L. Karlsson, P. Berander, B. Regnell and C. Wohlin, "Requirements Prioritisation: An Experiment on Exhaustive Pair-Wise Comparisons versus Planning Game Partitioning", Proceedings 8th Conference on Empirical Assessment in Software Engineering, Edinburgh, UK, 2004.

Requirements Prioritisation: An Experiment on Exhaustive Pair-Wise Comparisons versus Planning Game Partitioning

Lena Karlsson[1], Patrik Berander[2], Björn Regnell[1], Claes Wohlin[2]

[1]*Department of Communication Systems,*

Lund University, Sweden
{lena.karlsson, bjorn.regnell}@telecom.lth.se

[2]*Department of Software Engineering and Computer Science,*
Blekinge Institute of Technology, Sweden
{patrik.berander, claes.wohlin}@bth.se

Abstract

The process of selecting the right set of requirements for a product release is highly dependent on how well we succeed in prioritising the requirements candidates. There are different techniques available for requirements prioritisation, some more elaborate than others. In order to compare different techniques, a controlled experiment was conducted with the objective of understanding differences regarding time consumption, ease of use, and accuracy. The requirements prioritisation techniques compared in the experiment are the Analytical Hierarchy Process (AHP) and a variation of the Planning Game (PG), isolated from Extreme Programming. The subjects were 15 Ph.D. students and one professor, who prioritised mobile phone features using both methods. It was found that the straightforward and intuitive PG was less time consuming, and considered by the subjects as easier to use, and more accurate than AHP.

1. Introduction

Software requirements need to be prioritised when the elicitation process has yielded more requirements than can be implemented at once. There exist a number of different techniques and tools to use for requirements prioritisation. However, some software organisations may not have enough resources to buy or develop a tool and therefore it is interesting to investigate techniques that do not need computer support.

This paper describes an experiment aimed at comparing two requirements prioritisation techniques. The intention with the experiment is to compare a rudimentary prioritisation technique (Planning Game) with a more elaborate one (Analytical Hierarchy Process). The main variables that were investigated were the difference in time-consumption, accuracy, and ease of use. The experiment was performed during a one-day session with 15 Ph.D. students and one professor as subjects. Instead of real requirements, the subjects prioritised features of mobile phones, which is a well-known product with a range of features to choose from.

In order to investigate the trade-off between low price and high value, the prioritisation was performed with respect to both Price and Value. The experiment also aimed at investigating if the preferred choice of prioritisation technique depended on the number of features involved.

As expected, the results indicate that the more rudimentary technique was less time-consuming and a majority of the subjects found it easier to use. Most subjects also found the results from the rudimentary technique more accurate, which is a bit surprising.

The paper is structured as follows. Section 2 explains and discusses the matter of requirements prioritisation in general and the two compared techniques in particular. Section 3 describes the design of the experiment and brings up some validity issues. Further, Section 4 presents the results discovered in the experiment while Section 5 discusses what the results may imply. Finally, the paper is concluded in Section 6.

2. Requirements Prioritisation

The ultimate goal of any software organisation is to create systems that meet the stakeholder demands. Since there are usually more requirements than can be implemented, decision makers must face the dilemma of selecting the right set of requirements for their next product release. In order to select the correct set of requirements, the decision makers must understand the relative priorities of the requested requirements [19]. 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 techniques are based on more or less structured sorting algorithms, while others use pair-wise comparisons or numeral assignment [7].

The two techniques compared in this paper are (1) the Analytical Hierarchy Process (AHP) that is based on exhaustive pair-wise comparisons [15], and (2) the Planning Game (PG) [1] that uses a sorting algorithm to partition the requirements. The two techniques are further described below.

2.1 Analytical Hierarchy Process (AHP)

AHP is a decision-making method that involves comparing all possible pairs of requirements, in order to determine which of the two 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.* 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, AHP is rather insensitive to judgement errors. Furthermore, AHP includes a *consistency check* where judgement errors can be identified and a *consistency ratio* can be calculated.

In AHP, any system structure can be abstracted into a hierarchy that explains the system's components and their functions. Hence, AHP takes the whole system into account during decision-making since it prioritises the components on each level in the hierarchy [15].

