K. Henningsson and C. Wohlin, "Assuring Fault Classification Agreement - An Empirical Evaluation", IEEE Conference Proceedings International Symposium on Empirical Software Engineering, pp. 95-104, Redondo Beach, California, USA, 2004.
Distinguished paper award.

Assuring Fault Classification Agreement – An Empirical Evaluation

Kennet Henningsson
Blekinge Institute of Technology
PO Box 520
SE-375 25 RONNEBY
Kennet.Henningsson@bth.se

Claes Wohlin
Blekinge Institute of Technology
PO Box 520
SE-375 25 RONNEBY
Claes.Wohlin@bth.se

Abstract

Inter-rater agreement is a well-known challenge and is a key issue when discussing fault classification. Fault classification is, by nature, a subjective task since it highly depends on the people performing the classification. Measures are required to hinder the subjective nature of fault classification to propagate through the fault classification process and onto subsequent activities using the classified faults, for example process improvement. One approach to prevent the subjective nature of fault classification is to use multiple raters and measure inter-rater agreement.

In this paper, we evaluate the possibility to have an independent group of people classifying faults. The objective is to evaluate whether such a group could be used in a process improvement initiative. An empirical study is conducted with eight persons classifying 30 faults independently. The study concludes that the provided material were unsatisfactory to obtain inter-rater agreement.

Key words: Fault classification, empirical evaluation, classifier agreement, orthogonal defect classification, Serial studies.

1. Introduction

The fault classification process is important, since actions such as fault correction and process improvement depends on the classification outcome. A well-documented fault classification is the Orthogonal Defect Classification (ODC) scheme [Chillerage92], which then may form the basis for process improvement.

If the fault classification is done subjectively from only one person's perspective, it is not possible to assure that the fault classification is correct; however this is the normal scenario. An incorrect classification is likely to affect the subsequent process improvement activity negatively and there is a risk that erroneous decisions are taken based on the information.

Moreover, each fault is often classified from several different perspectives. It may include the type of fault and the severity of a fault. Whatever classification is used correctness is an important issue.

By this reasoning, the need of having multiple individuals classifying the faults and assuring agreement of the classification is evident.

The question addressed in this paper is whether it is possible for a separate group of software engineers, i.e. separate from the developers, to correctly classify a number of faults based on only fault descriptions.

The question is empirically evaluated through a study, including determining the Kappa statistic to evaluate the agreement between different subjects. The empirical study is discussed in detail in Section 3.

The next subsection discusses the context of the study, and how the study presented in this paper fits into a larger context.

1.1 Context

This study is the first in a series with the intention to build empirical knowledge concerning fault classification and assure the correct classification of faults for an industry partner. The series consists of four steps, as shown in Figure 1

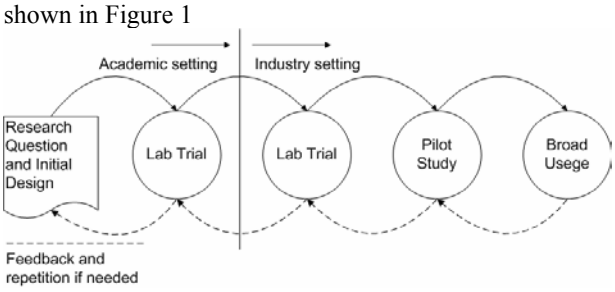

Figure 1: Study series assuring a correct fault classification.

The first step answers the following research question: *"Is it possible for a separate group to correctly classify a set of faults given the fault description, where correctly means that each individual classify faults in the same way?"*

All previous steps influence the consecutive steps in Figure 1. The long-term objective is to have a fault classification on which a fault-driven process improvement approach can be applied. The second study, to be performed in an industry setting, intends to incorporate the findings from the first study. Study number three intends to take the form of a pilot project assuring conformance in a real project. The final step means implementation in the organization and broad usage. This approach is similar to the approach described by Linkman and Rombach [Linkman97].

1.2 Research focus

This paper address the research question, stated in the previous section, empirically by evaluating if a set of classifiers are assigning the right classification and thus agreeing. The evaluation is a formative evaluation with the overall intention to form the process of correctly classifying faults [Robson00].

In the same time as addressing the research question additional empirical knowledge is gathered through a questionnaire aiming at two areas:

- After completing the classification, the classifiers are asked to answer questions related to the classification. The questions are concerned with: type of information needed to increase the likelihood of correctly classifying faults, issues related to the environment and classification of faults in retrospect, and appreciation and other issues related to ODC.
- What are the recommended improvements of the study design and preparation prior to the industry evaluation?

