Regression testing goals - View of practitioners and researchers

Nasir Mehmood Minhas
Department of Software Engineering
Blekinge Institute of Technology
Karlskrona, Sweden
Email: nasir.mehmood.minhas@bth.se

Kai Petersen
Department of Software Engineering
Blekinge Institute of Technology
Karlskrona, Sweden
Email: kai.petersen@bth.se

Nauman bin Ali
Department of Software Engineering
Blekinge Institute of Technology
Karlskrona, Sweden
Email: nauman.ali@bth.se

Krzysztof Wnuk
Department of Software Engineering
Blekinge Institute of Technology
Karlskrona, Sweden
Email: krzysztof.wnuk@bth.se

Abstract

Context: Regression testing is a well-researched area. However, the majority regression testing techniques proposed by the researchers are not getting the attention of the practitioners. Communication gaps between industry and academia, and disparity in the regression testing goals are the main reasons. Close collaboration can help in bridging the communication gaps and resolving the disparities.

Objective: The study aims at exploring the views of academics and practitioners about the goals of regression testing. The purpose is to investigate the commonalities and differences in their viewpoints and defining some common goals for the success of regression testing.

Method: We conducted a focus group study, with 7 testing experts from industry and academia. 4 testing practitioners from 2 companies and 3 researchers from 2 universities participated in the study. We followed GQM approach, to elicit the regression testing goals, information needs, and measures.

Results: 43 regression testing goals were identified by the participants, which were reduced to 10 on the basis of similarity among the identified goals. Later during the priority assignment process, 5 goals were discarded, because the priority assigned to these goals was very low. Participants identified 47 information needs/questions required to evaluate the success of regression testing with reference to goal G5 (confidence). Which were then reduced to 10 on the basis of similarity. Finally, we identified measures to gauge those information needs/questions, which were corresponding to the goal (G5).

Conclusions: We observed that participation level of practitioners and researchers during the elicitation of goals and questions was same. We found a certain level of agreement between the participants regarding the regression testing definitions and goals. But there was some level of disagreement regarding the priorities of the goals. We also identified the need to implement a regression testing evaluation framework in the participating companies.

Keywords: Regression testing, Regression testing goals, GQM, Focus group

I. Introduction

A close collaboration between industry and academia is important to both sides, and this collaboration should be based on similar views of the studied problems and their importance [18]. Setting common goals and achieving a shared understanding is important for successful industry-academia collaboration. Having consensus on goals for collaborative research is a real challenge [11].

The key constraint of regression testing is the maintenance of the regression test suite (adding new test cases or updating or deleting obsolete test cases) [20], [26]. Test suite maintenance is not an easy task and if not done in a correct manner, utility of the test suite will be decreased and associated risks will be amplified [12]. To measure the success of regression testing, we need to define the regression testing goals. Chernak [4] emphasizes that test suite evaluation is the basis for the improvement of the overall testing process.

In earlier work Engström et al. [7] investigated regression testing practices and challenges using the focus group meeting and an online questionnaire with the industry practitioners. We complement these findings by exploring the value for practitioners and researchers alike. The objective is to reflect on how to evaluate regression testing. By choosing the right measures for the goals of a successful regression testing.

From the EASE (Embedded Applications Software Engineering) project platform, together with the testing practitioners, we identified 7 software testing challenges in 3 companies. These companies operate in mobile-communications, surveillance, and embedded software systems. To identify the testing challenges at the companies, we utilized the SERP-test taxonomy. The SERP-test is designed to support the industry-academia collaboration [6]. The identified challenges were related to test planning, test design, and test execution. Out of these challenges, 3 were related to regression test selection, regression test prioritization, and test suite minimization. With the consultation of companies' representatives, we find that companies were more interested to cope with the regression testing challenges. This study is a step forward in the identified direction, with a focus on understanding the regression testing goals. To determine the scope of the study, we have formulated the following research question:

RQ: *What are the views of academics and practitioners about regression testing?*

We conducted a focus group study with industry and academic participants. 7 experts participated in the study. Among the participants, 4 were representatives of testing practitioners from 2 large companies, and 3 were senior researchers from 2 universities. The contributions of this study could be listed as, a) regression testing definition, b) success goals, c) information needed (questions) to evaluate the success and d) measures to answer the questions.

The reminder of this paper is structured as follows: Section II presents the related work, Section III presents the detail about the

978-1-5386-2649-8/17 $31.00 © 2017 IEEE
DOI 10.1109/APSECW.2017.23

methodology (i.e. planning, design, and conduct of the focus group). Threats to validity have been discussed in Section IV. Study results have been discussed in Section V, and conclusions on key findings have been presented in Section VI.