Karlsson *et al.* [11] performed an evaluation of six different prioritisation techniques based on pair-wise comparisons, including AHP. The authors concluded that AHP was the most promising approach because it is based on a ratio scale, is fault tolerant, and includes a consistency check. AHP was the only technique in the evaluation that satisfied all these criteria. Furthermore, it includes a priority distance, i.e. a *ratio scale*, while the other approaches only provided the preferred order. However, because of the rigour of the technique, it was also the most time-consuming one in the investigation.

Since the major disadvantage of AHP 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 [3, 5, 8, 16]. The results of these have been that it is possible to reduce the number of comparisons with as much as 75 % [8]. However, when reducing the number of comparisons, the number of comparisons that are redundant is also reduced, and hence the possibility to identify inconsistent judgements [11].

2.2 Planning Game (PG)

In the last years, there have been an increased use and interest in agile methodologies, such as Extreme Programming (XP). Agile methodologies are based on streamlined processes, attempting to reduce overhead such as unnecessary documentation. The interest and use of agile methodologies have been both from industry and academia. Tom De Marco has aligned to this interest and have expressed that *"XP is the most important movement in our field today"* [2].

XP is composed of 12 fundamental practices, of which Planning Game (PG) is one. For the purpose of this experiment we have isolated PG despite that the practices likely affect each other [1].

PG is used in planning and deciding what to develop in a 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 [1].

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 [13].

The result of this easy and straightforward technique is a sorted vector of requirements. This means that the requirements are represented as a ranking on an *ordinal scale* without the possibility to see how much more important one requirement is than another.

2.3 Cost-Value Trade-Off

When prioritising requirements, it is often not enough to just prioritise 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 [12]. 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. Or as Wiegers puts it: *"Prioritisation means balancing the business benefit of each requirement against its cost and any implications it has for the architectural foundation and future evolution of the product"* [19].

Karlsson and Ryan [9] use AHP as an approach for prioritising both Value and Cost in order to implement those requirements that give most value for the money. The data can be further used to provide graphs to visualise the Value to Cost ratio between the requirements.

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 [9] with the difference that the result from PG is based on an ordinal scale instead of a ratio scale.

3. Experiment Design

This section describes the experiment approach and execution as well as the analysis performed by the research-

ers[1]. Finally, it is concluded with a number of validity issues.

3.1 Experiment Approach

The experiment was carried out with a *repeated measures design, using counter-balancing* [14, 20]. The 16 subjects in the convenient sample included 15 Ph.D. Students in their first or second year, and one professor. 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, before the experiment a pre-test was performed, and a few weeks after the experiment a second post-test was conducted.

The two requirements prioritisation techniques described above (Section 2) were used as input to the experiment, but were modified in order to be further comparable. The system aspect of AHP was not considered, and thus there is only one level of the hierarchy in this investigation [15]. Neither did we use 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 Value and Price: (1) Necessary, (2) Adds to the value and (3) Unnecessary, and (1) Very high price, (2) Reasonable price and (3) Low price, respectively. In practice, PG is performed by a customer representative and a developer, but in this experiment each subject had to play both roles.

3.1.1 Research Hypotheses. The goal of the experiment is to compare two prioritisation techniques and to investigate the following hypotheses:

1. The average time to conclude the prioritisations is larger when using AHP.
2. The ease of use is considered higher for PG.
3. AHP reflects the subjects' views more accurately.

The objective dependent variable *average time to conclude the prioritisations* was captured by measuring each subject's time to conclude the tasks. The subjective dependent variables *ease of use* and *reflecting the subjects' views* were captured by questionnaires after the experiment.

3.1.2 Pilot Experiment. A pilot experiment was performed before the main study to evaluate the design. Six colleagues participated and they prioritised ten features each, with both techniques. After this pilot experiment, it was concluded that the experiment should be extended to 8 and 16 features 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 AHP sheets included to remove the scale and instead use the "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.

3.1.3 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 opinion 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 features. Nine of the 16 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.

3.1.4 Experiment Execution. 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 features, for example SMS, Games, WAP, Calendar, etc. In this experiment, the prioritisation criteria were *Value for me*, which corresponds to how important and interesting the subject find the feature, and *Added price on the phone*, which is an estimation of how much the feature might add to the actual mobile phone price. Note that this is not the same as development cost, which would be difficult for laymen to estimate.

