K. Henningsson and C. Wohlin, "Monitoring Fault Classification Agreement in an Industrial Context", Proceedings of 9th Conference on Empirical Assessment in Software Engineering, Keele, UK, 2005.

Monitoring Fault Classification Agreement in an Industrial Context

Kennet Henningsson, Claes Wohlin
Blekinge Institute of Technology
[Kennet.Henningsson||Claes.Wohlin]@bth.se

Abstract

Based on prior investigations and the request from a collaborative research partner, UIQ Technology, an investigation to develop an improved and more informative fault classification scheme was launched. The study investigates the level of agreement, a prerequisite for using a fault classification, between classifiers in an industrial setting.

The method used is an experimental approach performed in an industrial setting for determining the agreement among classifiers facilitating for example Kappa statistics for determining the agreement.

From the study it is concluded that the agreement within the industrial setting is higher than obtained in a previous study within an academic setting, but it is still in need of improvement.

This leads to the conclusion that the experience within industry as well as the improved information structure in relation to the previous study aids agreement, but to reach a higher level of agreement, additional education is believed to be needed at the company.

1. Introduction

It is common for software engineering companies and projects to collect fault information, typically for fault tracking and for retrospective analyses. The typical information collected and stored for a fault contains a classification of the fault and the classification can represent several perspectives. The perspectives may represent user impact, frequency of fault occurrence, and fault type, for example functional or algorithmic fault [1].

In our experience, the main purpose of fault management systems is to track faults from detection through analysis and correction, and then ending with verification. But, the fault classification has no imminent impact or benefit to the fault tracking and correction process.

However, the fault classification contains vital information for process improvement, if used wisely. The connection between the existence of faults in the system and lapses in the process is evident, since no process , with the exception of fault seeding methods [2, 3], intentionally supports insertion of faults into the developed software.

The usage of fault information, including fault classification, for process improvement is not novel as such. However, inter-rater agreement or classifier agreement and the appropriateness of the fault classification is not that frequently discussed.

The viewpoint in this paper represents two important features of a fault classification. 1) What does the classification represent, typically, what question is answered by the classification? 2) Are the classifiers agreeing, meaning that the classification assigned to a fault is trustworthy and that it is not depending on the person classifying the fault?

If adhering to the importance of fault information and fault classification for process improvement, then the importance of these two questions is evident. If the fault classification does not focus on the right question, the process improvements based on the classification will not deliver as hoped for. Additionally, although the fault classification is aligned with the question posed, the subjective nature of the fault classification threatens the result by supplying skewed information as a basis for process improvements. This may lead to that the process improvement misses its intention.

This paper addresses the two features mentioned above by presenting a tailored fault classification for UIQ Technology (subsidiary of Symbian), hereafter referred to as the company, and also presenting an initial evaluation of the developed fault classification in experimental style in an industrial context. In summary, the objective of the paper is to present a fault classification tailored for the company, and in particular to evaluate the classification empirically.

The paper is structured as follows. Section 2 describes the background, and Section 3 discusses the research methods applied. In Section 4, the study result is presented followed by an analysis in Section 5. Section 6 addresses the conclusions and Section 7 states suggestions for future research.

2. Background

The main reason for developing a tailored fault classification came from a previous improvement study, [4]. The improvement effort showed that a number of faults have similar origin and reason for existence, and that this is not captured by the existing fault classification.

The improvement process worked and delivered result, but is typically not run on regular basis and does not cover all types of faults. It is desirable to have a continuous tracking of faults within the company's projects to identify common reasons for faults. The fault classification introduced contains information about type and time, both for insertion and detection. By the continuous classification of all faults, the base for process improvement would be broad and complete both in terms of monitored faults and in time. Continuous fault classification also makes it possible to determine the impact of a process improvement effort as well. The faults targeted by the process improvement should decrease for the coming projects, if the improvement effort is successful.