II. Related Work

Researchers believe that industry-academia collaboration in software engineering is very low [9], [10], [11]. Garousi et al. [9] emphasize the importance of collaboration between industry and academia to support improvement and innovation in the industry. Ramler et al. [21], suggest the collaboration between industry and academia for the improvement and innovation of software testing in the industry. This collaboration could be the basis for transferable and empirically evaluated results. To facilitate the collaboration between industry and academia, Engström et al. [6] proposed a taxonomy of testing. The taxonomy can assist to improve communication between practitioners and researchers. It can work for both types of communication (i.e. direct communication and indirect communication).

Kapfhammer [14] pointed out the limited adoption of regression testing techniques, the reason identified is the lack of empirical evaluations. Chernak [4] stresses the importance of test suite evaluation as a basis for improving the test process. Chernak emphasizes that objective measures should be defined and built into the testing process to improve the overall quality of testing. Rothermel & Harrold [22], [23], proposed a 5 step framework to evaluate the regression testing techniques.

Engström et al. [8] suggested that more empirical evaluations conducted in industrial settings are required to facilitate the adoption of RT research in practice. The authors concluded that in order to enable practitioners to utilize the outcomes of research on testing, these outcomes must be evaluated in the actual environment. Through a focus group and an online questionnaire, Engström & Runeson [7] conducted a survey on regression testing practices, authors investigated what is considered regression testing by practitioners i.e. the definition, purpose and scope of it. They further investigated the challenges practitioners face with respect to regression testing. Our work complements the results of [7], as our subjects are the representatives of both sides (i.e. industry and academia). It is comparatively more focused, as purpose was to identify the regression testing goals.

We conducted an exploratory study to systematically elicit the goals, information needs and measures. We are focusing on industry-academia collaboration within regression testing challenges. The current focus is on regression test suite evaluation, as the first step in this study we tried to establish the regression testing goals.

III. Methodology

Focus groups are used to acquire the viewpoints of a group on some defined topic, which is a common area of interest for all group members. The key role in the focus groups is the moderator, who is responsible for guiding, facilitating and making sure that the discussion stays focused. Different guidelines are available for focus groups, Kontio et al. [15], [16] have deduced software engineering specific guidelines for conducting focus groups. Our approach to conducting the focus group was aligned with [15], a brief description about each step is given in the following sub sections.

A. Planning the research.

It is essential to make sure, that the focus group is suitable for the planned work [15]. Considering the research question presented in Section I, our intention was to know the viewpoints of academics and practitioners about regression testing. Focus group was selected as it facilitates discussion, immediate reflection and it helps find the depth of the problem and some potential ideas for future research. As part of planning, we acquired the informed consent of the participants. We did also inform all participants about the purpose of the activity.

B. Designing the focus groups.

Focus group can comprise 3 to 12 participants, but a suitable number is between 4 and 8 [15]. We invited 7 participants from 2 Sweden based companies and 2 Swedish universities. Among the invited participants, 4 were testing practitioners from the companies (2 from each). 3 participants were senior academics from 2 universities. It is important to mention that all 3 academics are actively involved in software testing research. A brief description of the participants is shown in Table I.

We followed the GQM approach for the focus group. GQM is an established method for planning and executing software engineering research and capturing software engineering related phenomena [3]. We phrased the questions using the interview guide formulated by Petersen et al. [19]. Table II shows the GQM template for the evaluation of regression testing, the template is divided into 5 activities (i.e. A1. A2, A3, A4, & A5). The purpose of A1 and A2 was to identify and prioritize the regression testing goals respectively, whereas A3 was to elicit the information needs (questions) corresponding to the identified goals. A4 was to capture the measures that could be used to answer the questions of related goal(s), while the objective of A5 was to know about the measures that the industry experts are actually using for the evaluation of test suites.

TABLE I
Focus Group Participants

P.Id.	Organization	Participant's Expertise
P1.	Sony Mobile Communications	Testing/ Development
P2.	Sony Mobile Communications	Testing/Development
P3.	Axis Communications	Testing/Development
P4.	Axis Communications	Testing
P5.	Blekinge Institute of Techology	SE & Testing Research
P6.	Lund University	RE & Testing Research
P7.	Lund University	Regression Testing Research

TABLE II
GQM-Template for Evaluation of Regression Testing

GQM Activity	Question	Rational
A1	Regression Testing is successful when a), b), c)... Complete the sentence (e.g. Regression testing is successful if it is a) efficient.)	Capture the goals
A2	Which success factors/goals are most important to you? Prioritize.	Prioritize success factors and goals and hence determine which measures should really be collected and whether this matches to what is collected today.
A3	What information is needed to evaluate success? Formulate as a question (e.g. How complex are our test cases, How many test cases are we running in a given test period?)	Capture the information needs (questions)
A4	What measures do we need to collect to answer the questions? (e.g. #test-steps for complexity)	Capture the measures that allow to quantify (and hence automate) the analysis of results
A5	What are you collecting today (measurements) of what has been identified in #4	Learn about input we already have available for evaluating test suites