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

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 features, while the other half prioritised 16 features. 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 AHP and half started with PG. The order of the Value and Price 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 Table 1.

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

For AHP, one sheet per criterion and person had been prepared, with all possible pair-wise combinations of the features to compare. For the purpose of eliminating order effects, the order of the pairs was randomly distributed so

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

Table 1. Experiment using counter-balancing design

Subject	Nbr of Features	Tech 1	Tech 2	Criteria 1	Criteria 2
A	8	AHP	PG	Price	Value
B	8	AHP	PG	Price	Value
C	16	AHP	PG	Price	Value
D	16	AHP	PG	Price	Value
E	8	AHP	PG	Value	Price
F	8	AHP	PG	Value	Price
G	16	AHP	PG	Value	Price
H	16	AHP	PG	Value	Price
I	8	PG	AHP	Price	Value
J	8	PG	AHP	Price	Value
K	16	PG	AHP	Price	Value
L	16	PG	AHP	Price	Value
M	8	PG	AHP	Value	Price
N	8	PG	AHP	Value	Price
O	16	PG	AHP	Value	Price
P	16	PG	AHP	Value	Price

every subject got a different order of the comparisons. With 16 features to compare, there was 16(16-1)/2 = 120 pair-wise comparisons for Value and Price, respectively. With 8 features, there was 8(8-1)/2 = 28 pair-wise comparisons for both Value and Price. In between each pair in the sheets there was a scale where the difference of the requirements' Value or Price was circled, see Figure 1. In order 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.

Which of the two features are most valuable to you?
Alarm <<<< <<< << < = > >> >>> >>>> Timer
WAP <<<< <<< << < = > >> >>> >>>> SMS

Fig. 1. Example of AHP sheet

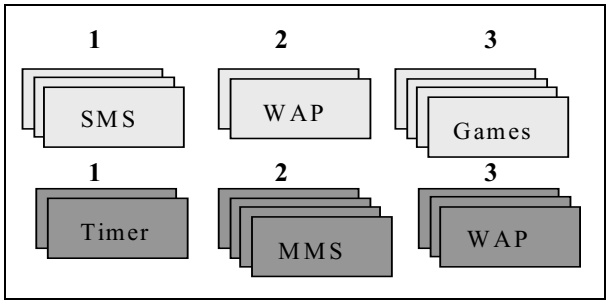

Fig. 2. Example of PG cards

The further to the left a symbol was circled, the more valuable (or expensive) was the left feature than the right one. If the features were equally valuable (or expensive) the "equal" symbol was circled.

For PG, the subjects were given two sets of cards (one for Value and one for Price) with one mobile phone feature written on each. The cards were partitioned into three piles, separately for the Value criterion and the Price criterion, see Figure 2. 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 is at the top of the pile and the less valuable (or expensive) are 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 features 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 of clarification.

3.1.5 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 the 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?
3. Which technique do you think is most sensitive to judgemental errors?

3.1.6 Post-Test 2. After completing the analysis, the subjects were, in a second post-test, asked to state which technique they thought gave the most accurate result. They were sent two sheets (one for Value and one for Price) with two different lists of features, corresponding to the results from the PG and AHP 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 agreed the best with their views. The ratio scale from AHP was not taken into consideration, and neither was the pile distribution from PG. This was necessary in order to get comparable lists.

3.2 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™ and the computing tool MATLAB™.

Two different scales were tried for the AHP analysis: 1~5 and 1~9. According to Zhang [21] the scale 1~5 is better than 1~9 at expressing human views and therefore the scale 1~5 was used when compiling the prioritisation ranking lists.

Furthermore, Saaty [15] has 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. However, the RI scale was extrapolated and the RI for 16 factors was set to 1.61.

3.3 Validity

The experimental design involves some threats to validity, which we have tried to prevent. Using the counter-balancing design, the order effects have been balanced out since the subjects were randomly given different orders to perform the techniques and using the criteria. Therefore, we believe that the order of the techniques and criterion will not affect the results.

