Using the ISO/IEC 9126 product quality model to classify defects : a Controlled Experiment

Using the ISO/IEC 9126 product quality model to classify defects : a Controlled Experiment

Antonio Vetro'[1,2], Nico Zazworka[1], Carolyn Seaman[1,3], Forrest Shull[1]

[1]Fraunhofer CESE
College Park, MD, USA
nzazworka@fc-md.umd.edu
fshull@fc-md.umd.edu

[2]Automatics and Informatics Dept.
Politecnico di Torino
Torino, Italy
antonio.vetro@polito.it

[3]UMBC
Department of Information Systems
Baltimore, MD, USA
cseaman@umbc.edu

***Abstract*—Background: Existing software defect classification schemes support multiple tasks, such as root cause analysis and process improvement guidance. However, existing schemes do not assist in assigning defects to a broad range of high level software goals, such as software quality characteristics like functionality, maintainability, and usability.**

Aim: We investigate whether a classification based on the ISO/IEC 9126 software product quality model is reliable and useful to link defects to quality aspects impacted.

Method: Six different subjects, divided in two groups with respect to their expertise, classified 78 defects from an industrial web application using the ISO/IEC 9126 quality main characteristics and sub-characteristics, and a set of proposed extended guidelines.

Results: The ISO/IEC 9126 model is reasonably reliable when used to classify defects, even using incomplete defect reports. Reliability and variability is better for the six high level main characteristics of the model than for the 22 sub-characteristics.

Conclusions: The ISO/IEC 9126 software quality model provides a solid foundation for defect classification. We also recommend, based on the follow up qualitative analysis performed, to use more complete defect reports and tailor the quality model to the context of use.

Keywords:defect classification, software quality

I. DEFECT CLASSIFICATIONS

Defect classification is used in software processes with various objectives: provide feedback to developers [1], generate better tests [2], assess the quality of software and improve software processes [3] [4]. However, existing schemes do not assist in assigning defects to a broad range of high level software goals, such as software quality characteristics like functionality, maintainability, and usability. This study that investigates whether a classification based on the ISO/IEC 9126 software product quality model is reliable and useful to link defects to quality aspects impacted.

So far, one of the most widely used defect classification scheme is Orthogonal Defect Classification (ODC), introduced in 1992 by Chillarege et al. [5]. ODC was presented as the bridge between statistical defect models, whose goal is to predict the reliability of the software and its components, and root cause analysis, which aims at identifying the cause of defects. The semantic information provided by ODC permits developers to link causes of defects with their effects on process or product. Chillarege et al. [5] identified eight different defect types and mapped them to processes (e.g., low level design, code), enabling process feedback to developers and identifying the existence of measurable cause and effect relationships in the software development process. Although Chillarege et al. focused their work on a subset of defect effects (e.g., reliability growth), the effect of a defect can be measured on other product or process attributes. They provide as example the severity of defects and the impact of field problems at a customer organization through CUPRIMD [6] (capability, usability, performance, reliability, installability, maintainability and documentation), which is a quality procedure to control the different aspects of software quality during development and software lifetime. CUPRIMD introduces the idea of this work, because we investigate whether defects can be categorized according to the software qualities that they affect, which take perspectives from different stakeholders (e.g., manager, end user) into account.

A. Need for a comprehensive classification scheme

Understanding the linkage between defects and their effect on the overall software quality can help with several tasks:

- Defect severity and priority can be better understood depending on which quality attributes are more important in a project. For example, focus on the ease of which users can interact with the software product will penalize defects that affect usability.
- Testing techniques can be tailored towards specific important quality characteristics. For example, if portability is a major concern then defects falling into this category will help with the analysis of shortcomings of applied testing techniques (i.e. in form of an in-depth analysis why a defect was not detected by the technique).
- Measurement of process improvement activities will be supported. The linkage between defects

TABLE I. ISO/IEC 9126 SOFTWARE QUALITY CHARACTERISTICS (ADAPTED FROM [14])

Characteristic	ID	Description	Sub Characteristics
Functionality	F	The capability of the software to provide functions which meet the stated and implied needs of users under specified conditions of usage (what the software does to meet needs)	Suitability, Accuracy, Interoperability, Compliance, Security
Reliability	R	The capability of the software product to maintain its level of performance under stated conditions for a stated period of time.	Maturity, Fault Tolerance, Recoverability
Usability	U	The capability of the software product to be understood, learned, used and provide visual appeal, under specified conditions of usage (the effort needed for use)	Understandability, Learnability, Operability, Attractiveness
Efficiency	E	The capability of the software product to provide desired performance, relative to the amount of resources used, under stated conditions.	Time Behavior, Resource Utilization
Maintainability	M	The capability of the software product to be modified which may include corrections, improvements or adaptations of the software to changes in the environment and in the requirements and functional specifications (the effort needed for modification)	Analyzability, Changeability, Stability, Testability
Portability	P	The capability of the software product to be 'transferred from one environment to another. The environment may include organizational, hardware or software.'	Adaptability, Installability, Conformance, Replaceability

and software quality allows a better understanding if a process improvement affects the distribution of defects among quality characteristics. Questions such as: "Do additional code inspections lead to less functionality defects?" can be investigated.

With these objectives in mind it is crucial that a classification scheme provides a lean and repeatable way of assigning defects to quality characteristics. Assuming that humans carry out the classification task, the results should lead to similar results, independent of the person performing the task. Further, besides high level of agreement, the goodness of the classification scheme will depend on how well the classes map to the real-world concerns of stakeholders. For example, high agreement might be achieved if a classification consists of only a few, but in practice incomplete number of classes. Last, but not least, a good classification scheme should be seen as an instrument that will allow to effectively and measurably improve supporting the above listed tasks.

