Available online at www.sciencedirect.com

ScienceDirect

Information and Software Technology 51 (2009) 150–161

INFORMATION AND SOFTWARE TECHNOLOGY

www.elsevier.com/locate/infsof

Improving the effectiveness of root cause analysis in post mortem analysis: A controlled experiment

Finn Olav Bjørnson [a,*], Alf Inge Wang [a,1], Erik Arisholm [b,c,2]

[a] Norwegian University of Science and Technology, Department of Computer and Information Science, Sem Sælandsvei 7-9, 7491 Trondheim, Norway
[b] Simula Research Laboratory, Department of Software Engineering, P.O. Box 134, 1325 Lysaker, Norway
[c] Department of Informatics, University of Oslo, P.O. Box 1080, Blindern, N-0316 Oslo, Norway

Received 15 August 2007; received in revised form 17 January 2008; accepted 7 February 2008
Available online 19 February 2008

Abstract

Retrospective analysis is a way to share knowledge following the completion of a project or major milestone. However, in the busy workday of a software project, there is rarely time for such reviews and there is a need for effective methods that will yield good results quickly without the need for external consultants or experts. Building on an existing method for retrospective analysis and theories of group involvement, we propose improvements to the root cause analysis phase of a lightweight retrospective analysis method known as post mortem analysis (PMA). In particular, to facilitate brainstorming during the root cause analysis phase of the PMA, we propose certain processual changes to facilitate more active individual participation and the use of less rigidly structured diagrams. We conducted a controlled experiment to compare this new variation of the method with the existing one, and conclude that in our setting of small software teams with no access to an experienced facilitator, the new variation is more effective when it comes to identifying possible root causes of problems and successes. The modified method also produced more specific starting points for improving the software development process.

Keywords: Retrospective method; Software process improvement; Controlled experiment; Knowledge management; Post mortem analysis

1. Introduction

In today's software engineering industry, it is critical to improve software development processes. In this context, one lesson that may be learned from general efforts to improve processes, such as total quality management and standardisation, is that the ability to learn from past success and failure is a central factor for success [8]. Learning from the past involves knowledge management, or creating a "learning software organisation", which is defined by Dybå [11] as "A software organisation that promotes improved actions through better knowledge and understanding".

Keegan and Turner [16] claim that, in general, software development is conducted at too fast a pace. In 2001, they performed a study on project-based learning practices in 19 European software development companies. They found that project team members frequently did not have time for meetings to review lessons learned. Where recommended process models did exist, these were seldom used. In an editorial in IEEE software in 2002, Glass [14] stated that the software engineering field is so busy that there is rarely time to think of how development could go better, not just faster. He further claimed that companies should pause from time to learn the lessons they had been through. He recommended reviewing performances on completed projects (project retrospectives) as a good way of learning.

There is a principle in agile software development that states that "At regular intervals, the team reflects on how

[*] Corresponding author. Tel.: +47 73 59 3440; fax: +47 73 59 4466.
E-mail addresses: Finn.Olav.Bjornson@idi.ntnu.no (F.O. Bjørnson), Alf.Inge.Wang@idi.ntnu.no (A.I. Wang), erika@simula.no (E. Arisholm).
[1] Tel.: +47 73 59 3440; fax: +47 73 59 4466.
[2] Tel.: +47 67 82 8200; fax: +47 67 82 8201.

0950-5849/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.infsof.2008.02.003

to become more effective, then tunes and adjusts its behaviour accordingly." [1]. In accordance with this principle, iterative and light retrospective sessions have been suggested for use in agile projects [4,9,18]. Myllyaho et al. state that the small teams and short iterations of extreme programming will affect how retrospective workshops can be conducted [20]: "The workshops needs to be short and effective, i.e., not taking too much effort from the project team, yet yielding immediate and visible outcomes to motivate the project team for further such activity."

In this paper, we take as our starting point an existing, lightweight retrospective method known as the post mortem analysis (PMA) [2]. We propose a modified method that exploits theories on brainstorming and group performance combined with the notation of causal maps. The effectiveness of the original PMA and our revised PMA is compared in a controlled experiment, using a quantitative measure. We also assess qualitative differences in the results of the two approaches. The main research questions we address are these:

(1) Is the revised PMA method more effective than the original PMA method?
(2) How do the two methods differ in their result?

The remainder of this paper is structured as follows. Section 2 discusses related work on retrospectives in software engineering. Section 3 presents the two lightweight methods that were used in the experiment. Section 4 describes the design of the experiment. In Sections 5 and 6, quantitative and qualitative results, respectively, are presented. Section 7 contains a discussion of the results. Section 8 concludes and suggests avenues for further research.

2. Related work

According to Rising et al. [21], retrospective analysis as a method for learning from work experience was identified in 1988 by Joseph Juran and named "Santayana review" in homage to the philosopher George Santayana. Since then, many organisations have used many variations of the method and under many different names. We adopt Dingsøyr's definition [8], such that retrospective analysis is a "collective learning activity, which can be organised for projects either when they end a phase or are terminated. The main motivation is to reflect on what happened in a project in order to improve future practice – for the individuals that have participated in the project and for the organisation as a whole." Dingsøyr lists the most common names for retrospective analysis in [8]: "project retrospectives", "post mortem analysis", "postproject review", "project analysis review", "quality improvement review", "autopsy review", "after action review", and "touch down meetings". For the remainder of this paper, we use the term 'retrospective analysis' to denote the corpus of these methods and the term 'post mortem analysis' (PMA) to refer to the specific method we investigated.

Myllyaho et al. [20] conducted an extensive literature review within the software engineering and management literature, with the aim of reviewing retrospective analysis as a project-based learning technique. The results suggest that the use of retrospective analysis is well worth the effort, and that a simplified or 'lightweight' version of PMA can be beneficial when time is a factor.

Dingsøyr [8] discusses the importance of retrospective analysis as a method for sharing knowledge in software projects and gives an overview of the methods of retrospective analysis that are employed in the field of software engineering. In particular, Dingsøyr presents three lightweight methods of retrospective analysis, which are presented in Whitten [26], Collison and Parcell [6], and Birk et al. [2]. To give an overview of key differences in retrospectives, we present his comparison of the three methods (Table 1).

Desouza et al. [7] compared two kinds of output from retrospective analysis: traditional reports and stories. The comparison can be found in Table 2. They also identified four factors that should affect the choice of writing the result of the PMA as a report or as a story: (1) the nature of the project, (2) the cost you are willing to bear, (3) how much organizational impact is desired, and (4) what lessons you wish to convey.