It is also possible that the subjects could become fatigued during the experiment. Especially the subjects who perform the tasks with 16 features may get tired or bored, which in turn may affect the concentration. This has been tested during the analysis, by calculating the consistency for AHP and the results indicate that there is no significant difference in consistency depending on the number of features (see Table 8).

Another possibility 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. Especially when using PG first, it may affect the AHP performance. This is not the case. Although the mean values in Table 10 indicate a difference in the consistency, the hypothesis tests show that the difference is not significant.

Group pressure and the measure of each subject's time to complete the task might impose time-pressure, which can affect the results. However, it may not be a large problem since there is no major correlation between the time and the consistency in the results (see Table 9). Therefore we can argue that time-pressure will not affect the performance of the prioritisation.

The number of subjects was only 16, which reduces the generalisability, i.e. there is a threat that the findings are specific to this particular group or context. On the other hand, Ph.D. students may have similar views as the requirements engineers and customers who are intended to use the techniques in practice [6]. It is also likely that the subjects are not taking the prioritisation as seriously as a requirements engineer or customers would in a real project (see Section 4.7).

Unfortunately, the scales with "more than" and "less than" in the AHP sheets were accidentally switched so that it could be interpreted in the opposite way than was intended (see Figure 1). This caused some confusion during the experiment. However, the interpretation was explained and clarified and therefore this should not be considered a threat to validity.

It would have been valuable to start the session with an introduction explaining each feature in the prioritisation to clarify their meaning. However, the subjects had their own interpretation of the features, which was the same throughout the experiment and therefore this should not affect the result.

4. Results

This section presents some of the results found during analysis. First, the three hypotheses are discussed, then some other interesting findings are described.

4.1 Hypothesis 1: The average time to conclude the prioritisations is larger when using AHP.

As expected, the time to conclude the prioritisation is larger with AHP than with PG, for both criteria. As Table 2 shows, the difference in time between the two techniques is 6.1 minutes for 8 features and 14.7 minutes for 16 features. The time increase in percent from 8 to 16 features for AHP is 88 %, while the same for PG is only 48 %. Thus, a larger number of objects to prioritise affect the time-consumption for AHP more than for PG, at least when using 8 and 16 features.

Table 2. Average time consumption for the prioritisation

Nbr of features	Criteria	AHP	PG	Diff.
8	Value	7.8 min	3.6 min	4.2 min
	Price	6.4 min	4.5 min	1.9 min
Total		**14.2 min**	**8.1 min**	**6.1 min**
16	Value	12.6 min	6.5 min	6.1 min
	Price	14.1 min	5.5 min	8.6 min
Total		**26.7 min**	**12.0 min**	**14.7 min**
% increase		88 %	48 %	

This can also be seen in Figure 3 where the median values are higher for AHP than for PG, and the difference between 8 and 16 features is larger for AHP than for PG. Additionally, the boxplot indicates that the subjects' time to conclude the prioritisation with AHP are more dispersed.

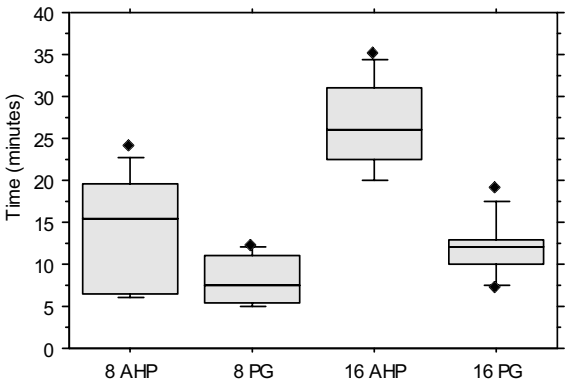

Fig. 3. Box plots of the time spent on prioritisation

As Table 3 suggests, the subjects have in average used less time per feature when they had more features to prioritise. It is particularly interesting to see that it takes less time per feature to perform PG partitioning with 16 features than with 8. One could expect that it should be more complex to perform PG with more features but this result show that it is even faster with more features. However, there might be a breakpoint when the number of features is too great and it becomes hard to obtain an overview.

