Pair-wise comparisons versus planning game partitioning-experiments on requirements prioritisation techniques

Karlsson, Lena; Thelin, Thomas; Regnell, Björn; Berander, Patrik; Wohlin, Claes

Published in:
Empirical Software Engineering

DOI:
10.1007/s10664-006-7240-4

2007

Empir Software Eng (2007) 12:3–33
DOI 10.1007/s10664-006-7240-4

Pair-wise comparisons versus planning game partitioning—experiments on requirements prioritisation techniques

Lena Karlsson · Thomas Thelin · Björn Regnell · Patrik Berander · Claes Wohlin

Published online: 22 March 2006

Editor: Daniel Berry

Abstract The process of selecting the right set of requirements for a product release is dependent on how well the organisation succeeds in prioritising the requirements candidates. This paper describes two consecutive controlled experiments comparing different requirements prioritisation techniques with the objective of understanding differences in time-consumption, ease of use and accuracy. The first experiment evaluates Pair-wise comparisons and a variation of the Planning game. As the Planning game turned out as superior, the second experiment was designed to compare the Planning game to Tool-supported pair-wise comparisons. The results indicate that the manual pair-wise comparisons is the most time-consuming of the techniques, and also the least easy to use. Tool-supported pair-wise comparisons is the fastest technique and it is as easy to use as the Planning game. The techniques do not differ significantly regarding accuracy.

Keywords Requirements engineering · Requirements prioritisation · Release planning · Decision making · Controlled experiment

L. Karlsson (✉) · T. Thelin · B. Regnell
Department of Communication Systems, Lund University, Sweden
e-mail: lena.karlsson@telecom.lth.se

T. Thelin
e-mail: thomas.thelin@telecom.lth.se

B. Regnell
e-mail: bjorn.regnell@telecom.lth.se

P. Berander · C. Wohlin
Department of Systems and Software Engineering, School of Engineering,
Blekinge Institute of Technology, Sweden
e-mail: patrik.berander@bth.se

C. Wohlin
e-mail: claes.wohlin@bth.se

1 Introduction

In market-driven software development, products are developed in several consecutive releases intended for an open market. Market-driven development does not have easily identifiable customers and the requirements often need to be *invented* based on the needs of several potential users (Sawyer, 2000). When requirements are elicited from several stakeholders, it often yields more requirements than can be implemented at once. The requirements need to be prioritised so that the most significant ones are met by the earliest product releases (Wiegers, 1999; Siddiqi and Shekaran, 1996).

During a project, decision makers in software development need to make many different decisions regarding the release plan. Issues such as available resources, milestones, conflicting stakeholder views, available market opportunity, risks, product strategies, and costs need to be taken into consideration when planning future releases. Unfortunately, there is a lack of simple and effective techniques for requirements prioritisation, which could be used for release planning (Karlsson and Ryan, 1997).

The software literature includes many sources that state the importance of prioritising requirements. In the field study by Lubars et al. (1992), several companies expressed a need for guidance in assigning, modifying and communicating requirements priorities. Siddiqi and Shekaran (1996) identified requirements prioritisation as an important, though disregarded, issue in RE research at that point in time. Yourdon (1999) states that one reason why projects often exceed deadlines and budgets is that people are not used to the idea of not implementing *all* the functionality requested by the user. The key is to focus on the 20 percent of the requirements that deliver 80 percent of the benefit. Thus, requirements prioritisation is a crucial project activity.

Our goal is to *analyse and compare requirements prioritisation techniques for the purpose of gaining increased understanding of the techniques with respect to their time-consumption, ease of use, and accuracy from the point of view of the decision maker.* The paper describes two consecutive experiments aimed at comparing requirements prioritisation techniques. The first experiment[1] compares a rudimentary prioritisation technique (Planning game) with a more elaborate one (Pair-wise comparisons) and is described in Section 3.

As the Pair-wise comparisons turned out to be very time-consuming, a majority of the subjects found it less easy to use and most subjects even found it less accurate, the second experiment was designed to investigate if the technique would benefit from tool-support. In the second experiment, prioritisation with a commercial requirements management tool (www.telelogic.com/corp/products/focalpoint/overview.cfm) was compared to prioritisation with the manual Planning game, which is described in Section 4. The results from the second experiment indicate that the Tool-supported pair-wise comparisons is a faster technique than the Planning game while the ease of use and accuracy are equally high.

The paper is structured as follows. Section 2 explains and discusses the matter of requirements prioritisation in general and the compared techniques in particular. Section 3 describes the first of the two experiments, including the planning,

[1] The first experiment was presented at EASE 2004, see Karlsson et al. (2004).

operation and analysis. Section 4 describes the planning, operation and analysis of the second experiment. Section 5 discusses the results and compares the two experiments. Finally, the paper is concluded in Section 6.

2 Requirements Prioritisation

In order to select the correct set of requirements, the decision makers must understand the relative priorities of the requested requirements (Wiegers, 1999). By selecting a subset of the requirements that are valuable for the customers, and can be implemented within budget, organisations can become more successful on the market.

There are several different techniques to choose from when prioritising requirements. Some are based on determining the *absolute* importance of the candidate requirements, by e.g., assigning each requirement a certain priority such as essential, conditional or optional (IEEE, 1998; Wiegers, 1999). Other techniques are *relative* and require a person to determine which requirement is more important. Thereby, all requirements get different priorities, whereas absolute techniques assign several requirements to the same priority. Relative approaches tend to be more accurate and informative than absolute ones (Karlsson, 1996). One relative technique is the $100-test presented in Leffingwell and Widrig (2000). Each person is given $100 of "idea money" to be spent on "purchasing ideas" among the elicited requirements. The technique is particularly useful for calculating a cumulative vote based on several participants' views. Another technique is Wiegers' method (Wiegers, 1999), which takes several criteria into consideration, such as benefit, penalty, cost, and risk, and calculates a priority value for each requirement. In addition, there are several techniques aimed at release planning, in particular when several stakeholders are involved, such as EVOLVE (Greer and Ruhe, 2004) and Quantitative WinWin (Ruhe et al., 2002). Both techniques are aimed at release planning of incremental software development. For a thorough review of these and other prioritisation techniques, see Berander and Andrews (2005), Lehtola and Kauppinen (2004) and Moisiadis (2002).

The three techniques compared in this paper are all relative techniques: Pair-wise comparisons (Karlsson, 1996; Saaty and Vargas, 2001), Planning game (Beck, 1999), and Tool-supported Pair-wise comparisons (www.telelogic.com/corp/products/focalpoint/overview.cfm; Karlsson et al., 1997), see Table 1. The techniques are further described below.

Table 1 Details about the three techniques compared in the experiments

Technique	Abbreviation	Prioritisation algorithm
Pair-wise comparisons	PWC	Exhaustive pair-wise comparisons between requirements
Planning game	PG	Sorting algorithm to partition and rank requirements
Tool-supported pair-wise comparisons	TPWC	Tool-support for PWC, reduced number of comparisons

2.1 Planning Game (PG)

PG is used in planning and deciding what to develop in an Extreme Programming (XP) project. In PG, requirements (written on so called story cards) are elicited from the customer. When the requirements have been elicited, they are prioritised by the customer into three different piles: (1) those without which the system will not function, (2) those that are less essential but provide significant business value, and (3) those that would be nice to have (Beck, 1999).

At the same time, the developers estimate the time required to implement each requirement and, furthermore, sort the requirements by risk into three piles: (1) those that they can estimate precisely, (2) those that they can estimate reasonably well, and (3) those that they cannot estimate at all.

Based on the time estimates, or by choosing the cards and then calculating the release date, the customers prioritise the requirements within the piles and then decide which requirements that should be planned for the next release (Newkirk and Martin, 2001). Thus, the technique uses a sorting algorithm, similar to numeral assignment (Karlsson, 1996), to partition the requirements into one of three piles. Then, the requirements within each pile are compared to each other in order to achieve a sorted list.

The result of the PG technique is an ordered list of requirements. This means that the requirements are represented as a ranking on an *ordinal scale,* without any information about how much more important one requirement is than another.

In the investigation performed by Karlsson et al. (1998) a similar technique, called Priority groups, was investigated. In the Priority groups technique, requirements are put into one of three groups, corresponding to high, medium and low priority. In groups with more than one requirement, three new subgroups are created until no group has more than one requirement. Thereby an ordered list of requirements is compiled. Priority groups was given the lowest subjective ranking (regarding ease of use, reliability and fault tolerance) of the six investigated prioritisation techniques in Karlsson et al. (1998). The technique was ranked as 4th of the six techniques regarding the objective measure total time-consumption.

2.2 Pair-Wise Comparisons (PWC)

Pair-wise comparisons involves comparing all possible pairs of requirements, in order to determine which of the two requirements is of higher priority, and to what extent. If there are n requirements to prioritise, the total number of comparisons to perform is $n(n-1)/2$. For each requirement pair the decision maker estimates the relation between the requirements on the scale {9, 7, 5, 3, 1} where 1 represent equal importance and 9 represent one requirement being much more important than the other.