Stålhane et al. [24] conducted an assessment of two retrospective methods. One was based on the PMA of Birk et al. and the other consisted of structured interviews. The main focus of their research was to determine whether there are situations in which one method performs better than the other. They found that this depends on whether a focused or broad analysis is desired. For a focused analysis, the semi-structured interviews worked better than the PMA. For a broad analysis, the PMA worked better and yielded more surprises.

3. The PMA methods used

In this section, we describe the methods that we adapted for the PMA. The original method we used was the one suggested by Birk et al. applied in [2,10,12,17,24] (see Table 1) with structured reports as output (see Table 2). In what follows, both the original and the modified method are described in detail.

3.1. PMA method 1: the original

The aim of this method is to bring together project participants and have them discuss what went well and what could be improved, and to analyse the root causes. Birk et al. use two techniques to carry out the PMA. To discover the positive and negative experiences, they use a focused brainstorm method called the KJ-method [22], resulting in affinity diagrams. To analyse the causes of these experiences, they performed root cause analysis using fishbone diagrams (also known as Ishikawa diagrams, in reference to their inventor Dr. Kaoru Ishikawa, a Japanese quality control statistician).

Table 1
Summary of selected differences among three methods for conducting retrospective analysis, taken from [8]

	Whitten	Collison and Parcell	Birk et al.
Whom to invite?	From each major participating organisation	All project members, possibly new project	All project members
Homework?	Yes	No	No
Type of discussion	Open	Open	Structured
Output	Recommendations	Guidelines, histories, names of people, key artefacts	Structured report on issues that went well and those that could be improved

Table 2
Reports vs. stories, taken from [7]

	Reports	Stories
Structure of knowledge	Highly structured	Semi-structured
Cost to prepare	Low	High
Richness of knowledge	Low	High
Ease of comprehension	Easy	Medium
Ease of recall	Difficult-medium	Easy

The post mortem meeting itself had the following four steps:

1. Introduce the PMA method and explain the purpose of the review.
2. KJ-session 1: elicit positive experience.
3. KJ-session 2: elicit negative experience.
4. Perform root cause analyses using fishbone diagrams for the most important positive and negative experiences.

3.1.1. The KJ-sessions

KJ-sessions are conducted as follows. Each participant receives a number of post-it notes and is asked to write down what they regard as the most significant experiences from the project. After everyone has finished writing, each participant puts a note on a whiteboard while explaining what he means by it. The process is repeated until all the notes have been presented, as illustrated in Fig 1a. Once all the notes have been placed on the whiteboard, the whole group discusses them and groups them according to similarity in concept. Each group of notes is then given a name, as illustrated in Fig. 1b. Possible connections between groups can be marked with arrows if required. In our study, each participant received five post-it notes and the entire process was repeated twice; first for positive experiences (KJ-session 1), then for negative experiences (KJ-session 2).

3.1.2. The root cause analysis method

The root cause analysis, or fishbone diagram method, needs a facilitator who takes control of the whiteboard. The group selects a (positive or negative) experience they want to analyse the cause of and the facilitator writes the name on a whiteboard and draws an arrow to it. The group then discusses what the cause of the experience might have been and as more causes are identified, the facilitator draws

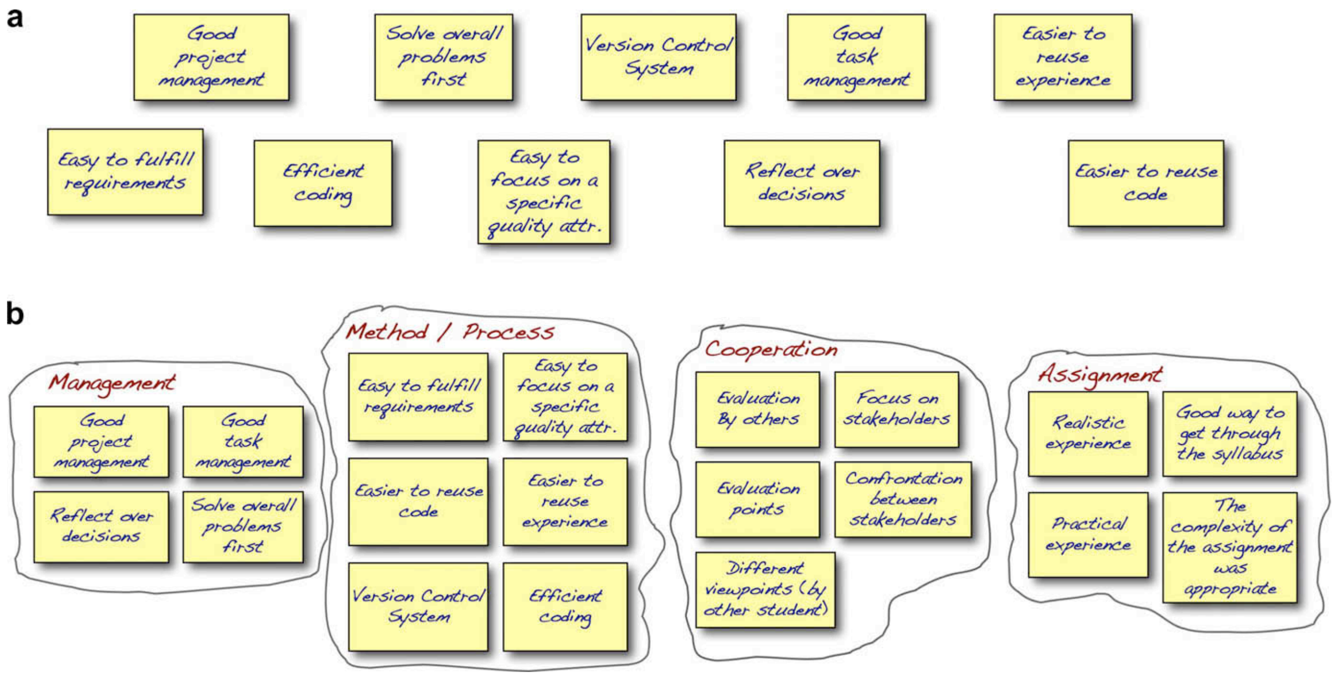
a
Good project management
Solve overall problems first
Version Control System
Good task management
Easier to reuse experience
Easy to fulfill requirements
Efficient coding
Easy to focus on a specific quality attr.
Reflect over decisions
Easier to reuse code
b
Management
Good project management
Good task management
Reflect over decisions
Solve overall problems first
Method / Process
Easy to fulfill requirements
Easy to focus on a specific quality attr.
Easier to reuse code
Easier to reuse experience
Version Control System
Efficient coding
Cooperation
Evaluation By others
Focus on stakeholders
Evaluation points
Confrontation between stakeholders
Different viewpoints (by other student)
Assignment
Realistic experience
Good way to get through the syllabus
Practical experience
The complexity of the assignment was appropriate

Fig. 1. KJ example.

arrows into the large arrow, writing in the causes. If a cause has several subcauses they are drawn as arrows into the minor arrows, as illustrated in Fig. 2. The figure shows a Fishbone diagram resulting from a positive root-cause analysis on good management. This example is a decomposition of one of the groups identified in the KJ-diagram shown in Fig. 1.

