Does the prioritization technique affect stakeholders’ selection of essential software product features?

Hans Christian Benestad  
Simula Research Laboratory  
Lysaker, Norway  
benestad@simula.no

Jo Erskine Hannay  
Simula Research Laboratory  
Lysaker, Norway  
johannay@simula.no

Abstract— Context: Being able to select the essential, non-negotiable product features is a key skill for stakeholders of software projects. Such selection relies on human judgment, sometimes supported by structured prioritization techniques and associated tools. Goal: Our goal was to investigate whether certain attributes of prioritization techniques affect stakeholders’ threshold for judging product features as essential. The four investigated techniques reflect four combinations of granularity (low, high) and cognitive support (low, high). Method: In one experiment, 94 subjects in four treatment groups indicated the features (from a list of 16) that would be essential in their decision to buy a new cell phone. With a similar setup in a controlled field experiment, 46 domain experts indicated the software product features that were essential for the fulfillment of the project’s vision. The effects of granularity and cognitive support on the number of essential ratings were analyzed and compared between the experiments. Result: With lower granularity, significantly more features were rated as essential. The effect was large in the first experiment and extreme (Cohen’s $d=2.40$) in the second. Added cognitive support had medium effect (Cohen’s $d=0.43$ and 0.50), but worked in opposite directions in the two experiments, and was not statistically significant in the second. Implications: The results of the study imply that software projects should avoid taking stakeholders’ judgments of essentiality at face value. Practices and tools should be designed to counteract potentially harmful biases; however, more empirical work is needed to obtain more insight into the causes of these biases.

Requirements, Prioritization techniques; Essential features; Stakeholders; Field experiment

I. INTRODUCTION

Making sound judgments about the importance of proposed product features is a key skill for stakeholders of software projects, from project inception and throughout construction and evolution. In this study we wanted to draw attention to essential product features (denoted essentials in this paper)—those product features that, according to the stakeholders, cannot be left out if the product vision is to be fulfilled.

Different stakeholders may have different perspectives and interests and may assess the importance of a given feature differently. With prioritization techniques, the views of individual stakeholders are collected in a structured manner, sometimes with tool support. The goal of such techniques is to support experts in making priorities that reflect their knowledge and best judgment.

It is well known that expert judgments are subject to contextual biases. For example, in software cost estimation, questioning format and irrelevant information have been shown to significantly affect the estimates [1, 2]. Due to the cognitive commonality between cost estimation and value judgments, it is reasonable to assume that assessments of value and priorities of product features are subject to similar biases as those found for cost estimation. However, the empirical evidence on the nature of such biases on value and priority assessments, when and how they occur, and how the biases can be managed, is scarce.

The goal of this study was to investigate how different prioritization techniques affect stakeholders’ threshold for judging product features as essential. With a lower threshold, more features would be rated as essential. Therefore, a natural outcome measure for the study is the number of features reported as essential from a list of candidate features presented to the subjects.

Prioritization techniques may be classified according to how they differ along the dimensions of measurement scale, granularity, and sophistication [3]. With regards to these three dimensions, our study focuses on techniques that elicit priorities on an ordinal measurement scale, according to varying granularity (low, high), and according to cognitive support (low, high). Cognitive support is a specific aspect of sophistication. We will elaborate on the precise meaning of these dimensions below. The research question for this study is:

Does granularity and cognitive support of prioritization techniques affect stakeholders’ threshold for judging software product features as essential?

In two controlled experiments we investigated four ordinal prioritization techniques representing the combinations of low/high granularity and low/high cognitive support. We denoted the techniques Simple dropdown, Drag into bins, Sortable table, and Pairwise comparisons&ranking.