This relation results in a dramatically increasing number of comparisons as the number of requirements increases. However, due to redundancy of the pair-wise comparisons, PWC is rather insensitive to judgement errors. Furthermore, PWC includes a *consistency check* where judgement errors can be identified and a *consistency ratio* can be calculated.

PWC is used in the Analytic Hierarchy Process (AHP) (Saaty and Vargas, 2001). In AHP it is possible to take the system perspective into account, so that a system structure of related requirements can be abstracted into a hierarchy that describes

requirements on different abstraction levels. Hence, AHP can take the whole system into account during decision making since it prioritises the requirements on each level in the hierarchy (Saaty and Vargas, 2001).

In the investigation by Karlsson et al. (1998), the authors conclude that PWC [called AHP in Karlsson et al. (1998)] was the most promising approach because they found it trustworthy and fault tolerant. It also includes a consistency check and it is based on a *ratio scale*, i.e., it includes the priority distance. PWC was the only technique in the evaluation that satisfied all these criteria. However, because of the rigour of the technique, it was also the most time-consuming in the investigation.

In another empirical investigation of prioritisation techniques performed by Lehtola and Kauppinen (2004), PWC was compared to Wiegers' method (Wiegers, 1999). The authors conclude that "users found it difficult to estimate how much more valuable one requirement is than another" and that "some users conceived pair-wise comparisons as pointless" as they felt it would have been easier for them to just select the most important requirements (Lehtola and Kauppinen, 2004).

2.3 Tool-Supported PWC (TPWC)

Since the major disadvantage of PWC is the time-consumption for large problems, different investigations have been performed in order to decrease the number of comparisons, and thus the time needed (Carmone et al., 1997; Harker, 1987; Karlsson et al., 1997; Shen et al., 1992). The results of these have been that it is possible to reduce the number of comparisons with as much as 75%. Techniques for reducing the number of comparisons are called Incomplete Pair-wise Comparisons (IPC). The techniques are based on providing *stopping rules*, indicating when additional pair-wise comparisons are no longer necessary (Karlsson et al., 1997). However, when reducing the number of comparisons, the number of redundant comparisons is also reduced. Thereby, the sensitivity for judgemental errors increases (Karlsson et al., 1998).

The PWC technique described in Section 2.2 has been built into a requirements management tool (www.telelogic.com/corp/products/focalpoint/overview.cfm). The tool guides the user to apply pair-wise comparisons between requirements in a similar manner as the PWC technique. The tool contains an IPC algorithm and stopping rules that indicate to the user when the necessary number of comparisons has been performed. The number of required comparisons is reduced to the approximate size $2n$, where n is the number of requirements. Thereby, the time-consumption is reduced radically in comparison with the manual PWC.

The tool displays one requirement pair at the time to the user, possibly including descriptions of the requirements. The prioritisation is based on a ratio scale, and applies pair-wise comparisons between requirements based on some criteria chosen by the user beforehand. The user selects one of the nine possible "more than," "equal" or "less than" symbols between the two requirements, as illustrated in Fig. 1. When the user clicks "ok," the next pair of requirements is displayed. In that manner the focus is retained, since only one task at the time is presented to the user. As the redundancy is reduced by the IPC algorithm, it affects the quality of the results. The tool includes a consistency check that identifies inconsistencies among the requirement priorities. The user may then revise the inconsistent comparisons until an acceptable consistency is achieved.

Fig. 1 Part of the user interface in the tool used for TPWC

The tool also incorporates solutions for requirements management and project portfolio management and has visualisation possibilities.

2.4 Cost-Value Trade-Off

When prioritising requirements, it is often not enough to prioritise only how much value the requirement has to the customers. Often other factors such as risk, time, cost and requirements interdependencies should be considered before deciding if a requirement should be implemented directly, later, or not at all. For example, if a high-priority requirement would cost a fortune, it might not be as important for the customer as the customer first thought (Lauesen, 2002). This means that it is important to find those requirements that provide much value for the customers at the same time as they cost as little as possible to develop.

Karlsson and Ryan (1997) use PWC as an approach for prioritising regarding both Value and Cost in order to implement those requirements that give most value for the money. The data can be used further to provide graphs to visualise the Value to Cost ratio between the requirements. The tool-supported PWC can visualise these Value to Cost ratios in different charts and diagrams.

In PG, a similar approach is taken when requirements are prioritised based on both customer value and implementation effort. The information that could be extracted from PG should hence be possible to use in the same way as it was used in (Karlsson and Ryan, 1997) with the difference that the result from PG is based on an ordinal scale instead of a ratio scale.

Wiegers (1999) suggests that the value of a requirement is balanced against not just its cost, but also any implications it has for the architectural foundation and future evolution of the product. He also proposes that the value is seen as being dependent both on the value it provides to the user and the penalty incurred if the requirement is absent.

3 Experiment 1

This section describes the first of the two experiments, the experiment planning and operation as well as the analysis.[2] Finally, it is concluded with a discussion.

The motivation for the experiment is that although requirements prioritisation is recognised as an important area, few research papers aim at finding superior prioritisation techniques that are accurate and usable. This experiment aims at comparing two of the available techniques in order to understand their differences. The PWC was pointed out as a superior technique in a comparison between

[2] For more information, see http://serg.telecom.lth.se/research/packages/ReqPrio

prioritisation techniques (Karlsson et al., 1998), while a technique similar to PG was ranked rather low. However, the PG technique is of current interest since it is used in the agile community. Therefore these two techniques are interesting to investigate.

The experiment design described in this section is to a large extent also used in the second experiment. Therefore, Section 4 is focused on describing the second experiment and the differences between the two experiment designs.

3.1 Hypotheses and Variables

The goal of the experiment is to compare two prioritisation techniques and to investigate the following null hypotheses:

H_o1 The average time to conclude the prioritisations is equal for both techniques, PG and PWC.
H_o2 The ease of use is equal for both techniques, PG and PWC.
H_o3 The accuracy is equal for both techniques, PG and PWC.

The alternative hypotheses are formulated below:

H_A1 The average time to conclude the prioritisations is not equal for both techniques, PG and PWC.
H_A2 The ease of use is not equal for both techniques, PG and PWC.
H_A3 The accuracy is not equal for both techniques, PG and PWC.

The independent variables are the techniques PG and PWC. The objective dependent variable *average time to conclude the prioritisations* was captured by each subject by noting their start and stop time for each task. The subjective dependent variable *ease of use* was measured by a questionnaire, which was filled out by all subjects after the experiment. The subjects were asked "Which technique did you find easiest to use?" The subjective dependent variable *accuracy* was measured by conducting a post-test a few weeks after the experiment. Each subject was sent four personal lists (two for each criterion), corresponding to the priority order compiled from the two techniques investigated during the experiment. The subjects were asked to mark the priority order that corresponded best to their views. The time-consumption and ease of use are very important measures since resources are limited and a fast and easy technique is more likely to be used than a more effort-demanding one. The third and probably most important variable is the accuracy, i.e., that the technique is trustworthy and that the resulting priority order reflects the decision maker's opinion. In a recent case study investigating prioritisation techniques, participants found the resulting priority order incorrect when using Wiegers' method. Some participants changed their estimates in order to get a better priority order, when the results given by the method seemed wrong (Lehtola and Kauppinen, 2004). This accuracy of the resulting priority order is interesting to investigate and therefore we compare the subjective accuracy of the techniques in this experiment.

3.2 Experiment Design

The experiment was carried out with a *repeated measures design, using counterbalancing* i.e., all subjects used both techniques (Robson, 1997; Wohlin et al., 2000). The 16 subjects in the convenient sample included 15 Ph.D. students (10 male and

5 female) in their first or second year, and one professor (male). The experiment was conducted as part of a research methodology course. Before the experiment, a pre-test was performed. The experiment was carried out during a one-day session, which included an introduction to the task, the experiment itself, a post-test, and finally a concluding discussion of the experiment implementation. In addition, a few weeks after the experiment a second post-test was conducted. Figure 2 outlines the activities performed in Experiment 1.

The requirements used in the prioritisation were mobile phone features, which are requirements on a high level of abstraction and rather independent. The prioritisation was performed without taking requirements dependencies into account.

The trade-off between cost and value, often faced by a development organisation, was difficult to investigate for our subjects, as the cost of developing a certain requirement is difficult for laymen to estimate. Therefore the criterion Price was selected instead, as the trade-off faced by consumers regards the Value of different functions in the phone and the Price of the phone. The criteria are defined as follows:

- The Value criterion corresponds to how important and valuable the subject find the requirement.
- The Price criterion corresponds to how much the subject thinks the requirement adds to the price of the mobile phone.

The Value criterion has probably been regarded by most subjects when buying or comparing mobile phones. The Price criterion may also be accounted for since buying or comparing mobile phones gives a clue of how the price differs depending on the included requirements. Thus, there is a trade-off between Value and Price when buying a mobile phone.