3.2. PMA method 2: the revised method

Our previous experiences of the PMA method with fishbone diagrams as a means of analysis had taught us that group activity tended to be high during the KJ phases, but that the activity seemed to dwindle as the groups proceeded with the analysis with fishbone diagrams. This tendency has also been observed by Stålhane et al. [24]. We wanted to increase the level of participation during the analysis phase, so we examined step four in the PMA process and proposed two main changes, which were inspired by theories on brainstorming and the notation of causal maps.

3.2.1. Theory for change

The setup of the original PMA method can be seen as the group working *nominally* in the KJ-session and *interactively* in the root-cause analysis phase. A group is defined as nominal if its members work independently, but in each other's presence. A group is defined as interactive if its members generate ideas in face-to-face discussions. According to Faure [13], evidence in the field suggests that nominal groups outperform interactive groups on the number of original ideas generated in a brainstorming session. Accordingly, we attempted to make phase four more nominal. We did this by using the same technique as in the KJ-sessions; namely, by using post-it notes and letting the group members come up with possible causes individually before coming together to discuss them.

In order to better accommodate the nominal brainstorm technique, we needed a more free form diagrammatic technique for presenting the results. For this, we examined the technique of causal mapping, which according to Hodgkinson [15] is one of the most popular methods for investigating individuals' cognitive representations in strategic decision making. Hodgkinson further observes that a growing number of researchers are employing one or more variants of causal mapping directly, as a means of eliciting actors' cognitions *in situ*, in an attempt to gain insights into the nature and significance of cognitive processes in organizational decision making. There exist many alternative elicitation procedures, but for our study we opted for a simple freehand mapping variety, using only the notation illustrated in Fig. 3 The figure shows the Causal map resulting from a positive root-cause analysis on good assignment, which is an identified group from the KJ-diagram in Fig. 1. Here, every oval represents a concept, every arrow indicates a cause–effect relationship, and the whole map represents a specific situation.

3.2.2. Practical changes

The procedure for the post mortem meeting itself is the same as in the original method outlined in Section 3.1, except for step four, for which we substituted what we call "the causal map analysis".

4. Causal map analysis: On the most important positive experience and the most important negative experience.

The new causal map analysis works as follows. All participants are given post-it notes and are asked to write down the causes of the experience to be analysed. These notes are then presented and placed on the whiteboard, much in the same way as when using the KJ-method. The group then gathers at the whiteboard and groups the causes where applicable. Cause–effect relationships are then indicated by arrows. The members are then allowed to write new notes that state deeper causes, or if causes are seen to be missing, write those in and indicate them with arrows. When the new causes have been placed on the whiteboard, the process is iterated until the group is satisfied with the analysis.

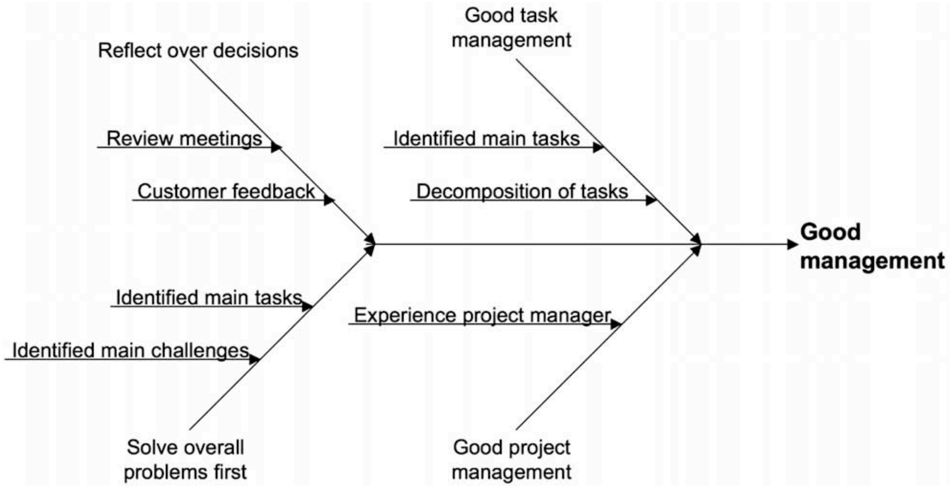

Fig. 2. Fishbone example.

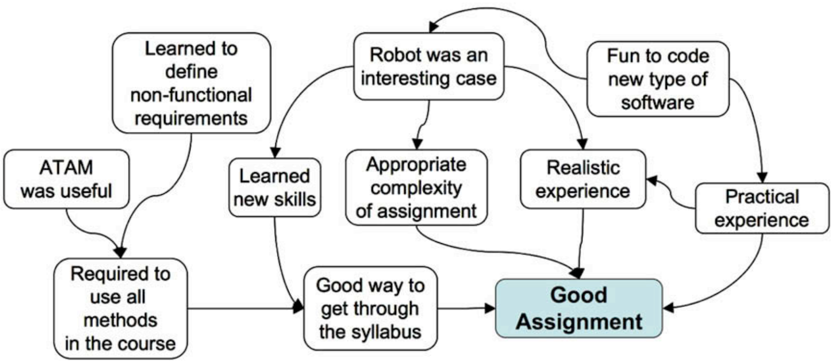
Learned to define non-functional requirements
Robot was an interesting case
Fun to code new type of software
ATAM was useful
Learned new skills
Appropriate complexity of assignment
Realistic experience
Practical experience
Required to use all methods in the course
Good way to get through the syllabus
Good Assignment

Fig. 3. Causal map example.

The main differences between the new and original analysis phase are:

- Forcing everyone to participate more actively by filling out the mandatory post-it notes.
- Allowing more freedom in the diagrams.

