The Effects of Request Formats on Judgment-based Effort Estimation

Magne Jørgensen
Simula Research Laboratory & Department of Informatics, University of Oslo, Norway
Torleif Halkjelsvik
Department of Psychology, University of Oslo, Norway

Abstract: *In this paper we study the effects of a change from the traditional request "How much effort is required to complete X?" to the alternative "How much can be completed in Y work-hours?". Studies 1 and 2 report that software professionals receiving the alternative format provided much lower, and presumably more optimistic, effort estimates of the same software development work than those receiving the traditional format. Studies 3 and 4 suggest that the effect belongs to the family of anchoring effects. An implication of our results is that project managers and clients should avoid the alternative estimation request format.*

Keywords*: Effort estimation, human judgment, anchoring*

1. Background

The main determinant of many types of software-related investments is the amount of development work effort required. Software clients' ability to base decisions on accurate cost estimates or prices are consequently strongly tied to the software providers' ability to estimate the effort accurately. Similarly, project managers' ability to plan a project and to ensure efficient development work frequently depends on accurate effort estimates. This importance of accurate effort estimates is illustrated by the finding in a recent (2007) survey of more than 1,000 IT-professionals.[1] The survey reports that two out of the three most important causes of IT-project failure were related to poor resource estimation, i.e., inaccurate effort estimates. The survey response is understandable. A review of estimation accuracy studies (Moløkken and Jørgensen 2003) reports that software projects expend, on average, 30-40% more effort than estimated. Software projects can experience severe delivery and management problems when plans are based on overoptimistic effort estimates.

The negative effects of overoptimism are accentuated by (i) software bidding rounds in which those companies that provide overoptimistic effort estimates are more likely to be selected, and (ii) strong overconfidence in the accuracy of the estimates; for example, 90% confidence effort prediction intervals only include the actual effort 60-70% of the time (Jorgensen, Teigen et al. 2004).

Software engineering researchers have been addressing the problems of inaccurate and overoptimistic effort estimation in software development projects since at least the 1960s (see e.g. (Farr 1964). The focus of this research has mainly been on the construction of formal software effort estimation models, and a number of estimation tools that implement these models are now commercially available. Yet, despite high awareness of these estimation tools, and the promotion of model-based estimation, software engineers typically use judgment-based methods ("expert estimation") to estimate effort (Heemstra and Kusters 1991; Hihn and Habib-Agahi 1991). One rational reason for the low use of models, supported by the review in (Jørgensen 2007), is that models do not on average produce more accurate effort estimates than expert judgment. Inherent, essential advantages of judgment-based effort estimation seem to include:

[1] See: www.informationweek.com/news/management/showArticle.jhtml?articleID=198000251

- Software development domain experts typically possess highly context-specific information that is not part of the model. This situation has been found to favour judgment-based estimation over model-based estimation (Webby and O'Connor 1996; Goodwin 2000).
- Essential software development relations, such as the effort-size relation, seem to be unstable (Dolado 2001). As pointed out in (Sanders and Ritzman 1991), judgment-based forecasts were better in unstable situations, while the models performed better during periods of stability.

These two advantages may compensate for the typical negative aspects of judgment-based estimates, such as a higher degree of inconsistency and improper weighting of variables.

If, through better knowledge about the judgmental processes, we could reduce their shortcomings while retaining the advantages, the improvement in accuracy when estimating software development projects may be substantial. This paper describes an attempt to explore a different format of the estimation request. Instead of the traditional request of work hours on a specified task, we rather asked for the amount of work to be completed within a specified time. Our goal is two-fold, first and foremost we want to provide helpful guidelines for how to formulate estimation requests, and second, we hope to gain insights into the judgmental processes at work.

In earlier studies we report how much changes in formats of the request influenced the software professionals' judgments in field and experimental settings (Jørgensen 2004; Jørgensen 2006; Jørgensen and Grimstad 2008), e.g., how estimating the completion of identical work described as a "minor extension", an "extension" or a "new functionality" affected the effort estimate. The results of these studies suggest that effort estimation processes may be strongly affected by a change in how requests are formulated.