The remainder of this paper is organized as follows. Section 2 presents related work and Section 3 presents the method used in the study. In Section 4, the outcome of the study is presented, and in Section 5, the analysis of the results is discussed. Section 6 presents the conclusions, and Section 7 describes further work.

2. Related Work

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

The area of fault classifications and in particular ODC is described in [Bhandair93, Chillarege92]. Some of the work covering fault classification and ODC makes use of the process related aspects of fault classification, i.e. the trigger concept within ODC. Leszak et al. [Leszak02] and Chillarege and Prasad [Chillarege02] describe this further.

The approach taken in this paper to determine the inter-rater agreement employs Kappa statistic, in addition to classifiers' confidence and coherency in the classifications. Leszak et al. [Leszak02] present a study using the Kappa statistic in conjunction with fault classification. Leszak et al. also perform the fault classification in retrospect in a mixed hardware and software environment, with the purpose of Root Cause Analysis and cost reduction. However, the agreement between raters was not investigated as described in this paper. Since retrospective classification is present within industry, it is important to evaluate the possibility to correctly classify the faults in retrospective based on the information provided by the fault description. This is particularly important if the fault data is further used as basis for process improvement.

El Emam and Wieczorek, [ElEmam98] present a study where the Kappa statistic is used, but the classification is 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. 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 evaluation method.

3. Method

This section presents three main aspects of the research method: design, operation, and analysis.

3.1 Design

The study design consists of four components, 1) Persons participating in the classification, i.e. classifiers (also denoted subjects) 2) Software faults and their characteristics, including fault description, 3) Fault classification scheme, and 4) Questionnaire with the purpose to extract information from the persons classifying the faults.

These four components are input to the classification study. The outcome is a classification from each classifier for all faults and completed questionnaires, illustrated in Figure 2 below.

The components of the evaluation are further described in the subsequent sections.

The study is run in an academic setting, with participants not currently working in industry.

The study is executed in four steps. Steps 1 through 3 are carried out in sequence when all participants were gathered, except one. The main author then completed the fourth step:

1. Introduction, the researcher presents the process of the study.
2. The subjects are presented with the material: fault descriptions, classification scheme, and logging forms. Then the classification is initiated. The subjects perform the classification individually.
3. When the completed classification log is handed in, the subject receives the questionnaire, which is completed in direct conjunction to the classification activity.
4. The fourth step analyses the result from both the classification of faults and the questionnaire.

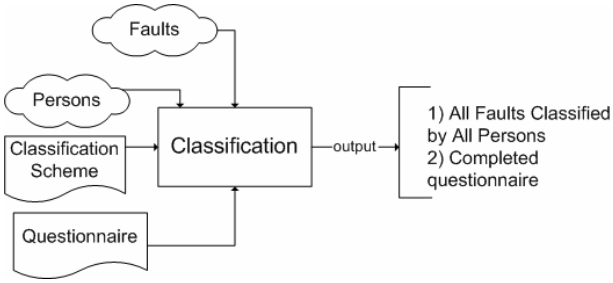

Figure 2: Design overview and output.

The four components of the study design are next described in further detail.

3.1.1 Fault/Faults Description. The faults selected originate from a graduate course teaching the Personal Software Process (PSP) [Humphrey95], i.e. fourth year students. A positive aspect is that the programs within the course are rather general and hence the fault descriptions are, most likely, understandable to the participants in the study.

The total number of faults in the database from the course is several hundreds. A first sample of faults was randomly selected consisting of 131 faults. From this sample the distribution of fault types was determined. However, 131 faults are impossible to handle within the evaluation and hence a final sample of 30 faults was randomly selected. The distribution of fault types was kept using blocking, i.e. faults were selected randomly from the different fault types keeping the proportions of fault types from the original sample of 131 faults.

Each fault is presented using a unique identifier, (consecutive numbers) and a verbal fault description. The developer, participating in the PSP course originally, wrote the fault description, i.e. the original descriptions are kept to mimic that different persons write the fault descriptions, as is the case in industry. It should be noted that the fault descriptions are written in the spirit of the PSP, i.e. they are primarily written to log your own performance. This may be a limitation in the sense that the fault descriptions are written for personal use. However, it is assumed that this situation is not that different from the situation in industry, where descriptions most often are written to handle fault corrections and not for subsequent fault classification as a basis for process improvement.