4. Research method

This section describes the design of the controlled experiment that investigated the effectiveness of using fishbone diagrams vs. using causal maps in the root-cause analysis phase of a PMA.

We performed PMA sessions in 2004 and 2005 in which we used fishbone diagrams and causal maps, respectively. The PMA reports from the 2 years were analysed, and we found that participants in the sessions produced a greater number of items when using fishbone than causal maps. However, when we looked at the content of the ideas generated we found that causal maps produced a greater number of *new* items than when using fishbone diagrams. These PMA sessions were, however, not planned intentionally as a controlled experiment and we did not have control of factors that could affect the results. On the basis of our experiences from the PMA sessions in 2004 and 2005, we planned a controlled experiment and performed it in 2006. The motivation for this experiment was to limit other factors that could threaten the experimental results, such as lack of randomization of subjects, different introductions to the two PMA methods (fishbone and causal), and different working conditions and time limits for the groups.

4.1. Experimental context

The experiment described in this paper was executed as a part of a software architecture course for Masters' students at the Norwegian University of Science and Technology. In this course, the students must carry out a software architecture project, the goal of which is to develop the software for a robot controller. The students work in groups of four to six. During the semester, the students must deliver a requirement specification, an architectural description, an architecture evaluation (using ATAM [3]) and an implementation of the robot controller according to the architecture. In the final phase of the project, the students perform a post mortem analysis of the robot project using PMA methods as described in Section 3. The students should spend half of the time on finding and analysing positive aspects of the project and the other half on the negative aspects [25]. The number of students taking this course varies from 60 to 100.

4.2. Study variables

This section defines the independent and dependent variables of the experiment and outlines how they were measured.

4.2.1. AnalysisMethod

The independent variable describes whether the subjects performed the second PMA phase using (1) fishbone diagrams with an interactive group process lead by a facilitator (PMA method 1 or 2) causal maps with a nominal brainstorming process (PMA method 2), as described in Section 3. Thus there are two factors that are controlled: the diagram technique and the group process. Although it would be possible to test all four combinations of these two factors independently (e.g., in a 2 × 2 factorial experiment design), we have in this experiment focused on what we consider the two most practical combinations, as defined by PMA method 1 and PMA method 2, respectively. Henceforth, when we refer to causal or fishbone analysis in this experiment, we are not only referring to the differences in diagrammatic technique but also the corresponding changes in the group process.

4.2.2. AnalysisEffectiveness

The dependent variable of the experiment attempts to measure the effectiveness of the PMA methods. To explain the *AnalysisEffectiveness* variable properly, we recapitulate briefly the PMA process, which consists of two main phases. In the first phase (steps 2 and 3), the participants elicit positive or negative aspects of the project and all the items found are represented as post-it notes in an affinity diagram. I_{PHASE1} is the number of items found in phase 1. In the second phase (step 4), the participants analyse one particular issue (positive or negative) to determine the reasons or background for this issue. The second phase generates a number of items, which are represented in a fishbone diagram or causal map. I_{PHASE2} is the number of items found in phase 2. To measure effectiveness, we compute how many of the items found in the second phase are new from the first phase. Thus, we can compute the Analysis-Effectiveness as

$$\text{AnalysisEffectiveness} = \frac{(I_{\text{PHASE2}} - (I_{\text{PHASE1}} \cap I_{\text{PHASE2}})) * 100}{I_{\text{PHASE2}}}$$

$I_{\text{PHASE1}} \cap I_{\text{PHASE2}}$ denotes the number of items that are common in phases 1 and 2. For example, if none of the items found in the second phase were found in the first, the effectiveness will be computed as 100%. If all of the items found in the second phase were also found in the first, the effectiveness will be computed as 0%. This means that the effectiveness will range from 0% to 100%.

When counting items, two or more items that describe exactly the same issue are regarded as duplicates and are removed. Items presented in the second phase are new if no items in the first phase state the exact same meaning.

The AnalysisEffectiveness variable was measured by going through the PMA reports of the subjects. The first step was to eliminate redundancy by removing duplicate items. Two or more items were considered to be duplicates if they had the exact same wording or the exact same meaning. The second step was to count items from the brainstorm phase and the items from the analysis phase. The third step was to find the items with the exact same wording or meaning from both phases, and mark these. The effectiveness was then computed by counting the number of unmarked items from the analysis phase divided by the total number of items from the same phase. To reduce the possible bias caused by subjective judgement, two researchers performed this process independently and later compared the results. In cases where there was disagreement, the items of concern were examined carefully before a decision was made.

4.3. Hypothesis formulation

Our hypothesis assesses whether the choice of analysis method (causal maps vs. fishbone) affects the percentage of new items found in the analysis phase (second phase) of a post mortem analysis, as quantified by the dependent variable *AnalysisEffectiveness*. Thus, we wanted to investigate whether one of the post mortem analysis methods is more effective than the other. The hypothesis was as follows:

$$\begin{aligned}\text{H0} &: \text{AnalysisEffectiveness(Causal maps)} \\ &= \text{AnalysisEffectiveness(Fishbone)} \\ \text{H1} &: \text{AnalysisEffectiveness(Causal maps)} \\ &> \text{AnalysisEffectiveness(Fishbone)}\end{aligned}$$

The test was one-tailed, to reflect our expectation that the causal maps would be more effective than fishbone, as suggested by our previous experiences and also justified theoretically in Section 3.

4.4. Group assignment

The population in this experiment consisted of 95% postgraduate and 5% last year bachelor software engineering students, where 20% of the population were females. Seventy percent of the students had prior theoretical knowledge of post mortem analysis methods from a software engineering course but none had any practical experience. A randomized experimental design was used in the controlled experiment. Each subject (group of students) was assigned randomly to either the fishbone diagram or causal map treatment. The groups were established at the beginning of the software architecture course. A list was made available for the students to sign up for a group before a specified deadline. Most of the students that signed this list knew each other beforehand. After the deadline, the remaining students were assigned to groups that had open slots or to new groups. The assignment to the fishbone and causal map treatments was distributed evenly in relation to groups that were joined by students and groups that were assigned by course staff. Table 3 describes the distribution of the number of subjects (groups) to the two PMA variants. The size of the groups varied from four to six students. A total of 142 students participated in the experiment.