The research question of this paper is as follows:

RQ: Does a change from the traditional request "How much effort is required to complete X?" to the alternative "How much can be completed in Y work-hours?" affect effort estimates?

Rationally speaking, the estimated effort of the same part of a development project should be the same regardless of which of these two formats that are applied. It is, however, also possible that different request formats trigger different judgmental processes and knowledge.

The estimation request: "How much effort is required to complete X?" is the format traditionally employed in software development effort estimation work. The alternative format: "How much can be completed in Y work-hours?" may however be more and more relevant, due to the increased use of incremental development. In incremental development the main question in many cases is how much a team is capable of producing as part of the next increment, i.e., a problem description close to that of the alternative format. Incremental development methods, such as agile methods, recommend the use of historical productivity ("velocity") of previous increments as input to the estimation process. The estimate should, however, not be mechanically derived from this historical productivity, since there may be learning effects and differences in complexity or development skill that need to be adjusted for from increment to increment. This means that even when there are historical data available, there will be judgment involved when applying the alternative format. Another situation where the alternative format may be of relevance is the situation where a client asks the project manager about how much of the functionality he can get before a certain date or within a given budget. If it turns out that the alternative format increases or reduces overoptimism, this knowledge can be used to develop better estimation practices.

Before we started our study, we made a rather extensive search for studies on time and effort estimation in various research domains. We were unable not find any study examining the rather fundamental differences in request format described in our research question.

The remaining part of this paper is organized as follows: Sections 2-4 describe the design and results of the Studies 1-4, respectively. Section 5 discusses possible explanations of the results. Section 6 concludes.

2. Study 1

2.1 Design

Fifty-four software professionals from the same web-development company participated in Study 1. Twenty-one of the participants were developers, 15 were designers (graphical designers and human interaction designers) and 18 were managers (project managers and general managers). We expected that nearly all of them had some experience in estimating software development effort, although those in non-technical roles would base their estimates on different estimation processes than those in technical roles (Moløkken and Jørgensen 2005).

The participants were randomly divided into two groups: Traditional and Alternative. The Traditional group received the traditional estimation request, while those in the Alternative group received the alternative estimation request, see Table 1. All participants received the same project to be estimated, see Appendix 1. We collected information about each participant's self-assessed general programming competence before the estimation work, and his/her self-assessed competence to estimate the enclosed project after the estimation work. We instructed the participants to assume an iterative (incremental) development model. This is a type of model the company had used for many of its projects.

Table 1: Estimation Instructions

Group: Traditional format	**Group: Alternative format**
The next page describes the system RDinner. Assume that your company is assumed to complete this project and work in accordance to an iterative development model with weekly deliveries. Each of these weekly deliveries should have production quality.	
How much effort do you think your company would most likely spend to develop RDinner? Assume medium productivity, that you aim at good quality of each of the deliveries, and, that you use development technology where your company has sufficient expertise.	*How many of the "user stories" (requirements) described on the next page do you think it will be possible to include in iteration one (the delivery of the first week), assuming that the developer(s) will spend about 30 work-hours on this project the first week? Assume that you will start with user story 1 (U1), continue with user story 2 (U2), etc.* *You should base the answer on medium productivity, that you aim at good quality of each of the deliveries, and, that you use development technology where your company has sufficient expertise.*
I believe that the most likely effort required by our company to develop this system is approx. _____ work-hours	*I believe that our company, spending about 30 work-hours, would be able to deliver the user stories _____ (specify on the format: U1-Ux) in their first delivery (first week).*
	How much effort is required for the remaining user stories?
	I believe that the remaining user stories (after delivery 1) will require approx. _____ work-hours.

As can be seen from Table 1, the estimates produced by those in the Traditional group should, normatively speaking, be the same as the sum of the estimated remaining effort and 30 work-hours in the Alternative group. Notice that those in the Alternative group had to estimate with the presupposition that user stories had to be completed successively, as listed in Appendix 1. This restriction may have a minor productivity decreasing effect, given that the sequence is not optimal for the development work. We did, however, not expect this to affect the estimates markedly and, if anything, in the direction of higher effort estimates. Any large differences in the estimates of these two groups would, as far as we could see, be caused by the difference in the estimation request format.