In addition to offering input to the reflective practice of software practitioners, this study offers input to empirical
software engineering research. First, the study contributes with a new objective outcome measure for prioritization studies. Second, the study uses underlying characteristics of prioritization techniques as independent variables, rather than the technique itself. This is important, because it facilitates a deeper understanding of the results by making it possible to draw lines from theories in psychophysics and behavioral decision science. Objective outcome measures are important for the validity of empirical studies investigating judgmental biases, and such measures have been scarce in earlier prioritization studies. We suggest these research design elements as complements to a proposed framework for prioritization studies [4]. Methodologically, the study demonstrates the viability of conducting a decently scaled controlled experiment in a live software engineering context, and illustrates the usefulness of comparing results (effect sizes in particular) with replications in a more artificial context.

The rest of this paper is organized as follows: Section II presents the investigated techniques, Section III reviews related work, Section IV proposes possible effects in play, Section V explains the experimental design, Section VI summarizes the results, Section VII discusses the results, Section VIII discusses validity issues, and Section IX presents our conclusions.

II. INVESTIGATED TECHNIQUES

Prioritization techniques can be characterized according to the measurement scale on which experts are to prioritize features, the granularity of the scale, and the degree of cognitive support for the technique [3]. Priorities given on an ordinal measurement scale hold simple information about the relative ordering of features. In contrast, interval or ratio scale priorities add information about the magnitude of differences between features. However, the available precision should match both the objective and the cognitive range of the prioritization task. For example, when determining essentials, it suffices to offer experts an ordinal scale. In an ordinal setting, high granularity means that the expert can choose from a larger set of possible ratings than with lower granularity.

There are many possible forms of cognitive support, but many of them are designed to help the expert prioritize more reliably. For example, with prioritization techniques based on the analytical hierarchy process (AHP) [5], the sensitivity for accidentally inaccurate assessments is reduced through redundant pairwise comparisons.

Based on the study’s objectives and the cognitive range of the associated prioritization tasks, we focused on ordinal scale techniques in this study. Each of the independent variables have two levels (high and low) leading to a 2x2 factorial design, as summarized in Table I.

The four techniques shown in Table I were instantiated using the tool EstimationWeb, a publicly available web-based tool for estimation, prioritization, and scheduling, developed by our research group [6]. The visual appearance and functionality of the techniques are presented below. Section V.C presents more details on the experimental material.

<table>
<thead>
<tr>
<th>TABLE I. INVESTIGATED PRIORITIZATION TECHNIQUES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>High granularity</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Low cognitive support</td>
</tr>
<tr>
<td>High cognitive support</td>
</tr>
</tbody>
</table>

A. Simple dropdown

With Simple dropdown (T1), shown in Figure 1, each line on the prioritization page contains a feature description and a dropdown box offering choices to give the priorities Essential, Significant, Limited, and Insignificant.

B. Drag into bins

With Drag into bins (T2), shown in Figure 2, the user drags features into or between categories. The categories and their descriptions in T2 were identical to those in T1, implying equal granularity for these techniques. However, the spatial grouping of features may lead the expert to repeatedly compare features of equal or adjacent ratings, possibly resulting in more reliable priorities. T2 is therefore said to offer more cognitive support than does T1.

C. Sortable table

With Sortable table (T3), shown in Figure 3, the features are presented in a table view in which the rows can be dragged and re-arranged according to priority. After sorting the table, the subjects in our experiments used a simple input field to indicate the number of essential features in the now prioritized list. In effect, this method classifies features into two categories, essential and not essential.

![Figure 1. Simple dropdown (T1)](image-url)
D. Pairwise comparisons & ranking

With Pairwise comparisons & ranking (T4) shown in Figure 4, the user is presented with pairs of features, and is prompted to indicate the difference in importance on a scale from 1 (equal) to 9 (extreme difference), in either direction of the two features. After these pairwise comparisons, the tool computes a global ranking, and displays the results in the table to the right. In this table, the user can adjust the calculated ranking. After any adjustments, the subjects in our experiments indicated the number of essentials in the now sorted table. As with T3, this method classifies features into two categories, essential and not essential. With the pairwise comparison step, some degree of redundancy is introduced, leading to reduced sensitivity for accidentally inaccurate or arbitrary assessment. Also, through the repeated contrasting of feature pairs, users can possibly uncover important criteria relevant for the prioritization. T4 is therefore assumed to offer more cognitive support than does T3.