4.5. Experiment tasks

The controlled experiment consisted of the following tasks:

- *Presentation of PMA method (30 min)*. The two variants of the PMA method (fishbone and causal) were presented simultaneously in two different rooms by two different lecturers. The content of the presentations was analysed before they were made, to ensure that they were similar in all respects except those that pertained to describing the two variants. The first part of the presentation was exactly the same, while in the second the presentations differed in that one described the fishbone

Table 3
Distribution of subjects in the controlled experiment

	Fishbone diagram	Causal map	Total
Number of groups	14	15	29

method and the other described the causal method. In the second part, the two methods were presented in a similar way and used the same number of slides. The participants asked roughly the same number of questions in each presentation, but the causal session lasted 5 min shorter than the fishbone session.

- *Positive brainstorm (30 min).* The participants brainstormed on positive aspects of the project and described the results in an affinity diagram. The result was recorded on a laptop PC or on paper.
- *Negative brainstorm (30 min).* The participants brainstormed on negative aspects of the project and described the results in an affinity diagram. The result was recorded on a laptop PC or on paper.
- *Positive root-cause analysis (20 min).* The issue that received the most votes from the brainstorming session on positive aspects was analysed using either a fishbone diagram or causal maps. The result was recorded on a laptop PC or on paper.
- *Negative root-cause analysis (20 min).* The issue that received the most votes from the brainstorming session on negative aspects was analysed using either a fishbone diagram or causal maps. The result was recorded on a laptop PC or on paper.
- *Write PMA-report (approx. 2 h).* All groups involved in the PMA had to write a report on the PMA. The report had to contain (i) four diagrams and a description from the brainstorm and analysis phase and (ii) a description of their experience of doing the PMA.

4.6. Analysis

4.6.1. Quantitative

The purpose of the quantitative test was to determine whether or not there was a difference between the use of the fishbone and causal methods. The hypothesis was tested using a standard two-sample one-tailed *t*-test assuming unequal variances. Although the *t*-test assumes a normal distribution, it is known to be relatively robust to mild deviations from this assumption. However, given our small sample size, it is not really possible to assess deviations from this assumption in a reliable way, due to lack of power to perform formal normality tests. Thus, to reduce potential threats to validity that might have resulted from violations of the *t*-test assumptions, a non-parametric Wilcoxon rank sum test was also performed. Given the small sample size, the Wilcoxon test was performed using the *Exact* option in the SAS statistical software package. The level of significance of the hypothesis test was set to $\alpha = 0.05$.

4.6.2. Qualitative

The purpose of running a qualitative analysis was to determine what the difference between the use of the fishbone and causal methods consisted of, if there was a difference. The qualitative analysis was performed after the results from the quantitative analysis were known.

Qualitative data were collected from three sources: (i) observation of the students by two researchers as they performed the different methods; (ii) the collection of the final reports; and (iii) a brief open-ended report that the students were told to write on their impression of the method and their experience with it. The data were analysed by hand, using a simple constant comparison method [19].

5. Quantitative results

This section describes the quantitative results and shows the results from the hypothesis test for the controlled experiment.

5.1. Descriptive statistics

Table 4 shows the descriptive statistics of the experiment, including the average (*Mean*), standard deviation (*Std*), minimum (*Min*), lower 25% quartile (*Q1*), median (*Med*), upper 75% quartile (*Q3*) and maximum (*Max*) values of AnalysisEffectiveness for fishbone diagrams and causal maps, respectively. The analysis effectiveness was 59.8% for fishbone diagrams and 78.4% for causal maps, which indicates a practically significant mean difference of 18.6%.

5.2. Formal hypothesis test

The two-sample, one-tailed *t*-test on the difference in means resulted in a *p*-value of 0.0062. The corresponding exact Wilcoxon rank sum test resulted in a *p*-value of 0.0041. Thus, for both the parametric and non-parametric tests, the *p*-value is well below the 0.05 level, which suggests that there is a statistically significant difference between the analysis effectiveness of the two methods of analysis.

5.3. Effect size

The sample's mean, data distribution, and 95% confidence interval of the mean for the dependent variable *AnalysisEffectiveness* are presented in a diamond plot, as a way to visualize the effect size of the two treatments (Fig. 4). The line across each diamond represents the group mean and the vertical span of each diamond the 95% confidence interval for each group. Overlap marks are drawn below

Table 4
Descriptive statistics of analysis method and effectiveness

AnalysisMethod	Mean (%)	Std (%)	Min (%)	Q1 (%)	Med (%)	Q3 (%)	Max (%)
Fishbone	59.8	19.8	20.0	50.0	68.3	73.3	83.3
Causal	78.4	15.6	46.2	73.9	80.0	89.2	100.0

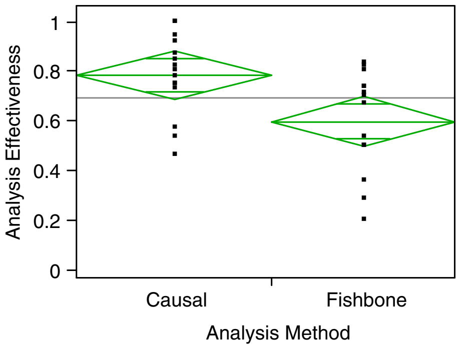

Fig. 4. Diamond plot of the effect of AnalysisMethod on AnalysisEffectiveness.

and above the means and an overlap represents a difference that is not significant at $\alpha = 0.05$. The line crossing the diagram is the entire sample mean.

To further quantify the difference between the two analysis methods, we calculated a standardized effect size measure known as Cohen's d [5]. In our case, Cohen's d was calculated by dividing the difference between the mean AnalysisEffectiveness of causal maps and fishbone diagrams with the pooled standard deviation, yielding $d = 1.05$. Cohen suggested that if d is greater than 0.8, the effect size can be considered to be large.

6. Qualitative results

The quantitative tests suggest that there is a difference in effectiveness between the two methods, but what that difference consists of remains an issue. To determine what the difference consists of, we made a qualitative analysis of the final reports.

The two diagram types differ importantly in the structure that they yield. One difference concerns the number and depth of the causal links stated. The fishbone diagrams usually contained three to four main causes, and subcauses varied from none to four. The average cause–effect chain was two links. By contrast, the causal maps contained from two to eight main causes and had cause–effect chains up to five links long, the average being about three links. The free form of the causal maps seems to support and encourage a greater degree of analysis of causes into their relevant subcauses.

Another difference concerns the way in which causes were analysed into subcauses. The students using the fishbone diagrams would put evenly distributed subcauses on all their main causes, whether they were particularly relevant or not. The students using the causal maps would typically select a few relevant causes and create longer cause–effect chains for these, and ignore the more irrelevant main causes, such as causes outside their control.