Table 3. Time consumption per feature

Nbr of features	AHP	PG
8	53.5 s/feature	30.5 s/feature
16	50.0 s/feature	22.5 s/feature

Four hypothesis tests (see Table 4) were performed, for 8 and 16 features 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*. The hypothesis tests show that on the 5 %-level there is a significant time difference for three of the four cases. In the fourth case, the Price criterion on 8 features, the test shows that the difference is only significant on a higher level. This is illustrated in Table 4, where the p-value is lower than 5 % in three of the four cases.

Table 4. Wilcoxon tests for the time difference

Nbr of features	Criteria	Wilcoxon p-values
8	Value	0.0117
	Price	**0.0781**
16	Value	0.0098
	Price	0.0039

4.2 Hypothesis 2: The ease of use is considered higher for PG

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 more or much more easy to use than AHP. Only 3 found them equally easy and 1 stated that AHP was more easy to use, see Table 5. Hence, 75 % of the subjects found PG easier to use.

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

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

Nbr of features	AHP Much more	More	Equal	More	PG Much more
8	0	0	1	3	4
16	0	1	2	1	4
Total	**0**	**1**	**3**	**4**	**8**

4.3 Hypothesis 3: AHP reflects the subjects' views more accurately

Right 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 AHP to be better.

Table 6. Results from the first post-test: Expected accuracy

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

For most subjects, the actual ranking that was captured in the analysis differed somewhat between the two prioritisation techniques. In order to evaluate which technique that gave the most accurate results, a second post-test was sent out to the subjects. This was done a few weeks after the experiment was performed, when the analysis was finished.

As Table 7 shows, the most common opinion was that PG reflects the subjects views more accurately. 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

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

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

not as good as the subjects expected even if it was clearly better than AHP.

4.4 Judgement Errors

Another question at the first post-test was which technique the subjects expected to be most sensitive to judgemental errors. The objective was to find out the subjects' views, although it has been shown that AHP is insensitive to judgemental errors due to the redundancy in the pair-wise comparisons [11, 15]. However, among the subjects 75 % expected AHP to be most sensitive. Perhaps this is because the AHP-technique "feels like pouring requirements into a black-box" as one of the subjects stated. It may be difficult to trust something that you are not in control of.

4.5 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 [15] has 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 is used frequently in practice [9].

The CR limit above is only valid for the scale 1~9, and in this experiment the scale 1~5 was used instead. Therefore, the limit for acceptable CR will be lower. The average consistency ratios for scale 1~5 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

Table 8. Mean consistency ratio and Wilcoxon test for the difference in consistency

Criteria	Nbr of features	Scale 1~5	Wilcoxon p-values
Value	8	0.11	0.3270
	16	0.08	
Price	8	0.10	0.6744
	16	0.12	

Table 9. Correlation between time and consistency

Pearson correlation coefficients	Value	Price
8 Features	0.06	-0.25
16 Features	0.26	-0.21

a non-parametric test. The Wilcoxon test shows that on the 5 %-level, there is no significant difference in consistency depending on the number of features prioritised. It was decided not to exclude any of the prioritisations, even though CR was high, in order to keep all available data.

In order to investigate if the time spent on each comparison affects the consistency, the correlation between the parameters was calculated. The Pearson correlation coefficients indicate an insignificant correlation between the time and the consistency, positive for the Value criterion and negative for the Price criterion, see Table 9. Generally, the absolute value of the correlation coefficient should be greater than 0.5 in order for the values to be considered correlated [17]. Thus, the conclusion is drawn that the consistency is not particularly influenced by the time consumption.

4.6 Order Effects

There is a chance that the order in which the two techniques are used can influence the result. Table 10 shows that the mean consistency ratio is a bit lower for the subjects who used PG before AHP. This may indicate that using PG can provide an image of ones preferences that are not possible to get from using AHP. Therefore it may be easier to be consistent when PG precedes AHP.

Table 10. Order effect on consistency

Mean consistency	AHP-PG	PG-AHP	Mann-Whitney p-values
Value	0.11	0.08	0.6773
Price	0.12	0.10	0.6773

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*. The p-values are all larger than 5 %, and therefore we can draw the conclusion that there is no significant difference depending on the order. This finding validates that the experiment analysis have not suffered from any order effects since there is no significant difference between the two groups.

