floor.

V. Conclusion

This paper successfully evaluates the inter-rater reliability of the technicians on the continuous quality improvement project before actual implementation. Rather than implementing the newly developed troubleshooting and repairs database with discrepancies between technicians' ratings, the finding of the study shows good to excellent agreement and high accuracy in problem identification. Overall, the technicians are capable in understanding the newly developed troubleshooting and repairs database and able to carry out the continuous improvement project effectively.

To the best of the authors' knowledge, this is the first study which demonstrates how Fleiss' Kappa for multiple raters can be easily implemented in Minitab software to evaluate Inter-rater reliability of the technicians in a smart manufacturing environment. The existing studies implement Fleiss' Kappa for medical and health science case studies. In addition, the dataset generated from the simulation was based on a real-life scenario in monitoring the troubleshooting and repairs of the equipment.

This case study only implements Fleiss' Kappa analysis on one continuous quality improvement project. Therefore, in future research, the application of Fleiss' Kappa analysis can be adopted in other continuous quality improvement projects and provide top management insights for evidence-based decision making to thoroughly execute the newly developed digital database in smart manufacturing.

Acknowledgments

This research is supported and funded by the Research University Grant (RUI) 1001/PMGT/8016033, UniversitiSains Malaysia (USM) and Collaborative Research in Engineering, Science and Technology (CREST) grant 304/PMATHS/650809/C121. Special thanks to Mr. Amir Hamzah Nawawi, Mr. Fuad Hisham and Mr. Ong Whee Wei for their relentless contribution of ideas and information.

VI. References

[1] J. Lee, H-A, Kao, S. Yang, Service innovation and smart analytics for Industry 4.0 and big data environment, Procedia CRIP, Vol. 16, pp. 3-8, 2014.

[2] B. Nadia, B. Amit, An overview of continuous improvement: From the past to the present, Manage. Decis., Vol. 43, No. 5/6, pp. 761-771, 2005.

[3] J. Oakland, Total Organizational Excellence – Achieving World-Class Performance. Butterworth-Heinemann, Oxford, 1999.

[4] S. Caffyn, Development of a continuous improvement self-assessment tools, Int. J. Oper. Prod. Man., Vol. 19, No. 11, pp. 1138-1153, 1999.

[5] M. Gallagher, S. Austin, S. Caffyn, Continuous Improvement in Action: The Journey of Eight Companies. Kogan Page, London, 1997.

[6] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas., Vol. 20, pp. 37-46, 1960.

[7] W. A. Scott, Reliability of Content Analysis: The Case of Nominal Scale Coding, Public Opinion Quarterly, Vol. XIX, pp. 321-325, 1955.

[8] J. Cohen, Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit, Psych. Bull., Vol. 70, pp. 213-220, 1968.

[9] B. S. Everitt, Moments of the statistics kappa and the weighted kappa, British J. Math. Statist. Psych., Vol. 21, pp. 97- 103, 1968.

[10] A. E. Maxwell, A. E. G. Pilliner, Deriving Coefficients of Reliability and Agreement for Ratings, Br. J. Math. Stat. Psychol., Vol. 21, pp. 105-116, 1968.

[11] J. L. Fleiss, Measuring nominal scale agreement among many raters, Psych. Bull., Vol. 76, pp. 378-382, 1971.

[12] J. L. Fleiss, Statistical Methods for Rates and Proportions. John Wiley & Sons, Inc., New York, 1981.

[13] S. I. Bangdiwala, H. E. Bryan, "Using SAS Software Graphical Procedures for the Observer Agreement Chart," in Proc. 12th Annu. SAS Users Group Int. Conf., Texas, 1987, pp. 1083-1088.

[14] W. Barlow, N. Y. Lai, S. P. Azen, A comparison of methods for calculating a stratified kappa, Statist. Med., vol. 10, pp. 1465-1472, 1991.

[15] M. L. McHugh, Interrater reliability: The Kappa statistic, Biochem. Medica., Vol. 22, No. 3, pp. 276-282, 2012.

[16] D. Michael, C. Frederick, G. Gregory, S. Steve, B. David, Measurement Systems Analysis (4th ed.). Automotive Industry Action Group (AIAG), Michigan, 2010.

[17] N. Gisev, J. S. Bell, T. F. Chen, Interrater agreement and interrater reliability: Key concepts, approaches, and applications, Res. Soc. Admin. Pharm., Vol. 9, No. 3, pp. 330-338, 2013.