After a study, in an academic setting, investigating the approach to assure fault classification agreement [5], it became clear what roles and responsibilities within the company should perform the fault classification. Further, the study also indicated when in the process a suitable fault classification is possible.

The objective of the fault classification evaluated here is three folded, i.e. it is designed to answer three important questions:

1. When is the fault inserted?
2. What is the plausible reason for fault insertion?
3. What type of fault is it?

This means that each fault is classified according to these three questions.

The focus of the fault classification is to aid in process improvements targeting project lead-time and the overall quality of the product, which is expressed as the number of faults detected.

2.1 Related research

The work described in this paper relates to a number of areas, typically fault classification, inter-rater or classifier agreement and also process improvement based on fault data. It should be noted that this paper uses the terms inter-rater agreement and classifier agreement interchangeably.

The area of fault classifications and in particular ODC is described in [1, 6]. The connections between fault classification and process improvements are stated previously by, for example, Leszak et al. [7], also ODC has a clear goal of supporting process improvement, both during and after project completion. The trigger concept discussed by Leszak et al. [7] and also by Chillarege and Prasad [8] connects to the process aspect of fault classification.

The approach taken in this paper to determine the inter-rater agreement employs two methods for classifier agreement calculation. The Kappa statistic and percentage agreement are further described by Altman among others [9] and in Section 3.4. Besides the classifier agreement the classifiers' confidence and consistency in the classifications is also monitored. Leszak et al. [7] present a study using the Kappa statistic in conjunction with fault classification. They also perform the fault classification in retrospect in a mixed hardware and software environment, with the purpose of Root Cause Analysis and cost reduction. The agreement between raters in the paper by Leszak et al. was not investigated as described in this paper. Retrospective classification is useful within industry as a means for process improvement. Thus, it is important to evaluate the possibility to correctly classify the faults retrospectively based on the information provided by the fault description. The classification is hence not assumed to be performed by the person detecting the fault.

El Emam and Wieczorek, [10] present a study where the Kappa statistic is used and the classification are done when the fault is detected. In their study, the classifiers are fewer in number and are locating the faults themselves, which rules out the necessity of good fault descriptions. This procedure simulates the classification when the fault is detected, i.e. the classification is not dependent on the fault description. This brings it closer to the operational solution suggested in this paper, though there is a difference between identifying and correcting the faults. However, the usage of the Kappa statistic is similar as well as the focus on the classifier agreement as a necessary condition for gaining correct information for subsequent process improvement activities. The next section presents the method applied in this study.

3. Method

3.1 Design

The study consists of one group of people determining the fault classification for a number of joint faults. The sample of persons comes from within the company and is selected based on four criteria: their current role, the experience within the company, their current working area, and project participation.

The study imitates the operational setting as much as possible. To fulfill this demand, all material, excluding the fault classification, is pre-existing within the company. By performing the study within the industrial setting, the study imitates the actual process and environment for conducting the fault classification as closely as possible. This means that participants, also known as classifiers, use familiar tools, documentation, software domain and fault documentation style.

Based on a prior study, [5] experience and domain knowledge are important for achieving a correct fault classification, which motivates the usage of industry personnel and running the study within an industrial setting.

For collecting information and knowledge gained by the participants a questionnaire is used. The questionnaire addresses issues concerning the fault classification scheme used and the process for determining the correct classification.

3.1.1. Threats. If this study is viewed upon as an experiment, the main concern is that no control group or any pre- or post-tests are used. However, this study should not be compared with a typical experiment; the study presented in this paper aims at gathering a snapshot view of the agreement given the fault classification at the current time within the company. The pre-study to this evaluation took place within an academic setting and not using the same faults or system domain [5].

In the reminder of this subsection, the identified threats are discussed [11]. Further, the approach used to address them is also described.

Subject bias – does not impact this study, the participants have not a specific interest in the result. However, participants will potentially (depending on the outcome) be affected by the result, as it may influence how they are expected to work, but the impact is motivated for the participants, and is an improvement. It is also so that the decision of implementing the result of the study is not up to the participants of the study, but rather a decision for management within the company.