3.1.2. Fault classification scheme. Given that the fault classification at the company is likely to be influenced by the Orthogonal Defect Classification (ODC), it is natural to use ODC in the evaluation. ODC is selected by both due to availability in literature [Bhandair93, Chillarege92, Chillarege02] and by usage within PSP [Humphrey95]. Moreover, ODC has shown to work well when evaluated for inter-rater agreement [ElEmam98]. Other fault classifications would have been possible to use, for example the IEEE Standard Classification for Software Anomalies, IEEE Std. 1044-1993. However, the choice fell on ODC.

ODC is described for the classifiers as in Table 1, showing a short id, the fault class name and its description.

3.1.3 Classifiers and Education. The participants in this study, acting as classifier, are sampled from the department of the authors. In total eight participants conducted the study. Eight participants provide the possibility to build groups of sizes between two and eight.

Participation in the study was on voluntary basis. The collegiality might be a liability and threat against internal validity of the findings. However, in this situation this is judged as omissible, based on that the participants have no stake in the result. It should also be noted that the subjects are unable to aim for a good result, since we are primarily interested in the agreement between subjects. Moreover, the participants all have at least a Master's degree in a relevant subject such as computer science or software engineering. In addition, a majority of the subjects has industrial experience.

In addition, information and the perception of the participants' experience after completing the fault classification were collected by having them completing a questionnaire, as described in the following section.

3.1.4. Questionnaire. The questionnaire gathers qualitative information from the classifiers in three major areas: Information and roles, Setting and involvement, and Experiences and apprehension of ODC. These three areas and questions related to them are described as follows.

- Information and roles: Based on the classification, what information do you need to do an accurate fault

classification? Additionally, what typical role within a project would be suited to perform an accurate fault classification?

Table 1: ODC description as distributed to classifiers.

ID	Defect Classification	Description
FU	Function	A function defect is one that affects significant capability, end-user features, product application programming interface (API), interface with hardware architecture, or global structure(s). It would require a formal design change.
AS	Assignment	Conversely, an assignment defect indicates a few lines of code, such as the initialization of control blocks or data structure.
IN	Interface	Corresponds to defects in interacting with other components, modules, device drivers via macros, call statements, control blocks, or parameter lists.
CH	Checking	Addresses program logic that has failed to properly validate data and values before they are used, loop conditions, etc.
TS	Timing/ Serialization	Timing/serialization defects are those that are corrected by improved management of shared and real-time resources.
BPM	Build/Package/ Merge	These terms describe defects that occur due to mistakes in library system, management of changes, or version control.
DO	Documentation	Defects can affect both publications and maintenance reports
AL	Algorithm	Defects include efficiency or correctness problems that affect the task and can be fixed by (re)-implementing an algorithm or local data structure without the need for requesting a design change.

- Setting and involvement: In what setting could the most accurate fault classification be performed, meaning what information and persons need to be present for facilitating accurate fault classification? Also if it is judged possible to classify faults found and logged by other persons, and if so what would be needed to do that?
- Experiences and apprehension of ODC: The third area was related to prior experiences of fault classification in general and ODC in particular. In addition, the ease of using ODC and the provided information were touched upon.

The questions were answered binary (yes/no), or ranked, and also complemented with additional comments and own suggestions. The data is primarily on an ordinal scale, and hence it limits the opportunities for statistical analyses.

3.2. Operation

3.2.1. Preparation. The preparation for the participants was minimal, i.e. only limited to allocating time for the joint meeting when classifying the faults.

The preparation for the researchers included printing material, allocating, and booking time with participants, arranging facilities, and other practical issues. Since the classifiers have different nationalities, and some are unable to read the descriptions in Swedish, the material was translated into English. The English version was not given to all participants, with the motivation to keep the original fault descriptions as far as possible. Only one subject used the English version.

In addition, the researcher prepared a short presentation setting the focus and frame for the classification meeting and explaining the procedure for the participants as well as answering any questions.

3.2.2. Instrumentation. The fault classification activity consists of two parts, first classifying the faults, and secondly completing the questionnaire. Initially, the researcher shortly introduced the process. The classification took place for all of the participants, except one, in the same room at the same time. This ensures that the participants are uninterrupted and focused on the task, this might not always be the case if the classification were done in their own working environment. The study objective was to see if different people can agree on a classification. Thus, we have tried to avoid confounding factors such as interruptions during the classification. Interruptions would very well happen in a real situation. However, in this study the research question was concerned with if agreement on classification is possible or not.

During the fault classification, the researcher is present, and available for questions, however, this option was not used. During the first part of classification of faults, two mandatory and one optional type of information were collected. The mandatory information is fault class (stated with the IDs given in Table 1) and the confidence of the classifier with respect to the classification. The confidence was assigned a value between 1 and 5, where a value of 1 means least confident and 5 most confident. The optional information was gathered as comments. The classifiers had the opportunity to comment on the classification of each fault. This information was filled in manually on paper and the result is not anonymous. The name of the classifiers should be stated on the defect classification log, making it possible to return to the classifier, if needed, for clarifying the answer.