The two requirements prioritisation techniques described in Section 2.1 and 2.2 were used as input to the experiment, but were modified in order to be more comparable. The PWC is conducted using the AHP for calculating requirements priorities. A flat requirements structure was used, i.e., the system aspect of AHP was not considered in our PWC technique (Saaty and Vargas, 2001). Neither did we use

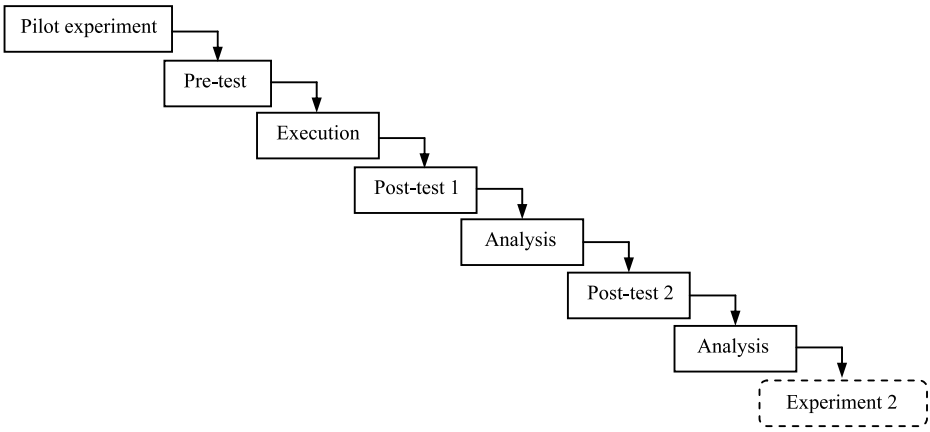

Fig. 2 Activities conducted in Experiment 1

any of the possible ways of reducing the number of comparisons, thus the pair-wise comparisons were exhaustive. PG was modified so that the piles were labelled according to the Value and Price criteria: (1) Necessary, (2) Adds to the value and (3) Unnecessary, and (1) Very high price, (2) Reasonable price and (3) Low price, respectively. Thus, the aspects of implementation cost and risk, which are emphasised in XP were substituted by Price in our experiment to make it reasonable for laymen to estimate.

3.2.1 Pilot Experiment

A pilot experiment was performed before the main study to evaluate the design. Six colleagues participated and they prioritised ten requirements each, with both techniques. After this pilot experiment, it was concluded that the experiment should be extended to 8 and 16 requirements in order to capture the difference depending on the number of factors to prioritise. Another change was to let the subjects use the techniques and criteria in different orders to eliminate order effects. Further, changes to the PWC sheets included to remove the scale and instead use "more than" and "less than" signs so that the participants would not focus on the numbers, and to arrange the pairs randomly on each sheet.

3.2.2 Pre-Test

Before the session, the subjects were exposed to a *pre-test* in order to get a foundation for sampling. A questionnaire was sent out by e-mail in order to capture the knowledge about mobile phones and the subjects' knowledge and opinions of the two prioritisation techniques. The pre-test was used to divide the subjects into groups with as similar characteristics as possible.

Another objective with the pre-test was to investigate how well the subjects could apprehend the price of mobile phone requirements. A majority of the subjects stated that they consider buying a new mobile phone at least every second year, and therefore we believe that their knowledge of mobile phone prices is fairly good.

3.2.3 Execution

The experiment took place in an ordinary lecture room during a one-day session. Data was mainly collected through questionnaires where the subjects filled out the time spent on each task and their opinions on the techniques.

The domain in this experiment was mobile phones and according to the pre-test, all subjects were familiar with this context. The factors to prioritise were mobile phone requirements, for example SMS, Games, WAP, Calendar, etc. (see Appendix for complete list).

One intention of the experiment was to investigate if a different number of requirements would affect the choice of preferred technique. Therefore, half of the subjects were asked to prioritise 8 requirements, while the other half prioritised 16 requirements. Another intention was to investigate if the order in which the techniques were used would affect the choice of preferred technique. Therefore, half of the subjects started with PWC and half started with PG. The order of the Value and Price criteria was also distributed within the groups in order to eliminate

order effects. Thus, the experiment was performed using a counter-balancing design, as shown in Appendix.

The experiment was conducted in a classroom with the subjects spread out. Each subject was given an experiment kit consisting of the PWC sheets and the PG cards.

For PWC, one sheet per criterion and person had been prepared, with all possible pair-wise combinations of the requirements to compare. For the purpose of eliminating order effects, the order of the pairs was randomly distributed so every subject received a different order of the comparisons. With 16 requirements to compare, there was $16(16-1)/2 = 120$ pair-wise comparisons for Value and Price, respectively. With 8 requirements, there was $8(8-1)/2 = 28$ pair-wise comparisons for Value and Price, respectively. In between each pair in the sheets there was a scale where the difference of the requirements' Value or Price was circled, see Fig. 3. To be able to try different scales, no scale numbers were written on the sheets. Instead, a scale with 9 different "more than," "equal" and "less than" symbols was used. The further to the left a symbol was circled, the more valuable (or expensive) was the left requirement than the right one. If the requirements were regarded equally valuable (or expensive) the "equal" symbol was circled.

For PG, the subjects were given two sets of cards (one set for Value and one for Price) with one mobile phone requirement written on each. The cards were partitioned into three piles, separately for the Value criterion and the Price criterion, see Fig. 4. The piles represent (1) Necessary, (2) Adds to the value and (3) Unnecessary, for the Value criterion, and (1) Very high price, (2) Reasonable price and (3) Low price, for the Price criterion.

Within the piles, the cards were then arranged so that the most valuable (or expensive) one was at the top of the pile and the less valuable (or expensive) were put underneath. Then the three piles were put together and numbered from 1 to 8 and 1 to 16 so that a single list of prioritised requirements was constructed for each criterion.

The subjects were given approximately 2 hours to conclude the tasks, which was enough time to avoid time-pressure. During the experiment, the subjects were instructed to note the time-consumption for each prioritisation. Further, the subjects had the possibility to ask questions for clarification.

3.2.4 Post-Test 1

The subjects handed in their experiment kit after finishing the tasks and were then asked to fill out a post-test. This was made in order to capture the subjects' opinions right after the experiment. The test included the questions below, as well as some optional questions capturing opinions about the techniques and the experiment as a whole. The questions were answered by circling one of the symbols "more than," "equal" or "less than."

1. Which technique did you find easiest to use?
2. Which technique do you think gives the most accurate result?

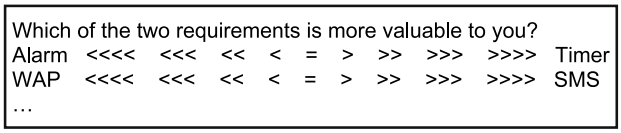

Fig. 3 Example of PWC sheet

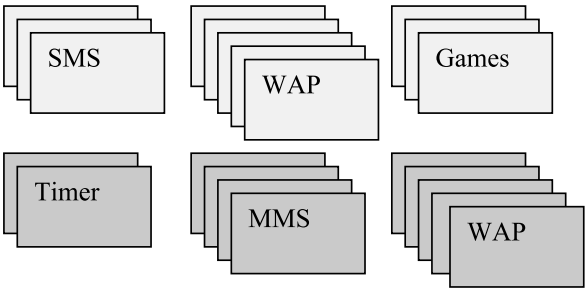

Fig. 4 Example of PG cards

3.2.5 Post-Test 2

After completing the analysis, the subjects were, in a second post-test, asked to state which technique that, in their opinion, gave the most accurate result. They were sent two sheets (one for Value and one for Price) with two different lists of requirements, corresponding to the results from the PG and PWC prioritisations. The post-test was designed as a *blind-test*, thus the subjects did not know which list corresponded to which technique, but were asked to select the list they felt reflected their opinions the most. In order to get comparable lists, the ratio scale from PWC was not shown, and neither was the pile distribution from PG.

3.3 Threats to Validity

In this section, the threats to validity in the experiment are analysed. The validity areas considered are conclusion, internal, construct and external, according to Wohlin et al. (2000).

Conclusion validity concerns the relationship between the treatment and the outcome. Robust statistical techniques are used, measures and treatment implementation are considered reliable. The data was plotted and tested to check if it was normally distributed. In all cases, the data could not be concluded to be normally distributed and, thus, non-parametric tests were used. However, a threat is low statistical power, since only 16 subjects were used.

Furthermore, we have tried to increase the reliability of measures by conducting a pilot experiment and thereafter adjusting the wording and instrumentation. Another issue is that objective measures, e.g., time-consumption, are more reliable than subjective ones, e.g., ease of use and accuracy. However, the subjective measures are very important in this experiment and therefore we have chosen to include them. The experiment took place during one single occasion and therefore the implementation and setting are not a threat in this case.

Internal validity concerns the relationship between the treatments and the outcome of the experiment. The internal threats that may have affected the experiment are the fatigued effect, testing and group pressure. The subjects could become fatigued during the experiment, which may affect the concentration. In particular, the subjects who perform the tasks with 16 requirements may get tired or bored. This has been checked in the analysis, by calculating the consistency index for PWC. There is no significant difference in consistency for groups using different

number of requirements (see Table 8). Hence, we draw the conclusion that the threat to the fatigue is low.

The testing threat is that the subjects get practice during the experiment and unconsciously get an opinion on the context using the first technique, which will affect the result for the second technique. At least when using PG first, it may affect the PWC performance. In Table 9, the order effect on consistency is analysed. There is no statistical difference in consistency depending on the order. Hence, this indicates that learning effects have not affected the experiment.