Observer error – this threat is judged as omissible, the classification is rather straightforward to monitor and is also collected in digital format. The classification analysis is designed to monitor the agreement between classifiers.

Construct validity – through straightforward and quantitative collection of information and accepted calculations for determining the classifier agreement, this threat is addressed. Construct validity for the fault classification is addressed in two ways. First, results and experiences from a previous study were used [4]. Secondly, the fault classification was developed jointly between the researchers and company representatives.

Instrumentation – by supplying digital preformatted submission forms and instructing the participants how to use these forms during the introduction meeting, this threat is evaluated as handled.

Regression – through leaving the sample to management within the company, the influence over this threat is out of the researchers' control, thus not being able to influence the selection and sampling process of participants. Management has been instructed that the participants must be a representative sample from the company.

Setting – by adjusting the evaluation to mimic the working environment as closely as possible, it avoids the problem of letting the setting influence the result negatively.

When discussing generalizability, the results are per se not generalizable outside the organization, depending on that this represents the agreement for the company and for the sampled faults, the used fault classification, and for the sampled participants. However, the sampling approach used, further described in Section 3.2.1, strengthens the generalizability within the organization. Moreover, the approach taken here can be used in other companies to evaluate a new fault classification before introduction.

The next section presents the planning of this study.

3.2. Planning

3.2.1. Sampling. Two parts are subject for sampling: participants and faults selected for classification. The sampling aims for a representative sample concerning the faults and developers within the company.

The participants in this study are employees of the company with sufficient experience in software development and fault correction on the company's products. The lower experience limit is set to one year of employment within the company as a developer. In addition, it was requested that the participants should have participated in the project from which the faults used in the study originated. The intention being thar the participants should have a good understanding of the project and the faults.

For the participants, the total sample size is eight persons. The sampling of the actual persons was placed upon four project managers that have the time planning responsibility and knowledge concerning the experience of their personnel. The project managers decided which persons that should participate, but the selection process was partly on a voluntary basis, i.e. no one was forced to participate. Each project manager represents a subproject, meaning that the participants are not all involved with the same subproject, and hence the participants represent a number of different areas within the project. The industry setting, in which this study is performed, is considered as an asset but also requires special considerations to availability and parallel activities.

The sampling of faults was done in a stepwise manner.

1. Extract the total number of faults from the relevant project.
2. Determine the distribution of the faults over a number of categories, e.g. component adherence.

3. Randomly select a large sample, maintaining the distribution.
4. Clean the large sample from non-correct faults. Typically rejected faults, and faults that are not being corrected.
5. Setting the sample size, in this case 30 faults.
6. Sampling 30 faults from the larger sample, maintaining the distribution.

These steps results in a sample with the same distribution as for the totality of faults, that does not contain any incorrect faults, and that originates from the selected project.

3.2.2. Instrumentation. The use of digital information and forms assures easy distribution of information as well as simple and accurate data collection. The instrumentation consists of thee parts: the information used by the classifiers for classifying the faults, the fault log collecting the classifications, and the questionnaire gathering experience from the classifiers.

For information distribution, two alternatives were used. First, digital copies of the documentation are distributed. The digital documents are extractions from the common document base connected to the monitored project. For decreasing the amount of documentation presented to the classifier, only the documents relating to the components monitored were presented. The second way for distributing the information is direct access to the document base, there was no restrictions posed for accessing the document base directly.

Further, the information contained in the fault logging system, typically the fault reports and fault tracking information, were available on line.

The fault classification log is distributed digitally, completed, and collected digitally. The classifiers are presented with a spreadsheet where each fault is linked to the fault management system, to simplify the lookup of the fault and to avoid mistakes. In the fault classification log, the correct labeling of the fault classifications is defined for easy selection when logging the classification. The classification log is designed for simple completion and guidance to avoid mistakes as well as assuring correct interpretation during analysis.