The calculations of ranks uses the algorithms of AHP [7], complemented with Harker’s method to calculate weights and ranks from incomplete pairwise comparisons [8]. The tool was configured to fix the number of pairs to $1.5n$, with $n$ candidate features, i.e., the subjects were asked to perform 24 pairwise comparisons with 16 features. With complete pairwise comparisons, 120 comparisons would have been needed. The chosen number was a compromise between obtaining more stable rankings by way of more
comparisons, and avoiding fatigue and experiment drop-outs due to too many comparisons. Simulations done by Harker showed stable ordinal ranks even below 1.5n [8].

III. RELATED WORK

A number of case studies [9-14] and controlled experiments [15-18] have been conducted to evaluate prioritization techniques with respect to outcome variables such as time usage and accuracy, but none of the studies directly address our research question. However, because our interest is in the accuracy of essential ratings, studies on the effects of prioritization techniques on the overall accuracy of priorities are still relevant.

To date, the results from such studies are inconclusive. For example, techniques based on pairwise comparisons performed well in some studies, e.g., [12], while in other studies simpler techniques performed equally well [17]. One problem with comparing results for prioritization techniques is the possible confounding impact of tool support [17]. Another problem is that no objective measures of accuracy have been employed in the studies. Subjective measures of accuracy would fall short when measuring effects of factors of which subjects are not consciously aware.

A source of insight related to our research question is the study of judgment and choice within behavioral decision science, see, e.g., [19], with links to psychophysics, a discipline within psychology that investigates the relationship between physical stimuli and perceptions, see, e.g., [20]. The next section discusses possible effects based on such insight in more depth.

IV. POSSIBLE CONTEXTUAL EFFECTS

The experimental material was prepared so that subjects across treatments received identical definitions of what it is that a feature is essential. Hence, if the properties of the prioritization technique did not affect the subjects in their assessment, there should be no difference in the reported number of essential features between the techniques. However, we postulate that unconscious biases do affect prioritization, and, further, that various prioritization techniques promote such biases. In the following sections we propose biases that may be inherent in the variation of granularity and cognitive support.

A. Possible effects of granularity

Equal frequency bias. A contextual effect that might affect the threshold for judging features as essential is the tendency in people in a rating situation to distribute stimuli equally over available categories. This equal frequency bias [19] might affect the results of the experiments in two ways: Consider a distribution of features according to a person’s unbiased view of importance. With a positively skewed distribution Figure 5 is an example of such a distribution), the rating of a given feature would be higher than if the same feature occurred in the context of a balanced or negatively skewed distribution. This contextual skewing [21] affects the priorities of the subjects but would not in itself imply group differences because the same set of features is applied across treatments.

However, when increasing the number of categories, the equal frequency bias predicts that the count of features in the highest category (essential) would decrease. This effect is directly relevant in our study because the number of available categories differs between the treatment groups.

If subjects distributed features uniformly across categories, twice as many essentials for two-category techniques (T3 and T4) than for four-category techniques (T1 and T2) would be measured. We note that in situations where categories are described with relative labels, e.g., 1 to 4 or low to very high, such a result could be considered rational, rather than biased. In contrast, our assumption is that in many software engineering contexts the attempt is to assign a precise definition to categories, particularly to the category for essentials. Assigning specific meanings to ordinal categories actually gives the categories a nominal flavor. The tendency to distribute features equally over the available categories in such cases is correctly denoted a bias.

