Evaluating the effectiveness of data quality framework in software engineering

Marshima Mohd Rosli [1,2], Nor Shahida Mohamad Yusop[1]
[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia
[2]Institute for Pathology, Laboratory and Forensic Medicine (I-PPerForM), University Teknologi MARA, Sungai Buloh, Malaysia

Article Info

Article history:

Received Jul 27, 2021
Revised Jun 6, 2022
Accepted Jul 1, 2022

Keywords:

Data quality framework
Datasets
Empirical studies
Quality of datasets
Software engineering

ABSTRACT

The quality of data is important in research working with data sets because poor data quality may lead to invalid results. Data sets contain measurements that are associated with metrics and entities; however, in some data sets, it is not always clear which entities have been measured and exactly which metrics have been used. This means that measurements could be misinterpreted. In this study, we develop a framework for data quality assessment that determines whether a data set has sufficient information to support the correct interpretation of data for analysis in empirical research. The framework incorporates a dataset metamodel and a quality assessment process to evaluate the data set quality. To evaluate the effectiveness of our framework, we conducted a user study. We used observations, a questionnaire and think aloud approach to provide insights into the framework through participant thought processes while applying the framework. The results of our study provide evidence that most participants successfully applied the definitions of dataset category elements and the formal definitions of data quality issues to the datasets. Further work is needed to reproduce our results with more participants, and to determine whether the data quality framework is generalizable to other types of data sets.

Corresponding Author:

Marshima Mohd Rosli
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA
40450 UiTM, Shah Alam, Selangor, Malaysia
Email: marshima@fskm.uitm.edu.my

1. INTRODUCTION

Research indicates that the quality of data sets is critical to the results of empirical studies [1]–[4]. Although data sets play a central role in empirical research, few studies consider the quality of the data sets used [5]–[9]. If the quality of the data is poor, then the results of empirical studies cannot be trusted, and any models or conclusions based on the data sets are questionable. Ensuring data quality is therefore fundamentally important to the field of empirical software engineering.

Data quality is important because poor data quality may lead to invalid conclusions. This issue has been mentioned in previous studies, but there have been small number of serious discussions on the validity of results [2], [10]–[13]. For example, Gray *et al.* [2] demonstrated that empirical studies based on data sets that contain duplicate data may lead to erroneous results. Shepperd *et al.* [10] confirmed this finding by investigating data integrity and inconsistencies between different versions of defect data sets from the National Aeronautics and Space Administration (NASA) metrics data program (MDP). Gray *et al.* [2] and Shepperd *et al.* [10] urged research communities to take action to establish good-quality data sets in order to ensure the validity of results.

To conduct research with good-quality data sets, we need a way to evaluate the quality of the data sets. A number of researchers have proposed techniques for dealing with data quality issues in data sets [14]–[21]. For example, Abnane *et al.* [14], Mockus [8] and Kim *et al.* [19] proposed imputation techniques to handle missing data problems. Further, Wu *et al.* [20] and Khoshgoftaar and Hulse [22] developed noise identification and filtering techniques to improve data quality. Although many techniques have been proposed to resolve the challenges associated with data quality [15], [20], [23], there is a lack of research that evaluates the quality of data sets-particularly those from public data repositories.

Public data repositories such as NASA MDP [11], PROMISE [24], software artifact infrastructure repository (SIR) and The International Software Benchmarking Standards Group [25] contain a large number of data sets for research purposes. For some data repositories, known quality issues are not clearly documented in the existing data sets [11]. As a result, inexperienced researchers may assume that the available data sets are of reasonable quality for analysis in empirical research [26].

In our previous work, we developed a dataset metamodel that provides a standard terminology for describing a data set and classifies the elements of the data set into three levels: physical structure, dataset category and dataset concept. The metamodel will help researchers to understand the contents of data sets and to identify what information may be missing in the datasets [27]. We evaluated datasets from real data repositories that have a variety of formats and structures to demonstrate the dataset metamodel utility. We then developed a framework for data quality assessment that incorporate the dataset metamodel and a quality assessment process [28] for evaluating the quality of dataset.

We also reported on an online survey we conducted as our preliminary evaluation to determine the effectiveness of dataset category element definitions [28]. The survey results are less meaningful because of the small number of participants who responded. However, the results suggested that participants with relevant background knowledge and experience in research, particularly in analyzing datasets, were able to apply the definitions of dataset category elements correctly. As the findings also indicate there were some problems with the definitions of dataset category elements, we improved these after conducting an analysis of the survey. Therefore, in this paper, we report on a user study we conducted to test the new definitions of dataset category elements and the formal definitions of data quality issues, aiming to observe how researchers apply the new definitions of dataset category elements and the formal definitions of data quality issues. We also provide evidence to answer the question: whether the data quality framework is useful for evaluating the quality of data sets.

2. DATA QUALITY FRAMEWORK

As mentioned earlier, we have developed a data quality framework that aim to determine whether a data set contains sufficient information to facilitate the correct interpretation of data for analysis in empirical research [28]. The framework may identify when there is uncertainty as to which entities have been measured and which metrics have been used. The framework may also help researchers to identify not only common data quality issues (e.g., missing and duplicate data), but also quality issues related to the interpretation of data.

Figure 1 shows the data quality framework that consists of two processes: the modelling of the data set and quality assessment. The framework requires a data set from a data repository. The first process aims to model the data set using a formal procedure based on the dataset metamodel in the knowledge platform. The second process aims to evaluate the quality of data set using the formal definitions of data quality issues in the knowledge platform. The framework produces a data quality report that describes which data set elements existed in the model of the data set and which did not, along with both the identified and unidentified data quality issues related to the interpretation of data.

The framework also suggests ways to improve the quality of the data sets. Moreover, researchers who want to use the existing data sets from data repositories can also use our framework to determine whether or not the data sets are usable for their research purpose. In this case, they need to model the data sets and assess their quality. The framework will allow researchers to better understand the quality of the data sets and choose the appropriate data sets for their research.

Using a formal definition is a suitable approach to clearly describe data quality issues in data sets. As a first step in the construction of the formal definition of data quality issues, as mentioned earlier, we have developed a dataset metamodel that describes the structure and concepts in a data set, as well as the relationships between each concept. Each concept in the metamodel is defined using standard terminology, which allows researchers to specify data quality issues consistently using these standard definitions. Detailed discussion of the metamodel can be found in the paper by Rosli *et al.* [27].