Reflecting on the second argument of finding the right concerns of stakeholders, current classification schemes proposed in the literature are often limited by focusing on a very specific set of concerns. For instance, several taxonomies have been developed specifically for security concerns (e.g., [7], [8] and [9]) ignoring other stakeholder interests. Other defect classifications [10] [11], derived from ODC, are specifically designed from the point of view of software reliability. Leszac et al. [12] used their classification scheme to improve both, reliability (defect density) and maintainability (code size and complexity). Even though their scheme considers multiple attributes, many other perspectives are neglected.

This work aims to build a first step to link defect classification schemes to a comprehensive range of software quality goals. We consider the well-known and widely adopted product quality model of the ISO/IEC 9126 International Standard [13] as our initial attempt to fill the gap. We conduct a first experiment to evaluate the reliability with respect to human classification of a defect classification scheme based on ISO/IEC 9126 quality attributes. If this classification scheme were found to be reliable then this would motivate for conducting further research on the goodness of the classification with respect to supporting tasks such as defect prioritization or tailoring of testing techniques.

The paper is structured as follows: we introduce the ISO/IEC 9126 Product Quality model and the classification scheme in Section II; then we describe the experiment conducted in Section III; Section IV shows results and Section V discussion. Threats to validity and conclusions are presented in sections VI and VII.

II. ISO/IEC 9126 PRODUCT QUALITY MODEL (NOW ISO/IEC 25010) AND PROPOSED DEFECT CLASSIFICATION SCHEME

ISO/IEC 9126 Software engineering-Product quality is an international standard for the evaluation of software quality. It defines a quality model with six *main characteristics*, namely: functionality, reliability, usability, efficiency, maintainability, and portability, which are further broken down into 22 *sub-characteristics*. Table 1 (adapted from [14]) provides descriptions of the six main characteristics and the 22 sub characteristics. The standard was revised in March 2011 by the ISO/IEC 25010 committee [15]. The new standard added a new main characteristic (Compatibility), and moved Security from a sub-characteristic to a main characteristic with its own set of sub-characteristics. Some other sub-characteristics were added in the 2011 revision (confidentiality, integrity, non-repudiation, accountability and authenticity, functional completeness, capacity, user error protection, accessibility,

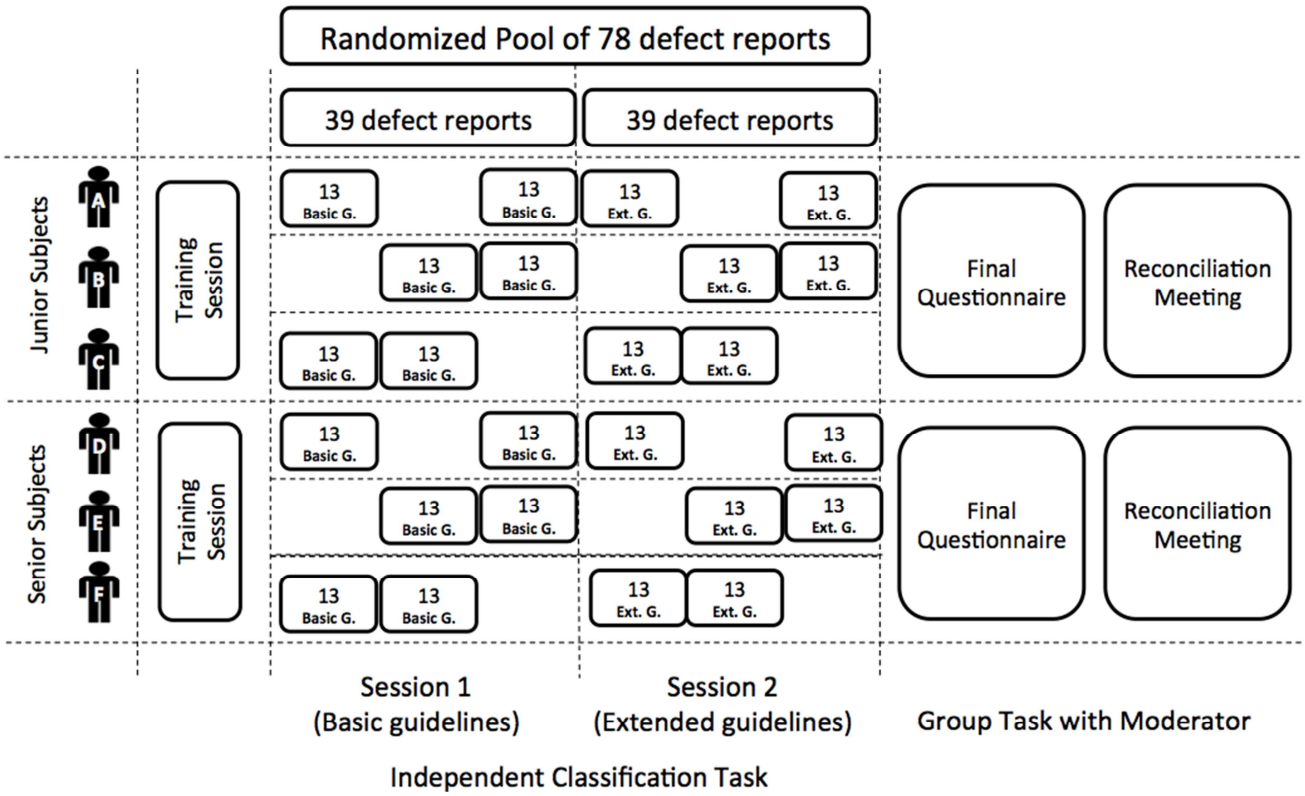

Figure 1. Experiment design

availability, modularity and reusability), while compliance was removed.