The third internal threat is the group pressure that may affect the subjects to rush through the task. In Section 3.4.4, there is an analysis of the correlation between the time used by the subjects and the consistency. The data indicates that the time-consumption has not affected the consistency of the prioritisation.

Construct validity concerns the relation between theory and observation. One threat in the design has been observed. It would have been valuable to start the session with an introduction explaining each requirement in the prioritisation to clarify their meaning. However, the subjects had their own interpretation of the requirements, which was the same throughout the experiment and therefore this should not affect the result.

External validity concerns whether the outcome of the experiment can be generalized to the population. Threats to external validity limit the generalisability of the experiment to industrial practice. The subjects are sampled from software engineering PhD students. Hence, the outcome of the experiment can be generalized to this group. In addition, for this experimental context it is likely that this group would perform equally to the requirements engineers and product managers who are intended to use the techniques in practice. The subjects are familiar with the application domain (mobile phone requirements) and several of the participants had prior working experience. The difference between industrial professionals and students in their final years has been considered small in other studies (Höst et al., 2000; Runeson, 2003). Furthermore, if a student experiment shows that one technique is better than another it is rather unlikely that professionals would come to the opposite conclusion (Tichy, 2000).

As most experimental conditions, the time is an important factor. In order to reduce the time needed for the experiment, the number of prioritised requirements is rather few. In most real cases, the total number of requirements is higher and therefore the results found in this paper may be valid if the prioritisation is performed on a subset of the requirements. This may be the case e.g., if only the newly arrived requirements are prioritised or only the requirements for a certain sub system. It is difficult to judge whether extending the number of requirements would lead to the same result. Therefore, future replications and case studies have to be made in order to draw conclusions when more requirements are used.

As the requirements used in this experiment are rather independent, they may have been easier to prioritise than is usually the case in industry. For example, the time required to perform the prioritisation would probably be larger in an industrial case due to more difficult trade-offs and dependencies between requirements. Requirements dependencies can require a group of requirements to be selected for a release instead of individual ones. This has not been investigated in the experiment.

A recent study investigated different criteria for selecting requirements for a certain release (Wohlin and Aurum, 2005). The results indicate that technical

concerns, such as requirements dependencies, are less important than management-oriented criteria when deciding which requirements to select for a project or release. Therefore it is likely that requirements dependencies would have a relatively small effect on the results in an industrial case. We believe that these results may be used as a pilot for identifying trends before conducting a study in industry (Berander, 2004).

In summary, the main threats to the validity are that fewer, and more independent, requirements were used than in most industry cases. Hence, future replications are needed in order to reduce these threats. We believe that the other threats are under control. However, one mistake was made during the experiment. The scales "more than" and "less than" in the PWC sheets were accidentally switched so that it could be interpreted in the opposite way than was intended (see Fig. 3). This caused some confusion during the experiment. However, the interpretation was explained and clarified and therefore this should not be considered as a threat to validity.

3.4 Data Analysis

The analysis of the experiment was divided between two independent researchers, in order to save time and to perform spot checks so that the validity could be further improved. The analysis was performed with Microsoft Excel™, the computing tool MATLAB™ and the statistical analysis tool StatView™.

Two different scales were tried for the PWC analysis: $1 \sim 5$ and $1 \sim 9$. According to Zhang and Nishimura (1996) the scale $1 \sim 5$ is better than $1 \sim 9$ at expressing human views and therefore the scale $1 \sim 5$ was used when compiling the prioritisation ranking lists.

Furthermore, Saaty and Vargas (2001) have calculated *random indices* (RI) that are used in the calculation of the consistency ratios. Unfortunately, this calculation only includes 15 factors while this experiment included as many as 16 factors, i.e., requirements. Therefore, the RI scale was extrapolated and the RI for 16 requirements was set to 1.61.

3.4.1 H1: Time-Consumption

The time to conclude the prioritisation is larger with PWC than with PG, for both criteria. As Table 2 shows, the difference in time between the two techniques is 6.1 minutes for 8 requirements, which corresponds to an increase of 43%, and 14.7

Table 2 Average time-consumption for the prioritisation

Nbr of requirements	Criteria	PG	PWC	Difference
8	Value	3.6 min	7.8 min	4.2 min
	Price	4.5 min	6.4 min	1.9 min
Total		**8.1 min**	**14.2 min**	**6.1 min**
%				**43%**
16	Value	6.5 min	12.6 min	6.1 min
	Price	5.5 min	14.1 min	8.6 min
Total		**12.0 min**	**26.7 min**	**14.7 min**
%				**55%**
% increase		48%	88%	

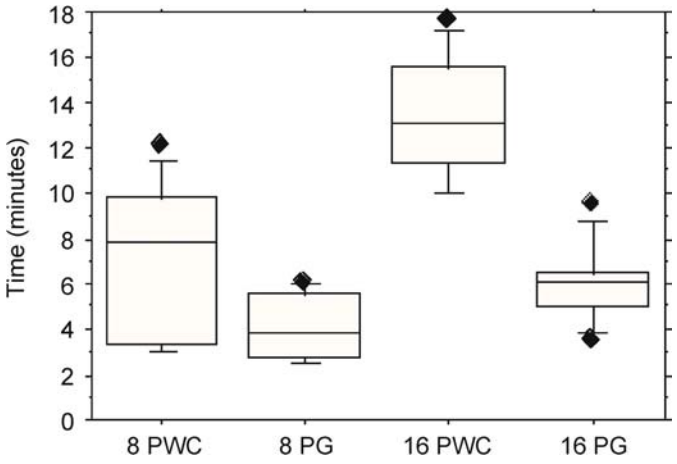

Fig. 5 Box plots of the time spent on prioritisation

minutes for 16 requirements, which corresponds to an increase of 55%. Thus, for 16 requirements, it takes more than twice as much time to use the PWC compared to the PG, while for 8 requirements, the difference is a bit smaller.

The time increase in percent from 8 to 16 requirements for PWC is 88%, while the same for PG is only 48%. Thus, a larger number of objects to prioritise affect the time-consumption for PWC more than for PG, at least when using 8 and 16 requirements.

This can also be seen in Fig. 5, where the median values are higher for PWC than for PG, and the difference between 8 and 16 requirements is larger for PWC than for PG. Additionally, the box plot indicates that the subjects' time to conclude the prioritisation with PWC are more dispersed.

As Table 3 shows, the subjects have in average used less time per requirement when they had more requirements to prioritise. It is particularly interesting to see that it takes less time per requirement to perform PG partitioning with 16 requirements than with 8. One could expect that it should be more complex to perform PG with more requirements but this result show that more requirements tend to speed up the prioritisation per requirement. However, there might be a breakpoint when the number of requirements is too great and it becomes hard to get the valuable overview of the PG cards.

Four hypothesis tests were performed, for 8 and 16 requirements respectively, and one for each criterion. The frequency distribution was plotted in histograms to check the distribution. Due to the not normally distributed sample, we chose a non-parametric test, the *Wilcoxon test* (Siegel and Castellan, 1988). The hypothesis tests show that on the 5%-level there is a significant time difference for three of the four cases. This is illustrated in Table 4, where the p-value is lower

Table 3 Time-consumption per requirement

Nbr of requirements	PG	PWC
8	30.5 s/requirement	53.5 s/requirement
16	22.5 s/requirement	50.0 s/requirement

Table 4 Wilcoxon tests for the time difference

Nbr of requirements	Criteria	Wilcoxon p-values
8	Value	0.0251
	Price	0.1159
16	Value	0.0209
	Price	0.0117

than 5% in three of the four cases. Thus, the first null hypothesis is rejected for these cases.

3.4.2 H2: Ease of Use

Immediately after the experiment, the subjects filled out the first post-test that, among other things, captured the opinions of the techniques' ease of use. Among the 16 subjects, 12 found PG easier or much easier to use than PWC. Only 3 found them equally easy and 1 stated that PWC was easier to use, see Table 5. Hence, 75% of the subjects found PG easier to use.

This was tested in a Chi-2 test (Siegel and Castellan, 1988) by comparing the number of answers in favour of PWC to the number of answers in favour of PG. It turned out that there is a statistically significant difference, as $p = 0.0023$. Thus, the second null hypothesis is rejected.

It seems as if the subjects prioritising 16 requirements are a bit more sceptical to PG than those prioritising 8 requirements. This could indicate that the more requirements the more difficult to keep them all in mind.

3.4.3 H3: Accuracy

Directly after the experiment, the subjects performed the first post-test that captured which technique the subjects *expected* to be the most accurate. As Table 6 illustrates, a majority of the subjects expected PG to be better, while less than a fifth expected PWC to be better.

In order to evaluate which technique that gave the most accurate results, a second post-test was filled out by the subjects. This was done a few weeks after the experiment was performed, when the analysis was finished.