C. Conducting the focus group session.

A focus group may last for 2 to 3 hours and it should have a predefined schedule. Within one session, the number of issues to be focused should be limited so that participants can have sufficient time to give their opinion on every aspect of the topic [15]. We allocated 2 hours for the activity, 30 minutes were assigned for setting up the focus group environment and introducing the basic purpose to the participants, although the overall objective was already communicated. We used the following schedule in the focus group:

1) Introduction: Short introduction to the goals of the workshop.
2) Round-the-table: What is regression testing in your view? Describe in one to 2 sentences.
3) Summarizing, presenting and verifying.
4) GQM-Activity (Table II).
5) Summarizing, presenting and verifying (after every GQM, i.e. A1....A5).
6) Closing (Any other reflection or discussion points? Next steps).

We used color stickers (green and yellow) for data collection, green stickers were used by the practitioners and yellow stickers were used by the researchers. Discussion points were recorded by 2 of the authors. We took several breaks in between to collect the answers (to gather the sticky notes), cluster similar answers, put logical labels on clusters. Reflect on the names of the clusters and also whether individual sticky notes belong in it. Finally, we presented the initial results and asked the participants to verify the labels according to their given options.

D. Analyzing the data and reporting the results.

We followed the inductive approach for data analysis. It is a systematic approach for analyzing qualitative data [25], [17].

According to Thomas [25],*"inductive analysis refers to approaches that primarily use detailed readings of raw data to derive concepts, themes, or a model through interpretations made from the raw data by an evaluator or researcher"*.

The inductive approach allows the findings to emerge from the raw data without imposing any restrictions, the approach revolves around 3 steps: 1) data reduction, 2) data visualization and 3) conclusions and verifications .

We collected the sticky notes from the participants and made the groups of the responses along with the labels (reduction). We displayed the results to the participants and asked them to verify the labels with reference to their options. For example, we received the 43 options for regression testing goals, we reduced the options to 10 by making the clusters of the options on the basis of similarities. After the clustering of the data, results were displayed and the participants were invited to verify the labels according to their given options. In the second phase, together with the authors, results were reviewed by all participants in a separate review meeting, resultantly identified anomalies were fixed in the results.

The inductive approach provided us with the required flexibility to understand the viewpoints of the experts. The outcomes of focus group study are presented in Section V.

IV. Threats to Validity

This study presents the viewpoints of academics and practitioners about the goals, information needs and measures of regression testing. The results presented here are of an exploratory nature. We addressed the threats to validity according to guidelines of Runeson and Host [24].

Construct Validity: This aspect of validity is regarding the underlying operational measures, concepts and terms of the study. One potential threat to construct validity for our study is the subjects of the study representing 2 different working environments (i.e. academics and industry). Potentially they can have different understanding of concepts and terms. To mitigate the threats to this aspect of validity, we started with the exploration of the perception of participants about regression testing. To ensure the common understanding about the concept and terms during the entire focus group meeting.

Internal Validity: This aspect of validity threat is important if causal relations are examined. Generally, we can state that studying causal relationships was not in the scope of this study. It is a descriptive/interpretive study, as it presents the viewpoints of the participants. We created a mapping between information needs and measures, that is the only causal relationship presented in the study. The mapping created between information needs and measures requires empirical evaluation to determine the validity of relationships between information needs and measures.

External Validity: This aspect of the validity refers to the generalization of findings. We selected subjects of the study from academics and industry, to ensure the acceptability of results for both communities (i.e. practitioners and researchers). But as the practitioners were representing only 2 companies, so acceptability of results cannot be ensured in all companies working in the field of telecommunication. Further analytical generalization of results is possible, to support this we have reported the information of the participants in Table I.

Reliability: To ensure the reliability of the study, we triangulated the results, as we presented and verified the initial results to the participants during the focus group meeting. Later after the complete analysis, results were presented to all participants in a review meeting. For detail please refer to the Section III-D. Goals and measures identified in this study have not been verified through actual implementations.

V. Results and Analysis

A. Defining Regression Testing.

As a kick-off activity, we asked the experts to give their opinion about, *[What is regression testing in your view?]*. The purpose was to elicit the definition of regression testing with respect to participants' perception/experience. 5 out of 7 people came up with their definitions, presented in Table III. Here an interesting fact that can be drawn from the individual definitions is the agreement between the views of academics and practitioners. We find that, the definitions presented at S.No. 1, 2 and 5 are almost the same and could be grouped together. Similarly, definitions at 3 and 4 are on same lines and we can group these 2 as well. After collecting the 5 definitions, we presented the definitions to the participants. Participants were agreed with our grouping scheme i.e. to group 1,2, & 5 and 3 & 4 in the form of the following 2 definitions:

1) *Regression testing is an activity which gives us confidence in what we have done and a trust that we have not broken anything.*
2) *Regression testing is an activity which makes sure that everything is working correctly after the changes in the system and it is a guarantee to continue in future.*