The experiment was designed two months after the new standard was released, but the authors decided to keep the old standard because of its wide adoption. The large overlap between the two versions of the standard encourages the generalizability of the findings of this experiment to the new standard.

We use the main characteristics and sub-characteristics listed by ISO/IEC 9126 as attributes in defect classification schemes. The underlying idea is that each defect is reducing the capability and quality of the software in one or more of the model's main characteristics. Therefore, we propose a classification scheme that is based on the following guidelines:

- A defect impacts a software quality main characteristic if its effects reduce the associated capability described in Table I
- A defect impacts a software quality sub-characteristic if its effects reduce the associated capability (see Table I, [13] and [14])

A defect can often be related to more than one quality main characteristics and sub-characteristics. As a consequence, the defect classification scheme proposed is not orthogonal. We will discuss in following sections how this property must be taken into account in data analysis and results interpretation.

III. EXPERIMENT

A. Goal and Questions

The goal of the main experiment is to assess the *reliability* of the defect categorization scheme based on the ISO/IEC 9126 main characteristics and sub-characteristics, with respect to human classification, i.e. a human decides which quality attributes are impacted by a defect. We define *reliability* as a measure to show how well a group of human classifiers agree in mapping a set of defects to the categories of the classification. We analyze multiple aspects and formulate our aims in the following research questions (RQ):

- RQ1 - Reliability: How reliable is the defect classification based on ISO/IEC 9126?
- RQ2 – Expertise: Is the reliability of the classification dependent on the level of expertise of the human classifiers?
- RQ3 – Main vs. Sub-characteristics: Is the reliability of the classification dependent on the main characteristics/sub-characteristics?
- RQ4 – Extended Guidelines: Can the adoption of extended guidelines raise reliability?

- RQ5 – Human Perception: How do human classifiers perceive the ISO/IEC 9126 scheme when used to classify defects?

B. Context

We collected a set of defects from a project in active development from an industrial partner. The defects were extracted from a JIRA bug tracking system[1]. The industrial partner (appraised at CMMI® Maturity Level 3) has about 40 employees and develops web applications in C# (using .NET and Visual Studio). The software application has a size of about 35,000 lines of code and has been active in production since November 2009, with four developers working on it. At the time of the experiment, the JIRA system contained 78 fixed defects, and all of these were used in the study. Each defect in JIRA is a report that has been completed by developers or customers. Each report contained the following data:

- **Defect Report Identifier and Summary:** each report has a unique label, a title and a short description.
- **Location**: component/s of the software affected by the defect, e.g. "Authentication/Security" or "Database Development"
- **Version and Time Information**: affected and fixed versions: the version(s) of the software the defect affected (i.e. in which the defect was present), and the version of the software in which the defect was fixed. Further: creation and resolution date of the defect report, as well as estimated and actual time spent to fix the defect.
- **People**: reporter (who reported the defect), and assignee (the person in charge of fixing it)
- **Description and supporting information**: more detailed description of the defect, file attachment(s) (e.g., screenshots, documents), and a comment thread (containing discussions between reporter and assignee)
- **Category**: a set of categories was defined by the company. The categories were different from the ISO/IEC 9126 quality main characteristics.
- **Other main characteristics**: priority, severity (filled in by the software end user), phase of the development cycle on which the defect was detected

We initially conducted a pilot study in which the first author classified all the defect reports with the guidelines specified in Section II. Based on this first experience, the first author defined a set of extended guidelines to be added to the initial set. The aim of the extended guidelines is to clarify and simplify the classification task. The extended guidelines are:

- A defect impacts Functionality if its effects reduce the capability of what the system does (eg. "does not work", "is wrong", "error")
- A defect impacts other characteristics other than Functionality if its effects reduce the capability of how the system performs its tasks (eg. "it should be faster", "must be easier", "the popup is annoying")
- The evidence criterion must be adopted in the classification process: evidence about information should be provided on the defect report, otherwise the information must be considered as missing. For instance, if the title of a defect report is not clear, the bind between defect and quality characteristic/sub-characteristic judgment should not rely on it but only on other information (comment, picture). Or, if a link to a requirement is missing, there is no enough evidence to define that the defect was related to a particular requirement.
- A reported stop/crash of the system impacts its Reliability
- Any relation to standard U508 impact the Usability of the system

The base and extended guidelines are evaluated separately in the experiment.

C. Demographics

The six participants of the experiment were divided into two groups, based on their level of experience in software engineering. Participants of the first group were three students completing their Master of Science degree in computer science (subjects A, B, C) with 2 to 6 years of programming experience and little a-priori knowledge of the ISO/IEC 9126. We call them in the following *junior subjects*. The second group included expert software engineers (subjects D, E, F, that are also the co-authors of this paper[2]) with 10 to 20 years expertise and with some familiarity but little working knowledge on the standard. We will refer to them as *senior subjects* in the following.

D. Experiment Design

The subjects of the study were divided into two equally sized groups and had to complete two major tasks after training. Figure 1 depicts and summarizes this design. The first task (independent classification task) was to read a set of defect reports and to assign each report to one or more ISO/IEC 9126 quality main characteristics and sub-characteristics. The second task was a group task with moderator that included a questionnaire and reconciliation meeting for resolving disagreement between subjects.

[1] http://www.atlassian.com/software/jira

[2] The three co-authors were not involved in preparation of the experiment or other pre-study activities. They solely participated in data analysis and interpretation and paper authoring after completing their roles as expert subjects.