The most common opinion among the subjects was that PG reflects their views more accurately than PWC. This is shown in Table 7 where 47% of the subjects were in favour of PG and only 28% were in favour of the PWC. This is, however, *not* statistically significant, $p = 0.2200$ with a Chi-2 test (Siegel and Castellan, 1988),

Table 5 Results from the first post-test: Ease of use

Nbr of requirements	PG much easier	Easier	Equally easy	Easier	PWC much easier
8	4	3	1	0	0
16	4	1	2	1	0
Total	8	4	3	1	0
Total %	**50%**	**25%**	**19%**	**6%**	**0%**

Table 6 Results from the first post-test: expected accuracy

Nbr of requirements	Favour PG	Equal	Favour PWC
8	4	3	1
16	5	1	2
Total	9	4	3
Total %	**56%**	**25%**	**19%**

so it cannot be determined if there is a difference between the techniques' accuracy. Thus, the null hypothesis is not rejected. Half of the ones that have stated that both techniques are equally accurate actually had the same order in the lists.

An interesting observation is that this implies that PG was actually not as good as the subjects expected even if most subjects preferred PG to PWC.

3.4.4 Consistency Ratio

The consistency ratio (CR) describes the amount of judgement errors that is imposed during the pair-wise comparisons. The CR is described with a value between 0 and 1 and the lower CR value, the higher consistency. Saaty and Vargas (2001) have recommended that CR should be lower than 0.10 in order for the prioritisation to be considered trustworthy. However, CR exceeding the limit 0.10 occurs frequently in practice (Karlsson and Ryan, 1997).

The CR limit above is only valid for the scale $1 \sim 9$, and in this experiment the scale $1 \sim 5$ was used instead. Therefore, the limit for acceptable CR will be lower. The average consistency ratios for scale $1 \sim 5$ are presented in Table 8.

The frequency distribution for the consistency was plotted in histograms to check the distribution. The data was not normally distributed and therefore we chose a non-parametric test. The Wilcoxon test resulted in $p > 0.30$ for both criteria. Therefore, it cannot be proved, on the 5%-level, to be a significant difference in consistency depending on the number of requirements prioritised.

In order to investigate if the time spent on each comparison affects the consistency, the correlation between these parameters was calculated. The Spearman rank-order correlation coefficients indicate no correlation between the time and the consistency, as the correlation varies between -0.40 and 0.20. According to Siegel and Castellan (1988), the absolute value of the correlation coefficient should be greater than 0.738 in order for the correlation to be considered significant in this case. Hence, the consistency is not particularly influenced by the time spent on prioritisation.

Table 7 Results from the second post-test: Perceived accuracy

Nbr of requirements	Criteria	Favour PG	Equal	Favour PWC
8	Value	6	2	0
	Price	1	3	4
16	Value	4	1	3
	Price	4	2	2
Total		15	8	9
Total %		**47%**	**25%**	**28%**

Table 8 Mean consistency ratios

Criteria	Nbr of requirements	Scale 1 ~ 5
Value	8	0.106
	16	0.082
Price	8	0.101
	16	0.120

3.4.5 Order Effects

There is a chance that the order in which the two techniques are used can influence the result. Table 9 shows that the mean consistency ratio is a bit lower for the subjects who used PG before PWC. This may indicate that using PG can provide an image of ones preferences that are not possible to get from using PWC. Therefore it may be easier to be consistent when PG precedes PWC.

However, the hypothesis tests show that the difference is not significant on the 5%-level. Due to the not normally distributed sample, we chose a non-parametric test, the *Mann-Whitney test* (Siegel and Castellan, 1988). The p-values are larger than 0.6, and therefore we cannot confirm a significant difference depending on the order.

A set of significance tests was also conducted investigating the order effect on time-consumption. Neither of the cases (with 8 or 16 requirements or with Value or Price criterion) showed a significant difference in time depending on the order in which the techniques were used.

This finding validates that the experiment analysis has not suffered from any order effects, neither regarding time nor consistency.

3.4.6 Qualitative Answers

In the post-test performed right after the experiment, the subjects had the opportunity to answer some optional questions about their general opinion. Opinions about PWC include "effort demanding but nice," "it feels like a black-box wherein you pour requirements," "good but boring," "it feels like you lose control over the prioritisation process," and "straightforward." Opinions about PG are for example "fast and easy," "lets the respondent be creative," "intuitive," "prone to errors," "good overview," and "logical and simple." These opinions correspond well to the results of the captured subjective dependent variables: ease of use and expected accuracy, discussed in prior sections.

3.5 Results

The main results are that the PG technique is superior to the PWC regarding the two variables time-consumption and ease of use, while it could not be determined which technique that has the highest accuracy.

Table 9 Order effect on consistency

Mean consistency	PWC-PG	PG-PWC
Value	0.107	0.082
Price	0.119	0.102

Two groups prioritised 8 and 16 requirements, respectively, in order to investigate if there is a breakpoint between 8 and 16 where one of the methods is more efficient than the other. It was suspected that a greater number of requirements would eliminate the valuable overview in PG, since it would be difficult to keep all requirements in mind. However, this experiment only shows an insignificant tendency of less overview affecting the ease of use when prioritising 16 requirements (see Table 5). Therefore, it is suspected that the breakpoint is at an even higher number of requirements.

Another interesting observation in this experiment was that the time-consumption did not affect the consistency in PWC (see Section 3.4.4). One could assume that if someone rushes through the comparisons, the consistency would be poor. However, these are only initial results and with another set of objects to prioritise, the results might be different.

The objective measure *total time-consumption* is higher for PWC than for PG both in our study and in the one by Karlsson et al. (1998). On the other hand, PWC was given a higher rank than PG regarding the subjective measures ease of use and fault tolerance in the study by Karlsson et al. (1998). Our experiment shows that PG is easier to use than PWC. This difference in result may be due to differences in methodology. While our study is a controlled experiment with 16 participants, Karlsson et al. (1998) is based on an evaluation by three individuals who discussed their opinions. The result regarding time-consumption is considered reliable, while the difference regarding ease of use indicates that additional studies need to be performed in order to further understand the strengths and weaknesses of these techniques.

Karlsson et al. (1998) suggested a combination of the two techniques Priority groups and PWC, in order to use the PWC with a reasonable amount of effort. Using PWC on the three priority groups, separately, would decrease the number of comparisons. Another possibility is to use PWC only on those requirements that end up in the middle priority pile. This would imply that PG, or Priority groups, is used first, to divide the requirements into three groups. The high priority group of requirements will most certainly be implemented, the low priority group will be postponed and looked into in a following release, while the ones in the middle need special treatment to determine the outcome.

This approach agrees with what Davis (2003) has written about the *requirements triage* where he recommends requirements engineers to focus on the difficult requirements and skip the ones that will either be implemented or rejected anyway. In this manner, PWC can be used on the requirements that are difficult to estimate and need a more precise scale for determining its cost and value. The technique's ratio scale and fault tolerance would then come to its right.

4 Experiment 2

This section describes the second of the two experiments, the experiment planning, operation and analysis. Finally, the section is concluded by a discussion. Much of the design in the first experiment have been reused in the second one, therefore several references are made to Section 3.

The motivation for the second experiment is that although the first experiment indicates that PG is superior to PWC, we suspect that PWC with tool-support may have certain benefits for practitioners. With tool-support it is possible to reduce the number of comparisons and to visualise the priorities. It may also be easier to use, as it guides the decision maker during the prioritisation process. We believe that the PWC would benefit more than PG from tool-support, and therefore we chose to investigate the tool-supported PWC (TPWC) and compare it with PG.

4.1 Hypotheses and Variables

The goal of the second experiment is to compare two prioritisation techniques and to investigate the following null hypotheses:

H_o1 The average time to conclude the prioritisations is equal for both techniques, PG and TPWC.
H_o2 The ease of use is equal for both techniques, PG and TPWC.
H_o3 The accuracy is equal for both techniques, PG and TPWC.

The alternative hypotheses are formulated below:

H_A1 The average time to conclude the prioritisations is not equal for both techniques, PG and TPWC.
H_A2 The ease of use is not equal for both techniques, PG and TPWC.
H_A3 The accuracy is not equal for both techniques, PG and TPWC.

The independent variables are the techniques PG and TPWC and the dependent variables are the same as in the first experiment, i.e., *average time to conclude the prioritisations*, *ease of use* and *accuracy*.

The time-consumption was captured by each subject by noting their start and stop time for each task, the ease of use was measured by a questionnaire which was filled out by all subjects after the experiment, and the accuracy was measured by conducting a post-test a few weeks after the experiment similarly to Experiment 1.

4.2 Experiment Design

The second experiment was also carried out with a *repeated measures design, using counter-balancing,* i.e., all subjects used both techniques. The subjects were 30 MSc students (25 male and 5 female) in their final year, taking an optional requirements engineering course. The experiment was conducted within a compulsory laboratory session in the area of requirements prioritisation. The session was conducted for teaching purposes and gave the students an opportunity to try out and compare two commonly known prioritisation techniques. No pre-test was performed, so the participants were randomly assigned to perform the tasks in a certain order.

The experiment was divided into two separate occasions with 20 subjects at the first session and 10 at the second. Both sessions were guided by two teachers. Before the experiment the participants were given an introduction to the tool by conducting a comprehensive tutorial.

