Increasing the Reliability of FMEA Evaluation by Modifying Rating Scales and Applying Pairwise Comparison Method

Edina Kulcsár¹, István Gábor Gyurika¹, Tamás Csiszér²

¹Department of Mechanics, University of Pannonia, H-8200 Veszprém, Hungary
²Sándor Rejtő Faculty of Light Industry and Environmental Protection Engineering, Öbuda University, H-1034, Budapest, Hungary

kulcsar.edina@mk.uni-pannon.hu

Abstract. Failure Mode and Effect Analysis (FMEA) is a frequently applied risk assessment methodology in different fields of industry. It provides the opportunity for deep analysis of risky events, connecting to their potential causes and effects. Due to its robust method, risk can be identified and managed but, because of its complexity, the assessment can be time-consuming, and the answers can be inconsistent. In the authors' experience, the scale used to quantify the priority number can influence these outcomes. To identify and measure the significance level of this connection, a survey was organized with experts of digital manufacturing, who were asked to rank the probability of risk occurrence in planning a production process based on digital manufacturing principles. Participants had to use three assessment methods: 1) traditional 'linguistic' scale; 2) 'ratio scale; 3) pairwise comparison. In the case of the linguistic scale, probabilities were defined by their descriptions and an integer between number 1 and 7. These numbers did not equal the probability percentage. In contrast, the ranks of the ratio scale equalled the ranges of probability percentage. In the pairwise comparison, risks were compared by the frequencies of their occurrences. The assessment results were compared by time-need, difficulty, priority list, and consistency. Time and difficulty were ranked by the survey participants on a 4-level scale. The priority list was created by the summarized rank values. In the consistency assessment, we looked at how the amended scale affects the assessment uncertainty. In addition, pairwise comparisons were used to examine the consistency of each expert's evaluation and filter out those that were not accurate enough.

1. Introduction
Failure Mode and Effects Analysis (FMEA) is a widely used risk assessment method. Its purpose is to identify and quantify the risks associated with the manufacture or use of a product to properly address them [1]. The FMEA can be considered as a "measurement system" for risk assessment, the elements and characteristics of which significantly influence the outcome of the study. As in any measurement system, we distinguish between the object to be measured (risk), the measuring device (FMEA), the measurement method (evaluation scales), the person performing the measurement (evaluator), and the environment (historical data, assumptions, etc.). For the measuring system to give adequate results, the following must be ensured. The sensitivity can determine the value to be measured with adequate sensitivity. Accuracy ensured that the measured value shows the actual value. Repeatability means that successive measurements on the same batch must give the same result if all the other variables in the measurement system are identical. Reproducibility is that successive measurements on the same batch
should give the same result if another of the other variables in the measurement system, usually the person performing the measurement, changes.

Compliance with expectations is affected by all variables in the measurement system. Our research aims to examine how the expectations can be interpreted in the case of FMEA, and in the case of the method as a measurement system, which variable affects the extent to which it meets the criteria related to the outcome of the risk assessment. In this dissertation, we examine the effect of the scale used to assess risk factors on FMEA reproducibility (R&R).

During the Fourth Industrial Revolution, customer requirements for product manufacturing changed significantly. The most important element of the change is that cyclical, in many cases extreme, changes in customer requirements can be observed in the entire design and production planning process of the product. Companies producing products need to respond with sufficient flexibility to the frequent changes in customer demand generated during the product manufacturing process to remain successful market players. These new property requirements provided the basis for the creation of digital manufacturing science, one of the main pillars of which is design and manufacturing design based on a digital manufacturing system. The basic idea of a digital manufacturing system is that companies and manufacturing associations working together to produce products to meet customer needs are independent of each other in time and space, and that information flows between members of a manufacturing association in a digital environment: needs, conceptual designs, material requirements, design and manufacturing design in a virtual environment, and simulation. Given the importance of this topic area, the risk assessment focuses on the risks of design based on a digital manufacturing system [2].

In this context, after a brief literature review, we describe the theoretical background of the FMEA and R&R methodology and then present the impact of the traditionally used scale and a possible alternative scale and a pairwise comparison method on each outcome of the risk assessment by presenting a case study.

2. Background

2.1 Subjectivity of the method

We hypothesize that an improperly designed scale reduces the objectivity of evaluation, which is supported by various researches [3]. The combination of the FMEA method with Monte Carlo simulation does not ensure the elimination of subjectivity either [4]. It was performed severity assessment by varying the amount of error-specific input information. The results of the study showed that participants’ understanding was influenced by the availability of information, and therefore the risk classification became inconsistent [5]. The appropriate scale was selected for the use of the AHP (Analytic Hierarchy Process) methodology, and the consistency test that can be performed with it have been proposed [6]. Other research has used the R&R method to measure FMEA consistency. In one study, two different but equally trained teams assessed the risks, producing different results [7].