1) Training

To train subjects on the matter of quality main characteristics, instructions were given to them before the start of the study, which included:

- short description and goal of the experiment;
- high level description of ISO/IEC 9126 as shown in Table I;
- examples and descriptions of a defect report;
- classification guidelines and instructions to execute the experiment;

The full text of the instructions is available online[3].

2) Classification Task

The 78 defect reports were assigned to participants in the following way. Each participant categorized 52 defects in two sessions: in session one, 26 defect reports were classified with the base guidelines. Subsequently, in session two, the other 26 defect reports were classified with the extended guidelines by the same subjects.

Participants received defect reports on paper (with possible linked documents like screenshots) and the list of the main and sub characteristics with a short description (available online for replication[3]). Then, they used their expert knowledge to classify each defect report with one or more main characteristics and sub-characteristics which impacted by such defect in their opinion. Classification was recorded electronically using a spreadsheet to avoid possible transcribing errors (the spreadsheet is also available online[3] to ease replication). There was no time limit given to the participants for completing the classification tasks.

The defects were assigned randomly to subjects to reduce threats due to the temporal order in which the defect reports were entered into JIRA: for instance, it could be possible that defects detected earlier (and so coming earlier in the list) are different from defects reported later. The defects were assigned in a manner such that each defect was classified by exactly four subjects.

Moreover, participants were asked to report, for each defect classification, a confidence level with which they assigned ISO/IEC 9126 main characteristics to the defect reports. They used the following scale: 1= "I'm not sure", 2="I'm quite sure", 3= "I'm very sure".

3) Group Task with Moderator

A reconciliation meeting for each group followed a few days after the classification task, where all participants in each group met to answer moderator questions and resolve conflicts. The role of moderator was done by the first author, who asked participants, for each conflict, to answer the following questions:

- Why did you classify this defect report as …?
- Do you think that you could add some of main characteristics that the other rater selected?
- Which of the main characteristics that you selected would you keep?
- Do you have any other comment on this defect report?

Beyond conflict resolution questions, subjects answered also 4 further questions:

- **[General Observation]** Do you have any general observation on the classification experience?
- **[Classification methodology]** How did you use the guideline tables to classify defect reports?
- **[Extended guidelines]** Did you think the extended guidelines were useful/useless or do they add confusion/uncertainty?
- **[More information]** If it could be possible to buy new information with a small amount of money, would you spend that money to get additional information on defect reports or on extended guidelines?

E. Analysis methodology

The main metric used to answer research questions 1 to 4 is Cohen's Kappa [16], widely used in the assessment of defect classification schemes [4] [17]. The assumption is that a classification scheme is considered "reliable" if multiple humans classify items the same way when using the classification independently, thus achieving a high Kappa value. The Kappa statistic is computed as:

$$K = \frac{(Po - Pe)}{(1 - Pe)}$$

in which P_O is the percentage of classifications that are the same between two subjects, and Pe is the probability of agreement among coders due to chance, whose computation is based on marginal probabilities under the assumption of complete statistical independence of raters. Pe estimates the proportion of times raters would agree if they guessed completely on every case and with probabilities that match the marginal proportions of the observed classifications. The values of K are constrained to the interval [-1,+ 1]. A K value of one means perfect agreement, a K value of zero means that agreement is equal to chance, and a K value of negative one means "perfect" disagreement.

Despite its widespread use in the literature, the Kappa coefficient has two notorious problems [18] [19]: bias and prevalence. Prevalence occurs when the distribution of categories is skewed and labels are concentrated in one or a few categories; in those cases Kappa tends to be lower. The bias problem occurs when raters' individual classifications are very different, leading to the paradox that Kappa increases as they are less similar. For these reasons and according to [19], in addition to Kappa we report two other metrics: the proportion of agreement P_O[20] and the Kappa adjusted for prevalence [18], that is equivalent to $2P_O$-1 [19]. The first measure is useful to understand the possible effect of the bias problem, because it does not take into account the chance agreement and it is usually higher than Cohen's Kappa. A P_O similar to K is a clue for a possible bias problem. The second measure

[3] http://softeng.polito.it/vetro/confs/ease2012/data.zip

takes out the effect of prevalence from Kappa, but must be taken into account only if prevalence occurs.

Moreover, since the proposed classification scheme is not orthogonal (i.e. subjects can select multiple main characteristics for one defect report), we adopt the weighted versions of the three metrics (WK, WP_O, $W2P_O$-1) as defined in [16] [18] [20]:

$$WK = \frac{(WPo - WPe)}{(1 - WPe)}$$

whereby P_O and Pe are computed with a weight W that is a similarity distance between two overlapping rates. An example of an overlapping pair of rates for a given defect report is: Subject1:{FU}, Subject2:{FR}, where the first subject labeled the defect as impacting Functionality (F) and Usability (U), while the second one classified it as Functionality (F) and Reliability (R). In this case, the agreement is only partial, i.e. on Functionality, and since one in three main characteristics is in common between the two subjects, the weight (i.e. the agreement) is ⅓, i.e. 0.33. An example of perfect agreement is: Subject1:{FU}, Subject2:{FU}, and the weight is 1. An example of disagreement is instead: Subject1:{F}, Subject2:{R}, and the weight is 0. The same weighting criterion is applied to sub-characteristic classifications.

We now discuss in more detail the research questions listed in section III A, translating them, when appropriate, into a set of testable hypotheses. All hypotheses are tested comparing pairs of subjects because different set of defect reports were classified by different pairs of subjects (see Figure 1). For example: A and C have in common defect reports from 0 to 13. B and C from 14 to 26, and so on.

RQ1 - Reliability: How reliable is the defect classification based on ISO/IEC 9126?