TABLE III
Defining Regression Testing

S.No.	Perspective	Definition
1.	Academic	Make sure that we have not broken anything.
2.	Academic	Trust on what you have done
3.	Industry	To make sure that everything else work correctly
4.	Industry	To make future work possible, it is a guarantee to continue in future
5.	Industry	Trust on what you have done and make sure that we have not broken anything

1) Regression Testing Definitions Presented in the Literature: We selected 3 definitions from the literature to compare with the definitions presented by our experts. First definition was selected from a study presented by Engström and Runeson [7]. We selected this definition as it represents the practitioners' perspective and it could be regarded closer to our concept. Second and third are the standard definitions taken from IEEE software Engineering terminology [5], and IEEE, Systems and software engineering – vocabulary [13] respectively.

1) *“Regression testing involves repetitive tests and aims to verify that previously working software still works after changes to other parts. Regression testing shall ensure that nothing has been affected or destroyed” [7].*

2) *“Regression testing is defined as retesting a system or component to confirm that changes cannot introduce any new bugs or causes other code errors, and the system or component can still follow the prescribed requirement specification” [5].*

3) *“Regression testing is the selective retesting of a system or component to verify that modifications have not caused unintended effects and that the system or component still complies with its specified requirements” [13].*

We observed that the definitions finalized in our study are closer to the definition presented by Engström and Runeson [7]. The distinctive factor of the definition proposed in our study is that it presents the viewpoints of both practitioners and researchers, while Engström's definition presents the viewpoints of practitioners only. On the other hand IEEE standard definitions is about that after the modification modified system or component still conforms to the specified requirements. That is system or component still works correctly after the changes. Our second definition conforms with the standard definitions. If we look at the individuals' definitions presented in Table III 3 words (*make sure, guarantee, and trust*) are prominent. This indicates that through regression testing our experts are seeking some assurance about the system's correctness, a guarantee that future work is possible and a trust on what they have done. Moreover, the results indicate that practitioners and researchers have similar viewpoints on the definition of regression testing that addresses one of the challenges highlighted in Section II.

2) Regression Testing Definition Adopted: During the second phase of the study (i.e. presentation of results and obtaining feedback from the participants), it was decided to adopt a single definition for the study. The agreed upon opinion was to merge the two definitions into a single definition in a way that it should represent the viewpoint of all participants. Later on, we combined the both definitions and created the following definition:

> *Regression testing is an activity which makes sure that everything is working correctly after the changes to the system. It builds the trust, that nothing has broken in the system and it guarantees to continue work in the future.*

B. GQM Activity.

To execute the GQM (Goal, Question, Metric) theme, we used the GQM template, the template of questions used here are the inspiration from [19]. We divided this activity as A1, A2, ..., A5 as listed in Table II. The purpose of A1 was to elicit the goals and A2 was to prioritize the goals. A3 was for the elicitation of information needs (questions) to achieve the regression testing goals. With the A4 we intended to collect measures for answering the questions related to information needs. Finally, with the A5 intention was to know about the measures that are currently used by the practitioners. The concept followed in the study is represented in Figure 1.

1) A1-Goals Identification: To identify the goals, participants were asked to complete this, *[Regression Testing is successful when a), b), c) ?]*. Here a), b), c) indicate that the participants can state more than one goal regarding the success of regression testing. A total of 43 different options for goals were identified by the experts, we find that majority of the options were similar. For example we had the following options for G1 (Fault slippage to the customer):

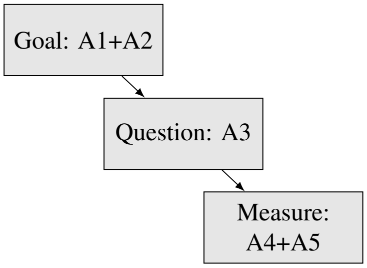

Fig. 1. GQM Representation

1) No fault slippage through.
2) The customer/user finds no further bugs/flaws.
3) No issue leakage.
4) Ensure that the system fulfills the desired properties and no stopping bug slipped away.
5) We have no issue leakage on release.

With the consent of participants, we decided to group the identified goals on the basis of similarities and assign an appropriate label for each group. Hence the identified goals were restricted into 10 regression testing goals. The final list of the goals along with the description about each goal is shown in Table IV.

There are some goals identified by the participants, which are either irrelevant or too generic in scope. For example, visual overview could be taken as irrelevant or too broad in scope. Similarly, automation could be subsumed in efficiency. Achieving desired pass fail rate has been highlighted by 4 participants, if we see the goal description it can be subsumed by the effectiveness goal. It is important to highlight that visual overview and automation were identified by only one participant.