Randomness. Even when influenced by the equal frequency bias, subjects may be able to reliably rank the features on an ordinal scale. In contrast, imagine a situation where features were allocated to available categories by a random process. The predicted result of such a process would be a uniform distribution over the available categories. Hence, randomness and the equal frequency bias are distinct effects that coincide. It is reasonable to assume that more randomness will be present when people have less knowledge or weaker opinions about the subject matter. In software projects, assuming they know their product goals and the features needed for their fulfillment, indications of substantial randomness in priorities would be a cause for concern.

Category effect. Another bias relevant to granularity is related to range-frequency theory. Briefly, the theory describes that judgments of a set of stimuli are a compromise between two principles: Distributing equally over categories and distributing according to equally sized ranges. Somewhat simplistically described, Parducci showed that for skewed distributions and with few available categories, more weight was put on the first of these principles [21]. It is difficult to tell upfront how this will affect the results because the skewness with respect to feature importance is not known. Assuming a positively skewed distribution, the category effect would lead to more essential ratings.

B. Possible effects of cognitive support

With a constant definition of essential features, a neutral result would be that the degree of cognitive support does not affect the reported number of essential features. The cognitive support added by T2 and T4 is assumed to help the subject make more reliable ratings. For example, with
T2 the subject can more easily compare a feature with features of equal ranking. It is possible that this will help the subjects in removing some of the randomness described above. With less randomness, it can be predicted that the average number of essentials will decrease in a bell-shaped or positively skewed distribution.

C. Summary and propositions

Figure 5 summarizes the discussed effects. The graph shows the proportion of tasks as a function of feature importance; the latter measured by some objective measure, such as the return of investment for implementing a feature. Prior to implementation, the subject must make judgments about this feature importance.

When a subject sets a threshold for what constitutes an essential feature with a low-granularity technique (the dashed line), the equal frequency bias, the category effect, and randomness pull in the direction of more essentials, i.e., a larger area under the curve, when compared to a judgment made with a high-granularity technique (the rightmost solid line). These effects are to an unknown extent counteracted by the cognitive support of the techniques and by the experimental condition that the definition of essential is kept constant.

The figure is not intended to suggest a precise position for the dashed line; rather it illustrates that its actual position is some context-dependent compromise between the effects illustrated as arrows, with a higher barrier at the rightmost solid line and a lower barrier at a position corresponding to a uniform distribution over categories (the leftmost solid line).

Figure 5. Summary of proposed effects

V. EXPERIMENTAL DESIGN

A. Overview

Two experiments were conducted to investigate the effect of granularity and cognitive support on the number of product features reported as essential. In the first experiment, 94 subjects were recruited from our research institution and from a Polish software house. The subjects were asked to rate the importance of 16 cell phone features on their buying decision for a new phone. In the second experiment, 44 subjects were recruited from the pool of functional experts at a large development project in the public sector in Norway. The subjects were asked to assess the contribution of 16 proposed features to the fulfillment of the project’s vision. In both experiments the subjects were randomly allocated to one of four treatments groups, corresponding to the four prioritization techniques investigated.

The experiments followed a 2×2 factorial design, with Granularity (low, high) and Cognitive support (low, high) as independent variables and the count of features reported as essential as the dependent variable. Two-way ANOVA on the rank-transformed dependent variable was used to analyze the data.

The first, more artificial experiment can perhaps appear to be superfluous in the presence of the more realistic setting of the second experiment. However, artificiality in experiments can be useful to establish and isolate the existence of phenomena and mechanisms, enable generalization from particularly unfavorable to more favorable conditions, and to relate experiments to theory [22]. This study exemplifies such advantages.

B. Recruiting subjects

Experiment 1. To recruit subjects for the first experiment, we explained the goal and the basics of the experimental design to the head of administration of our research institution. All 120 employees were invited by email, explaining the overall goal and the relevant procedures. This population represents employees in the domain of ICT research. The average age of the invited employees was 39 years, and they represented 25 nationalities. Fifty-nine colleagues agreed to participate. Concurrently we contacted a Polish software house with which our research group has collaborated. The same procedures were followed. The average age of the invited developers was slightly below 30 years, and all of them were Polish citizens. Thirty-five developers agreed to participate. We offered compensation based on standard hourly rates and an estimate of 20 minutes time usage per subject.