After completing the fault classification, the second part took place, i.e. the questionnaire was handed out for completion. The questionnaire was completed in the same room and directly after the fault classification. There was still the possibility to ask the researcher questions.

The data gathered are qualitative in nature. However, it is quantified through the questionnaire, using ordinal scales. This eases the analysis of the data. The data analysis used is briefly introduced in the following section.

3.3. Analysis method

The fault classifications of the individuals are analyzed using the Kappa statistic [Altman91]. 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 [ElEmam99] and for evaluating ODC [ElEmam98]. It has also been used to evaluate the goodness of a model in comparison with the actual outcome as discussed in for example [Wohlin00].

The agreement in terms of classification can be measured by an agreement index, often referred to as Kappa statistic [Altman91]. Briefly, the Kappa statistic can be explained as follows for the simple case with two raters (or classifiers) and two fault classifications (A or B). Table 2 illustrates this. The cells state the proportions of the faults with a given rating according to faults of Type A and Type B. For example, $p_{11} = 0.20$ means that 20% of the faults are considered to be of Type A by both classifiers. The columns and rows are summarized (last column and last row respectively in Table 2), which is indicated with p_{01}, p_{02} p_{10} and p_{20}.

Table 2: A diffusion matrix for fault classification.

		Rater A		
		Type A	Type B	Sum
Rater B	Type A	p_{11}	p_{12}	p_{10}
	Type B	p_{21}	p_{22}	p_{20}
	Sum	p_{01}	p_{02}	

The entries in Table 2 are used to derive an agreement index. Let P_A be the proportion in which there is agreement. Then, P_A becomes

$$P_A = \sum_{i=1}^{2} p_{ii}$$

This agreement includes cases in which the agreement is obtained by chance. To remove the effect of chance behavior, the extent of agreement that is expected by chance is defined as

$$P_E = \sum_{i=1}^{2} p_{i0} \times p_{0i}$$

The agreement index is then defined as

$$\kappa = \frac{P_A - P_E}{1 - P_E}$$

To be able to understand the degree of agreement, the Kappa statistic is usually mapped into a rank order scale describing the strength of agreement. Several such scales exist, although they are by and large minor variations of each other. Three scales are presented in [ElEmam99]. Here the scale suggested by Altman [Altman91] is used. It is shown in Table 3.

Table 3: 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. An eight by eight matrix is created for each pair as shown in Table 4.

The rows and columns represent the fault classification stated by each classifier. When the classifiers agree, i.e.

they have assigned the same fault class to a fault, then the value of the diagonal increases with one. An example is provided in Table 4 for classifiers 4 and 7.

The Kappa statistic provides information about inter-rater agreement between two classifiers. However, the Kappa statistic does not provide information about the most frequent or commonly selected fault classification.

Table 4: An example of an eight by eight matrix presenting the classifications of two classifiers.

Classifier 7

Classifier 4	FU	AS	IN	CH	TS	BPM	DO	AL	*Tot*
FU									*0*
AS	1	**8**		1			1		*11*
IN			**4**						*4*
CH		3		**3**					*6*
TS									*0*
BPM									*0*
DO									*0*
AL		3		1			1	**4**	*9*
Tot	*1*	*14*	*4*	*5*	*0*	*0*	*2*	*4*	**19**

Thus, the Kappa analysis is complemented with frequency analysis of different fault types to see whether the classifiers use some fault types in particular. The classifications by the subjects are also compared with the initial classification given by the person describing and classifying the fault originally.

In addition, the confidence values given by the classifiers are analyzed using descriptive statistics such as median, minimum, and maximum values. Further, to identify if some specific faults were easier to agree upon, i.e. one fault type is dominant for a specific fault, histograms are used to visually display the relations.

4. Result

4.1. Kappa Statistic

The Kappa value for each pair of classifiers is presented in Table 5. The Kappa analysis indicates low and inadequate agreements in general according to the interpretation in Table 3. However, no negative values are achieved; there is only one pair that agrees moderately and none with good or very good agreement, i.e. achieves Kappa values above or equal to 0.41. The average Kappa value is 0.16, which is considered as a very low overall agreement, poor according to Table 3. By having eight participants, group sizes between two and eight are possible to create. However, since the smallest group, two, does not agree, according to Kappa calculations, larger groups were not further investigated.