PG was used in the same manner as in the first experiment. The TPWC used a requirements management tool with pair-wise comparisons as prioritisation technique. The participants conducted the number of comparisons required by the stopping rules in the tool (approximate size $2n$), and could revise comparisons when inconsistency was indicated by the tool. Note that the tool was used only as an approach to prioritisation, i.e., the visualisation possibilities in the tool were not investigated.

Two post-tests, which are described below, were performed similarly to the first experiment in order to capture the dependent variables. Figure 6 illustrates the activities performed in the second experiment.

4.2.1 Execution

The experiment took place in a computer laboratory room during a half-day session. The manual technique PG was used in the same room but the students could move to empty desks.

For each subject an experiment kit had been prepared, consisting of the PG cards and a personal instruction regarding the order to perform the tasks. Each subject also had a personal login to the prioritisation tool.

Data was mainly collected through post-tests. The PG priority piles were attached with a paper clip and handed in, while the TPWC lists were compiled by the researcher after the session by extracting the information needed from the requirements management tool. Each subject noted the start and stop time in the post-test conducted right after the experiment, as well as their opinion on ease of use. Then, the second post-test captured the accuracy through a blind-test a few weeks later. The subjects were given 2 hours to perform the tasks, including the introductory tutorial for the tool.

The design of the second experiment is very similar to the first one, since it was intended to investigate the same hypotheses. Thus, the main difference was that the PWC was tool-supported in the second experiment.

Furthermore, since the first experiment showed that the number of requirements did not affect the outcome of the first experiment, it was decided to have all participants prioritise between 16 requirements. The same mobile phone requirements were used, as well as the same criteria Value and Price. The counterbalancing design is illustrated in Appendix.

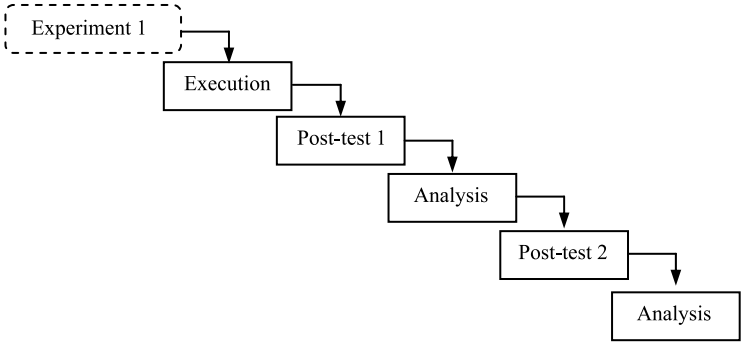

Fig. 6 Activities conducted in Experiment 2

4.2.2 Post-Test 1

The subjects were asked to fill out the same post-test as in experiment 1 after they had handed in their experiment kit. This was made in order to capture the subjects' opinions right after the experiment.

4.2.3 Post-Test 2

A few weeks after the experiment, the subjects were, in a second post-test, asked to state which technique they found most accurate. This was conducted as a blind-test in the same way as for the first experiment, but the lists corresponded to the results from the two techniques PG and TPWC.

4.3 Threats to Validity

This section discusses the threats to validity for the second experiment. The same four classes of validity threats as for the first experiment are considered in this section.

Conclusion validity. As in the first experiment, the statistical techniques, measures and treatment implementation are considered reliable. Both objective and subjective measures are used. The student group is a homogeneous group, with similar background and education.

Internal validity. The internal threats in the second experiment is the fatigue effect, mortality and the instrumentation. The fatigue threat is present since one of the sessions took place after office hours. However, there was no disturbance during the performance.

Furthermore, the subjects could be influenced by the first priority list, and unconsciously prioritise in a similar manner when producing the second priority list. On the other hand, when conducting the pair-wise comparisons, it is difficult to use knowledge from another prioritisation, which reduce the threat. In addition, in the first experiment, the threat was estimated as low, and there are no indications that it would be higher in the second experiment.

Another threat is the mortality effect. This is small, but present since one of the subjects was absent during the second post-test and therefore one data point is missing.

There is also a potential instrumentation threat. The TPWC technique was not used to visualise the priority list since we intended to conduct the second post-test with a comparison of the lists from the two techniques. However, the PG technique directly results in a priority list and it can therefore not be hidden. Therefore, some subjects may have remembered the priority order from the PG and could thereby identify which of the lists in the second post-test that correspond to which technique. This may also have been the case in the first experiment, but then the time between the experiment and the second post-test was longer, which reduces the risk of remembering. However, we believe that the subjects chose a list based on perceived accuracy and not based on remembering which list the priorities come from.

Construct validity. The experiment would need another set of requirements to perform the prioritisation on in order to be able to discover if the results are the same, or if the set of requirements have affected the results. Testing and treatment

may interact. When the subjects know that the time is measured, it is possible that they get more aware of the time they spend, and thus the time-consumption is affected. However, they were not aware of the other two measures when conducting the experiment, so only the time can have been affected.

External validity. The subjects in the experiment is a homogenous group. This improves the conclusion validity, but makes it more difficult to generalise the result to a broader population. This issue was discussed in Section 3.3 for the subjects in the first experiment and the same discussion is valid for the subjects in this experiment.

As discussed in Section 3.3, the small number of requirements decreases the possibility to generalise to cases where a higher number of requirements is prioritised.

In summary, the main threat in this experiment is the instrumentation threat and that it is difficult to generalise to situations where a larger set of requirements are prioritised.

4.4 Data Analysis

This section presents the analysis and results from the second experiment. The analysis was performed by two researchers using Microsoft Excel™, the computing tool MATLAB™ and the statistical analysis tool StatView™.

4.4.1 H1: Time-Consumption

The first hypothesis regards the time needed to perform the prioritisation. As can be seen in Table 10 the average time required is lower for both criteria when using the TPWC. In fact, TPWC required 17% less time than PG.

As can be seen in the box plots in Fig. 7, where the times for both criteria are added, the median values are higher for PG than for TPWC. The times for PG are also more dispersed.

Normal probability plots indicated that the data was not normally distributed. Therefore, it was decided to use non-parametric tests during analysis. The difference in time is significant on the 5%-level as the *Wilcoxon test* (Siegel and Castellan, 1988) results in p-values below 0.04. Therefore we can draw the conclusion that the TPWC technique is a faster technique than the PG, i.e., the null hypothesis is rejected.

4.4.2 H2: Ease of Use

After using both techniques, the participants handed in a post-test answering the question "Which techniques did you find easiest to use?" In total, 10 of the 30 subjects found the PG easier or much easier to use, while 16 pointed out TPWC as

Table 10 Average time-consumption (in minutes)

Criteria	PG	TPWC	Difference
Value	5.8	4.8	1.0
Price	5.5	4.6	0.9
Total	11.3	9.4	1.9
%			**17%**

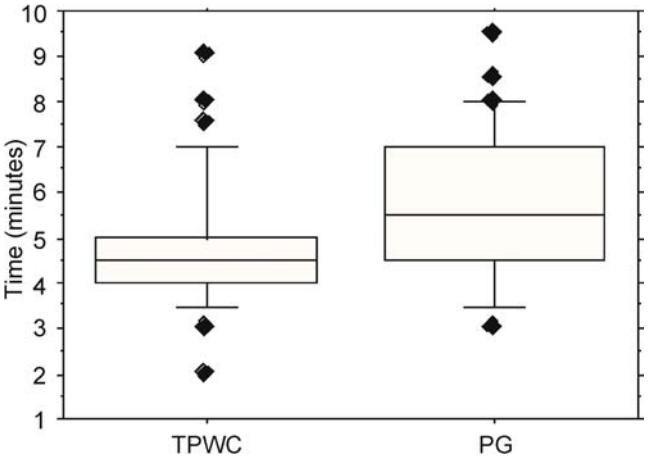

Fig. 7 Box plots for the time spent on prioritisation

easier or much easier. As can be seen in Table 11, this corresponds to that 33% found PG easier, while 53% found TPWC easier. 4 of the subjects, i.e., 13%, found the techniques equally easy to use.

A Chi-2 test (Siegel and Castellan, 1988) shows that the difference between the number of subjects that found PG easier and the number of subjects that found TPWC easier is not significant, p = 0.2393. Thus, the null hypothesis cannot be rejected.

4.4.3 H3: Accuracy

The first post-test also captured which technique the participants *expected* to be the most accurate. As can be seen in Table 12, 77% of the subjects expected the TPWC to be more accurate than the PG. Thus, a majority of the subjects found the tool trustworthy after using it.

However, the second post-test investigated which of the techniques the subjects found most accurate by conducting a blind-test where the subjects were given their priority lists from both techniques. Due to absence, only 29 of the 30 participants filled out the second post-test. As can be seen in Table 13, where both criteria are added, 50% found the PG lists more accurate, while 37% found the TPWC lists more accurate. 12% found the priority lists equally accurate. This difference is not statistically significant, as the p-value turned out to be 0.3270 in a Chi-2 test.

Thus, the TPWC did not get as high accuracy as expected, while PG turned out to be more accurate than expected. The null hypotheses cannot be rejected.