2.2 Results

2.2.1 The Quality of the Estimation Work

The participants did not spend more than 5-10 minutes on the estimation work and some of the participants did not have much technical competence. We therefore conducted a simple analysis of the quality of the estimation work to assess the meaningfulness of the estimates.

In a previous study (Jørgensen and Grimstad 2008), effort estimates from forty-six outsourcing companies in various countries were collected applying almost the same requirement specification as in the current study. The estimation request to those companies was based on the traditional format. The only differences between the requirement specification of the previous and the current study are that some estimation irrelevant information and one functional requirement were removed from the specification in the current study. The removed requirement is a functionality that, by most of the outsourcing companies, was estimated to be a rather small task, typically less than 10 work-hours. Irrelevant contextual information was removed to shorten the reading time for the participants in the current study.

The companies participating in the previous field study applied their ordinary estimation processes and were paid on an ordinary basis, i.e., they were not part of an experiment but completed ordinary estimation work. If the estimates from the current study are similar to those of the forty-six outsourcing companies, this suggests that the estimates of the software professionals in the current study are of acceptable quality in spite of the short time usage. Table 2 shows that the effort estimates of those exposed to the same request format were indeed quite similar.

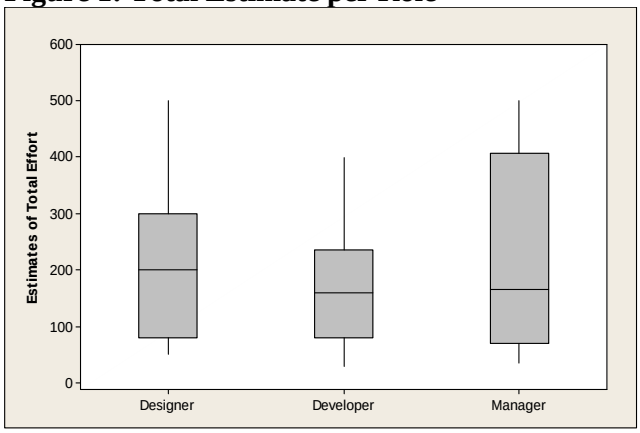
Table 2: Estimates of the Field Study and the Current Study

Study	n	Q1	Md	Q3	M	SD
Previous field study	46	119	190	339	273	229
Current study (traditional format)	28	160	220	375	287	218

2.2.2 Role and Level of Competence

Self-reported programming and estimation competence was coded as 1 = "Poor", 2 = "Acceptable", 3 = "Good" and 4 = "Very good". There were no differences between the experimental groups on level of estimation competence, Kruskal-Wallis Test, $\chi^2 = .62$, $p = .43$, or programming competence, $\chi^2 = .64$, $p = .43$. The proportion of designers, developers and managers were approximately equal in the two groups, $\chi^2 = .26$, $p = .88$. Those not working as developers typically rated their competence as "Poor". Their estimates were nevertheless similar to those working as developers, see boxplot in Figure 1. Based on these observations we found it meaningful to include all roles and competence levels in our analysis.

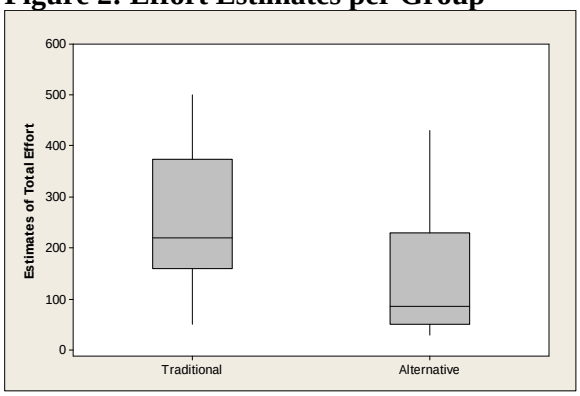
Figure 1: Total Estimate per Role



2.2.3 Effects of Format

Our research question is, as stated in Section 1, whether the alternative format affects the effort estimates. The collected data show that there could be a critical difference between the traditional and the alternative format, see boxplot in Figure 2. The median total estimate of those in the Traditional group was 220 work-hours compared to only 85 work-hours in the Alternative group. A Kruskal-Wallis Test indicates a statistically significant difference, $\chi^2 = 12.84$, $p < .001$.