Table 5: Kappa values for the 28 pairs of classifiers, the highest value is marked in bold.

Pair	***K***	***Pair***	***K***	***Pair***	***K***	***Pair***	***K***
1-2	0.10	2-3	0.07	3-5	0.21	4-8	0.14
1-3	0.22	2-4	0.15	3-6	0.11	5-6	0.21
1-4	0.31	2-5	0.01	3-7	0.17	5-7	0.25
1-5	0.05	2-6	0.07	3-8	0.08	5-8	0.04
1-6	0.11	2-7	0.03	4-5	0.25	6-7	0.04
1-7	0.31	2-8	0-30	4-6	0.12	6-8	0.16
1-8	0.11	3-4	0.21	**4-7**	**0.50**	7-8	0.15

4.2. Confidence Degree

As a measure of the confidence in the fault classification, the classifiers were asked to grade their confidence on a scale from 1 to 5. The confidence degree reflects how strongly the classifier believes in the fault class assigned to the fault. High values indicate strong confidence and vice versa. The median confidence value is 4 (average 3.51). This indicates reasonable confidence by the classifiers. This is rather surprising given the low agreement represented by the Kappa statistic.

By analyzing the relations between confidence value and the coherency of classes' three interesting classes were found. Coherency refers to how many classifiers agreed regarding the fault classification, the coherency is further discussed in Section 4.3. However, three sets of faults are of interest in this discussion, namely:

- Faults with high confidence and high coherency, HH.
- Faults with high confidence and low coherency, HL.
- Faults with low confidence and high coherency, LH.

The fourth possible set, containing faults with low confidence and low coherency is not of interest, since it represents faults that the classifiers do not agree upon and do not feel confident in classifying. The three sets are presented in Section 4.3 after illustrating the coherency.

The set of faults with the highest confidence, confidence values on or above the median, is: CON = {1, 4, 5, 6, 7, 10, 11, 12, 14, 16, 18, 19, 21, 23, 24, 25, 26}.

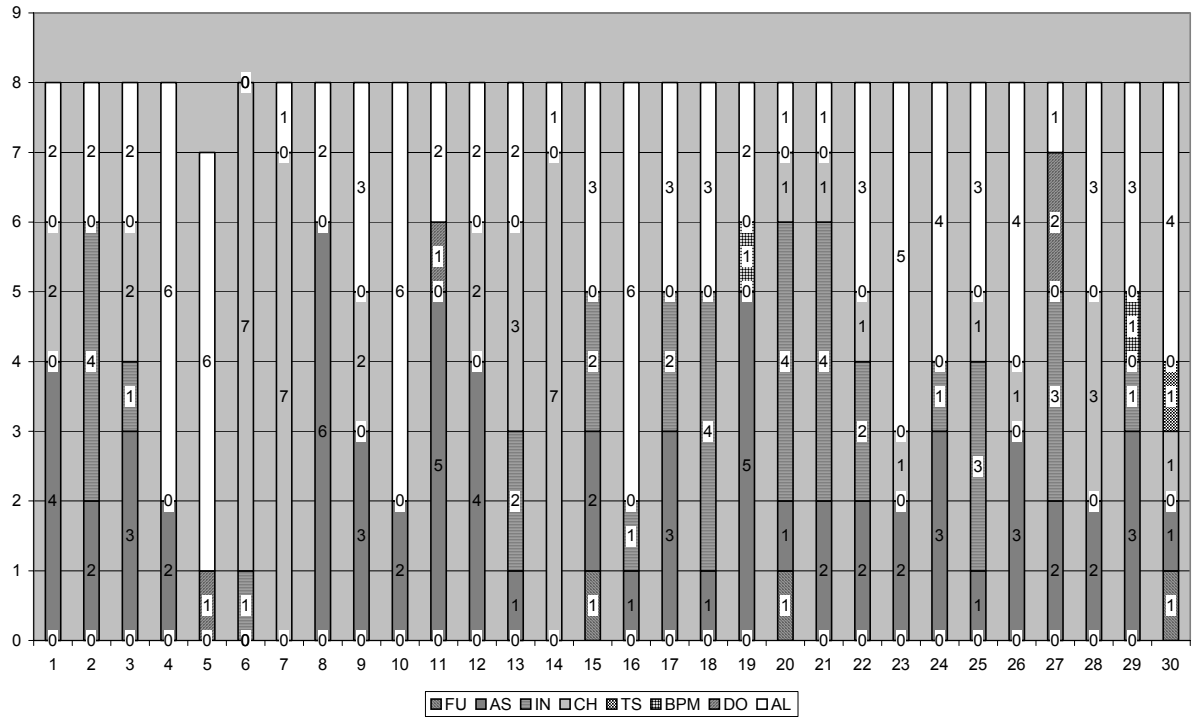

Figure 3: Visualization of fault classifications.