4.7 Distribution in Piles

For PG the subjects were asked to distribute the features in three different piles, dependent on Value and Price. In average, the respondents distributed 41 % of the features

in the middle pile (independent of criterion). This is a result that might not correspond well to how the features would have been distributed in a real case. One could assume that customers would put most of the features in the highest priority pile, which is often the case when customers need to prioritise between their wishes [12, 18, 19]. Therefore, this result might be somewhat misleading and further studies should clarify this condition.

4.8 Prioritising the Price Criterion

One of the problems that was identified before the experiment, was that the respondents may find it difficult to prioritise the Price criterion, since it is hard to know the price of different features. However, the results show that the mean standard deviation in PG was lower when prioritising the Price criterion than the Value criterion, see Table 11. This result shows that the respondents have been more united when prioritising Price than Value, which is a rather expected result since the Price is a somewhat more objective criterion. Therefore, it is concluded that the Price criterion is not considered a threat to validity.

Table 11. Mean standard deviation

	8 Features	16 Features
Value	1.73	3.02
Price	1.25	2.79

4.9 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 AHP include “effort demanding but nice”, “it feels like a black-box wherein you pour requirements”, “good but boring”, “it feels like you loose 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.

4.10 Price-Value Graphs

In order to illustrate the possibility of using the Cost-Value approach for requirements selection, two examples of Cost-Value graphs are available in Figure 4 and 5 (PG and AHP with 8 features). However, in this experiment, we use the term Price instead of Cost. The graphs are made in order to visualise the results from the experiment and to see how much the two techniques differ regarding Price-Value graphs.

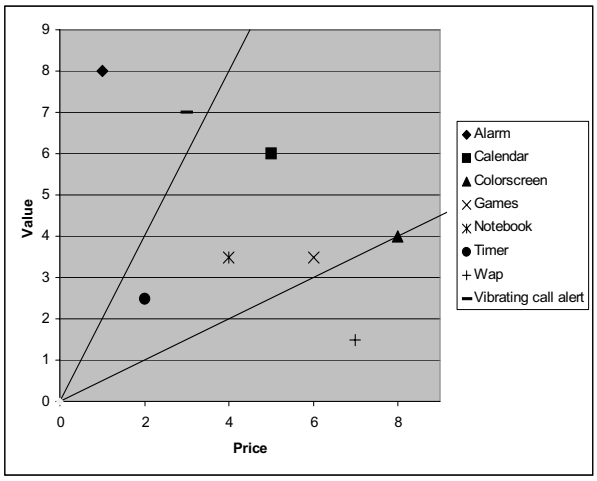

Fig. 4. Price-Value graph for PG with 8 features

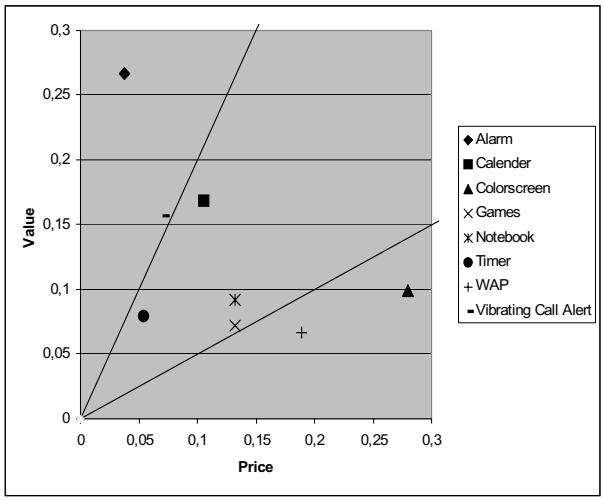

Fig. 5. Price-Value graph for AHP with 8 features

The three areas in the graphs represent different grades of contribution [10] and the lines visualise which Value to Price ratio each requirement has, as explained in [9]. The upper line in each graph divides those features that had more than 2 in Value to Price ratio from those that had between 2 and 0.5. The lower line in each graph divides those features that had between 2 and 0.5 from those with a ratio below 0.5 [9]. The Price and Value markings for AHP are based on the mean of the subjects’ relative weight of the features. In PG, the markings are based on the median of the subjects’ ranking number.