One of the major goals of the PMA is to learn from experience and improve performance for future projects. The cause–effect chains in causal maps are longer than those in fishbone diagrams, and the causes noted seem to be more specific. It is thus easier to think of courses of action to improve performance. The longer chains yielded by the causal map approach tended to paint a more nuanced picture of the situation in the project, with general causes being stated first and more specific causes being stated deeper in the chain as the general causes are analysed.

We also observed the formation of what we called *hubs* in causal maps. Since a node can be the cause of several other nodes and also the effect of several subcauses, sometimes we observed nodes with several arrows going in and out. These nodes were very easy to identify in the diagrams and typically marked major problem spots in the projects.

However, whether the causal or fishbone method was used is not the only factor that affected the result. One

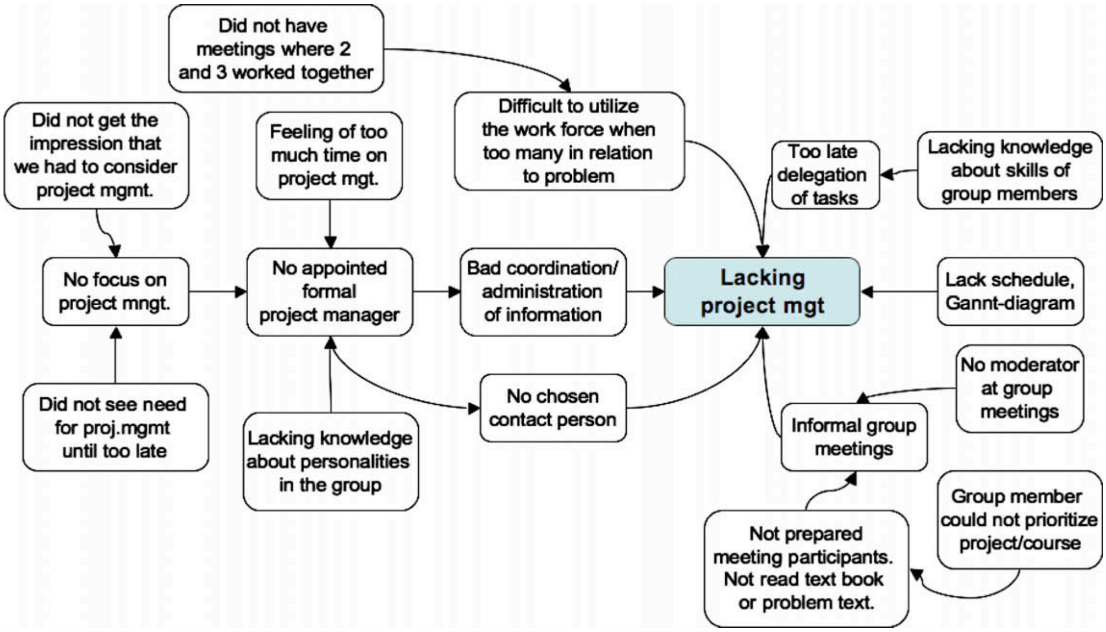
Did not have meetings where 2 and 3 worked together
Did not get the impression that we had to consider project mgmt.
Feeling of too much time on project mgt.
Difficult to utilize the work force when too many in relation to problem
Too late delegation of tasks
Lacking knowledge about skills of group members
No focus on project mngt.
No appointed formal project manager
Bad coordination/ administration of information
Lacking project mgt
Lack schedule, Gannt-diagram
No moderator at group meetings
Did not see need for proj.mgmt until too late
Lacking knowledge about personalities in the group
No chosen contact person
Informal group meetings
Group member could not prioritize project/course
Not prepared meeting participants. Not read text book or problem text.

Fig. 5. Example of a causal map of a problem.

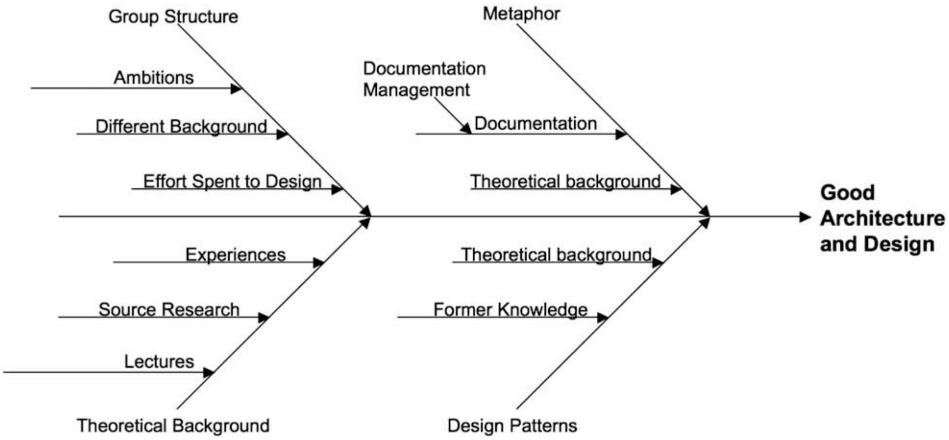

Fig. 6. Example of a fishbone diagram on a success.

observation from the qualitative analysis was that if the students chose a topic for analysis that was outside their control, the quality of the analysis was often low with regards to useful experience that could be transferred to new projects, regardless of the method they used for the cause–effect analysis.

Figs. 5 and 6 show two examples from the PMA experiment that illustrate the qualitative difference we found between the resulting causal maps and the fishbone diagrams. In the causal map shown in Fig. 5, the greatest number of links from a cause to the problem of focus is four. The causal map also contains hubs where one node is affected by several other nodes, e.g., the boxes “No appointed formal project manager” and “Informal group meetings”. Such hubs usually indicate a cause that relates to many problems. Also note that many of the causes in the diagram are specific and can be addressed so that performance in the next project can be improved; by, e.g., assigning a project manager and enforcing more formal group meetings.

Fig. 6 shows a typical fishbone diagram. Compared to the causal map diagram, fewer causes are analysed into subcauses, the causes are not analysed to a depth of more than three levels, and the causes are more general.

In addition to the final reports, we observed the groups' behaviour during the PMA sessions. Our qualitative observation was that the groups using the causal map technique participated more during the analysis phase than the groups using the fishbone diagrams.

The reports the students delivered about their experience with the methods were very much unison. All of them expressed a positive impression of the method, whether it was the original or revised one. They particularly liked the idea of focusing on what went well as well as what did not. About 70% of the students had previous theoretical knowledge of the original method, as it was taught in an optional course at the university, but none had practical experience with using it. The students with theoretical knowledge of the method were evenly distributed among the groups.