Several tables of how Kappa values can be interpreted into strength of agreement can be found in the literature [20] [21] [22] [23]. We observe that a threshold 0.60 corresponds to a "good"/"substantial" agreement in all the proposed ranks, while a Kappa in the range 0.21-0.60 includes the adjectives "fair" and "moderate" in the majority of the tables. Given the exploratory nature of this work, we report the different agreement metrics and we do not test for a particular hypothesis.

RQ2 – Expertise: Is the reliability of the classification dependent on the level of expertise of subjects?

The indexes domain in the formula below are i={A-B, A-C,B-C} and k={D-E, D-F, E-F} indicating the subject pairs, and j={1,2} indicating the session. Moreover, the hypotheses are tested both at main characteristic level (H1) and sub-characteristic level (H2).

- $H1_0$(RQ2): $WK_{i,j,\,char} > WR_{k,j,\,char}$
- $H1_A$(RQ2): $WK_{i,j,char} \leqslant WR_{k,j,char}$
- $H2_0$(RQ2): $WK_{i,j,\,subchar} > WR_{k,j,\,subchar}$
- $H2_A$(RQ2): $WK_{i,j,subchar} \leqslant WR_{k,j,subchar}$

TABLE I. AGREEMENT METRICS FOR JUNIOR CLASSIFICATIONS

Round	*Subj.*	**Characteristics**			**Sub-characteristics**		
		WP$_O$	***W2P$_O$-1***	***WK***	***WP$_O$***	***2WP$_O$-1***	***WK***
Base G.	A-C	0.73	0.47	0.55	0.47	-0.05	0.36
	B-C	0.60	0.20	0.37	0.30	-0.4	0.25
	A-B	0.52	0.04	0.32	0.26	-0.47	0.24
Extra G.	A-C	0.88	0.77	0.61	0.54	0.09	0.23
	B-C	0.43	-0.13	0.21	0.18	-0.63	0.13
	A-B	0.49	-0.03	0.25	0.26	-0.49	0.17

TABLE II. AGREMENT METRICS FOR SENIOR CLASSIFICATIONS

Round	*Subj.*	**Characteristics**			**Sub-characteristics**		
		WP$_O$	***W2P$_O$-1***	***WK***	***WP$_O$***	***2WP$_O$-1***	***WK***
Base G.	D-F	0.72	0.44	0.51	0.55	0.11	0.49
	E-F	0.63	0.26	0.34	0.44	-0.11	0.33
	D-E	0.64	0.28	0.50	0.42	-0.17	0.38
Extra G.	D-F	0.63	0.26	0.35	0.55	0.1	0.44
	E-F	0.59	0.18	0.32	0.35	-0.3	0.29
	D-E	0.53	0.05	0.28	0.40	-0.2	0.32

RQ3 – Main vs. Sub-characteristics: Is the reliability of the classification dependent on the main characteristics/sub-characteristics?

The subject pairs are i={A-B, A-C, B-C, D-E, D-F, E-F}and session index j={1,2}:

- H_0(RQ3): $WK_{i,j,\,char} = WK_{i,j,\,subchar}$
- H_A(RQ3): $WK_{i,j,\,char} \neq WK_{i,j,\,subchar}$

RQ4 – Extended Guidelines: Can the adoption of the extended guidelines improve the reliability?

The subject pairs are i={A-B, A-C, B-C, D-E, D-F, E-F}and session index j={1,2}. The hypotheses are tested both at main characteristic level (H1) and sub-characteristic level (H2):

- $H1_0$(RQ4): $WK_{i,1,char} \geqslant WK_{i,2,char}$
- $H1_A$(RQ4): $WK_{i,1,char} < WK_{i,2,char}$
- $H2_0$(RQ4): $WK_{i,1,subchar} \geqslant WK_{i,2,char}$
- $H2_A$(RQ4): $WK_{i,1,subchar} < WK_{i,2,char}$

We test the hypotheses related to RQ2-RQ4 by applying the Mann Whitney test [24] to the two sets of WK of each question. We apply a confidence interval of 95% and, given the small number of samples (6 for each set), we also provide boxplots for qualitative comparisons.

RQ5 – Human Perception: How do human classifiers perceive the ISO/IEC 9126 scheme when used to classify defects*?*

The answer to RQ 5 is addressed through qualitative analysis of reconciliation meetings records.

IV. RESULTS

A. Descriptive statistics

Functionality and Usability (F and U) were the dominant classifications in the junior subjects' classifications. Functionality was selected in 122 out of 156 classifications (78.2%), Usability in 74 classifications (47.4%), and Reliability (R) in 35, corresponding to 22.4%. Functionality was also dominant in seniors' classifications, with 78.8% of classifications, followed by Usability (44.9%) and Reliability (23.7%). The other three main characteristics obtained negligible frequencies in both groups. These figures suggest that classifications did not concentrate only on one main characteristic or sub-characteristic, therefore the prevalence problem did not occur in the experiment. Moreover, Usability and Functionality co-occurred more than any other pair of main characteristics in both groups.

Finally, junior subjects selected on average 1.51 main characteristics and 2.05 sub-characteristics for each classification, while senior subjects selected on average 1.48 main characteristic and 1.85 sub-characteristics.

Five conflicts (0% agreement) and 2 partial agreements (weights 0.33 and 0.20) were discussed in the reconciliation meeting for junior subjects. Participants resolved all conflicts by changing their classifications on the basis of the discussion. Further, 4 conflicts (0% agreement) were discussed in the seniors' reconciliation meeting where participants solved 3 conflicts by changing their classifications on the basis of the discussion.