In the case with 8 features, the two methods provide the same result when it comes to which feature that are located in which area of the graph. The features Alarm and Vibrating call alert have in average a high Value to Price ratio (above 2) and therefore they would give high contribution to the fictive product. The features Colorscreen and WAP have a low Value to Price ratio (below 0.5), and would bring low contribution to the product. Finally, Calendar,

Games, Notebook and Timer bring medium contribution (between 0.5 and 2 in Value to Price ratio).

The results indicate that it is possible to provide Price-Value (or Cost-Value) graphs with both PG and AHP. However, further studies are needed in order to validate if this result applies to other prioritisations.

In practice, the Cost-Value diagram would be used to guide the decision-maker in the difficult requirements selection. Other factors such as market segmentation, product focus and time constraints, will also influence the requirements selection.

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 strict time and budget pressure. Therefore, a strong argument for PG is that the time consumption is reasonable and the usage easy and intuitive.

In this experiment two groups prioritised 8 and 16 features, 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 features in mind. However, this experiment only shows a slight tendency of less overview when prioritising 16 features (see Table 5). Therefore, it is suspected that the breakpoint is at an even higher number of features.

Another interesting observation in this experiment was that the time-consumption did not affect the consistency in AHP. One could assume that if someone stresses through the comparisons, the consistency would be worse. However, this is only initial results and with more difficult features to prioritise, the results might be different.

In practice, it is common that a larger number of requirements need to be prioritised. When the number of requirements grow, it is hard to get an overview. Therefore, visualisation is very important in order to share information. This experiment showed that it should be possible to visualise the result of both AHP and PG. However, it should be further evaluated how the ordinal scale in PG affects the visualisation.

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. However, without tool support, AHP will be very time-consuming with a greater number of requirements, both to perform and to analyse.

Due to the small sample and the specific domain it is questionable if the results can be generalised to an industrial situation. Although the subjects may have opinions similar to decision-makers in industry, the context of mobile phone features is a bit too simplistic. The main weakness is that mobile phone features are on a high level and rather independent, while requirements in a real case often have interdependencies. It is also possible that industrial experience would affect the results, although we believe that in a relative comparison between these two techniques, it is likely that the rudimentary PG technique still would be preferred.

In the experiment performed by Karlsson *et al.* [11], AHP was ranked as the superior technique in relation to the others. The main reasons were that AHP had reliable results, was easy to use, was fault tolerant and was based on a ratio scale. This experiment shows that PG is superior to AHP on all of these criteria except for that it is not based on a ratio scale. Therefore, it is interesting to imagine a combination of the two techniques.

In order to decrease the number of comparisons, AHP could be used on the three piles, separately. Another possibility is to use AHP only on those requirements that end up in the middle pile in PG. This would imply that PG is used first, to divide the requirements into three groups according to the PG approach described earlier. 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 [4] 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, AHP 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.

The discussion above is based on the assumption that most requirements are not put into the same pile, which might be common in an industrial situation. Therefore, some constraints might be needed in order to force the piles to be rather evenly distributed. With three piles, this could for example mean that no pile is allowed to have less than 25 % of the requirements.

Based on the results from this experiment, it could not be concluded if a combination of the two techniques is efficient or not, or how such a combination should look like. However, we strongly believe that such a combination could be valuable and that it is worth evaluating. Therefore, it is recommended that a combination is tried in a separate experiment or case study, with more data points.

6. Conclusions

This paper describes an experiment aimed at comparing two requirements prioritisation techniques regarding time consumption, ease of use and accuracy in the result. The investigated techniques are the elaborate Analytical Hierarchy Process (AHP), which is based on pair-wise comparisons and has a ratio scale, and the elementary Planning Game (PG), which is based on pile partitioning and has an ordinal scale.

The results reveal that the intuitive and quick PG technique is superior with regard to time consumption, ease of use, and accuracy. The mean time consumption was higher when using AHP and the result was statistically significant in three of four cases. PG was considered easier to use by 75 % of the subjects, although the results indicate that AHP is more preferred by those who prioritised a greater number of requirements. A blind-test performed after the experiment showed that 47 % found the priority order from PG more accurate, while 28 % favoured the order from AHP and 25 % found both priority orders equally accurate. However, it was concluded that a combination of the two techniques would further improve prioritisation. By first using PG to get an overall picture of the problem and then use AHP for the most difficult decisions, you would, with reasonable effort, get an accurate priority list.