2.2 Failure Mode and Effects Analysis

The FMEA is a risk assessment tool that provides a structured, transparent method for collecting and evaluating potential errors, identifying measures to reduce the likelihood of error, and documenting the entire process [8]. The experts assess the risks according to three aspects, which are the severity of the effect of the error (Severity - S) and the frequency (Occurrence - O) and detectability (Detection - D) of the error. For each aspect, potential errors are rated with a predetermined scale [9]. These scales typically range from 1 to 5 or 1 to 10, with the highest value representing the worst case. The product of the three scale values (S, O, D) gives the RPN (Risk Priority Number), which indicates the priority of the risk: the higher the value of the RPN, the riskier the event under study. Measures should normally be set for risks with a product of more than 125 for the 10-point scale with the highest RPN.
The new AIAG manual replaces the RPN with the Action Priority (AP) value to prioritize risks, according to which risks can be divided into three groups based on the values of the factors: low, medium, and high priority. In the latter case, it is mandatory to specify the appropriate measures [10].

2.3 Repeatability and Reproducibility examination
R&R is a procedure used to jointly evaluate the repeatability and reproducibility of measurements. [11]. The adequacy of the measurement system is assessed by fluctuating the results of tests performed several times while keeping its elements constant (repeatability) or changing (reproducibility). In both cases, the goal is to consistently produce values that can be considered identical based on expectations while ensuring the robustness of the measurement system [12].

2.4 Pairwise comparison based FMEA
The application of pairwise comparison in the ranking of problems is already possible when using the FMEA methodology. The aim of the method is to reduce the uncertainty and subjectivity of the evaluation by pairwise comparison. The pairwise comparison is based on a customized version of the basic scale of the AHP method. The consistency of the result of the pairwise comparison can also be evaluated by the method. The FMEA team performs a comparison of \( \frac{n(n-1)}{2} \) in a pairwise comparison matrix. To avoid human fatigue and increased inaccuracy, it is advisable to ensure that the number of assessments is kept to a minimum [13].

3. Case study
The study created an FMEA team of 6 people who are experts in design based on a digital manufacturing system. For the assessment, they identified 17 risks, which are included in table 1. This ensures that the risks are clear to everyone. Three scales were developed for evaluation. A scale of ratios that indicates the probability of occurrence of each risk in percentage terms. A linguistic scale with no numerical values or a scale for pairwise comparison, which is a modified version of the AHP methodology scale. All three scales range from 1 to 7, with 0 as the same occurrence value for pairwise comparison. During the risk assessment, participants had to estimate the probability of the risks occurring. The scales are shown in table 2. To estimate the incidence of risks, three questionnaires were prepared according to the three scales, which were completed several weeks apart, thus ensuring that team members did not remember their previous answers. During the evaluation, participants were also asked how long it took to perform the evaluation with the method and how difficult it was found to use it. The risk assessment was performed separately by the participants so that we could also measure the accuracy relative to themselves.
| Identification of the risks | Identified Risks                                                                                                                                                                                                                                                                                                                                 |
|----------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1                          | Inadequate in-depth assessment of customer needs due to lack of precision. Possible Consequence: The customer needs assessment process is therefore not fast enough. The customer may need to be asked several times during the product design process for clarifications.                                                                   |
| 2                          | When developing your own product, the market survey is not as detailed as is necessary for proper planning.                                                                                                                                                                                                                                                                                   |
| 3                          | When improperly defining your own competencies (at the development company) required for a product manufacturing process before searching for and selecting members of a virtual manufacturing association.                                                                                                               |
| 4                          | Improperly defining your own competencies (at the development company) required for a product manufacturing process before searching for and selecting members of a virtual manufacturing association.                                                                                                               |
| 5                          | Inadequate definition of competencies when creating alternative competency teams within a virtual manufacturing alliance. Explanation: Alternative competency teams are assembled within the virtual manufacturing alliance, selected from professionals in the federal companies according to the specific competency.                                  |
| 6                          | Selection of inappropriate professionals for the creation of alternative competency teams in the distribution of competencies within a virtual manufacturing alliance.                                                                                                                                                                                      |
| 7                          | Inadequate consideration of customer needs for collaborative design. Stricter manufacturing requirements are set by designers in construction plans than the customer expects. The consequence may be that both manufacturing and production may be more expensive.                                                                                                                                 |
| 8                          | In collaborative manufacturing design, you do not have the necessary expertise to properly apply the designs in structural design. The consequence may be that the wrong product to meet customer needs is produced.                                                                                                                                  |
| 9                          | In collaborative production planning, product production parameters that do not correspond to the production environment are determined due to a lack of expertise.                                                                                                                                                                                                                     |
| 10                         | Selection of an inappropriate quality control method to verify the conformity of the product during production and final inspection.                                                                                                                                                                                                                                                                  |
| 11                         | Non-compliant product parameters are checked during production and final inspection to verify the conformity of the product.                                                                                                                                                                                                                                                                                           |
| 12                         | Incomplete assessment of risks related to product non-conformity.                                                                                                                                                                                                                                                                                                                                              |
| 13                         | Lack of adequate spare parts during product support.                                                                                                                                                                                                                                                                                                                                                       |
Table 2. Scales used for risk assessment