4.3. Histogram

A histogram is used to visualize the coherency and point out the faults where there is most and least agreement.

In Figure 3 the histogram is shown, the X-axis represent the faults numbered, on the Y-axis the number of different fault classifications is assigned. For example, if looking at fault number 4, six of the classifiers agree that the fault class should be CH (Checking), on the other hand, if looking at fault number 3, there are four different opinions of fault classification, and only three of the classifiers agree. The limit for creating the set of faults with coherent classification is set to six agreeing classifiers, representing 75% of the classifiers. This creates the following set: COH = {4, 5, 6, 7, 8, 10, 14, 16}.

Now returning to the three sets, HH (High confidence and High coherency), HL (High confidence and Low coherency), and LH, (Low confidence and High coherency), the following is obtained:

- HH is the intersection between CON and COH, which results in: HH = {4, 5, 6, 7, 10, 14, 16}
- HL is the intersection between CON and the complement of COH, which results in: HL = {1, 11, 12, 18, 19, 21, 23, 24, 25, 26}.
- LH is the intersection between the complement of CON and COH, which results in: LH = {8}.

The results in Section 4 can be summarized as that the agreement in the classification was much lower than expected when comparing with, for example, the study by El Emam and Wieczorek [ElEmam98].

When comparing the confidence with the coherency, the figures are not that high, only every fourth fault has both high confidence and high coherency. To compare the classifications in the study with the original classification by the developer, the seven faults in HH were used. Here, the classifiers agree the most and they are also most confident. However, when comparing the majority classification from the study with the classification of the fault originator it shows that only one out of seven faults has the same classification. This means that even if the classifiers in this study agree, they still disagree with the original classifier, i.e. the developer.

The diagram and the findings here indicate that the agreement between classifiers is less than hoped for. To increase the understanding of the differences, some more in-depth analysis is presented in the following section.

5. Analysis

5.1. Introduction

The empirical study is focused on two variables: the assigned fault classification and the confidence degree. These are further analyzed in this section.

By analyzing the gathered data, it is obvious that the agreement between classifiers is low; only one pair

achieves a moderate Kappa value. However, the classifiers are still confident in their decisions with a median confidence value of 4 and an average of 3.51, where a score of 5 represents the highest confidence and 1 represents the lowest confidence. Further, the histogram in Figure 3 shows that for a number of faults, 1 out of 3, a majority of classifiers agrees, meaning that at least five classifiers assign the same fault classification.

As described in Figure 2, there are three parts in the fault classification: Fault/Fault Description, Classification Scheme, and Classifiers and Education. In this case, the inadequate agreement probably depends on the Fault/Fault Description and/or Classifiers and Education. The fault classification scheme, ODC, is not indicated as the reason for the low agreement, mainly due to previous successful studies and research indicating the applicability and repeatability of ODC as described in [ElEmam98, Chillarege92].

5.2. Analysis of Low Agreement

During the analysis, it became apparent that the agreement between classifiers was very low. The reason for the low agreement is further dissected in the following subsections.

5.2.1. Fault/Fault Description. This analysis investigates the characteristics of the fault descriptions with the objective to identify the reason for higher agreement for some particular faults. The set of faults with the highest agreement among classifiers was shown in Section 4.3.

The basis for the analysis is the two characteristics that are available and concrete: fault class assigned and the length of the fault description. The fault class with the highest accuracy is AL, i.e. it is most frequently assigned in this comparison. However, no conclusions can be drawn since it only differs with one from the CH fault category.

The length of the fault description is classified into two classes: minimum and others. Minimum implies that the fault description is just containing one or two lines of text with no expanding information. It is minimal in the sense that the description is written only for correction purposes (primarily for correction by the developer herself/himself). The other category includes all fault descriptions with more than a couple of lines of descriptions. These typically provide a little more information about the faults and they ought to be better suited for fault classifications in retrospect.

However, this division does not give any clarification, there is a weak indication that the more extensive fault descriptions provide higher confidence, but it does not result in higher coherency between classifiers. Thus, it is concluded that it is not sufficient to measure the length of the fault description to understand why certain faults are easier to classify or at least agree upon the classification.

Finally, the average time for fault classification is worth highlighting. The average time was 1.5 minutes per fault.

5.2.2. Classification Education. The next issue influencing the agreement of fault classification is the fault classification education.