As mentioned in the introduction, we improved our metamodel with new definitions for five dataset category elements. The five new definitions are for ancillary, ancillary value, ancillary label, ancillary metadata, and data type metadata. Figure 2 shows our new version of the metamodel for a dataset, using unified modeling language (UML) notation.

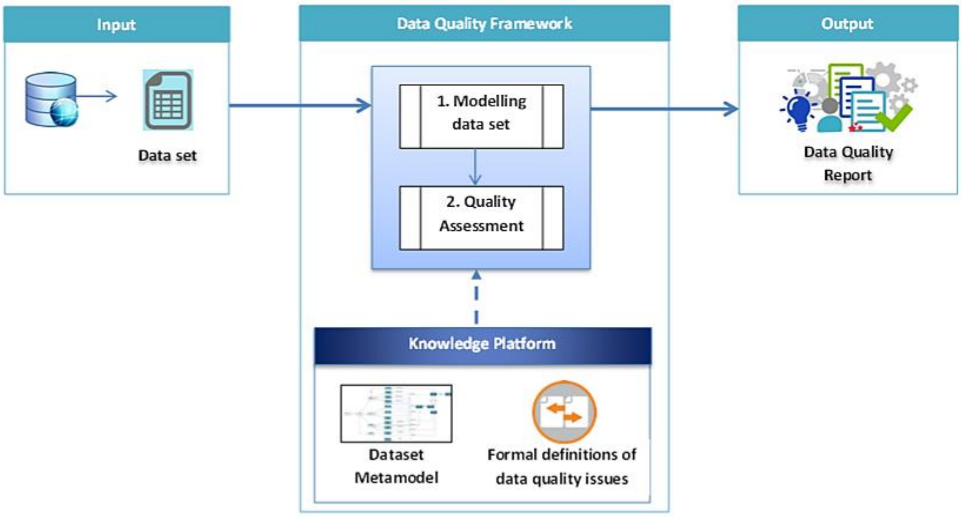


Figure 1. Data quality framework

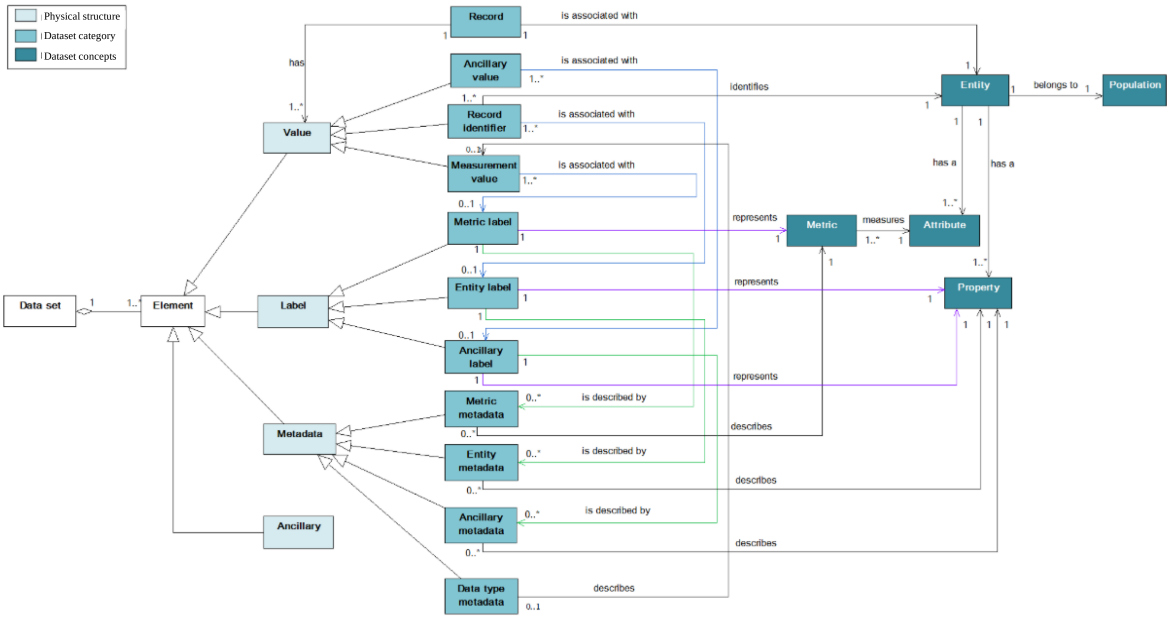


Figure 2. Dataset metamodel

Figure 2 shows a data set consists of a collection of elements. We classified elements in the dataset metamodel into three levels: physical structure, dataset category and dataset concepts. The dataset category consists of 11 elements: record, measurement value, ancillary value, record identifier, entity label, metric label, ancillary label, entity metadata, metric metadata, ancillary metadata, and data type metadata.

For the user study, we selected seven dataset category elements: record, record identifier, measurement value, entity label, metric label, metric metadata, and entity metadata. These elements were selected because they are essential terminology that we used for interpretation of data in the datasets. We use the seven dataset elements from the dataset metamodel to construct the formal definition of data quality issues. These seven dataset category elements are important for identifying the structure of a data set that may have quality issues. We provide the following formal definition for six data quality issues: i) duplicate data: two or more records that have the same measurement values associated with the same metric for the same entity; ii) inconsistent data: two or more records that have different measurement values associated with the same metric for the same entity; iii) missing data: a record that does not have a measurement value for a given metric; iv) incorrect data: a record that has an invalid measurement value for a given metric; v) incomplete metadata for a metric: a metric label that does not have metric metadata; and vi) incomplete metadata for an entity: an entity label that does not have entity metadata.

3. METHOD

3.1. Study design

The study was designed to determine the effectiveness of the application of definitions of dataset category elements and the effectiveness of the application of formal definitions for data quality issues to a range of datasets. In our study, we measure effectiveness by evaluating the extent to which participants can answer questions correctly. We constructed two research questions (RQs) for the study:

- RQ1: Can most software engineering researchers correctly apply the new definitions of dataset category elements?
- RQ2: Can most software engineering researchers correctly identify the quality issues in data sets using the formal definition of data quality issues?