Experiment 2. To recruit participants from the large development project, we sent a request to the project manager, explaining the goal and the basics of the experimental design. The authors already had a research collaboration established with this project [23, 24].
request was forwarded to one of the project’s three product owners who suggested a list of 60 project members who participate in prioritization activities on a regular basis. An email was sent to the potential subjects, explaining the purpose of the experiment and the procedures to follow. We explained that neither agreeing to, declining, or ignoring the invitation would have negative consequences for the invited person. We did not offer incentives for participation other than contribution to project-relevant research and experience with prioritization techniques and tools. Eventually, 44 people agreed to participate in the experiment. The subjects were free to perform the task at the time and location of their own choice within 8 days.

C. Experimental material

For Experiment 1, 16 features of modern cell phones were identified by the authors based on the authors’ judgment of what might be important to the experimental population in a buying situation. The actual features can be viewed in Figure 1.

For Experiment 2, we selected 16 proposed features from the product queue documented in the project’s issue tracker. The product owner helped us in making the selection based on two criteria:

- Expected effort is in the same order of magnitude for all features because it makes less sense to prioritize between features at different abstraction levels [3].
- Possible for the subjects to understand the basics of the feature through a brief description. This was to satisfy the experimental condition that limited time could be expended and interaction between project members was not allowed.

The short description already given in the issue tracker could be used almost unchanged, but some of the texts were improved with respect to clarity and consistency. Three examples follow, translated from the Norwegian counterparts used in the experiment:

As a user/employer, I can execute control of the salary file for logical errors so that errors can be corrected in the file before it is submitted to the system

As an agency official, I can reconstruct NAV- and AORD information so that I can see which data was registered at a certain point in time

As a customer service operator I want to have phone calls automatically recorded in the person log when members call SPK and identify themselves with a social security number, so that I spend less time recording the call

Table II outlines the instructions given to the subject for the four techniques in the two experiments.

D. Execution

The subjects were randomly allocated to four equally sized treatment groups, corresponding to each of the four techniques under investigation. The study administrator sent an email containing a brief general description of the study and a personalized link to the web form containing the experimental material.

[intro] ... The purpose is to investigate whether the format of various prioritization techniques affects the priorities. The participants have been divided into four groups, and through random allocation you have been assigned the technique [Simple dropdown|Drag into bins|Sortable table|Pairwise comparisons&ranking]. Once you have started, it is important that you complete without interruptions. Please spend the time needed for you to feel comfortable about your answers. We ask that you work individually and not use other sources than the material we present. Click the link below to start.

| TABLE II. TEXT PROVIDED (TEXT FOR EXPERIMENT 2 HAS BEEN TRANSLATED FROM NORWEGIAN VERSIONS) |
| Technique | Experiment | Prioritization instruction | Categorization |
| T1 and T2 | Experiment 1 | Categorize the features according to their influence on your buying decision for a new cell phone | Essential – I would not buy the phone without it |
| | | | Significant – It would be difficult for me to buy a phone without it |
| | | | Limited – Useful but I can buy a phone without it |
| | | | Insignificant – Does not influence my buying decision |
| T3 and T4 | Experiment 1 | Rank the features according to their influence on your buying decision for a new cell phone | “How many of the above features are essential – you would not buy the phone without it. Please type in one number between 0 and 16” |
| T1 and T2 | Experiment 2 | Categorize the tasks according to their contribution to the project’s vision, the way you perceive the vision | Essential – The vision will not be fulfilled without it |
| | | | Significant – The vision will be hard to fulfill without it |
| | | | Limited – Useful, but the vision can be fulfilled without it |
| | | | Insignificant – Does not influence the fulfillment of the vision |
| T3 and T4 | Experiment 2 | Rank the tasks according to their contribution to the project’s vision, the way you perceive the vision | How many of the above elements are essential – the vision will not be fulfilled without it. Please type in one number between 0 and 16 |

In Experiment 1, all 94 subjects submitted their priorities within 8 days. Reminders were sent by e-mail to nonrespondents after 3 and 7 days. For Experiment 2, 44 out of 60 invited subjects replied within 8 days. One reminder was sent to nonrespondents after 6 days. Sixteen invited subjects did not submit their responses.