The data analysis is initiated by manually inserting the eight classifiers responses to a common spread sheet document for further calculation according to the analysis methods described in Section 3.4. Also a sanity check was performed by manual treatment of the information.

3.3. Operation

Prior to the operation of the study, planning in terms of arranging facilities and scheduling took place. Since the study is carried out in the industrial setting, and during normal working hours, considerations and concern are taken to the operation at the company.

The fault classification evaluation takes place in three steps: introduction, fault classification, and questionnaire completion.

The introduction addresses the procedure for the evaluation, presenting the material and instruments used for classifying the faults. A presentation of the fault classifications is included in the introduction as well as a demonstration of the instruments to be used by the participants. The introduction was scheduled to be held for all the participants to establish a common view of the task and what is expected of the participants, and remove unnecessary questions and mistakes. Additionally the rules that apply for the participants in this study are expressed, typically that the classification should be done individually and completed during the time frame given. During the introduction, the participants did not express uncertainty concerning the task. The introduction meeting consumed one hour.

Though the introduction was booked in advance, one participant missed the start of the introduction and arrived when the meeting ended. However, by arranging an additional session, the same information was provided to this participant as well.

The complete information and instrumentation to be used for classification was distributed via e-mail to the participants, just prior to the introduction meeting.

The classification activity takes place at the participants’ work place, providing a familiar environment and tools, i.e imitating the normal handling of a fault as much as possible, though classifying the fault on available information, and not on actually correcting the fault.

The third part of the evaluation is the completion of a questionnaire. The questionnaire was distributed to the classifiers upon submission of a completed fault log.

During the execution of study, the researcher was present at the company, and available by phone, e-mail, and face-to-face contact. Though a few contacts occurred, no major issues were raised during the evaluation. However, when the analyses started, the suspicion arose that two of the participants misunderstood how to apply the classification concerning when the fault was inserted. A dialogue with these two participants confirmed the suspicion. The misunderstanding was the reason for the deviating results. The dialogue was carried out trying not to bias the participants with the suspicion. It was agreed that it was a misunderstanding. The misunderstanding resulted in a too frequent usage of one fault classification relating to when the fault is inserted into the system.

3.4. Analysis methods

3.4.1. Percentage agreement. Measuring the percentage agreement is the basic approach without any correction for agreement by chance. However, there is no additional overcompensation if the data does not have an even spread over the classification categories [9]. The calculation for percentage agreement requires the insertion of the classifications into an *n* by *n* table, *n* representing the number of available classifications. Table 1 shows an example table.

Table 1: Example table for five classifications of two classifiers.

	Classifier n						
	Classification	*Alfa*	*Beta*	*Gamma*	*Delta*	*Epsilon*	Total
Classifier m	*Alfa*	**2**			2		4
	Beta		**2**				2
	Gamma		3	**1**		4	8
	Delta			7			7
	Epsilon				4	**6**	10
	Total:	2	5	8	6	10	**11**

Table 1 shows an illustration of fault classifications in italic, denoted with Alfa, Beta and so forth. The numbers represent the number of occasions the classifiers marked down the specific combination. For example on two occasions, classifier m and n agreed upon the classification Alfa for a fault. It is also possible to see that in four cases, classifier n classified faults as being of type Delta while classifier m regarded the faults as being of type Epsilon.

The percentage calculation takes the number of occurrences of agreement, the diagonal in Table 1, divided with the total number on classifications, the total number in the table.

3.4.2. Kappa statistics. The fault classifications of the individuals are analyzed using the Kappa statistic [9]. This type of statistics is a standard method to evaluate inter-rater reliability. In a software engineering context, it has been used in, for example, process assessment [12] and for evaluating ODC [10].

The Kappa statistics results in an agreement index, the interpretation of this agreement index is provided by Altman [9], and stated in Table 2.

For a more detailed description of Kappa statistic see the above references, [5] or [9].