The study involved the use of a task list with observation and a questionnaire. The task list contained a set of tasks that participants had to complete, and the observation contained a set of observation questions to be answered by the researcher. For the task list, we asked participants to perform four different tasks relating to the data quality framework. These tasks were constructed to allow the participant to explore the data quality framework and to apply the new definitions of dataset category elements and the formal definitions of data quality issues to the datasets.

For the observation, we applied a combination of two methods for the observation: i) unobtrusive observation and ii) obtrusive observation. With unobtrusive observation, we observed the participant’s use of the data quality framework while performing the given tasks. This method was performed in the early stage of the user study.

In the questionnaire, we constructed five multiple-choice questions and two open-ended questions. The five multiple-choice questions were about the participants’ background in research and their experience with datasets from data repositories. In the two open ended questions, we asked participants if they had any comments on quality issues in datasets and we allow participants to give comment on ways to improve the data quality framework.

3.1.1. Task list

The task list contained a set of tasks that participants had to complete, and the observation contained a set of observation questions to be answered by the researcher. We asked participants to perform three different tasks relating to the data quality framework. The tasks are i) explore the data quality framework, ii) identify the dataset category elements, and iii) identify data quality issues.

3.1.2. Observations

In the obtrusive observation, we asked each participant what they thought about the framework after they had completed the given tasks. With both observation methods, no personal information about the participants was collected, as participation in this study was treated anonymously. To collect the observation data, we observed the participant’s activity while undertaking each task, according to the following observation questions (OQs):

- Task 1: Explore the data quality framework

  OQ1: Does participant look at the page containing the definitions of dataset category elements, the definitions of data quality issues and the examples of applying part of the data quality framework? (Yes/No)
- Task 2: Identify the dataset category elements

  OQ2a: Does the participant communicate with the researcher? (Yes/No)

  OQ2b: Does the participant look at the page containing the definitions of dataset category elements while performing the task? (Yes/No)

  OQ2c: Does the participant complete the given task? (Yes/No)
- Task 3: Identify data quality issues

  OQ3a: Does the participant communicate with the researcher? (Yes/No)

  OQ3b: Does the participant look at the page containing the formal definitions of data quality issues while performing the task? (Yes/No)

  OQ3c: Does the participant complete the given task? (Yes/No).

3.2. Execution

The user study was conducted on a one-to-one basis in selected rooms within the Department of Computer Science. As previously indicated, the study involved participants completing tasks on a list, observation of participants during completing the tasks, and a participant questionnaire. Before participants started performing the study, they were asked to follow the instructions written on the task list and the questionnaire. Participants were also encouraged to ask any questions during the study session. During the study, we observed how the participants performed the tasks on the list and answered the questionnaire. Participants were asked to send back the task list and the questionnaire when they completed.

4. RESULTS

By the end of the recruitment period, we had recruited 25 participants'-20 postgraduate students and five academic researchers. All participants took part in the one-to-one study. All participants managed to complete the two parts.

In the study, we asked participants to answer seven questions related to their background information. The aim of this part of the study was to find out whether the participant had experience in analyzing datasets and had encountered data quality issues with datasets. Five academic researchers and 20 postgraduate students participated in the study. Of the five academic researchers, two had more than 10 years of experience (8%) and three had five to 10 years of experience (12%) doing research in computer science or software engineering. All postgraduate students had less than five years of experience (80%) doing research in computer science or software engineering. Sixteen participants had experienced analyzing datasets in their research, 12 of the experienced participants had less than five years of experience doing research, 2 of the experienced participants had between 5 and 10 years of experience and more than 10 years of experience doing research respectively. However, there are 9 participants that do not have experience in analyzing datasets, 8 of participants had less than five years of experience doing research and one of them had between 5 and 10 years of experience doing research.

In the questionnaire, participants were also asked about dataset quality issues that they had encountered. This question allowed participants to select more than one quality issue. Although some participants reported that they did not have experience in analyzing datasets, all participants reported encountering data quality issues with datasets-possibly, these participants encountered the data quality issues with datasets not in their research, but in their work or assignments. Sixteen participants reported encountering 'missing data', 10 'inconsistent data', 12 'duplicate data' and six 'incorrect data'. Participants who indicated more than one quality issue were counted for each quality issue they selected; therefore, the sum of the participant responses is greater than 25.

4.1. Task list and observation

In part one of the study, we observed how participants applied the definitions of datasets category elements and used the formal definitions of data quality issues when carrying out the four tasks that we structured in the study. All participants managed to complete part one. Table 1 shows the results of the observation.

Table 1. The result of observation

Observation Questions (OQ)	Yes	No	Total
OQ1	25	0	25
OQ2a	9	16	25
OQ2b	25	0	25
OQ2c	25	0	25
OQ3a	6	19	25
OQ3b	25	0	25
OQ3c	25	0	25
OQ4a	7	18	25
OQ4b	25	0	25
OQ4c	25	0	25

4.1.1. Task 1: Explore the data quality framework

The results for OQ1, shown in Table 1, indicate that all participants looked at the related data quality framework documents (the definitions of dataset category elements, the formal definitions of data quality issues and the examples of applying part of the data quality framework). With the think-aloud approach, most of the participants communicated with the researcher to gain an understanding of the definitions of dataset category elements and the formal definitions of data quality issues. For example, some of participants asked one or two questions related to the definitions of dataset category elements, and after the researcher explained the examples of applying the definitions of dataset category elements, all participants managed to understand the definitions of dataset category elements.

4.1.2. Task 2: Identify the dataset category elements

The results for OQ2a, shown in Table 1, indicate that nine participants communicated with the researcher, through the think-aloud approach, while performing this task. Table 2 shows the number of participants who gave the correct answer and the number who gave the incorrect answer when applying the definitions of dataset category elements to the datasets. Most participants were able to apply the definitions of the dataset category elements by identifying these elements correctly across the two different datasets. This can be seen in Table 2, which indicates that all participants provided correct answers for measurement value, and there were 45 correct responses for metric label. This result can be explained by the most common elements of datasets being measurement values and these elements being associated with metric labels. Some participants with experience in research may have used measurement values and metric labels frequently in datasets.

We noticed that few participants provided incorrect answers for entity metadata across the two different datasets. In particular, seven responses for dataset A and eight responses for dataset B. All responses are from postgraduate students. These participants probably had no experience in using entity metadata in the datasets, because many existing datasets in data repositories do not contain metadata. We also noticed that seven participants six postgraduate students and one academic researcher provided incorrect answers for record identifier in dataset A. They may not have been familiar with a record identifier and misunderstood the definition for it.