A web-based experiment that gives subjects freedom to choose time and place for completing the experimental tasks imposes some threats to validity, as further discussed in Section VII. Such validity threats would have been reduced or eliminated had all subjects been under our supervision. However, having 44 domain experts from one project meet
E. Analysis

The experiments are designed with two independent variables: gr (granularity) and cogn (cognitive support). The variables have two levels, low and high, giving a 2x2 factorial design (see Table I). The dependent variable Ess is the number of features reported as essential. Two-way ANOVA is used to identify effects of the independent variables on Ess and interaction effects. A rank-converted measure of Ess was used to avoid sensitivity to the ANOVA normality requirements. A post-hoc Normal Q–Q plot indicated a non-normal distribution of residuals in Experiment 1 but not in Experiment 2. However, we employed rank-converted measures of Ess in the ANOVA for both experiments. The analysis was executed using the statistical analysis package R version 2.10.1.

F. Deviations

In three cases (Experiment 1) and four cases (Experiment 2), the subject misunderstood the instruction of providing the number of essential features. We noticed this misunderstanding on reception of the web forms (within a few minutes after submission) and asked the subject by e-mail to update the response.

VI. Results

Descriptive statistics for variable Ess in the two experiments are shown in Table III and Table V, while the ANOVA results are shown in Table IV and Table VI.

In Experiment 1, the average number of essential ratings increased by 56% (statistically significant) with low-granularity techniques (T3 and T4) compared to high-granularity techniques (T1 and T2). The effect size measured by Cohen’s d is 0.73, which is in the medium category of effect sizes reported in software engineering experiments [25].

In Experiment 1, the average number of essential ratings increased by 32% (statistically significant) with low cognitive support (T1 and T3) compared to high cognitive support (T2 and T4). The effect size measured by Cohen’s d is 0.43, which is in the lower end of medium effect sizes reported in software engineering experiments [25].

In Experiment 2, the average number of essential ratings increased by 195% (statistically significant) with low-granularity techniques (T3 and T4) compared to high-granularity techniques (T1 and T2). The effect size measured by Cohen’s d is 2.40, which is in the highest category of effect sizes reported in software engineering experiments [25].

In Experiment 2, the average number of essential ratings increased by 33% (not statistically significant) with high cognitive support (T1 and T3) compared to high cognitive support (T2 and T4). The effect size measured by Cohen’s d is 0.50.

In summary, the results showed that the number of essential ratings significantly increased with lower granularity. The effect size was greater in Experiment 2 (the software engineering context) than in Experiment 1. Added cognitive support affected the number of essential ratings in both experiments but in opposite directions and with statistical significance only in Experiment 1. The results showed no interaction effect between granularity and cognitive support. Re-doing the analysis with the original outcome measure Ess (the ANOVA analyses used a rank-converted measure) did not change any of the significance levels.

VII. Discussion

The results showed that with lower granularity significantly more features are rated as essential, even with a constant definition of essential. In Experiment 1, the effect on essential ratings was of a magnitude that could be predicted from the discussion in Section IV and summarized in Figure 5. Halving the granularity increased the essential ratings but not by 100% as a pure uniform distribution
strategy would have predicted. In this case it seems the
definition of essential (I would not buy the phone without it)
helped in limiting the effects that pull in the direction of
more reported essentials. Also, a higher level of cognitive
support had the predicted direction in Experiment 1.