Table 2: The Altman Kappa scale.

Kappa statistic	Strength of agreement
< 0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Good
0.81-1.00	Very good

All unique pairs of participants (classifiers) were analyzed. The reason is that it is important to see whether some individuals agreed while others did not. Thus, in total 28 pairs may be generated from the eight participants, i.e. n*(n-1)/2 pairs where n is the number of participants.

For using the Kappa statistics, the classifiers' results need to be structured in the way illustrated in Table 1.

The rows and columns represent the fault classification stated by each classifier.

3.4.3. Confidence. In addition to collecting the classification for each fault, the classification log also collects the confidence for each classification. The confidence ranking represents the classifiers' confidence at the time of classification, for that the assigned classification of the specific fault is correct. The confidence ranking is a five point ordinal scale with one representing the highest level of confidence and five representing the lowest level of confidence.

It is of interest to monitor the confidence of the classifiers. A low confidence would indicate a problem with either the fault classification or the fault descriptions. It is also interesting to track the confidence in relation to the agreement between classifiers. A low agreement and high confidence is a good reason for further investigations.

3.4.4. Fault distribution over classifications. Total agreement would imply that each individual fault is only allocated to one classification by all the classifiers. On the opposite, total disagreement would imply that each fault is distributed to a different class by each classifier. The number of classifiers and fault classes influences the later argument. In other words, a perceived agreement may be obtained due to having more classifiers than fault classes.

However, the monitoring of how each fault is distributed over the available fault classes by the classifiers provides input to how well the classification works. For example, if there is a majority of agreement or not. A distribution only involving one or a few classifications per fault classified is desired.

4. Result

The result illustrates the agreement based on the two methods for analysis. The analysis is done for all three perspectives of fault classification: time for insertion, reason for insertion, and type of fault.

4.1. Percentage agreement

Table 3 presents the three perspectives of the fault classification, along with the result from the percentage agreement calculation. In Table 3, corrections, according to the misunderstanding described in Section 3.3, are written in italic. The correction is only needed for the time for insertion classification. The table presents minimum, maximum, mean, and median values in percent along with required corrections for the three classification perspectives. For example the mean value for Time Insertion classification is 35% agreement. This means that the classifiers agree in 35% of the cases. For the mean value, it is noteworthy that the agreement in average is between 32-36% for all three classifications.

Table 3: Percentage agreement for classifiers for each classification perspective, values given in percent.

Percentage agreement				
Calculations	**Perspectives**	**Time Insertion**	**Reason for Insertion**	**Fault Type**
	Minimum	0	7	10
	Corrected Min	*40*		
	Maximum	87	53	57
	Corrected Max	*80*		
	Mean	35	32	36
	Corrected Mean	*56*		
	Median	43	35	37
	Corrected Median	*57*		

4.2. Kappa statistics

Table 4 presents the result achieved in terms of agreement using the Kappa statistics by the classifiers. The result is reported for the three perspectives of the fault classification. The misunderstanding described in Section 3.3 is also accounted for in Table 4. The same values are reported in Table 4 as in the previous table, although the Kappa statistics is shown here. From the table, it can for example be seen that the mean agreement using the Kappa statistics varies between 0.186 and 0.258 for the three classifications. These values can be compared with the interpretations suggested in Table 2.

Table 4: Kappa value agreement for the classifiers for each classification perspective.

Kappa statistics agreement				
Calculations	**Kappa agreement**	**Time Insertion**	**Reason for Insertion**	**Fault Type**
	Minimum	-0.017	-0.001	0.077
	Corrected Min	*0.126*		
	Maximum	0.660	0.471	0.470
	Corrected Max	*0.660*		
	Mean	0.187	0.220	0.258
	Corrected Mean	*0.308*		
	Median	0.186	0.241	0.254
	Corrected Median	*0.301*		

4.3. Confidence

Also for the confidence calculations the results are presented for each of the fault classification perspectives individually.