Table 2. Numbers of participants who correctly and incorrectly applied the definitions of dataset category elements

Dataset category element	Dataset	Participant responses		Total
		Correct	Incorrect	
Measurement value	Dataset A	25	0	25
	Dataset B	25	0	25
Metric label	Dataset A	24	1	25
	Dataset B	25	0	25
Record identifier	Dataset A	18	7	25
	Dataset B	22	3	25
Entity label	Dataset A	23	2	25
	Dataset B	22	3	25
Metric metadata	Dataset A	22	3	25
	Dataset B	24	1	25
Entity metadata	Dataset A	18	7	25
	Dataset B	17	8	25
Total	Dataset A	130	20	150
	Dataset B	135	15	150

4.1.3. Task 3: identify the data quality issues

Table 3 shows the numbers of participants who correctly and incorrectly applied the formal definitions of data quality to the given datasets. Most participants were able to identify one or more data quality issues using the formal definitions of data quality issues in the two datasets. This can be seen in Table 3, which indicates all participants provided correct answers for identifying missing data in dataset B. This result could be because the most common quality issue in datasets is missing data. Participants who had experience in analyzing datasets may have encountered missing data in datasets frequently.

Ten participants nine students and one academic researcher did not correctly identify inconsistent data in dataset B. These participants may not have encountered inconsistent data in datasets before. Thirteen postgraduate student participants provided incorrect answers for identifying the incomplete metadata for a metric in dataset A. This seems reasonable because the postgraduate students probably had less experience in analyzing datasets, so this could be why they were sometimes unable to identify the quality issue correctly.

4.2. Analysis of participants’ application of the new definitions of dataset category elements according to participants’ background

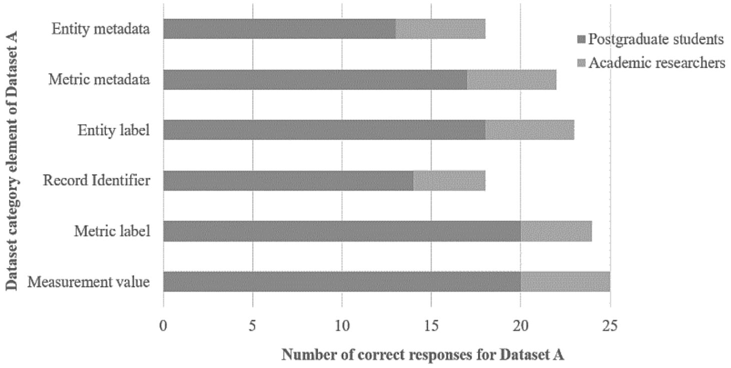
To answer RQ1, whether most surface-enhanced Raman spectroscopy (SERs) can correctly apply the new definitions of dataset category elements we focused on participant responses to the questions in Task 1. We analyzed participants’ correct responses grouped by participant background (academic researcher and postgraduate student). This analysis allowed us to explore which group provided more correct answers in the study. Figure 3 shows the results of the application of new definition analysis with correct responses for Dataset A in Figure 3(a) and correct responses for dataset B in Figure 3(b).

Figure 3(a) shows that most participants correctly applied the new definitions of dataset category elements with dataset A. In particular, all academic researchers correctly applied the new definitions for measurement value, entity label, entity metadata and metric metadata, and all postgraduate students correctly applied the new definitions for measurement value and metric label. Eighty per cent of academic researchers' applications of the new definitions for metric label and record identifier were correct. This shows most academic researchers responded with correct answers for most of the dataset category elements in dataset A.

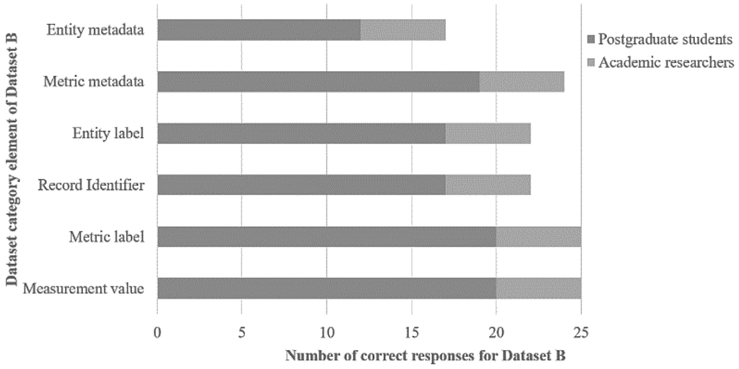
Figure 3(a) also shows that 70% of postgraduate students responded with correct answers for record identifier, 90% for entity label, 85% for metric metadata and 65% for entity metadata. They incorrectly applied the definitions for elements related to entity (record identifier, entity label and entity metadata). This might be because they were not familiar with the dataset elements related to entity.

Table 3. Numbers of participants who correctly and incorrectly applied the formal definitions of data quality issues

Data quality issues	Dataset	Participant responses		Total
		Correct	Incorrect	
Duplicate data	Dataset A	24	1	25
	Dataset B	22	3	25
Inconsistent data	Dataset A	22	3	25
	Dataset B	15	10	25
Missing data	Dataset A	22	3	25
	Dataset B	25	0	25
Incomplete metadata for a metric	Dataset A	12	13	25
	Dataset B	23	2	25
Incomplete metadata for an entity	Dataset A	10	15	25
	Dataset B	13	12	25
Total	Dataset A	80	35	115
	Dataset B	88	27	115



(a)



(b)

Figure 3. Number of correct responses for six dataset category elements in (a) dataset A and (b) dataset B