The result for granularity in Experiment 2 is extreme.
The effect of reducing granularity is larger than even a full
adherence to the principle of equal distribution over
categories would have accounted for. In this case it seems
that the definition of essential (The vision will not be
fulfilled without it) had little or no significance to the
subjects. We note that in the studied project the term vision
is central in the release planning and prioritization
processes. Project management assumes that the vision is
understood and shared between the key stakeholders; however, the present result indicates that this assumption
may not be met entirely. Indeed, in a post-hoc analysis, we
measured the intra-rater agreement of the priorities by
Kendall’s W, giving values of 0.39 and 0.33 for the two
experiments, respectively. These are both in the range of
low to medium correlation, and it is interesting to note that
individuals ranking their personal cell phone preferences
agree more than do software project stakeholders assumed
to share a common vision.

An important question is whether the subjects were able
to reliably assess the relative importance of features. If so,
the results can be explained by the equal frequency bias.
Alternatively, the results were influenced by a high degree
of randomness, which would be a larger cause of concern.
With reliable ordinal rankings, the most important features
would still be selected for development, but this would not
necessarily be the case with a high degree of randomness in
the priorities.

The differences in effect sizes between the two
experiments could possibly be explained by extensive
randomness in the priorities given in Experiment 2. Future
plans are to re-execute the experiment sessions, which
would allow us to measure the intra-rater agreements of the
essentiality ratings. A lower intra-rater agreement in the
field context than in the phone feature context would
support the proposition that more randomness was present
in the field.

At the outset, the implications of the study results seem
important. If stakeholders’ threshold for judging features as
essential is sensitive to the prioritization technique and
perhaps also to the questioning format in general, this could
be an important obstacle against collecting trustworthy
judgments as input to release planning. Indeed, it is possible
that the results point to a root cause for software failure
archetypes such as Unable to meet the user needs.

It is important to note that any direction of bias can be
harmful. Thresholds that are too strict for judging features as
essential can imply that required functionality is not
prioritized over more dispensable features or even gold-
plating features. Thresholds set too loosely could lead to
projects being prematurely stopped due to outlooks of
severe cost overruns or could mean that falsely reported
essentials take the place of true essentials when the project
budget becomes tight.

In practice, projects have mechanisms to compensate for
judgmental biases of the kind demonstrated in this study.
First and foremost, group discussions and other forms of
broad-banded communication can enable stakeholders to
counteract the biases through group discussions and
clarifications. However, such broad-banded communication
is more difficult to achieve in the largest projects and in
projects where stakeholders cannot meet frequently.
Unfortunately, these are exactly the kinds of projects that
are already considered at risk.

The results are also relevant for discussions on how to
handle change and feature requests in the context of a
commercial development contract. Such contracts
sometimes describe different procedures and conditions for
handling change or feature requests of different importance.
For example, a contractor might commit to expediting the
development of essentials or to develop them as part of a
fixed-price contract. Biases in judgments of essentiality
could therefore easily have commercial and legal
consequences.

Providing recommendations on the basis of the results
are not straightforward. Since biases in both directions can
be harmful, it is not possible to give recommendations, e.g.,
such as Use a high number of categories. Furthermore,
improvements in the accuracy of essential ratings are more
difficult to assess then in the context of cost estimation.
While cost estimates can be compared with an objective
measure of actual expenditure, it remains a matter of
subjective opinion to determine whether a requirement was
eventually essential. Currently, our best advice would be to
triangulate priorities by combining different prioritization
techniques. On major deviations between stakeholders or
between techniques, stakeholders should meet to clarify
their views so that the responsible product owner can make
the final decisions based on more and better information.
Being as precise as possible in the priority guidelines,
category definitions, and descriptions of features is likely to
help, but as the results from the present study demonstrate,
such measures are unlikely to fully remove the biases.