B. Answers to research questions

1) RQ1 - Reliability

All agreement indicators for both groups are listed in Tables II and III. On the main-characteristic level the weighted Kappa values range from 0.21 to 0.61, whereas on the sub-characteristic level values are lower, ranging from 0.14 to 0.49. Considering the bias problem with Kappa discussed previously, WP_O values are close to WK values in the sub-characteristics. This suggests that the bias problem could slightly affect results at the sub-characteristic level, where the variability is higher due to the large number of classification options.

At the main characteristics level, eleven out of twelve comparisons had $0.21 \leqslant WK \leqslant 0.60$, corresponding to a fair or moderate agreement [20] [21] [22] [23]. Only one classification (A-C session 2) had a good/substantial agreement (0.61). On the sub-characteristics side, two classifications were poor (WK < 0.21) and the other ten were moderate. Overall, we observe that the reliability of the classification is moderate. Since the descriptive statistics showed that the prevalence problem did not occur, we can ignore $W2P_O$-1. However the values of WP_0 suggest that the bias problem related to Kappa could have occurred on the sub-characteristics classifications.

2) RQ2 - Expertise

The boxplots in Figure 2 show that seniors clearly outperform juniors only at the sub-characteristic level (p-value<0.05) and that seniors have less variability. We do not observe differences at main characteristic level (p-value =1).

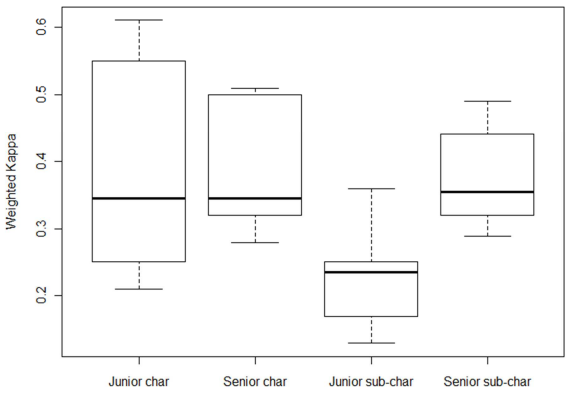

Figure 2. Level of agreement among Juniors vs Seniors at main characteristics and sub-characteristics levels.

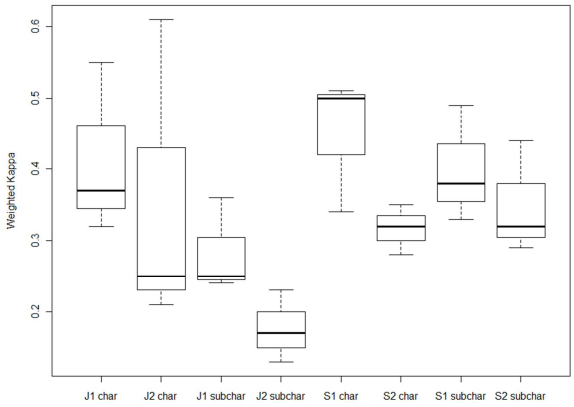

Figure 3. Level of agreement by classification sessions (1,2), subjects (Juniors J, Seniors S), classification level (main characteristic and sub-characteristic)

3) RQ3 – Main vs. Sub-characteristics

The boxplots in Figure 2 and 3 show that sub-characteristics have a lower agreement mainly for juniors (the null hypothesis is rejected with p-value = 0.037). In addition to the fact that possible bias could provoke a slight higher WK for sub-characteristics, we conclude that reliability is higher at the main characteristic level.

4) RQ4 – Extended Guidelines

The adoption of extended guidelines resulted in lower WK. However, the test can be rejected only for sub-characteristics (p-value=0.031), while for main characteristics is rejected only with confidence level 90% (p-val=0.094)

5) RQ5 – Human Perception

Question 5 is answered through qualitative analysis of reconciliation meeting records. We present findings for each section of the reconciliation meetings.

General Observations. Table IV summarizes the general observations of participants. The first author extracted themes by coding the recorded answer to the questions asked in the group meetings. Five out of ten identified themes indicate that subjects thought that the information provided in the defect reports was insufficient, or that a lack of information made it more difficult to classify a defect. This indicates a relationship between the ease of classification with ISO/IEC9126 and detail of defect information provided. The themes are:

- Difficult without the specifications of the software
- Maintainability information is hard to find in a defect report
- Little information in defect reports
- With little information, it is hard to distinguish between Functionality and Usability
- Pictures of defects reports are not useful

The remaining themes were dispersed across a range of topics. No single theme was mentioned by more than 3 subjects, which indicates that perceptions varied between subjects and that the classification scheme does not bear one common problem.

Classification methodology. All but one subject classified starting from sub-characteristics.

Extended guidelines. Comments about extended guidelines are summarized in Table V. Themes were diverse, however one theme: "Overall, not very useful" was mentioned by all six subjects, indicating that the extended guidelines are useless as the answer to RQ4 showed.

More information. All subject answered that they would buy more information on defect reports.

V. DISCUSSION

The major findings of this experiment are:

- The agreements between participants were moderate indicating that the classification is moderately reliable (RQ1)
- Classification performed by experts leads to less variability and higher reliability on sub-characteristics level (RQ2)
- Classification on main characteristic level is less variable and more reliable (RQ3)
- The extended guidelines adopted were not useful (RQ4)
- The quality of defect reports impacts reliability of the classification. (RQ5)

The WK agreements between subjects were "moderate" according to existing interpretation suggestions [20] [21] [22] [23], but in our opinion the classification can be considered reliable on the main-characteristic level for the following reasons. The lack of information in defect reports and the unfamiliarity of subjects with the inspected development project was identified as one major factor of the moderate WK in the qualitative analysis of the data.