Table 11 Results from the first post-test: ease of use

Ease of use	PG much easier	Easier	Equally easy	Easier	TPWC much easier
	1	9	4	11	5
%	**3%**	**30%**	**13%**	**37%**	**16%**

Table 12 Results from the first post-test: expected accuracy

Expected accuracy	PG much more accurate	More accurate	Equally accurate	More accurate	TPWC much more accurate
	0	5	2	14	9
%	**0%**	**17%**	**7%**	**47%**	**30%**

4.4.4 Order Effects

A set of significance tests was used to investigate whether or not there was a significant order effect on the time-consumption. The time-difference depending on the order in which the techniques were used, was investigated with a *Mann-Whitney test.* The test did not indicate a significant time-difference ($p > 0.10$ for Value and $p > 0.90$ for Price) and therefore we cannot show a significant order effect. Another set of tests investigated the effect on time-consumption depending on the order in which the criteria were used. In this case, a *Wilcoxon test* was used and could not show a significant order effect ($p > 0.60$ for both TPWC and PWC).

A third significance test was used to investigate if the occasion (afternoon or evening) affected the time-consumption. The Mann-Whitney test indicates no significant difference in time-consumption ($p > 0.40$ for Value and $p > 0.10$ for Price). Thus, we cannot determine any significant effect on the time-consumption depending on different orders.

4.4.5 Qualitative Answers

Some personal opinions on the experiment and the two techniques were also collected. Among the positive views on TPWC are "TPWC is probably better than PG for larger projects" and "TPWC is easy to use." There were also some opinions in favour of the PG, "TPWC makes it difficult to keep focus with many requirements" and "PG gives a better overview of the requirements."

4.5 Results

One of the main reasons for conducting the second experiment was that it was suspected that the manual PWC in the first experiment would benefit from tool support so that the drawbacks of e.g., high time-consumption could be reduced. The main result from the second experiment is that the Tool-supported PWC is a faster technique than the PG. Thus, the first null hypothesis could be rejected. However, although there are more subjects finding TPWC easier to use than PG, the

Table 13 Results from the second post-test: perceived accuracy

Perceived accuracy	PG much more accurate	More accurate	Equally accurate	More accurate	TPWC much more accurate
Value	0	15	1	12	1
Price	1	13	6	8	1
Total	1	28	7	20	2
%	**2%**	**48%**	**12%**	**34%**	**3%**

difference is not statistically significant. A difference in accuracy could not be determined either. Thus, the second and third null hypotheses could not be rejected.

There were no significant order effects depending on e.g., the order in which the techniques were used. However, using PG first and then TPWC on the separate piles would still decrease the necessary time-consumption, although the time reduction would not be as large as in the case with PG and manual PWC. This is due to the fact that TPWC only require approximately *2n* comparisons.

5 Discussion

Prioritisation is a very important activity in requirements engineering because it lays the foundation for release planning. However, it is also a difficult task since it requires domain knowledge and estimation skills in order to be successful. The inability to estimate implementation effort and predict customer value may be one of the reasons why organisations use ad hoc methods when prioritising requirements. For a prioritisation technique to be used it has to be fast and easy to manage since projects often have limited time and budget resources.

The experiments presented in this paper have investigated the time-consumption, ease of use and accuracy for different prioritisation techniques. But when deciding which prioritisation technique to use in an organisation there are several other aspects to take into consideration. The technique has to be supported by the software process and other project activities. For example, the PG may be successful if the overall development approach is agile and requirements are already written on e.g., story cards (Beck, 1999). On the other hand, if a requirements management tool is used and requirements are already stored in the tool, it is evidently natural to use it as a means for prioritisation as well. Thus, it is necessary to consider methods and tools for requirements prioritisation to be aligned with other methods and tools used in the organisation.

Another related issue is the necessary analysis effort that must be used in order to get a priority list from the conducted prioritisation. In PG, the result is in the form of ranked piles of cards, which need to be transformed into e.g., a Cost-value diagram in order to sufficiently visualise the trade-off between cost and value. The manual PWC requires plenty of analysis and matrix calculations before a priority list can be extracted and visualised. This analysis is not realistic to perform manually when the number of requirements grows. The PWC can provide additional information compared to PG, such as the consistency, and the data is on a ratio scale. The Tool-supported PWC has several different visualisation possibilities and the tool takes care of all calculations and displays the prioritisation in charts and diagrams. Thus, the required analysis differs between the techniques and it needs to be taken into consideration before deciding on a prioritisation technique.

Furthermore, minor investments are needed for the manual techniques since all resources required are pen and paper, while on the other hand more staff resources are needed to analyse the outcome, as discussed above. Commercial tools may be expensive and can be risky to rely on since anything from computer crashes to vendor bankruptcy can occur. On the other hand, it visualises priorities without any extra effort.

The generalisability of the study is limited due to the rather small sample and the specific context. Although the subjects may have opinions similar to decision

makers in industry, the context of mobile phone requirements may be a bit too simplistic. The main weakness is that mobile phone requirements are on a high level and rather independent, while requirements in a real case often have inter-dependencies. Industrial projects also have time and budget pressure to consider, which complicates the decision making. It is possible that industrial experience would affect the results, although we believe that in a relative comparison between the techniques, it is likely that the results would be similar.

The validity of controlled experiments using students as subjects is often debated. In Carver et al. (2003) it is acknowledged that the higher level of control in controlled experiments compared to e.g., case studies may help researchers ascertain whether a phenomenon of interest has statistically significant effect. This may not be possible in an industrial case study, as the environment itself decreases the possibility to generalise to other companies. Furthermore, hardly any industrial software developer can afford to use two different technologies to evaluate which one is more effective. Instead, this kind of study can be carried out in an empirical study with students (Carver et al., 2003).

The first part of PG is based on numeral assignment as each requirement is assigned to one of the three piles. This approach is similar to the manner used in many organisations, i.e., classifying each requirement as having high, medium or low priority. In an industrial situation it is common that most requirements are classified as high (Karlsson, 1996). To avoid that, some constraints might be needed, such as imposing each classification to include at least 25% of the requirements. It is, however, rarely sufficient to use only numeral assignment since the difference in importance of requirements assigned the same priority can be larger than the difference in importance of requirements assigned different priorities (Karlsson, 1996).

In practice, it is common that a larger number of requirements need to be prioritised. The results presented in this paper may be valid when a sub-set of the requirements is prioritised. When the number of requirements grow, it is hard to get an overview. Therefore, visualisation becomes very important in order to share information. In a real project, it may also be more valuable to use the ratio scale in order to, in more detail, differentiate requirements from each other. Thus, it may not be sufficient to determine which requirement that is of higher priority, without knowing to what extent. This would speak in favour of the PWC techniques.

It is interesting to explore a possible extension to PG, providing it with a ratio scale. When the requirements have been ordered in a priority list using PG it would be possible to compare each requirement to the one below it in the list and assign a number to their internal relation. For example, one requirement can be estimated as being twice as important as the one below it in the priority list, and thereby their relation is set to two, and so on. In this manner, it would be possible to, with a reasonable amount of effort, provide PG with a ratio scale. More research needs to be conducted in order to determine the validity of this extension.

6 Conclusions

The main conclusion that can be drawn from both experiments is that the TPWC is superior to both PG and PWC regarding time-consumption. This may be due to the reduced number of comparisons in the tool compared to in the manual techniques. It can also be an effect of the increased support for the user as only one pair is

displayed and it is therefore easier to stay focused. It would also be interesting to investigate how tool-support for PG would affect the results.

PG was regarded as easier to use than the manual PWC in the first experiment, while it could not be determined if either of the techniques TPWC or PG is easier to use, although a majority found the TPWC easier. This may be due to most subjects enjoying to use a tool-based technique more than a manual one.

A difference in accuracy could not be confirmed in either of the experiments although PG was preferred by most of the subjects in both experiments.

Although the generalisation of the presented experiments to industrial practice is not straightforward, the results are an important basis for the planning of industrial case studies. When companies want to find a prioritisation technique that suits their needs they can take the presented results into account when planning situated trials.

The presented experiment design could also be used on more subjects to get a larger data set and thereby a stronger basis for conclusions. There are, as discussed, several other prioritisation techniques that would be interesting to look into and compare to the presented techniques as well.