VIII. VALIDITY ISSUES

Conclusion validity. Unreliable measurement induced by
the nonsupervised experimental context can be a threat to
conclusion validity. With nonsupervised execution, subjects
may more easily break the rules by collaborating, answering
by random, answering destructively, or consulting
information outside of the experimental material. We have
no explicit reason to believe such problems were prevalent,
given the well-willingness to participate and the
professionalism of the participants. To some degree the use
of robust methods of analysis would have counteracted
effects of outliers in the data due to such problems.
**Internal validity.** We do not see any large threats to establishing internally valid cause–effect relationships from the experiment. However, the identification of causes is defined by the construct used for the independent variables, and we cannot obtain more information about underlying causes and mechanisms from the experimental results. We would have liked to have gathered qualitative data to complement the quantitative analyses; however, we preferred to spend our allowed time of access to subjects on collecting the best possible quantitative and objective data.

**Construct validity.** There are many ways of adding cognitive support to prioritization techniques. Effects might differ between different operationalizations of cognitive support; hence the results cannot be immediately generalized beyond the specific operationalizations used in this study. For granularity, we have shown effects of varying between two and four categories. Whether similar effects occur for other specific levels of granularity remain to be investigated. We do not believe the experimental material or context have provoked or exaggerated the results. On the contrary, we paid significant attention to articulating the instructions, the category definitions, and the feature descriptions precisely and in an understandable manner.

**External validity.** For Experiment 2, the population was restricted to one specific development project. Hence statistical inference can be used to generalize to this population but not automatically to other software development projects. However, combined with the results from Experiment 1, this indicates that the effects are indeed domain independent and possibly instances of robust psychological effects. Hence similar effects are likely to occur in other development projects.

It is also possible to use the logic of case studies to discuss external validity [26]. Earlier, we have argued that the investigated project can be seen as a critical case in the class of large and agile software development projects [23]. The project is a prestige project in the Norwegian public sector, attracting the best skilled workers, both on the client and contractor side. Great attention has been put on sharply defining the scope and vision for the project. Critical case logic implies that other, less fortunate projects in the same class are likely to face similar or more severe challenges.

IX. CONCLUSIONS

We have conducted two controlled experiments to investigate whether certain attributes, granularity and cognitive support, of prioritization techniques affect stakeholders’ threshold for judging product features as essential. In an experiment asking subjects to pick essential mobile phone features, the number of reported essentials increased by around 50% when granularity decreased from four to two categories. The effect was extreme—195% in the experiment conducted in a realistic software engineering context.

It seems that subjects in both experiments had a tendency to distribute features equally across available categories (known as the equal frequency bias), despite a clear and constant definition of essential across treatments. The extreme effect in the software engineering context indicates that the stakeholders were able to make absolute judgments of essentiality only to a very limited degree and instead resorted to, at best, ordinal ranking. Additionally, pure randomness in assigned priorities can explain such results.

For cognitive support, the results are less conclusive. In the phone feature experiment, adding cognitive support resulted in a statistically significant decrease in the number of essential ratings. An opposite, but not statistically significant, effect occurred in the field experiment. The most immediate explanation is that the potential effect of cognitive support was overshadowed by the equal frequency bias and randomness in the latter context.

Incorrectly picking essential product features can have harmful effects for software projects. This study has shown that contextual biases can have large effects when stakeholders assign priorities to software product features. We believe this study has shown the importance of designing and employing practices that counteract or manage such effects.

ACKNOWLEDGMENT

We are grateful to Krzysztof Rolak, Mona Hegreberg, and Ottar Hovind for accepting and helping us to conduct the experiment in their organizations and to the participants from PGS Software, the SPK/Perform project, and Simula Research Laboratory, respectively. We thank Torleif Halkjelsvik at the Department of Psychology, University of Oslo, for pointing out relevant empirical studies within psychology. The work was partly funded by the Simula School of Research and Innovation.

REFERENCES