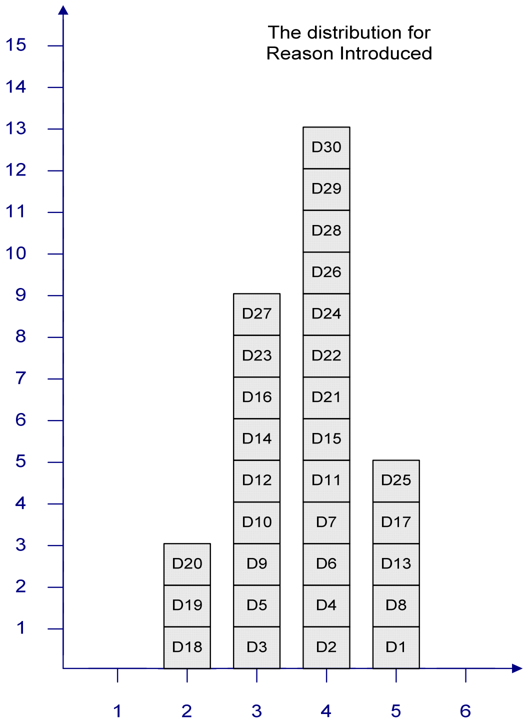
For the time when a fault is introduced the confidence calculations resulted in a mean of 1.5, with the median being 1. For the reason for fault introduction the confidence calculation resulted in a mean of 1.8, and median became 2, additionally for the fault type classification the mean is 1.7, and the median is 2. In general the classifiers are rather confident regarding their classifications.

4.4. Fault distribution over classifications.

The distribution analysis is presented in Figure 1. The numbered squares represent individual faults, and the bars on the x-axis represent the number of classifications the fault is placed in. For example, D30 is attributed to four different reasons for fault introduction by the classifiers.

Figure 1 describes the distribution of the Reason Introduced classification. In Figure 1, it is evident that the distribution is not as desired, it is clear that the majority of faults is distributed over three and four classes. The situation is similar for the type of fault and time for introduction though not as severe and not illustrated.

Figure 1: Distribution for Reason Introduced classification.



4.5. Questionnaire result

In this study the questionnaire provided additional information concerning the classification and the opinions of the classifiers.

The focus of the questionnaire is threefold, 1) the information usage and familiarity, 2) what is the key information needed to provide a correct fault classification, and 3) opinions about the fault classification scheme as such.

The result indicates that most of the classifiers were satisfied with the provided material, also that they were familiar with the material used. The most valuable material for determining the correct fault classification is the fault descriptions (ranking 1), accompanied with the source code listings and specifications. The familiarity ranking shows, that the classifiers are familiar with the material. They are about equally familiar with the fault descriptions, source code listings, and specifications. In an operational situation when the faults are classified, the fault descriptions are not completed to begin with, since the recommendation is to have the person correcting the fault classifying the fault. However, the information and knowledge existing in the fault descriptions are what the developer classifying the fault gains by correcting the fault.

The classifiers were asked to state in which situations they were most likely to correctly classify a set of faults, with several scenarios as possible replies. The highest ranked answers are when they were taking part in the correction activity of the fault. This was strengthened by the claim that most of the classifiers do not think it is possible to correctly classify faults purely based on the textual descriptions, additional information is needed. Moreover, experience and detailed knowledge gained through correcting the fault is also valued high.

When discussing how to improve the understanding of the classifications, two issues are highlighted. First the issue of example faults is mentioned, and secondly a consensus discussion among classifiers would be beneficial. The example faults should help in both education and operation of the classification. The consensus discussion typically focuses on educational activities, when a number of classifiers are gathered.

The classifiers were also asked about three attributes regarding the fault classification scheme. The three attributes, and their definitions, are:

- Usability – Meaning the usefulness of the fault classification in relation to the faults being classified..
- Comprehension – Meaning how understandable the fault classification scheme is.
- Applicability – Meaning how easy the classification is to apply during every day work.