As earlier described, the fault classification education in this case was minimal. Naturally, this affects the result.

The belief is that missing education and experience would, for example, lead to interchanging one or more fault classes. There are patterns reoccurring in the eight by eight tables that indicate frequent interchange of fault classifications that additional education may prevent.

When analyzing the interchanges between fault classifications in the eight by eight tables, it stands clear that the most common interchange is between fault classes AL and AS. A calculation is performed by summing up the cells in all 28 tables for the unique pairs of fault classes also summing up to 28 pairs, since the number of fault classes is eight. The AL-AS pair equals AS-AL in this perspective.

The five interchanges with the highest sums, according to the calculations are presented in Table 6.

Table 6: Confusions between fault classifications.

Confused fault classes	***Occasions***
AS – AL (Assignment and Algorithm)	184
IN – AL (Interface and Algorithm)	77
CH – AL (Checking and Algorithm)	68
AS – IN (Assignment and Interface)	59
AS – CH (Assignment and Checking)	49

The other common interchanges between fault classes have less frequency, typically from rare occasions to up to 30. The most frequent mix-up is the interchange between AS and AL, in total, 22 of the 28 pairs mixed these two classes up a number of times.

To avoid the interchange between different fault classes, two things can be improved, i.e. more education in the used fault classification scheme, and improved fault descriptions. It is judged hard to estimate the effect of improved fault descriptions based on the information available from this study. However, it is possible to study the potential educational effect by combining the AS and AL columns and rows, and hence creating seven by seven tables instead. Next the Kappa values are recalculated for these seven by seven tables. However, this only resulted in one more Kappa value above 0.41, i.e. 0.53, and increased the average Kappa value from 0.16 to 0.24, which was not a drastic change. According to the scale by Altman [Altman91] presented in Table 3 the average agreement changed from Poor to Fair. This indicates that

in this particular setting, education alone is not the explanation to the low Kappa values.

By combining AL and AS, all interchanges between these classes are removed, which would not be achieved by education alone. To find the solution of how to improve the agreement it is necessary to analyze the result of the questionnaire, and analyze the answers from the participants about what they believe is required in terms of information and setting for assuring an accurate fault classification. In the next section, the responses to the questionnaire are examined.

5.3. Questionnaire

As mentioned in Section 3.1.4 the questionnaire is divided into three main parts: Information and Roles, Setting and Involvement and finally Experience and Apprehension. This section distills the answers from the respondents (previously denoted subjects or classifiers – based on their role at the time) and presents the information in the structure of the three parts.

All except one respondent thought that the information was insufficient for making correct fault classifications. The next question to the respondents was what information would be helpful. The respondents were asked to rank the suggested additional information, and if needed add their own suggestions. The ranking values possible are between 1 and 9, where 1 is the highest rank. In Table 7, the alternatives and their median ranking is shown. The question was: "If you were missing any information, which information of the following would help you the most?"

As Table 7 shows, the most desired information, for making a correct fault classification, is source code, change information, and conversation with the developer.

Further questioning sought for comments on what role would be suitable for doing fault classification. No coherent view was given, but, according to the respondents, the classifier is desired to have fair knowledge about the system on both a general level and a specific level.

Questions were also directed towards in what situations the subjects would like to assign the fault classifications. Four alternatives were given: 1) when found during execution, 2) during a meeting where the fault is textually described, 3) when the fault is corrected in the system, and 4) when the correction is verified.

Table 7: Alternatives, median ranking and times selected.

Alternative	*Median*	*Selected*
Source Code	**3**	**5**
Test Cases	4	5
Requirement Specification	4	4
Operational System	7.5	4
Software Design	4.5	4
Change Information, changes solving the problem.	**3**	**5**
Talk to the developer.	**3**	**7**
Optional 1: Better fault descriptions. (Suggested by two respondents)	-	2
Optional 2: Better system knowledge (Suggested by one respondent)	-	1
Optional 3: More information about how to use the fault classifications. (Suggested by one respondent)	-	1

The two most selected options are when the fault is detected during execution, and when the fault is corrected. One additional comment recommended a meeting with the developer. Still, the respondents think that it is possible to correctly classify faults that are detected, reported, and/or corrected by another party.

However, though the respondents' experience and knowledge concerning fault classification and ODC were stated as low, five out of eight thought that ODC were easy to use and understandable. The latter was based on the brief description supplied. To aid understanding of the fault classification and ODC the respondents ranked the four suggestions in the following order, with the highest priority first: 1) Example faults, 2) Exhaustive description of the fault classifications, 3) Consensus discussions in training purposes, and 4) Feedback of performance along with a correct answer.