7. Discussion

In this section we discuss our findings and possible threats to the validity of our experiment.

7.1. Our results

The quantitative results presented in Section 5 showed that, in our setting with small software teams with no access to professional facilitators, using causal maps is more effective than fishbone diagrams for analysing root causes of problems or successes in PMAs. This result can be explained by the fact that the groups that made causal maps used a nominal brainstorming technique when generating their initial ideas on causes, whereas the groups that made fishbone diagrams used an interactive technique. The observation that the nominal group technique outperformed the interactive one, is a result that is in line with earlier research on brainstorming [13].

Another possible explanation for the significant difference in effectiveness between the two approaches is that we used untrained facilitators in our PMA sessions. The difference might have been less, had we used professional facilitators to properly steer the conversations. We know that the fishbone method will benefit from an experienced facilitator who can coax the underlying causes from the participants, but there have been no tests to suggest how much the causal map method would benefit from having such a facilitator. Our observations from several PMA sessions do, however, indicate that the motivation and level of activity is generally higher when making causal maps than when making fishbone diagrams, as the former approach enforces active participation of all involved. Also, the facilitator and form of discussion will still be a bottleneck in terms of productivity. This leads us to conclude that the proposed method of causal maps is less dependent on a professional facilitator, and as such, is more suited for companies who are new to retrospective methods, or where experienced facilitators are not readily available.

The qualitative results presented in Section 6 show that the quality of the analysis when using causal maps is higher than when using fishbone diagrams, in the following respects: the analysis of causes had greater depth; the issues identified were more specific and practical; and the analysis of the cause into subcauses was more varied. We believe that some of these differences are due to the limitations of the structure of the fishbone diagram. It is impractical to analyse fishbone diagrams to a depth of more than three levels. Further, variations in the depth of analysis (from 1 to 3 levels) are possible but not very practical. Most groups in our experiment conducted their analysis at a depth of two levels for all issues identified. In fishbone diagrams, less relevant issues will be analysed into their component parts, simply to "complete the fishbone structure". In contrast to this, when using causal maps, the structure is constructed after the issues have been identified. Issues that are not very relevant will not be analysed any further, whereas issues that are very relevant will be subject to a more thorough analysis to a depth of several levels. Such analysis will often result in the identification of specific issues that can be addressed with a view to improving performance in future projects. In addition, the construction of causal maps will often yield hubs, which constitute central issues that have several inputs and outputs.

One could argue that there are benefits to using methods that imposing more restrictions on the user, like the fishbone diagram. After all the method has been in successful use for a long time. In the process of creating a more restrictive diagram, the user is forced to ask questions, interact and refine their thinking. However, as has been pointed out in previous research [8] and as we have seen in this experiment, this is dependent on an experienced facilitator to properly steer the discussion. When no such facilitators are available, a more freeform technique seems to yield better results.

Another argument often raised against the causal maps, is the concern for "spaghetti diagrams", with no clear structure and the option to connect every item on the map, the diagram might become unreadable and not provide a good starting point for improvement. Fishbone diagrams on the other hand, has a clear structure that makes main causes readily identifiable. In our experiment, however, we did not observe these effects. The causal maps provided good overviews and often had the so called "hubs" which indicated strong causes. The fishbone diagrams on the other hand often did not provide any clear cause, since every bone was filled out "to complete the structure". This is another indication that an experienced facilitator was needed in this variation.

Each group in the experiment consisted of four to six persons. We believe that the difference in effectiveness between the two approaches would be even more significant for larger groups. The main reason for this is the form of brainstorming used. A large group using interactive discussion will suffer more from the effect of "production blocking" (impossibility for subjects to speak simultaneously), "evaluation apprehension" (fear of negative evaluation from other group members), and "free riding" (reduced effort exerted when individual contribution is not identifiable) [13] than a group using a nominal technique. With many subjects present, it is easier to fall silent and leave the discussion to the others. There is a greater risk of the analysis losing focus without coordination of a professional facilitator. The waiting time could also result in a drop of motivation that could hurt the end result.

7.2. Threats to validity

We now discuss, in a systematic way, possible threats to the validity of our experiment according to the taxonomy provided in [23].

7.2.1. Validity of statistical conclusions

The hypothesis was tested using both the non-parametric exact Wilcoxon rank sum test and the parametric two-sample *t*-test. The tests yielded consistent and significant results. Hence, in light of the simplicity of the experiment design and the straightforward statistical analyses, we do not believe that there are major threats to the validity of our statistical conclusions.

7.2.2. Internal validity

The primary means to address threats to internal validity in this experiment was randomization. In addition, we observed all the PMA sessions to make sure that they conformed fully to the prescribed processes. However, due to practical considerations, once the subjects had been assigned to one of the two treatments, they received different training (on either causal maps or fishbone diagrams, by two different instructors). As explained in Section 4.5, we took several precautions to ensure that the training was as similar as practically possible in quality and quantity, but we cannot completely rule out the possibility that a bias was introduced as a result of this differential training, e.g., that one of the groups became more motivated or better trained in their respective technique than the other group.

7.2.3. Construct validity

The dependent variable of the experiment was "AnalysisEffectiveness". According to Faure [13], originality of the ideas generated is the most commonly used measure when measuring creative techniques like brainstorming. Note also that the qualitative analyses triangulated the quantitative analysis by offering complementary insights on other aspects of "quality": the qualitative analysis explained and justified the quantitative result.

7.2.4. External validity

The most prominent threat to external validity is that the experiment was carried out by students for a student project, which is not necessarily representative of industrial settings. However, the students are part of a five-year

Masters programme and at the end of their fourth year, when they take the course, many of them have already gained industrial experience as software developers. The project itself was also designed to be as close to a real project as possible, engaging teams of four to six developers for a period of four months.

As explained in Section 7.1, we believe that the difference in effectiveness between the two approaches depends on the size of the teams. Each team in the experiment consisted of four to six persons. We expect that causal maps would be even more beneficial for larger teams than this, but less effective for even smaller teams. This expectation is due to several, competing underlying mechanisms: having a facilitator to structure discussions and results should provide benefits to the fishbone approach, but as the size of the team increases, production blocking, evaluation comprehension and free riding might counteract those benefits. This needs to be verified in future experiments.