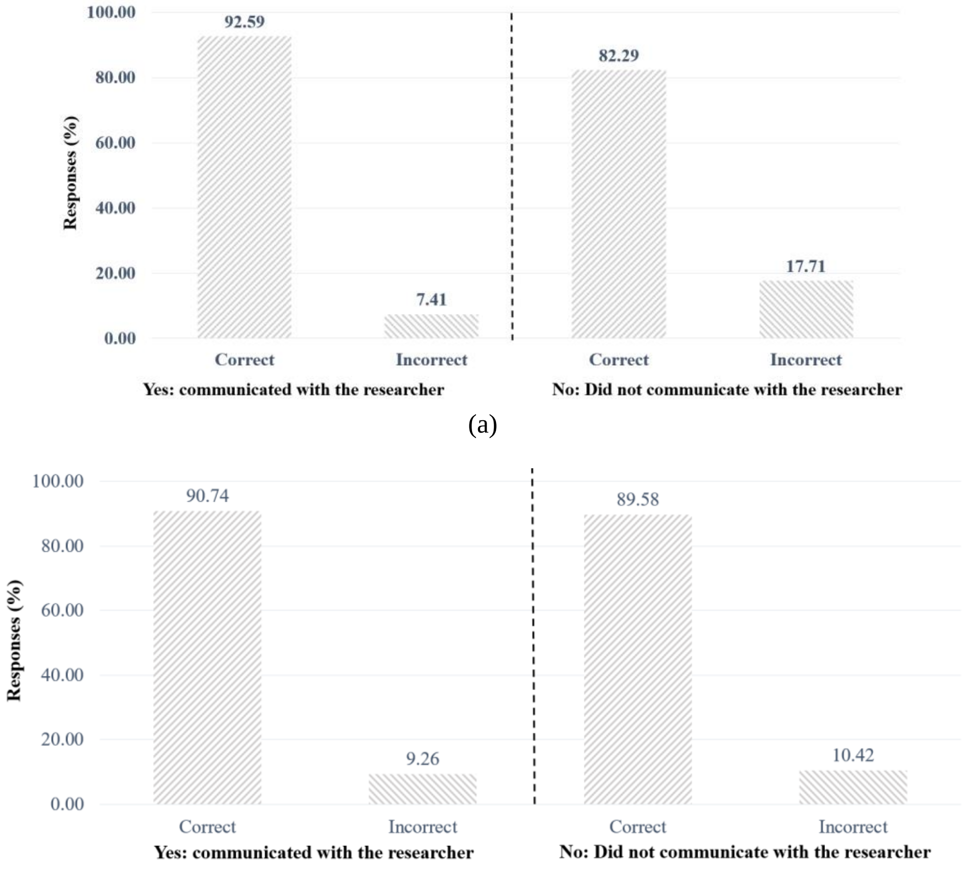
Figure 3(b) shows that most participants correctly applied the new definitions of dataset category elements with dataset B. All academic researchers correctly applied the new definitions for all dataset category elements, while all postgraduate students correctly applied the new definitions for measurement value and metric label. This shows more academic researchers responded with correct answers than did postgraduate students.

In Figure 3(b), we can see that 85% of postgraduate students responded correctly for both record identifier and entity label, 95% for metric metadata and 60% for entity metadata. As with dataset A, some postgraduate students responded incorrectly for elements related to entity. This might be because they were not familiar with the dataset elements or misunderstood some of the definitions of dataset category elements.

Figures 3(a) and 3(b) show that most of the academic researchers correctly applied the new definitions of dataset category elements to datasets. As described earlier, all academic researchers who participated in this study had more than five years' experience in research. This suggests that the academic researchers who participated in this study had a strong background of knowledge and experience in research that assisted them to apply correctly the new definitions of dataset category elements.

4.3. Analysis of participants' application of the new definitions of dataset category elements according to observation data

We analyzed participants' correct responses grouped by the observation data that indicated nine participants communicated with the researcher 'Yes' and 16 participants did not 'No'. This analysis allowed us to explore which group i) participants who communicated with the researcher or ii) participants who did not communicate with the researcher) provided the most correct answers in the study. We calculated the percentage of total correct responses for six dataset category elements for the two groups, i) participants who communicated with the researcher and ii) participants who did not communicate with the researcher. Figure 4 shows the results of the participant responses for dataset A in Figure 4(a) and the participants responses for dataset B in Figure 4(b).



(a)

(b)

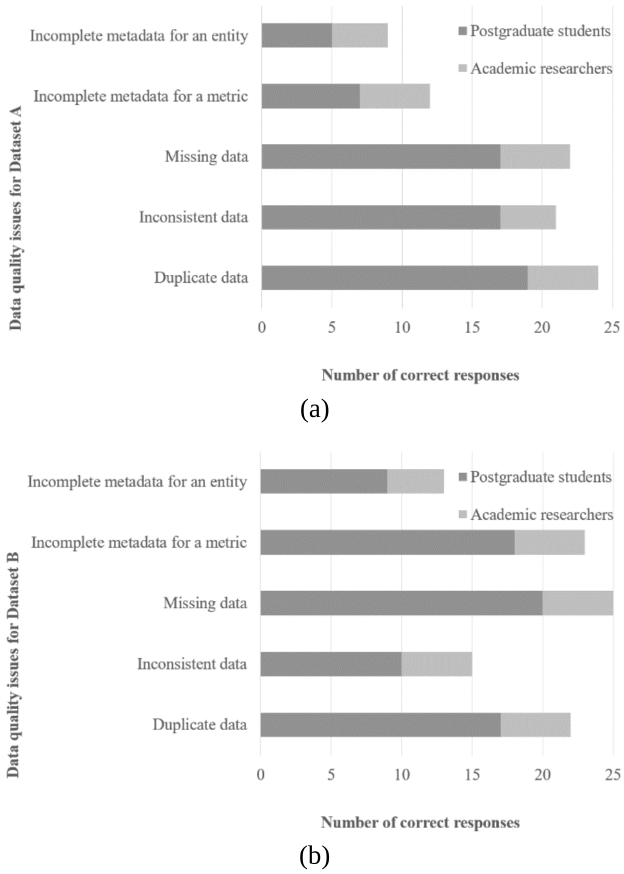
Figure 4. Percentage of responses for six dataset category elements in (a) dataset A and (b) dataset B

Figure 4(a) shows that 92.59% of the correct responses were made by participants who communicated with the researcher while performing Task 2 with dataset A, while 82.29% were made by participants who did not communicate with the researcher. This reveals that participants who communicated with the researcher responded correctly more often than participants who did not. It seems that the participants who communicated with the researcher gained a better understanding of the definitions of dataset category elements than those who did not communicate with the researcher.

Figure 4(b) shows that 90.74% of correct responses were made by participants who communicated with the researcher while performing Task 2 with dataset B, while 89.58% were made by participants who did not communicate with the researcher. This highlights a small difference in correct responses between participants who communicated with the researcher and participants who did not. Some of the participants who did not communicate with the researcher but responded correctly may have read carefully the definitions of dataset category elements and the examples of applying the definitions.