| Occurrence scale (percent) | Occurrence scale (linguistic) | Occurrence (pairwise comparison) |
|----------------------------|--------------------------------|----------------------------------|
| 1  | O=0%                          | 1                               | 0                               |
| 2  | 0%<O<=20%                     | 2                               | 1                               |
| 3  | 20%<O<=40%                    | 3                               | 2                               |
| 4  | 40%<O<=60%                    | 4                               | 3                               |
| 5  | 60%<O<=80%                    | 5                               | 4                               |
| 6  | 80%<O<100%                    | 6                               | 5                               |
| 7  | O=100%                        | 7                               | 6                               |

4. Results and discussions

The assessment was performed by all 6 group members on three different scales for all 17 risks. The results are shown in table 3, where P is the result of the answers on the percentage scale, L is the result of the answers on the linguistic scale, and PW is the result of the pairwise comparison method. In the table, the value of the standard deviation is marked in the table with S, and the value of the mean for the answers to the same risks is calculated for all respondents. Identical answers were marked in green for a filler using a different scale or method.

Table 3. Result of the risk assessment with the three scales, P as the percentage scale, L as the linguistic scale and PW as the pairwise comparison method. The green columns mark the same occurrence result on different scale.

The results can be examined from several perspectives based on what has changed in our evaluation system. One such case is when we examine how different the results of different respondents on the same scale are, thus examining the robustness of the scale. The standard deviation values show that different respondents did not give very different answers to the same risks using the same scale or method. 1-2 scale value deviations are the most typical. This can also come from the different opinions.
of the team members, which can be resolved by consensus in a group discussion. Figure 1 shows the distribution of the occurrence values for each respondent in case of the three scale.

Figure 1. Distribution of occurrence values for the three scale and the standard deviation.

The other variable in our system is the scale, which also changed for each participant. The distribution of the standard deviation given on each scale shows which scale was used to scatter the data the most. The results can be seen in figure 2. Based on the sum of the values of the standard deviation, the standard deviation of the values given on the linguistic scale is the smallest, and on the percentage scale the largest. In addition, it is worth examining the average and median values of the standard deviations, the results of which are shown in table 4. Average and median values show the lowest value when using the linguistic scale and the pairwise comparison method.

Table 4. Average and median values of the standard deviations.

|       | P       | L     | PW    |
|-------|---------|-------|-------|
| Average | 1.31    | 1.095 | 1.198 |
| Median  | 1.41    | 1.169 | 1.09  |
Figure 2. Distribution of the standard deviation of the occurrence values in case of the three scale.

From these results, it can be concluded that the design of the scales influences the final result of the evaluation even if the same content is displayed in different ways. This can already be deduced from figure 1, where the responses of different scales are often very different from each other.

Opinions on the scales and method are given in table 5. The notations for the scales and method are the same as those used in table 3. Based on the opinions, no significant difference in difficulty was observed when using each scale. One participant assessed that evaluation by pairwise comparison is very simple. In terms of lead time, each of its applications was considered long or very long. The difficulty of the percentage and language scales was different for the evaluators, with most responses coming from difficult or easy.
Table 5. Respondent's opinion about the scales and the pairwise comparison method.

| 1. | 2. | 3. | 4. | 5. | 6. |
|----|----|----|----|----|----|
| P | L | PW | P | L | PW | P | L | PW | P | L | PW |
| How easy was it to apply the scale in the risk assessment? |
| Very hard | x | x | x | x | x |
| Hard | x | x | x | x | x | x |
| Simple | x | x | x | x | x | x |
| It's very simple | x | x | |
| How long did the risk assessment take to use each method? |
| For a very short time (less than 10 minutes) | x | x | x | x |
| For a short time (between 10 and 20 minutes) | x | x | x | x | x |
| For a long time (between 20 and 30 lawsuits) | x | x | x | x | x |
| For a very long time (more than 30 minutes) | x | x | x | x |

The level of inconsistency in the pairwise comparison result was also evaluated. The methodology initially considers $RE(A)<0.1$ to be acceptable, but this can only be considered for a few factors. For 17 factors, it is difficult to avoid fatigue during evaluation [13]. The results of the evaluation are presented in table 6.