Confidence, Efficiency, and Effectiveness are the goals identified by the majority of participants. Here it is important to mention that a goal identified by more participants does not refer to its importance, rather it only shows how may subjects have pointed out a particular goal. G5 (i.e. confidence) was identified by all 7 participants, but with varying descriptions. For example, some of the perceptions can be summarized as, *"Stakeholders are confident with the reached quality and/or we can ensure that nothing is broken."* To measure the desired quality or to determine that nothing is broken, requires multiple testing metrics.

2) A2- Goals Prioritization: Next task was to assign the priority order to the elicited goals. The question asked to the participants was, *[Which success factors/goals are most important to you? Prioritize]*. The participants were asked to assign priorities against every goal, each participant was given 10 points to prioritize. We used colored markers for priority assignment, red for researchers and Black for practitioners. As the distribution of experts was not equal on both sides (i.e. 3 researchers and 4 practitioners), we decided to normalize the priority of both sides. For normalization we devised an equation presented at (1).

$$NP = AP/N * 4 \tag{1}$$

Here NP = Normalized Priority, AP = Actual Priority and N = No. of Experts.

The normalized priorities along with the total points are shown in Table V. G5 *(i.e. Confidence)* was marked with 21 total points, G2 *(i.e. High precision)* was given 17 points while G1 *(i.e. Fault slippage to customer)* was third in the list with 14 points. It was observed that in most cases there was a sort of agreement between researchers and practitioners. But there was a complete disagreement regarding the priority of some goals. We can see that for researchers

TABLE IV
REGRESSION TESTING GOALS

G.Id.	Options	Goal	Goal description
G1.	5	Fault slippage to customer	The customer/user finds no further bugs/flaws
G2.	3	High precision	Non affected test cases excluded
G3.	3	Inclusiveness	Have run the most effective tests first
G4.	5	Achieved desired coverage	All desired modules have been executed when a regression test suite runs
G5.	7	Confidence	Stakeholders are confident with the reached quality and/or We can ensure that nothing is broken
G6.	7	Efficiency	Finished retesting with the limited time and low cost
G7.	7	Effectiveness	Costly faults detected early and/or finding new defects in old code
G8.	1	Visual overview	Visualization of complete software is displayed
G9.	1	Automation	Being able to automate
G10.	4	Achieving desired pass fail rate	When the expected tests pass and/or fail

G1 & G5 are the highest priority goals with equal priority, whereas for Practitioners G5 is the highest priority. Similarly, G8 and G9 are somewhat important for practitioners but researchers assigned *zero* to both the goals. An interesting fact, that we think is important to mention here is that the participants on both sides marked *zero* priority for G7 (i.e. effectiveness). Although this goal was identified by all 7 participants. And it is among the goals which have been cited in the literature by different authors [2], [4]. We found similarity in views of both sides, regarding the top 3 goals (i.e G5, G2, & G1 respectively). As G5 **(Confidence)** was ranked as the highest priority goal by the participants, and considering its generic nature we decided to elicit the information needs for G5, in the next phase of the focus group (i.e. A3). During the final review meeting participants were agreed to consider G1, G2, G3, G5, & G6 as the final list of goals.

TABLE V
ALLOCATED PRIORITIES TO THE GOALS

G. Id	A Priority	I Priority	Total Priority
G1.	9	5	14
G2.	8	9	17
G3.	4	3	7
G4.	3	0	3
G5.	9	12	21
G6.	7	3	10
G7.	0	0	0
G8.	0	5	5
G9.	0	3	3
G10.	0	0	0

3) A3-Questions (Information Needs Elicitation): To elicit questions (information needs), participants were asked to answer the question, *[What information is needed to evaluate the success?]*. We decided to elicit information needs only for G5 (i.e. confidence), we took the decision because of the following reasons:

1) Because of the generic nature of the goal.
2) It was ranked as the highest priority goal by the participants.
3) It was highlighted that, to achieve this goal multiple metrics need to be evaluated.

47 questions (information needs) were identified by the participants. During analysis, we find that a majority questions are similar. On the basis of identified similarity, we grouped these 47 questions (information needs) into 10. The final list of information needs questions is shown in Table VI. The questions with most options were, *Have critical parts been covered? (16 options)*, and *What are the test outcomes? (10 options)*. A majority of information needs listed in Table VI are quantifiable, but some information needs are relatively generic in nature. Information need listed at Q2 (Team Experience) cannot be taken as a direct testing metric, but it is important with regard to confidence. Similarly, Q6 (Confidence perception) is not a specific metric, still it can affect the measure of other factors. Product characteristics listed as Q9 can determine the complexity of the product, this can also affect confidence perception. We can draw a correlation between Q2, Q6, and Q9. After finishing with the clustering, the final list of grouped information needs was presented to the participants for the verification of the clusters. Later in the results review meeting, all the stakeholders were agreed to consider Q1, Q3, Q4, Q5, and Q7 as the final list of information needs to achieve the confidence goal. Participants were agreed about the subjective importance of Q2 and Q7 with respect to the underlying goal of confidence.