4.4. Analysis of participants' application of the formal definitions of quality issues according to participants' background

To answer RQ2, whether most SERs can correctly apply the formal definitions of data quality issues we focused on participant responses to the questions in Task 3. We analyzed participants' correct responses grouped by participant background (academic researcher and postgraduate student). This analysis allowed us to explore which group provided the most correct answers in the study. We present the results of this analysis in Figure 5(a) dataset A and Figure 5(b) dataset B.



(a)

(b)

Figure 5. Number of correct responses for five quality issues in (a) dataset A and (b) dataset B

Figure 5(a) shows that most academic researchers correctly applied the formal definitions of data quality issues in dataset A. In particular, all academic researchers correctly applied the formal definitions for duplicate data, missing data, and incomplete metadata for a metric. Eighty per cent of all responses from academic researchers were correct for inconsistent data and incomplete metadata for an entity, with researchers correctly applying the formal definitions for these.

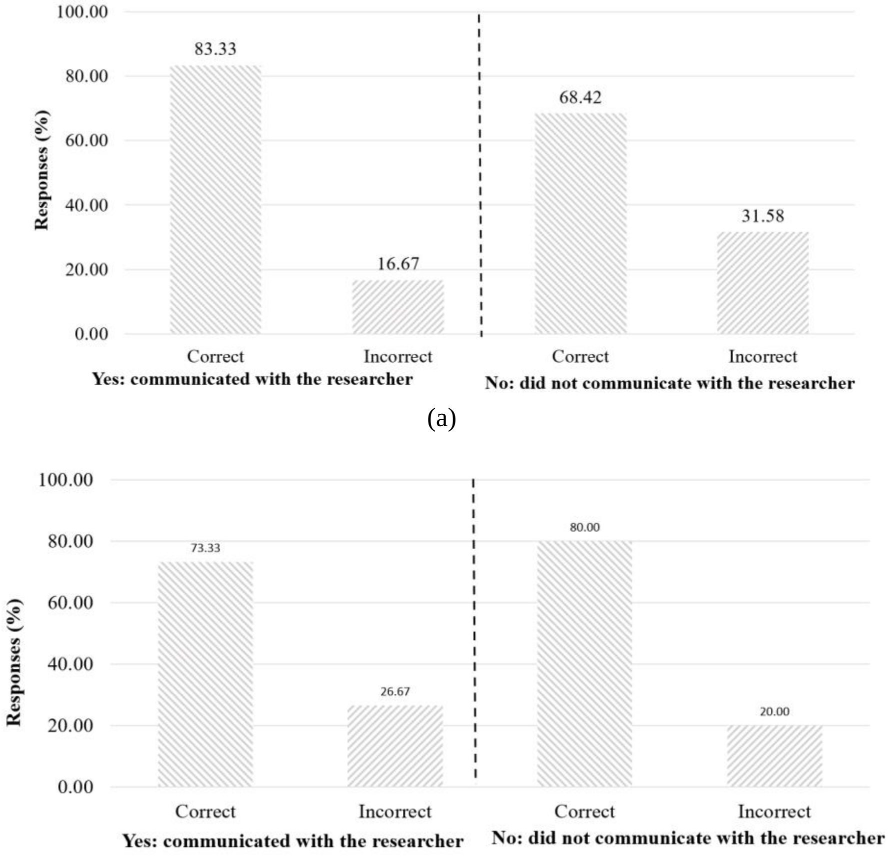
In Figure 5(b), we can see that 95% of postgraduate students' responses were correct for duplicate data, 85% for inconsistent data, 85% for missing data, 35% for incomplete metadata for a metric and 25% for incomplete metadata for an entity. They responded correctly less than 50% of the time for metadata-related quality issues (incomplete metadata for a metric). This could be because some of the postgraduate students may not have been familiar with metadata-related quality issues, so responded incorrectly.

Figure 5(b) shows that most academic researchers correctly applied the formal definitions of data quality issues in dataset B. In particular, all academic researchers correctly applied the formal definitions for duplicate data, inconsistent data, missing data, and incomplete metadata for a metric, while 80% of academic researchers correctly applied the formal definitions for incomplete metadata for an entity. This shows most academic researchers responded correctly for most of the data quality issues in dataset B.

In Figure 5(b), we can see that 100% of postgraduate students' responses were correct for missing data, 90% for incomplete metadata for a metric, 85% for duplicate data, 50% for inconsistent data and 45% for incomplete metadata for an entity. As with dataset A, some of the postgraduate students may have misunderstood the formal definitions for incomplete metadata for a metric, duplicate data, inconsistent data, and incomplete metadata for an entity because they might not have been familiar with these quality issues.

4.5. Analysis of participants' application of the formal definitions of data quality issues according to observation data

We analyzed participants' correct responses grouped by the observation data that indicated six participants communicated with the researcher 'Yes' and 19 participants did not communicate with the researcher 'No'. This analysis allowed us to explore which group (i) participants who communicated with the researcher, or ii) participants who did not communicate with the researcher) most often correctly applied the formal definitions of quality issues. We calculated the percentage of correct responses for five data quality issues across the two groups, i) participants who communicated with the researcher and ii) participants who did not communicate with the researcher. The results of this analysis are presented in Figure 6(a) dataset A and Figure 6(b) for dataset B.



(a)

(b)

Figure 6. Percentage of correct responses for five quality issues in (a) dataset A and (b) dataset B

Figure 6(a) shows that 83.33% of correct responses were made by participants who communicated with the researcher while performing Task 3 in dataset A, while 68.42% of the correct responses were made by participants who did not communicate with the researcher. This shows participants who communicated with the researcher responded more correctly than participants who did not. It seems that participants who communicated with the researcher gained a better understanding of the formal definitions of data quality issues than those who did not communicate with the researcher, allowing them to respond correctly in dataset A.

Figure 6(b) shows that 73.33% of correct responses were made by participants who communicated with the researcher while performing Task 3 in dataset B, and 80.00% of correct responses were made by participants who did not communicate with the researcher. Thus, participants who did not communicate with the researcher responded more correctly than participants who did communicate with the researcher. As described previously, most participants had encountered data quality issues in datasets. It could be that participants who did not communicate with the researcher responded correctly because they were familiar with data quality issues in datasets.