The generalisability of the study is limited due to the small sample and the specific context. A real project has requirements interdependencies, and time and budget pressure to consider, which cause the decision-making to be far more difficult. However, we believe that PG is valid as prioritisation technique, although it does not have the same elaborate and valuable attributes as AHP.

The main disadvantage of the experiment being the difficulty to generalise to industrial projects, it would be valuable to try the experiment out in a case study. The participating organisation would then get knowledge about prioritisation and perhaps find a technique that suits their needs.

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.

Acknowledgements

The authors would like to thank the subjects, without which this study would not have been possible and Daniel Karlström for giving valuable comments on an earlier version of the paper.

7. References

[1] Beck, K., *Extreme Programming Explained*, Addison-Wesley, 1999.

[2] Beck, K., Fowler, M., *Planning Extreme Programming*, Addison-Wesley, 2001.

[3] Carmone, F. J., Kara, A., Zanakis, S. H., "A Monte Carlo Investigation of Incomplete Pairwise Comparison Matrices in AHP", *European Journal of Operational Research*, Vol. 102, pp. 538-553, 1997.

[4] Davis, A., M., "The Art of Requirements Triage", *IEEE Computer*, Vol. 36, pp. 42-49, 2003.

[5] Harker, P. T., "Incomplete Pairwise Comparisons in the Analytical Hierarchy Process", *Mathl. Modelling*, Vol 9, pp. 837-848, 1987.

[6] Höst, M., Regnell, B., Wohlin, C., "Using Students as Subjects - A Comparative Study of Students and Professionals in Lead-Time Impact Assessment", *EASE´00 - Empirical Software Engineering*, Vol. 5, Issue 3, pp. 201-214, 2000.

[7] Karlsson, J., "Software Requirements Prioritizing", *Proceedings of ICRE*, pp. 110-116, 1996.

[8] Karlsson, J., Olsson, S., Ryan, K., "Improved Practical Support for Large-scale Requirements Prioritising", *Requirements Engineering*, Vol 2, pp. 51-60, 1997.

[9] Karlsson, J., Ryan, K. "A Cost-Value Approach for Prioritising Requirements", *IEEE Software*, pp. 67-74, Sept/Oct, 1997.

[10] Karlsson, J., Ryan, K. "Supporting the Selection of Software Requirements", *Proceedings of IWSSD*, pp. 146-149, 1996.

[11] Karlsson, J., Wohlin, C., Regnell, B., "An Evaluation of Methods for Prioritising Software Requirements", *Information and Software Technology*, Vol 39, pp. 939-947, 1998.

[12] Lauesen, S., *Software Requirements - Styles and Techniques*, Addison-Wesley, 2002.

[13] Newkirk, J. W., Martin, R. C., *Extreme Programming in Practice*, Addison-Wesley, 2001.

[14] Robson, C., *Real World Research*, Blackwell, 1997.

[15] Saaty, T. L., *The Analytical Hierarchy Process*, McGraw-Hill, 1980.

[16] Shen, Y., Hoerl, A. E., McConnell, W., "An Incomplete Design in the Analytical Hierarchy Process", *Mathl. Comput. Modelling*, Vol 16, pp.121-129, 1992.

[17] Siegel, S., Castellan, J. N., *Nonparametric Statistics for the Behavioral Sciences*, 2nd edition, McGraw-Hill International Editions, 1988.

[18] Sommerville, I., Sawyer, P., *Requirements Engineering - A Good Practice Guide*, John Wiley & Sons Ltd, 1997.

[19] Wiegers, K., *Software Requirements*, Microsoft Press, 1999.

[20] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., Wesslén, A., *Experimentation in Software Engineering - An Introduction*, Kluwer Academic Publishers, 2000.

[21] Zhang, Q., Nishimura, T., "A Method of Evaluation for Scaling in the Analytical Hierarchy Process", *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Vol. 3, pp. 1888-1893, 1996.