TABLE VI
G5. QUESTIONS (INFORMATION NEEDS)

Q.Id.	Question	Extracted from
Q1.	What are the changes to the system?	5 similar options
Q2.	What is the experience of development and testing?	3 similar options
Q3.	Have critical parts been covered?	16 similar options
Q4.	Have modifications been tested?	5 similar options
Q5.	What are the test outcomes?	10 similar options
Q6.	What is the perception of team about confidence?	3 similar options
Q7.	What is the nature of defects in the product?	2 similar options
Q8.	What is the balance between pass fail?	1 option
Q9.	What is the complexity of the product under test?	1 option
Q10.	What has been blocked by the tests?	1 option

4) A4-Measures Identification: The aim here was to identify the suitable measures to collect the information needs and ultimately achieve the defined goal (i.e. Confidence). We asked, *[What measures do we need to collect to answer the questions?]*. Our experts identified 5 measures presented in the Table VII. Later together with the experts we started a brainstorming activity to find the possible mapping between the questions and measures. We carried the activity in a step wise manner. That is, for every single goal we asked the experts to map it with possible measure(s). 4 measures (i.e. M1,M2,M3, & M4) were mapped to 7 questions (i.e. Q1, Q3, Q4, Q5, Q7, Q8, and Q10). The finalized GQM (goal-question-measure) mapping is shown in Figure 2.

TABLE VII
MEASURES

MID	Measure
M1	#LOC changed
M2	#Defect fixes from test
M3	#Changed LOC covered
M4	#defect history/change
M5	#Affected Non-changed LOC/Modules

5) A5-Available Measures: The last activity was to know about the actual measures that are being used in the companies. We asked the question, *[What are you collecting today? (measurements) of what*

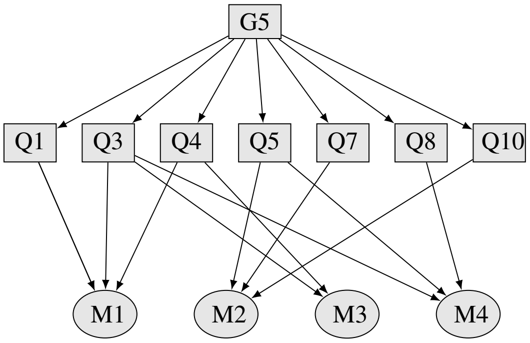

Fig. 2. Goal-Question-Measure Mapping

has been identified in A4]. This question was to know about the actual state of the implementation of measurement program regarding the evaluation of regression testing in the industry. Therefore we asked practitioners to answer this question. We requested the researchers to sit as observers and provide their feedback on the activity. Practitioners expressed, that they do not use any explicit measures. There is no predefined measurement mechanism regarding the evaluation of regression testing that could be used. Instead, they rely on their experience, to evaluate the success. Their agreed-upon statement about the measurement was, it is a gut feeling, that we have tested enough and we are successful. To further continue and to come up with some substantial outcome, we added another question.

PQ: Do, we actually need to evaluate the success of regression testing?

We asked the participants to provide their opinion about the need for measuring the regression testing. There was a consensus among the participants about the importance of measuring the success of regression testing. It was emphasized that suitable measures need to be defined and used in the companies. It was also highlighted, that participating companies are interested to implement an evaluation mechanism/framework to measure the success.

6) Related Measures Presented in the Literature: To make a way towards the identification/implementation of evaluation framework, we decided to identify the measures from literature and test the identified measures in the partner companies. As a starting point we identified some measures from literature to further strengthen our findings.

Rothermel and Harrold [22], [23] presented a complete framework for the evaluation of the regression testing techniques. They suggested *inclusiveness, precision, efficiency, generality, & accountability* as measures to evaluate the regression testing. Horváth et al. [12] used *code coverage & partition metrics* for measuring fault detection capability and fault localization. They defined coverage metric (Cov) as a ratio of the number of procedures in a code group P that are covered by test group T. Whereas they defined partition metric (Part) to express the average ratio of procedures that can be distinguished from any other procedures in terms of coverage. *output uniqueness* is defined by Alshahwan and Harman [1], who define the output uniqueness as if the 2 test cases yield different kinds of output. The authors believe that this metric can help in effective fault detection capability, it also works for fault finding consistency.

Vidacs et al. [26] uses the code coverage, efficiency & uniqueness for Assessing the Test suites of large system. The authors argue that better coverage or partitioning can be achieved using more test cases, provided test cases are different. But, in case if such test cases are added to the test suite, which covers the same code, they will increase the test suite size possibly with little additional benefit. They suggested measuring the efficiency, that (refer to the relative number of test cases in test suite), to measure efficiency, they defined *coverage efficiency* (EFFCOV) and *partitioning efficiency* (EFFPART). Coverage efficiency refers to the average number of procedures covered by a test case, while partitioning efficiency shows that on average, how much a single test contributes to the partitioning capability of whole functional unit. To measure uniqueness authors used 2 metrics (*specialization* metric SPEC and *uniqueness* metric UNIQ). SPEC shows how specialized a test group is to a code group, while the UNIQ metric measures what portion of the covered elements is covered only by a particular test group.