5. DISCUSSION

We conducted a user study to determine the effectiveness of the application of the new definitions of dataset category elements and the effectiveness of the application of the formal definitions of data quality issues to two datasets, A and B. We presented an analysis of the frequency of data in the user study to answer RQ1 and RQ2. The findings from the user study give us high confidence to answer our research questions. First, we find that the majority of participants applied the new definitions of dataset category elements in the two datasets correctly (RQ1). Second, we find that the majority of participants identified missing data and duplicate data using the formal definitions of data quality issues in the two datasets correctly (RQ2). To strengthen the findings, we analyzed participant responses based on the participants' background and their responses based on the observation data.

In summary, we found that the results of the user study provide more evidence than the results of the survey for the evaluation of part of the data quality framework, not only due to the number of participants who responded but also because of the design of the study. For example, we provided some general information about measurements in datasets in the online survey to guide the participants to answer the survey questions, whereas in the user study, we provided examples of applying the definitions of dataset category elements and the formal definitions of data quality issues to help them perform the given tasks. We also found that the results of the user study are better than the survey results. This could be because we allowed participants to ask the researcher questions, via the think-aloud approach, while performing the tasks in the study. We found that participants who communicated with the researcher made fewer errors. However, the observation data shows that some participants who did not communicate with the researcher also successfully applied the definitions of dataset category elements and the formal definitions of data quality issues. This suggests that our framework helped most of the participants to respond correctly in the user study.

The limitation of this study is the way the participants were recruited. We used convenience sampling for our target population of participants. We could not generalize our user study to the entire population engaged in our research, because our participants were limited to computer science researchers or SERs due to the specific content of the study. However, the analysis of results from the study indicates most participants were able to apply the definitions of the dataset category elements and the formal definitions of data quality issues to the datasets successfully. The results also give us high confidence to answer our research questions. Another limitation of this study is the random heterogeneity of participants. In our study, we focused on two types of participants, academic researchers, and postgraduate students. Their amount of research experience, whether in general research or analyzing datasets, may have affected how they answered the questions in the user study. However, from the analysis of the results, we did not find that the difference in experience in analyzing datasets had an influence on how effectively participants applied the definitions of dataset category elements. This was clear in that some participants, who had less than five years' experience in research and did not have experience in analyzing datasets, successfully identified the dataset category elements using the new definitions of dataset category elements and identified the quality issues using the formal definitions of data quality issues.

6. CONCLUSION

In this study, we have presented the design, execution, and results of our user study, as well as analyses based on the findings of these results. We have also described the threats to the validity of our user study. The user study was conducted to determine the effectiveness of the application of the new definitions of the dataset category elements and the effectiveness of the application of the formal definitions of data quality issues. The user study results show that most participants successfully applied the definitions of dataset category elements and the formal definitions of data quality issues to the datasets.

The study findings give us confidence to answer our research questions. First, the analysis of results for applying the new definitions of dataset category elements supports RQ1, which states the majority of SERs would correctly apply the new definitions of dataset category elements in datasets, and the results show most participants applied the new definitions of dataset category elements in the two datasets correctly. Second, the analysis of results for applying the formal definitions of data quality issues supports RQ2, that the majority of SERs would identify correctly one or more quality issues using the formal definitions of data quality issues. The results show that most participants identified correctly missing and duplicate data using the formal definitions of data quality issues.

In this study, we have demonstrated the usefulness of our part of data quality framework in evaluating the quality of software engineering data sets but not in other kinds of data sets such as social science data sets. Further research efforts are required to understand these kinds of data sets in order to apply our data quality framework to them. As suggested by the results of our user study, we believe that our data quality framework able to help researchers understand the quality of data sets. The framework incorporated a dataset metamodel to allow a common interpretation of the description of the structure of data sets, as well as an assessment process to evaluate quality. Further, it can help researchers to determine whether a data set has sufficient information to support the correct interpretation for analysis in empirical research.

ACKNOWLEDGEMENTS

The authors would like to thank the Ministry of Higher Education Malaysia and Universiti Teknologi MARA for their financial support to this project under UiTM Internal REI Grant 600-IRMI/REI 5/3 (001/2019).

REFERENCES

[1] M. F. Bosu and S. G. Macdonell, "Experience," *Journal of Data and Information Quality*, vol. 11, no. 4, pp. 1–38, Sep. 2019, doi: 10.1145/3328746.

[2] D. Gray, D. Bowes, N. Davey, Y. Sun, and B. Christianson, "Reflections on the NASA MDP data sets," *IET Software*, vol. 6, no. 6, 2012, doi: 10.1049/iet-sen.2011.0132.

[3] G. Liebchen and M. Shepperd, "Data sets and data quality in software engineering," in *Proceedings of the The 12th International Conference on Predictive Models and Data Analytics in Software Engineering*, Sep. 2016, pp. 1–4, doi: 10.1145/2972958.2972967.

[4] M. Shepperd and S. G. Macdonell, "New ideas and emerging research : evaluating prediction system accuracy," *Computing in Mathematics, Natural Science, Engineering and Medicine*, pp. 7–10, 2011.

[5] A. Idri, I. Abnane, and A. Abran, "Systematic mapping study of missing values techniques in software engineering data," *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2015-Proceedings*, 2015, doi: 10.1109/SNPD.2015.7176280.

[6] A. Bachmann, C. Bird, F. Rahman, P. Devanbu, and A. Bernstein, "The missing links," *Proceedings of the eighteenth ACM SIGSOFT international symposium on Foundations of software engineering-FSE '10*, 2010, doi: 10.1145/1882291.1882308.

[7] Y. Jiang, B. Cukic, and T. Menzies, "Fault prediction using early lifecycle data," in *The 18th IEEE International Symposium on Software Reliability (ISSRE '07)*, Nov. 2007, pp. 237–246, doi: 10.1109/ISSRE.2007.24.

[8] A. Mockus, "Missing data in software engineering," in *Guide to Advanced Empirical Software Engineering*, London: Springer London, 2008, pp. 185–200, doi: 10.1007/978-1-84800-044-5_7.

[9] R. Torkar, R. Feldt, and C. A. Furia, *Bayesian data analysis in empirical software engineering: the case of missing data*. Cham: Springer International Publishing, 2020, doi: 10.1007/978-3-030-32489-6_11.