6. Conclusion

Based on the achieved level of agreement provided by the Kappa calculations, it is safe to say that the hypothesis is rejected. For the case described in this study, it is not possible for a group of separate classifiers to assign the correct fault classification and by that agreeing.

The reason for this is related to four issues, namely:

- Fault description
- Education of classifiers in the classification scheme.
- Classification scheme as such.
- Classifiers initiation to the domain and system being classified.

In summary, the fault description is the likely cause for the low Kappa values achieved in this study. This is

supported by the analysis of fault classification and questionnaire data. The education part was partly simulated by recalculating the Kappa values for the two most interchanged fault classifications. However, this treatment did not result in a drastic improvement of the average Kappa value, and hence indicating that classifier education is not the single reason for the low agreement.

The classification scheme is also ruled out as a minor source for the low Kappa values, based on prior successful attempts such as [ElEmam98, Chillarage02].

The classifiers initiation is not discussed earlier in the paper, but the analysis of the questionnaire resulted in that the initiation into the domain and deeper knowledge of the system is an important factor. This conclusion is supported by the fact that desired information to the fault classification is source code, change information (what was done to solve the fault), and talking to the developer, i.e. more insight into the details of the faults and system. A further conclusion is that the developer or maintainer correcting the fault is best suited to correctly classify the fault, alternatively inserting more of the knowledge possessed by the developer or maintainer into the fault's description. By having a single developer, classifying the faults, raises the issue of being able to test the classification for correctness through Kappa statistic, which is still needed to assure a non-subjective correct fault classification.

7. Further work

As described in Section 1.1 the next step is to plan and perform a similar empirical evaluation within an industry setting. The conclusions and lessons learned from this study are incorporated into the design for the upcoming study, including:

- Using better fault descriptions, containing more of the information as indicated in this study, e.g. source code, change information, or equivalent.
- Supplying additional information as requested.
- Assuring that the subjects have higher understanding of the domain, system, and hence the faults that might reside within the system, indicating application in an industry environment.

Sampling the subjects from the industry partner and using logged faults from within their systems address the third issue.

8. Acknowledgements

Special thanks are addressed to the participating classifiers generating the data, for their interest and for allocating time for the evaluation. We are also grateful to the PSP course participants providing the fault descriptions and the original classification.

This work was partly funded by The Knowledge Foundation in Sweden under a research grant for the project "Blekinge - Engineering Software Qualities (BESQ)" (http://www.ipd.bth.se/besq).

9. References

[Altman91] D. Altman, "Practical Statistics for Medical Research", Chapman-Hall, 1991.

[Bhandair93] 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, Issue 12, pp. 1157 – 1171, 1993.

[Chillarege92] R. Chillarege, S. I. Bhandari, J. K. Chaar, M. J. Halliday, D. S. Moebus, K. Ray Bonnie and M-Y. Wong, "Orthogonal Defect Classification – A Concept for In-process Measurement", IEEE Transactions on Software Engineering, Vol. 18, Issue 11 pp. 943 – 957, 1992.

[Chillarege02] R. Chillarege and K. Prasad, "Test and Development Process Retrospective - A Case Study Using ODC Triggers", Proceedings of the International Conference on Dependable Systems and Networks, pp. 669 – 678, 2002.

[ElEmam99] K. El Emam, "Benchmarking Kappa: Interrater Agreement in Software Process Assessments", Empirical Software Engineering: An International Journal, Vol. 4, Issue 4, pp. 113-133, 1999.

[ElEmam98] K. El Emam and I. Wieczorek, "The Repeatability of Code Defect Classifications", Proceedings of International Symposium on Software Reliability Engineering, pp. 322-333, 1998.

[Humphrey95] W. S. Humphrey, "A Discipline for Software Engineering", Addison Wesley, 1995.

[Leszak02] M. Leszak, D. E. Perry and D. Stoll, "Classification and Evaluation of Defects in a Project Retrospective", Journal of Systems and Software, Vol. 61, Issue 3, pp. 173-187, 2002.

[Linkman97] S Linkman and H. D. Rombach, "Experimentation as a Vehicle for Software Technology Transfer – A Family of Software Reading Techniques", Information and Software Technology Vol. 39, Issue 11, pp. 777-780, 1997.

[Robson00] C. Robson, "Small-Scale Evaluation", SAGE Publications, 2002.

[Wohlin00] C. Wohlin and A. Amschler Andrews, A., "Assessing Project Success Using Subjective Evaluation Factors", Software Quality Journal, Vol. 9, Issue 1, pp. 43-70, 2000.