To measure the effectiveness, Chernak [4] named his measure as *defect*, which is the ratio between the number of defects covered by a test suite to the number of defects missed by the test suite. Athanasiou et al. [2] argued that test code quality has 3 dimensions completeness, effectiveness, and maintainability. They defined *assertion density* as a measure of calculating the effectiveness of test code to detect the defects. For the effectiveness of test code authors also suggested *directness* as measure, they defined directness as it measures the extent to which the production code is covered directly. Test suite evaluation metrics and corresponding measures selected from literature are presented in Table VIII.

TABLE VIII
MEASURES FOUND IN LITERATURE

S.No.	Metric	Measure	Reference
1.	Effectiveness	Defects, TestCaseEscaps, AssertionDensity, Directness	[2], [4]
2.	Fault Detection Capability	CodeCoverage, OutputUniqaueness	[1], [2], [12], [26]
3.	Fault localiztion	Partition	[12]
4.	Effeciency	EffCov, EffPart	[26]
5.	Uniqueness	UNIQ, SPEC	[26]

7) Mapping between Focus Group and Literature Findings: As we mentioned already, due to the time constraint, we investigated only one goal (G5) in the subsequent steps of the focus group session. Therefore we decided, to create a mapping between the goals presented in the Table IV, and metrics/measures we find in the literature. Majority goals listed in the Table IV are specific and measurable. Measures presented in the literature can be mapped to identified goals. For instance, G1 can be mapped to the metric “Fault detection capability” , related measures have been discussed in the following studies [1], [2], [12]. G2 & G3 can be mapped to the metrics “precision ” and “inclusiveness” defined in [22], [23]. Similarly, G6 can be linked to the metric “Efficiency” presented in [22], [23], [26]. Finally, G7 can be mapped to “effectiveness” metric discussed in [2], [4].

The measures identified from literature can also be mapped to some of the questions listed in Table VI. For example, Q5 could be mapped to *No. of Defects, TestCaseEscaps, & OutputUniqueness,* similarly Q7 can be mapped with *No. of Defects & CodeCoverage,* Q3 can be mapped with *AssertionDensity & Directness.*

VI. CONCLUSIONS

The study presented the definition of regression testing. The distinguishing factor of the presented definition is that it represents the viewpoint of testing experts from industry and academia. This study also presented the viewpoints of the participants about the goals of regression testing. Results were obtained in a focus group study, GQM approach was followed to elicit information from the participants. 10 regression testing goals were identified, which were further restricted to 5 goals by the participating experts after the priority assignment. G5 was regarded as the highest priority goal. We identified 10 questions (information needs) required to achieve the goal (G5). We also identified 5 measures, later 4 measures were mapped to the 7 questions (information needs).

Every participant contributed her/his share in the study and showed a commitment till the end. The contribution of experts from both sides (i.e. researchers and practitioners) during the GQM activity, was equal. We observed some level of agreement between the practitioners and researchers regarding the regression testing definition and goals. There were some points of disagreement as well, regarding the priority of goals. It has been revealed that at present, practitioners (participating) are not using explicit measures to evaluate success. Rather they rely on their experience to guess that they have tested enough and they are successful.

The study concludes, that by using such platforms industry academia can come closer to each other. Such collaborations can help in defining the concepts acceptable for both communities. This kind of studies can help the researchers to work on actual industrial problems. Resultantly, practitioners could be able to cope with the real challenges with the help of research.

This study identifies the need for test suite evaluation framework/mechanism for regression testing in a real industrial setting. We will continue with identification and implementation of evaluation measures, in collaboration with the partner companies and ultimately an acceptable evaluation framework would be identified and implemented.

ACKNOWLEDGMENT

The authors would like to thank the researchers from Lund University, Lund, Sweden, and Blekinge Institute of Technology, Karlskrona, Sweden for their participation in the focus group meeting and subsequent sessions of the study. Our sincere gratitude to the practitioners from Sony Mobile Communications AB and Axis Communications AB, who also participated in all sessions of the study. We are thankful to all the facilitators who helped us in organizing the focus group study. This work was funded by the project EASE (Embedded Applications Software Engineering, ease.cs.lth.se).

REFERENCES

[1] Nadia Alshahwan and Mark Harman. Coverage and fault detection of the output-uniqueness test selection criteria. In *Proceedings of the International Symposium on Software Testing and Analysis 2014*, pages 181–192. ACM, 2014.

[2] Dimitrios Athanasiou, Ariadi Nugroho, Joost Visser, and Andy Zaidman. Test code quality and its relation to issue handling performance. *IEEE Transactions on Software Engineering*, 40(11):1100–1125, 2014.