[10] M. Shepperd, Q. Song, Z. Sun, and C. Mair, "Data quality: some comments on the NASA software defect datasets," *IEEE Transactions on Software Engineering*, vol. 39, no. 9, pp. 1208–1215, Sep. 2013, doi: 10.1109/TSE.2013.11.

[11] D. Gray, D. Bowes, N. Davey, Y. Sun, and B. Christianson, "The misuse of the NASA metrics data program data sets for automated software defect prediction," in *15th Annual Conference on Evaluation and Assessment in Software Engineering (EASE 2011)*, 2011, pp. 96–103, doi: 10.1049/ic.2011.0012.

[12] C. Hwang, K. Lee, and H. Jung, "Improving data quality using a deep learning network," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 20, no. 1, pp. 306–312, Oct. 2020, doi: 10.11591/ijeecs.v20.i1.pp306-312.

[13] J. M. Z. Hoque *et al.*, "A survey on cleaning dirty data using machine learning paradigm for big data analytics," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 10, no. 3, pp. 1234–1243, Jun. 2018, doi: 10.11591/ijeecs.v10.i3.pp1234-1243.

[14] I. Abnane, A. Idri, and A. Abran, "Fuzzy case-based-reasoning-based imputation for incomplete data in software engineering repositories," *Journal of Software: Evolution and Process*, vol. 32, no. 9, Sep. 2020, doi: 10.1002/smr.2260.

[15] K. K. Bejjanki, J. Gyani, and N. Gugulothu, "Class imbalance reduction (CIR): a novel approach to software defect prediction in the presence of class imbalance," *Symmetry*, vol. 12, no. 3, Mar. 2020, doi: 10.3390/sym12030407.

[16] J. Huang *et al.*, "Cross-validation based K nearest neighbor imputation for software quality datasets: An empirical study," *Journal of Systems and Software*, vol. 132, pp. 226–252, Oct. 2017, doi: 10.1016/j.jss.2017.07.012.

[17] J. Huang and H. Sun, "Grey relational analysis based k nearest neighbor missing data imputation for software quality datasets," in *2016 IEEE International Conference on Software Quality, Reliability and Security (QRS)*, Aug. 2016, pp. 86–91, doi: 10.1109/QRS.2016.20.

[18] T. M. Khoshgoftaar and J. Van Hulse, "Empirical case studies in attribute noise detection," in *IRI-2005 IEEE International Conference on Information Reuse and Integration, Conf, 2005*, 2005, pp. 211–216, doi: 10.1109/IRI-05.2005.1506475.

[19] S. Kim, H. Zhang, R. Wu, and L. Gong, "Dealing with noise in defect prediction," in *Proceedings-International Conference on Software Engineering*, 2011, pp. 481–490, doi: 10.1145/1985793.1985859.

[20] R. Wu, H. Zhang, S. Kim, and S. C. Cheung, "ReLink: Recovering links between bugs and changes," in *SIGSOFT/FSE 2011-Proceedings of the 19th ACM SIGSOFT Symposium on Foundations of Software Engineering*, 2011, pp. 15–25, doi: 10.1145/2025113.2025120.

[21] S. M. Ghazali, N. Shaadan, and Z. Idrus, "Missing data exploration in air quality data set using R-package data visualisation tools," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 9, no. 2, pp. 755–763, Apr. 2020, doi: 10.11591/eei.v9i2.2088.

[22] T. M. Khoshgoftaar and J. Van Hulse, "Identifying noise in an attribute of interest," in *Proceedings-ICMLA 2005: Fourth International Conference on Machine Learning and Applications*, 2005, no. 561, pp. 55–60, doi: 10.1109/ICMLA.2005.39.

[23] C.-Y. Huang and T.-Y. Kuo, "Queueing-theory-based models for software reliability analysis and management," *IEEE Transactions on Emerging Topics in Computing*, vol. 5, no. 4, pp. 540–550, Oct. 2017, doi: 10.1109/TETC.2014.2388454.

[24] T. Menzies, E. Kocagüneli, L. Minku, F. Peters, and B. Turhan, "Using goals in model-based reasoning," in *Sharing Data and Models in Software Engineering*, no. 1, Elsevier, 2015, pp. 321–353, doi: 10.1016/B978-0-12-417295-1.00024-2.

[25] L. Cheikhi and A. Abran, "Promise and ISBSG software engineering data repositories: a survey," in *2013 Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement*, Oct. 2013, pp. 17–24, doi: 10.1109/IWSM-Mensura.2013.13.

[26] B. Kitchenham and E. Mendes, "Why comparative effort prediction studies may be invalid," *Proceedings of the 5th International Conference on Predictor Models in Software Engineering-PROMISE '09*, 2009, doi: 10.1145/1540438.1540444.

[27] M. M. Rosli, E. Tempero, and A. Luxton-Reilly, "What is in our datasets?," in *Proceedings of the Australasian Computer Science Week Multiconference*, Feb. 2016, pp. 1–10, doi: 10.1145/2843043.2843059.

[28] M. M. Rosli, E. Tempero, and A. Luxton-Reilly, "Evaluating the quality of datasets in software engineering," *Advanced Science Letters*, vol. 24, no. 10, pp. 7232–7239, Oct. 2018, doi: 10.1166/asl.2018.12920.

BIOGRAPHIES OF AUTHORS

Marshima Mohd Rosli is a senior lecturer at the Department of Computer Science, Universiti Teknologi MARA, Malaysia, where she has been a faculty member since 2007. Marshima graduated with B. Sc. (Hons) Information Technology from Universiti Utara Malaysia in 2001 and an M.Sc. in Real Time Software Engineering from Universiti Teknologi Malaysia in 2006. She completed her Ph.D. in Computer Science from The University of Auckland, New Zealand, in 2018. Her research interests are primarily in the area of software engineering, artificial intelligent and data analytics. She can be contacted at email: marshima@fskm.uitm.edu.my**.**

Nor Shahida Mohamad Yusop is a senior lecturer in the Department of Information System at the Universiti Teknologi MARA. She had a BA of eng. (computer) (hons) and M. SC. (comp. sc. real time software eng.) from Universiti Teknologi Malaysia in 2003 and 2004, respectively. She completed her Ph.D. in software engineering from Swinburne University of Technology, Melbourne in 2018. Her research interests are mostly in the area of human-centric software engineering, particularly on requirements engineering, software development and software testing. She can be contacted at email: nor_shahida@uitm.edu.my.