These attributes were assigned ranks on a five point ordinal scale, where 1 represents Good and 5 represents Bad. On average, all ranked a 2, implying satisfaction among the classifiers when it comes to using and handling the fault classification scheme.

5. Analysis

5.1. Percentage agreement

The percentage agreement and Kappa statistics are not directly comparable. Thus, the comparison is based on a value-based assessment using Kappa's strength of agreement levels stated by Altman [9]. The percentage intervals takes the same values as for the Kappa statistics as presented in Table 2, though adjusted to percentage values.

The percentage agreement indicates that two of the perspectives reached the level of Fair agreement, namely Reason for insertion, and Fault type, and Time insertion reached Moderate agreement. This is better that the results when using the Kappa statistics, which is discussed in the next subsection. However, this can be ascribed to the simplicity of the method and that there is no correction for chance agreement. The percentage agreement is likely to render in a too positive result.

5.2. Kappa statistics

For determining the strength of the agreement based on the kappa value, the levels stated by Altman is used for determining the level of agreement [9]. Table 2 presents the interpretation of agreement for the Kappa value.

For the Kappa calculation, presented in Table 4, all perspectives reached Fair agreement. Though percentage agreement delivers higher agreement, it is evident that the three perspectives vary in the same way. The strongest agreement is reached for the Time insertion classification.

It can be concluded by this evaluation that there is too low agreement to trust the classification, higher agreement is required to continue with subsequent activities, e.g. process improvement based on common faults. However, Kappa statistics corrects the chance agreement; so it might punish the overall agreement if the sample is not evenly distributed among the available fault classifications. The assumption is that in general a skewed distribution exists. In other words, that some classes are more common than others, which might influence the result of the Kappa statistics agreement negatively.

5.3. Questionnaire analysis

When analyzing the questionnaire, it shows that the most appreciated system, when classifying the faults, is the defect management system. This is however not surprising since the defect management system gathers the fault descriptions and information from testers and other checkpoints regarding the fault. Additionally, on occasions the corrected source code is copied into the fault description, which provides the classifiers with the full picture. The information available in the fault descriptions is partly produced by the developer during fault correction. Thus, in an operating situation the information in the fault description is produced by the developer, who is also classifying the fault.

5.4. Reflections

Based on the sampling of the classifiers, and through discussions with the classifiers in a debriefing meeting, it is possible to determine for which components that the classifiers possess the most experience.

To investigate impact of experience two steps are taken. First, identify the pairs based on the experience of the same component. Secondly, identify those faults that originate from the specific component.

The group of classifiers does not have experience in all of the components from which the faults originate. For that reason, the analysis focused on those components and classifiers where experience is determined.

The calculations did not indicate any increased agreement based on the experience of the classifiers, as was initially assumed, the reason for this is a subject for further studies.

5.5. Analysis summary

The prior study is aligned with this study, i.e. [5], both studies indicate the importance of domain knowledge, familiarity with the product and, sufficient fault descriptions. However, the study presented here addresses these three issues, and achieves higher agreement, but still it is not sufficient.

The questionnaire indicates the need for education, typically examples of classifications and consensus discussions. Additionally, the questionnaire confirmed the thought that the best situation for correctly classifying a fault is when the fault is corrected; during correction the information needed for a correct classification is gathered.

6. Conclusions

The conclusions from this study are that though the classifiers have experience within the domain, and that the fault descriptions and surrounding information is available and sufficient, the agreement among the classifiers is low. The low agreement and the responses to the questionnaire lead to the conclusion that other issues influence the agreement. Indications strongly suggest that a more in-depth education is requested for providing a sufficient agreement among classifiers. Further, requested parts of this education are consensus discussions and example of correctly classified faults.

The study supports the conclusions that the most suitable phase for correct fault classification is when the fault is corrected, and that the classification should be assigned by the person correcting the fault. During correction of the fault, the developer classifying the fault gains the information needed for correctly classify the fault.