The lack of information indicates that defect reports are filled insufficiently in practice (at least in our target project) to allow for a more reliable a-posteriori classification. This is to say that the reports are considered sufficient by the industrial partner in order to *fix the defect*, which is the primary purpose of writing a report. We suggest that an improvement of reliability in practice can be achieved by using subjects that can compensate for the

TABLE IV. SUBJECTS GENERAL OBSERVATIONS

Comment	A	B	C	D	E	F
Classification was difficult without the specifications of the software			X		X	
It was hard to distinguish between suitability and accurateness			X			
Extended-guidelines didn't make difference	X					
Maintainability information is hard to find in a defect report		X				
There was little information in defect reports	X			X		
Too many sub-characteristics	X					X
When there was few information, it was difficult to distinguish between Functionality and Usability			X		X	
It was better to classify at characteristic level		X	X			
Maturity is rather belonging to Reliability and Project Management than Maintainability	X	X	X			
Pictures on defect reports were not useful	X	X	X			

TABLE V. SUBJECTS COMMENTS ON EXTENDED GUIDELINES

Comments	A	B	C	D	E	F
Useful to distinguish suitability and accurateness	X					
Overall, not very useful	X	X	X	X	X	X
Even with better extended-guidelines, classification would have been difficult because of lack og information on defect reports	X		X			X
Useful for people who saw the standard the first time		X				
Would have prefered a better explanation on the differences between sub-characteristics			X			
Only criteria 1 and 2 were easy to understand				X	X	X
The fourth criterion was useful				X		

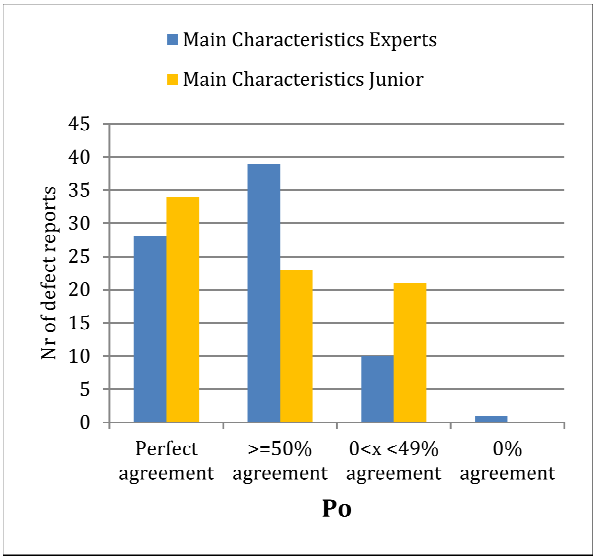

Figure 4. Agreements in classification (after reconciliation meeting)

lack of information with their own context knowledge (e.g. developers who reported and fixed the defects in first place). A follow-up experiment is required to confirm this hypothesis.

Moreover, Figure 4 reports the proportion of agreements (P_o) of both seniors and juniors at the main characteristic level and after the reconciliation. Eighty-eight percent of classifications of seniors and 75% of those of juniors had an agreement $\geqslant$ 50%, i.e. in accordance on at least half of their classification (e.g., FU and U). Only nine defects had a 0% agreement, and only one still had the full disagreement even after the reconciliation meeting. Looking at the figures from this perspective lets us conclude that agreement outweighs disagreement clearly.

Another finding of our analysis is that senior classifications were less variable and more reliable on the sub-characteristics level. We investigated the differences between juniors and seniors in depth and computed the most common conflicts. The most common conflict patterns were suitability/accuracy and suitability/operability; merging them would increase the P_o by 18%. Therefore, considering the quantitative and qualitative answers and also this follow-up analysis, our suggestion is to focus on the main characteristic level or, in order to achieve higher agreement, merge some of the sub-characteristics such as suitability, accuracy and operability.

Overall, we conclude that the standard builds a solid foundation in order to trace defects to quality goals when using the main-characteristics rather than sub-characteristics.

VI. THREATS TO VALIDITY

We classify threats as internal, external, construct, and conclusion, according to the taxonomy proposed by Wohlin et al. [25].

Internal threats. A first internal threat is introduced by some differences in the experiment operations, mainly two: 1) seniors performed their first session simultaneously in the same room and they could feel in competition in terms of task completion time; 2) seniors performed the second session without the presence of the experimenter, and several days after the first classification. We could not avoid these threats for practical reasons, but we believe their impact is negligible.

To remove possibility of bias, we let subjects classify independently (e.g. senior participants did not speak to each other during classification).

Yet another internal threat is the possibility of learning effects on the second session of the classification task, which might have masked an effect of the extended guidelines.

Conclusion threats. The adoption of Kappa statistics could lead to misinterpretation of results [18][19], in particular when classification distributions are skewed. However, we observed that prevalence did not occur and bias could only affect results at the sub-characteristics level. The small sample size (6 subjects) is yet another conclusion threat.

External threats. The threats derived from the selection of participants (partly academic setting) and of the case study (defects of a web application) must be taken into account in the generalization of results and recommendations. Yet another external threat regards the applicability of our findings on a defect classification based on the ISO/IEC 25010, which is the evolution of ISO/IEC 9126: since more sub-characteristics were added, we expect that the level of agreement with the new version of the standard would slightly decrease rather than increase. However, the high level of overlap between the two versions of the standard should make findings still generalizable to the new ISO/IEC 25010.