Table 6. Inconsistency of the measurements in the case of the six respondents.

| Inconsistency of the risk assessment |
|-------------------------------------|
| 1. 0.02753 | 3% |
| 2. 0.796235 | 80% |
| 3. 0.708084 | 71% |
| 4. 0.398368 | 40% |
| 5. 0.446742 | 45% |
| 6. 0.538912 | 54% |

For one team member, the level of inconsistency became very low at only 3%. The value of outstanding inconsistencies can be experienced in the case of the 2nd and 3rd fillers; here, the consistency and accuracy of the evaluation can be questioned.

5. Conclusions

The risks of the design process based on a digital manufacturing system were assessed on three different scales, where there was no difference in meaning between the scales. The evaluation was based on the change of 2 factors, whether the respondent or the scale used for the evaluation changed.

When changing the evaluator, the largest deviation was 2 scale values in terms of standard deviation. If we look at the difference of the standard deviations when changing the scales, it can be said that the
sum of the standard deviations became the largest on the percentage scale. Similarly, the average and median values were the highest on the percentage scale, while they were the lowest on the linguistic scale and in case of the pairwise comparison.

In conclusion, changes in scales and assessment methods affect the uncertainty of the assessment, with the same value given in few cases for the probability of occurrence. Regarding the standard deviation, average and median values, the language scale and the pairwise comparison seem to be a more accurate evaluation. The consistency of the pairwise comparison can also be measured, which was very high in two cases for the six evaluators and very low in one case. The most difficult and time-consuming method proved to be the pairwise comparison, but based on the evaluations, the evaluation on the other 2 scales was not so easy, but it took less time. This also follows from the fact that in a pairwise comparison, all pairs have to be compared, which means much more judgment.

Acknowledgement (s)
We would like to thank Dr Tibor Szalay, Dr Ádám Jacsó, Miklós Boleraczki, Dr Géza Haidegger and Dr Gábor Erdős for their participation in the evaluation and for their professional knowledge.

References
[1] L. Petrescu, M.C. Petrescu, E. Cazacu, V. Ionita. Risk Priority Number vs Action Priority in Electrical Systems. In 2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE) (pp. 1-6). IEEE, 2021.
[2] Z. Zhou, S.S. Xie, and D. Chen. Fundamentals of digital manufacturing science. Springer Science & Business Media, 2011.
[3] M. Banghart, et al. "Subjectivity in Failure Mode Effects Analysis (FMEA) Severity Classification within a Reliability Centered Maintenance (RCM) Context." International Journal of Aviation, Aeronautics, and Aerospace 5.1, 2018.
[4] S.J. Rhee, and I. Kosuke. "Life cost-based FMEA incorporating data uncertainty." ASME 2002 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers Digital Collection, 2002.
[5] M. Banghart, et al. "Subjectivity in Failure Mode Effects Analysis (FMEA) Severity Classification within a Reliability Centered Maintenance (RCM) Context." International Journal of Aviation, Aeronautics, and Aerospace 5.1, 2018.
[6] D.C. Aguiar, H. J. C. Souza, and V. A. Salomon. "AHP Application to Evaluate Scoring Criteria for Failure Mode and Effect Analysis (FMEA)." International Journal of Analytic Hierarchy Process 2, 3-13, 2010.
[7] M. T. Oldenhof, et al. "Consistency of FMEA used in the validation of analytical procedures." Journal of pharmaceutical and biomedical analysis 54.3, 592-595, 2011.
[8] K. Johnson, "It's fun to work with an FMEA." Quality Progress 35.1,152, 2002.
[9] R.J. Mikulak, R. McDermott, and M. Beauregard. The basics of FMEA. Productivity Press, 2008.
[10] Stamatis, D. H. Risk Management Using Failure Mode and Effect Analysis (FMEA). ASQ Quality Press, 2019.
[11] H. Urdhwareshe, "Validating measurement system using measurement systems analysis (MSA)." Quality World, 41-47, 2006.
[12] Ch. Dubois, and P. Courtois. "Choice of parameters for an R&R study and its impact on the measurement capability." 19th International Congress of Metrology (CIM2019). EDP Sciences, 2019.
[13] E. Kulcsár, T. Csiszér, J. Abonyi. Pairwise comparison based failure mode and effects analysis (FMEA). MethodsX, 7, 101007, 2020.