An initial assumption would be that detailed experience of a common component would strengthen agreement between two classifiers. However, the analysis from this study does not support that statement. A complicating factor is that there were only a small number of faults included in the analysis.

Based on the feedback through the questionnaire, it is shown that the fault classification scheme as such is not the major problem for reaching agreement. The rather high ranking on usability, applicability, and comprehension supports the conclusion; in addition, the classification scheme was discussed at a debriefing meeting, supporting this conclusion.

During the work with this study it became evident that it is hard to find a recommendation of what agreement level is feasible to expect, when monitoring fault classification agreement. It is probably not possible to reach full agreement or the level very good, according to Altman scale [9], presented in Table 2. However, the question is what level is to be expected? It is impossible to tell if more education or other efforts would have been sufficient to obtain higher agreement among classifiers..

7. Future research

The result from this study is intend to be used as the pre-study for further studies. The continuation intends to monitor the education effect, and then the result from this evaluation is used as a pre-test.

Additional updates to the fault classification scheme are also likely to occur based on the feedback and opinions from the classifiers participating in this study, though the changes would be minor.

The intention is to continue with monitoring classifier agreement after adding requested and feasible education to the group of classifiers. A continuation contains adding education and then re-evaluating the agreement in the same setting as within the study presented in this paper and thus monitoring the education impact.

A natural continuation, if the agreement after education is judged as sufficient, is to supervise the education and implementation of the fault classification process in an operating environment as a case study. Depending on the outcome of the case study, broad training and implementation would take place and it is also of interest to monitor the effect more long term on the impact of a tailored fault classification for providing the organization with background information for process improvements.

8. References

[1] R. Chillarege, I. S. Bhandari, J. K. Chaar, M. J. Halliday, D. S. Moebus, B. K. Ray, and M.-Y. Wong, "Orthogonal defect classification - a concept for in-process measurements," IEEE Transactions on Software Engineering, vol. 18, pp. 943-956, 1992.

[2] J. J. Marciniak, "Encyclopedia of Software Engineering," vol. 2003: Wiley, 2003.

[3] M. J. Harrold, A. J. Offutt, and K. Tewary, "An approach to fault modeling and fault seeding using the program dependence graph," The Journal of Systems and Software, vol. 36, pp. 273-295, 1997.

[4] K. Henningsson, T. Birath, and P. Molin, "A Fault-Driven Lightweight Process Improvement Approach," Proceedings Euromicro 2003, Belec Turkey, pp. 343-350, 2003.

[5] K. Henningsson and C. Wohlin, "Assuring Fault Classification Agreement - An Empirical Evaluation," Proceedings 2004 International Symposium on Empirical Software Engineering, Redondo Beach, California USA, pp. 95-104, 2004.

[6] I. Bhandari, M. Halliday, E. Tarver, D. Brown, J. Chaar, and R. Chillarege, "A case study of software process improvement during development," IEEE Transactions on Software Engineering, vol. 19, pp. 1157-1170, 1993.

[7] M. Leszak, D. E. Perry, and D. Stoll, "Classification and evaluation of defects in a project retrospective," Journal of Systems and Software, vol. 61, pp. 173-187, 2002.

[8] R. Chillarege and K. Ram Prasad, "Test and development process retrospective - a case study using ODC triggers," Proceedings International Conference on Dependable Systems and Networks, pp. 669-678, 2002.

[9] D. G. Altman, Practical statistics for medical research. London: Chapman and Hall, 1991.

[10] K. El Emam and I. Wieczorek, "The repeatability of code defect classifications," Proceedings Ninth International Symposium on Software Reliability Engineering, pp. 322-33, 1998.

[11] C. Robson, Real World Research. Cornwall: Blackwell Publishing, 1993.

[12] K. El Emam, "Benchmarking Kappa: Interrater Agreement in Software Process Assessments," Empirical Software Engineering, vol. 4, pp. 113-133, 1999.