VII. CONCLUSIONS AND FUTURE WORK

We conclude the paper with guidelines and recommendations to practitioners as well as suggestions for future work. We suggest adopting the defect classification scheme based on the ISO/IEC 9126 product quality model since it has shown reliable. Further research has to investigate if this classification can help with practical tasks, such as the prioritization of defects according to different stakeholders' perspectives, or the ease of process improvement measurement on specific quality dimensions. At this point in time, we recommend using the main characteristics that lead to good agreement results even on incomplete data, but to use the sub-characteristics with care. The level of experience does not affect the classification reliability at main characteristic level, but it does at sub-characteristic level.

Furthermore, future work should investigate whether further tailoring of the ISO/IEC 9126 quality model to the specific application and stakeholder context (e.g. by adding, removing or modifying the main characteristics and sub-characteristics) has a positive impact on reliability. We also encourage researchers to repeat this

experiment with a larger pool of subjects with (and without) a greater amount of contextual knowledge about the inspected defects.

Our own research agenda includes validating these results with developers at our industrial partner and to suggest the adoption of the scheme (in a possibly tailored form) in the company. In a second step we are interested how the usage of this model will support the various tasks and processes in the software development lifecycle that could benefit from a comprehensive understanding of the defect-quality relationship.

ACKNOWLEDGMENT

Authors are thankful to Till Ganzert, Christof Velte and Henning Femmer, whose voluntary participation in the experiment made this study possible. Authors are also thankful to the company that gave support and access to its JIRA database.

REFERENCES

[1] R. Chillarege, W.-L. Kao, and R. G. Condit, "Defect type and its impact on the growth curve," in *Proceedings of the 13th international conference on Software engineering*, ser. ICSE '91. Los Alamitos, CA, USA: IEEE Computer Society Press, 1991, pp. 246–255.

[2] B. Beizer, *Software testing techniques (2nd ed.)*. New York, NY, USA: Van Nostrand Reinhold Co., 1990.

[3] L. Mariani, "A fault taxonomy for component-based software," in *International Workshop on Test and Analysis of Component Based Systems (TACOS) satellite workshop at the European Joint Conferences on Theory and Practice of Software (ETAPS)*, ser. ENTCS, vol. 82, no. 6. Elsevier, 2003.

[4] B. Freimut, C. Denger, and M. Ketterer, "An industrial case study of implementing and validating defect classification for process improvement and quality management," in *Software Metrics, 2005. 11th IEEE International Symposium*, sept. 2005, pp. 10 pp. –19.

[5] R. Chillarege, I. Bhandari, J. Chaar, M. Halliday, D. Moebus, B. Ray, and M.-Y. Wong, "Orthogonal defect classification-a concept for in-process measurements," *Software Engineering, IEEE Transactions on*, vol. 18, no. 11, pp. 943 –956, nov 1992.

[6] R. Radice and R. Phillips, *Software engineering: an industrial approach*, ser. Software Engineering. Prentice-Hall, 1988.

[7] T. Aslam, I. Krsul, and E. H. Spafford, "Use of a taxonomy of security faults," in *Purdue University*, 1996, pp. 551–560.

[8] C. E. Landwehr, A. R. Bull, J. P. McDermott, and W. S. Choi, "A taxonomy of computer program security flaws," *ACM Comput. Surv.*, vol. 26, pp. 211–254, September 1994.

[9] S. Weber, P. A. Karger, and A. Paradkar, "A software flaw taxonomy: aiming tools at security," *SIGSOFT Softw. Eng. Notes*, vol. 30, pp. 1–7, May 2005.

[10] L. Ma and J. Tian, "Web error classification and analysis for reliability improvement," *Journal of Systems and Software*, vol. 80, no. 6, pp. 795 – 804, 2007.

[11] ——, "Analyzing errors and referral pairs to characterize common problems and improve web reliability," in *Proceedings of the 2003 international conference on Web engineering*, ser. ICWE'03. Berlin, Heidelberg: Springer-Verlag, 2003, pp. 314–323.

[12] M. Leszak, D. E. Perry, and D. Stoll, "Classification and evaluation of defects in a project retrospective," *Journal of Systems and Software*, vol. 61, no. 3, pp. 173 – 187, 2002.

[13] ISO/IEC, *ISO/IEC 9126. Software engineering – Product quality*, Std., 2001.

[14] I. Padayachee, "Iso 9126 external systems quality characteristics , sub- characteristics and domain specific criteria for evaluating e-learning systems," *Quality*, 2010.

[15] ISO/IEC, *ISO/IEC 25010. Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models*, Std., 2011.

[16] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[17] K. El Emam and I. Wieczorek, "The repeatability of code defect classifications," in *Software Reliability Engineering, 1998. Proceedings. The Ninth International Symposium on*, nov 1998, pp. 322 –333.

[18] T. Byrt, J. Bishop, and J. B. Carlin, "Bias, prevalence and kappa," *Journal of Clinical Epidemiology*, vol. 46, no. 5, pp. 423 – 429, 1993.

[19] B. Di Eugenio and M. Glass, "The kappa statistic: a second look," *Comput. Linguist.*, vol. 30, pp. 95–101, March 2004.

[20] J. L. Fleiss, *Statistical Methods for Rates and Proportions*, 2nd ed., ser. Wiley series in probability and mathematical statistics. New York: John Wiley & Sons, 1981.

[21] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977.

[22] D. G. Altman, *Practical Statistics for Medical Research (Statistics texts)*, 1st ed. Chapman & Hall/CRC, Nov. 1990.

[23] B. T., "How good is that agreement? (letter to editor)," p. 561, 1996.

[24] D. Sheskin, *Handbook of parametric and nonparametric statistical procedures*. Chapman & Hall/CRC, 2004.

[25] C. Wohlin, *Experimentation in software engineering: an introduction*, ser. Kluwer international series in software engineering. Kluwer Academic, 2000.