[3] VRBG Caldiera and H Dieter Rombach. The goal question metric approach. *Encyclopedia of software engineering*, 2(1994):528–532, 1994.

[4] Yuri Chernak. Validating and improving test-case effectiveness. *IEEE software*, 18(1):81–86, 2001.

[5] IEEE Standards Coordinating Committee et al. Ieee standard glossary of software engineering terminology (ieee std 610.12-1990). los alamitos. *CA: IEEE Computer Society*, 1990.

[6] Emelie Engström, Kai Petersen, Nauman bin Ali, and Elizabeth Bjarnason. Serp-test: a taxonomy for supporting industry–academia communication. *Software Quality Journal*, pages 1–37, 2016.

[7] Emelie Engström and Per Runeson. A qualitative survey of regression testing practices. In *Proceedings of the International Conference on Product Focused Software Process Improvement*, pages 3–16. Springer, 2010.

[8] Emelie Engström, Per Runeson, and Mats Skoglund. A systematic review on regression test selection techniques. *Information and Software Technology*, 52(1):14–30, 2010.

[9] Vahid Garousi, Matt M Eskandar, and Kadir Herkiloğlu. Industry–academia collaborations in software testing: experience and success stories from canada and turkey. *Software Quality Journal*, pages 1–53, 2016.

[10] Vahid Garousi and Kadir Herkiloglu. Selecting the right topics for industry-academia collaborations in software testing: an experience report. In *Proceedings of the IEEE International Conference on Software Testing, Verification and Validation (ICST), 2016*, pages 213–222. IEEE, 2016.

[11] Vahid Garousi, Kai Petersen, and Baris Ozkan. Challenges and best practices in industry-academia collaborations in software engineering: A systematic literature review. *Information and Software Technology*, 79:106–127, 2016.

[12] Ferenc Horváth, Béla Vancsics, László Vidács, Árpád Beszédes, Dávid Tengeri, Tamás Gergely, and Tibor Gyimóthy. Test suite evaluation using code coverage based metrics. pages 46–60, 2015.

[13] IEC ISO. Ieee, systems and software engineering–vocabulary. *ISO/IEC/IEEE 24765: 2010 (E)) Piscataway, NJ: IEEE computer society, Tech. Rep.*, 2010.

[14] Gregory M Kapfhammer. Empirically evaluating regression testing techniques: Challenges, solutions, and a potential way forward. In *Proceedings of the IEEE Fourth International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 2011*, pages 99–102. IEEE, 2011.

[15] Jyrki Kontio, Johanna Bragge, and Laura Lehtola. The focus group method as an empirical tool in software engineering. In *Guide to advanced empirical software engineering*, pages 93–116. Springer, 2008.

[16] Jyrki Kontio, Laura Lehtola, and Johanna Bragge. Using the focus group method in software engineering: obtaining practitioner and user experiences. In *Proceedings of the International Symposium on Empirical Software Engineering, ISESE'04. 2004.*, pages 271–280. IEEE, 2004.

[17] Lisha Liu. Using generic inductive approach in qualitative educational research: A case study analysis. *Journal of Education and Learning*, 5(2):129, 2016.

[18] Satoshi Masuda. Software testing in industry and academia: A view of both sides in japan. In *Proceedings of the IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 2017*, pages 40–41. IEEE, 2017.

[19] Kai Petersen, Cigdem Gencel, Negin Asghari, and Stefanie Betz. An elicitation instrument for operationalising gqm+ strategies (gqm+ s-ei). *Empirical Software Engineering*, 20(4):968–1005, 2015.

[20] Leandro Sales Pinto, Saurabh Sinha, and Alessandro Orso. Understanding myths and realities of test-suite evolution. In *Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering*, page 33. ACM, 2012.

[21] Rudolf Ramler, Michael Felderer, Takashi Kitamura, and Darko Marinov. Industry-academia collaboration in software testing: An overview of taic part 2016. In *Proceedings of the IEEE Ninth International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 2016*, pages 238–239. IEEE, 2016.

[22] Gregg Rothermel and Mary Jean Harrold. A framework for evaluating regression test selection techniques. In *Proceedings of the 16th International Conference on Software Engineering, ICSE-16., 1994.*, pages 201–210. IEEE, 1994.

[23] Gregg Rothermel and Mary Jean Harrold. Analyzing regression test selection techniques. *IEEE Transactions on software engineering*, 22(8):529–551, 1996.

[24] Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2):131, 2009.

[25] David R Thomas. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246, 2006.

[26] László Vidács, Ferenc Horváth, Dávid Tengeri, and Árpád Beszédes. Assessing the test suite of a large system based on code coverage, efficiency and uniqueness. In *Proceedings of the IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER), 2016*, volume 2, pages 13–16. IEEE, 2016.