There is also the factor of non-professional facilitators to consider. In previous research on PMA, it has been claimed that the facilitator plays a crucial role [8]. In this experiment, the students had to select a facilitator among themselves. Whether the results can be generalized to a setting with an experienced facilitator, for one or both variants of the method, is a matter for future experiments. We do, however, believe that our results can be generalized to settings in which experienced facilitators are not available.

8. Conclusion

The results of the experiment described in this paper show that when causal maps, rather than fishbone diagrams, are used to analyse successes and/or problems in a PMA, in a setting of small software engineering teams, with no experienced facilitator available, there is a significant increase in both effectiveness and quality. Thus, concerning our first research question: "Is the revised PMA method more effective than the original PMA method?", we base our answer on our quantitative analysis which states that there is a statistical significant difference between the two methods and that the effect of using the revised method compared to the original method is large. We must also consider the setting of the experiment in our answer, so the final answer is then: Yes, for a setting of small software teams where there is no experienced facilitator available, the revised method is more effective than the original.

To answer research question two: "How do the two methods differ in their result?", we used our qualitative observations as well as outlined theory. We conclude that the main explanation for the difference in the two methods is twofold. First, using a nominal brainstorming technique for causal maps will engage the whole evaluation group simultaneously and thus be more effective. This is in line with previous research on brainstorming. Second, the layout of fishbone diagrams limits the ways in which issues can be related and the PMA process can be carried out, and is as such much more dependent on an experienced facilitator to properly steer the discussion. Using fishbone diagrams forces the participants to analyse issues in a strict hierarchical manner and the diagram layout does not encourage deeper analysis into several levels or analysis of the relations between issues. Analysis using causal maps is not restricted in these ways.

The main difference in the use of the two methods was that the use of causal maps produced a more selective and deeper analysis of issues into their component parts that, in many cases, results in the identification of specific and practical issues that can be addressed in order to improve performance in future projects.

The results of our experiment may be extended by performing further experiments, in which the variables and environment are changed. For example, it should be determined how the group size and the usage of a professional facilitator will affect the effectiveness of the variants of the method. To reduce threats to external validity, we should also perform similar experiments in an industrial setting.

Acknowledgements

We are grateful to Torgeir Dingsøyr at SINTEF ICT, Reidar Conradi at the Department of Computer and Information Science, Norwegian University of Science and Technology, and the anonymous reviewers at IST for comments on earlier versions of this article. We would also like to thank Chris Wright for proofreading and providing useful comments.

References

[1] K. Beck, Principles behind the Agile Manifesto, 2001. Available from: <http://agilemanifesto.org/principles.html/>. Retrieved May 17th 2007.

[2] A. Birk, T. Dingsøyr, T. Stålhane, Postmortem: never leave a project without it, IEEE Software 19 (3) (2002) 43–45.

[3] P. Clements, R. Kazman, M. Klein, Evaluating Software Architectures: Methods and Case Studies, Addison-Wesley Longman Publishing Co., 2002.

[4] A. Cockburn, Agile Software Development, Addison-Wesley, 2002.

[5] J. Cohen, Statistical Power for the Behavioral Sciences, second ed., Erlbaum, Hillsdale, NJ, 1988.

[6] C. Collison, G. Parcell, Learning to Fly: Practical Lessons from one of the World's Leading Knowledge Companies, Capstone Publication, 2001.

[7] K.C. Desouza, T. Dingsøyr, Y. Awazu, Experiences with conducting project postmortems: reports versus stories, Software Process: Improvement and Practice 10 (2) (2005) 203–215.

[8] T. Dingsøyr, Postmortem reviews: purpose and approaches in software engineering, Information and Software Technology 47 (5) (2005) 293–303.

[9] T. Dingsøyr, G.K. Hanssen, Extending agile methods: postmortem reviews as extended feedback, in: Proceedings of the 4th International Workshop on Learning Software Organisations (LSO'02), 2002.

[10] T. Dingsøyr, N.B. Moe, Ø. Nytrø, Augmenting experience reports with lightweight postmortem reviews, Lecture Notes in Computer Science 2188 (2001) 167–181.

[11] T. Dybå, Enabling Software Process Improvement: An Investigation on the Importance of Organisational Issues, Dr. Ing. Thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, 2001.

[12] T. Dybå, T. Dingsøyr, N.B. Moe, Process Improvement in Practice – a Handbook for IT Companies, Kluwer, Boston, 2004.

[13] C. Faure, Beyond brainstorming: effects of different group procedures on selection of ideas and satisfaction with the process, Journal of Creative Behaviour 38 (1) (2004) 13–34.

[14] R.L. Glass, Project retrospectives and why they never happen, IEEE Software 19 (5) (2002) 111–112.

[15] G.P. Hodgkinson, A.J. Maule, N.J. Bown, Causal cognitive mapping in the organizational strategy field: a comparison of alternative elicitation procedures, Organizational Research Methods 7 (1) (2004) 3–26.

[16] A. Keegan, J.R. Turner, Quantity versus Quality in project-based learning practices, Management Learning 32 (2001) 77–98.

[17] N.L. Kerth, Project Retrospectives: a Handbook for Team Reviews, Dorset House Publishing, New York, 2001.

[18] D. Larsen, The manager's role in starting and ending projects: charters and retrospectives, in: Proceedings of the 21st Pacific Northwest Software Quality Conference, 2003.

[19] M.B. Miles, A.M. Huberman, Qualitative Data Analysis: an expanded sourcebook, second ed., SAGE publications, 1994.

[20] M. Myllyaho, O. Salo, J. Kääriäinen, J. Koskela, A review of small and large post-mortem analysis methods, in: Proceedings of the ICSSEA, Paris, 2004.

[21] L. Rising, E. Derby, Singing the songs of project experience: patterns and retrospectives, The Journal of Information Technology Management 16 (9) (2003) 27–33.

[22] R. Scupin, The KJ Method: a technique for analyzing data derived from Japanese ethnology, Human Organization 56 (1997) 233–237.

[23] W.R. Shadish, T.D. Cook, D.T. Campbell, Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Houghton-Mifflin, Boston, 2002.

[24] T. Stålhane, T. Dingsøyr, G.K. Hanssen, N.B. Moe, Post mortem – an assessment of two approaches, Lecture Notes in Computer Science 2765 (2003) 129–141.

[25] A.I. Wang, T. Stålhane, Using post mortem analysis to evaluate software architecture student projects, in: Proceedings of the 18th Conference on Software Engineering Education and Training, 2005, pp. 43–50.

[26] N. Whitten, Managing Software Development Projects: Formula for Success, Wiley, New York, 1995.