Appendix

Table A1 Experiment 1 using counter-balancing design

Subject	Nbr of requirements	Tech 1	Tech 2	Criterion 1	Criterion 2
1	8	PWC	PG	Price	Value
2	8	PWC	PG	Price	Value
3	16	PWC	PG	Price	Value
4	16	PWC	PG	Price	Value
5	8	PWC	PG	Value	Price
6	8	PWC	PG	Value	Price
7	16	PWC	PG	Value	Price
8	16	PWC	PG	Value	Price
9	8	PG	PWC	Price	Value
10	8	PG	PWC	Price	Value
11	16	PG	PWC	Price	Value
12	16	PG	PWC	Price	Value
13	8	PG	PWC	Value	Price
14	8	PG	PWC	Value	Price
15	16	PG	PWC	Value	Price
16	16	PG	PWC	Value	Price

Table A2 Experiment 2 using counter-balancing design

Subject	Occasion	Tech 1	Tech 2	Criterion 1	Criterion 2
1	PM	TPWC	PG	Value	Price
2	PM	TPWC	PG	Value	Price
3	PM	TPWC	PG	Value	Price
4	PM	TPWC	PG	Value	Price
5	EV	TPWC	PG	Value	Price
6	EV	TPWC	PG	Value	Price
7	EV	TPWC	PG	Value	Price

Table A2 (continued)

Subject	Occasion	Tech 1	Tech 2	Criterion 1	Criterion 2
8	PM	TPWC	PG	Price	Value
9	PM	TPWC	PG	Price	Value
10	PM	TPWC	PG	Price	Value
11	PM	TPWC	PG	Price	Value
12	PM	TPWC	PG	Price	Value
13	EV	TPWC	PG	Price	Value
14	EV	TPWC	PG	Price	Value
15	PM	TPWC	PG	Value	Price
16	PM	PG	TPWC	Value	Price
17	PM	PG	TPWC	Value	Price
18	PM	PG	TPWC	Value	Price
19	EV	PG	TPWC	Value	Price
20	EV	PG	TPWC	Value	Price
21	EV	PG	TPWC	Value	Price
22	PM	PG	TPWC	Price	Value
23	PM	PG	TPWC	Price	Value
24	PM	PG	TPWC	Price	Value
25	PM	PG	TPWC	Price	Value
26	PM	PG	TPWC	Price	Value
27	PM	PG	TPWC	Price	Value
28	PM	PG	TPWC	Price	Value
29	EV	PG	TPWC	Price	Value
30	EV	PG	TPWC	Price	Value

Table A3 Requirements prioritised in the experiments

Requirement	Selected for 8 requirements
Alarm	X
Bluetooth	
Calculator	
Calendar	X
Call alert creation	
Colorscreen	X
Games	X
IR	
MMS	
Notebook	X
Phonebook	
SMS	
Timer	X
WAP	X
Vibrating call alert	X
Voice control	

Acknowledgments The authors would like to thank all experiment participants for contributing with their time and effort.

References

Beck K (1999) Extreme programming explained. Addison-Wesley, Reading, MA

Berander P (2004) Using students as subjects in requirements prioritization. Proc Int Symp Empirical Software Engineering, Redondo Beach, CA, USA, pp 167–176

Berander P, Andrews A (2005) Requirements prioritization. In: Aurum A, Wohlin C (eds), Engineering and Managing Software Requirements. Springer-Verlag, Berlin, Germany

Carmone FJ, Kara A, Zanakis SH (1997) A Monte Carlo investigation of incomplete pairwise comparison matrices in AHP. European Journal of Operational Research 102:538–553

Carver J, Jaccheri L, Morasca S, Shull F (2003) Issues in using students in empirical studies in software engineering education. Proc Int Software Metrics Symp Sydney, Australia, pp 239–249

Davis AM (2003) The art of requirements triage. IEEE Computer 36:42–49

Greer D, Ruhe G (2004) Software release planning: an evolutionary and iterative approach. Information and Software Technology 46:243–253

Harker PT (1987) Incomplete pairwise comparisons in the analytic hierarchy process. Mathl. Modelling 9:837–848

Höst M, Regnell B, Wohlin C (2000) Using students as subjects—a comparative study of students and professionals in lead-time impact assessment. Empirical Software Engineering 5:201–214

IEEE Std 830–1998. (1998). IEEE recommended practice for software requirements specifications. IEEE

Karlsson J (1996) Software requirements prioritizing. Proc Int Conf Req Eng Colorado Springs, Colorado, USA, pp 110–116

Karlsson J, Ryan K (1997) A cost-value approach for prioritizing requirements. IEEE Software 14:67–74

Karlsson J, Olsson S, Ryan K (1997) Improved practical support for large-scale requirements prioritising. Journ Req Eng 2:51–60

Karlsson J, Wohlin C, Regnell B (1998) An evaluation of methods for prioritizing software requirements. Inf and Software Techn 39:939–947

Karlsson L, Berander P, Regnell B, Wohlin C (2004) Requirements prioritisation: an experiment on exhaustive pair-wise comparisons versus planning game partitioning. Proc Int Conf Empirical Assessment in Software Engineering. Edinburgh, United Kingdom, pp 145–154

Lauesen S (2002) Software requirements-styles and techniques. Addison-Wesley, Harlow

Leffingwell D, Widrig D (2000) Managing Software Requirements-A unified approach. Addison-Wesley

Lehtola L, Kauppinen M (2004) Empirical evaluation of two requirements prioritization methods in product development projects. Proc European Software Process Improvement Conf Trondheim, Norway, pp 161–170

Lubars M, Potts C, Richter C (1992) A review of the state of the practice in requirements modeling. Proc IEEE Int Symp Req Eng, pp 2–14

Moisiadis F (2002) The fundamentals of prioritising requirements. Proc Systems Engineering, Test & Evaluation Conf, Sydney, Australia, pp 108–119

Newkirk JW, Martin RC (2001) Extreme programming in practice. Addison-Wesley, Harlow

Robson C (1997) Real World Research. Blackwell, Oxford

Ruhe G, Eberlein A, Pfal D (2002) Quantitative WinWin: a new method for decision support in requirements negotiation. Proc of the Int Conf on Software Engineering and Knowledge Engineering, pp 159–166

Runeson P (2003) Using students as experiment subjects—an analysis on graduate and freshmen student data. Proc Int Conf Empirical Assessment and Evaluation in Software Engineering. Keele, United Kingdom, pp 95–102

Saaty TL, Vargas LG (2001) Models, methods, concepts & applications of the analytic hierarchy process. Kluwer Academic Publishers, Norwell, MA

Sawyer P (2000) Packaged software: challenges for RE. Proc Int Workshop on Req Eng: Foundations of Software Quality. Stockholm, Sweden, pp 137–142

Shen Y, Hoerl AE, McConnell W (1992) An incomplete design in the analytic hierarchy process. Mathl. Comput. Modelling 16:121–129

Siddiqi J, Shekaran MC (1996) Requirements engineering: the emerging wisdom. IEEE Software 13:15–19.

Siegel S, Castellan JN (1988) Nonparametric statistics for the behavioral sciences. 2nd ed. McGraw-Hill, New York

Tichy WF (2000) Hints for reviewing empirical work in software engineering. Empirical Software Engineering 5:309–312
Wiegers K (1999) Software requirements. Microsoft Press, Redmond, WA
Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2000) Experimentation in software engineering—an introduction. Kluwer Academic Publishers
Wohlin C, Aurum A (2005) What is important when deciding to include a software requirement in a project or release? Proc Int Symp on Empirical Software Engineering. Noosa Heads, Australia, pp 237–246
Yourdon E (1999) Death March. Prentice-Hall, Upper Saddle River, NJ
Zhang Q, Nishimura T (1996) A method of evaluation for scaling in the analytic hierarchy process. Proc Int Conf Systems, Man and Cybernetics. Beijing, China, pp. 1888–1893. http://www.telelogic.com/corp/products/focalpoint/overview.cfm, last visited 2005–12–18

Lic. Tech. Lena Karlsson is a Ph.D. Student at the Department of Communication Systems at Lund University, Sweden. She is a member of the Software Engineering Research Group and has a Master of Science and a Licentiate of Technology degree from Lund University. Her research interest includes empirical software engineering, requirements engineering, release planning decision-support and retrospective analysis.

Dr. Thomas Thelin is an associate professor of software engineering at Lund University. His research interests include empirical methods in software engineering; software quality, and verification and validation with emphasis on testing, inspections, and estimation methods. He received a Ph.D. in software engineering from Lund University.

Dr. Björn Regnell is an associate professor in software engineering at Lund University. His research interests include requirements engineering, empirical software engineering, software product management, process improvement, and market-driven software development. He received his Ph.D. in software engineering from Lund University. He is chair of the Swedish Requirements Engineering Research Network (SiREN).

Lic. Tech. Patrik Berander is a Ph.D. student in Software Engineering at the School of Engineering at Blekinge Institute of Technology in Sweden. He received his degree of Master of Science with a major in Software Engineering - specialized in Management in 2002. He further received his degree of Licentiate of Technology in 2004 with the licentiate thesis entitled Prioritization of Stakeholder Needs in Software Engineering Understanding and Evaluation. His research interests are requirements engineering in general and decisions related to requirements and products in particular. Further research interests include software product management, software quality, economic issues in software development, and software process management.

Dr. Claes Wohlin is a Professor of Software Engineering and Provost of Blekinge Institute of Technology, Sweden. Prior to joining BTH in 2000 he held professorships at Lund and Linköping Universities. He is currently Visiting Professor at Chalmers University of Technology in Göteborg. His research interests include empirical methods in software engineering, software metrics, software quality, requirements engineering and systematic improvement in software engineering. Since 2001 Professor Wohlin has been co-editor-in-chief of the journal of Information and Software Technology, published by Elsevier. He is on three other editorial boards: Empirical Software Engineering: An International Journal, Software Quality Journal and Requirements Engineering Journal. Claes Wohlin was the recipient of Telenor's Nordic Research Prize in 2004 for his achievements in software engineering and improvement of reliability for telecommunication systems.