Figure 2: Effort Estimates per Group



We do not know how much effort the completion of this system typically would require for the company. We find it nevertheless reasonable to assume that the estimates of those in the Alternative Group on average were overoptimistic. The two main reasons for this are that: i) We find it unlikely that the work should require substantial less than the effort estimated by the companies in the previous field study, ii) The average estimate of software professionals is 30-40% too low, when applying the traditional format.

Among those receiving the alternative format, the productivity assumed the first increment (first week) was similar to that of the remaining increments. The median estimated work-effort per user story was about 10 work-hours per user stories the first increment and 13 work-hours per user stories the remaining increments, Wilcoxon Signed Ranks Test, $Z = -.89$, $p = .38$. The median work-effort per user story was about 28 for those receiving the traditional format. This shows that those receiving the alternative request assumed a higher productivity not only for the first increment. The software professionals may have used the productivity of the first increment as input to estimate the effort of the remaining work.

The large effect was surprising to us and we felt a need for a replication with another population of software professionals and small changes in the estimation instructions to investigate its robustness.

3. Study 2

3.1 Design

We invited 77 software professionals to estimate the same development work as in Study 1. The participants, who were different from those in Study 1, were randomly divided into four groups. Two of these groups (n=35) was exposed to the traditional estimation request, and one group (n=19) to the alternative request. The fourth group (n=18) was exposed to a condition that was not part of this study.

Two changes were made to the estimation instructions from Study 1: i) The software professionals were instructed to assume that they would do all the work by themselves. In Study 1 they were instructed to base their estimates on the average productivity of their own company, ii) After predicting the number of user stories to be completed the first week, the software professionals in the Alternative group estimated the total effort for the project, not the remaining effort as in Study 1.

These changes enabled us to test the robustness of the findings in Study 1, e.g., whether some of the differences in estimates between the groups are related to estimation of own versus others' work, or to the estimation of remaining versus total effort.

3.2 Results

3.2.1 The Quality of the Estimation Work

Similarly to Study 1, we compared the estimates to those of the 46 outsourcing companies in the previous field study, see Table 3. As can be seen, the differences are not very large, which suggest meaningful estimation work of the participants in Study 2.

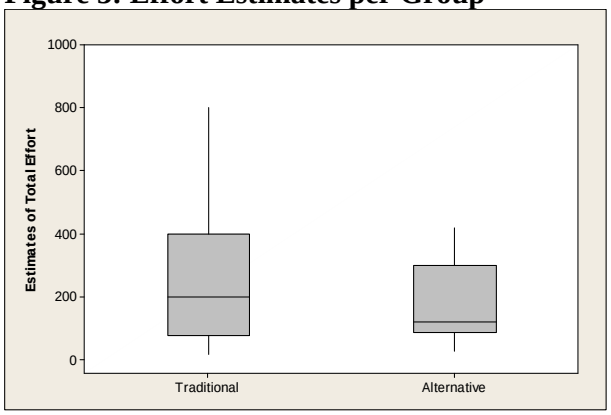
Table 3: Estimates of the Field Study and the Current Study

Study	n	Q1	Md	Q3	M	SD
Previous field study	46	119	190	339	273	229
Current study (traditional format)	35	80	200	400	343	460

3.2.2 Effects of Format

The median estimate of those in the Traditional group was 200 work-hours compared to 120 work-hours in the Alternative group, i.e., the difference between the group medians is only slightly smaller than in Study 1. A boxplot of the estimates per competence groups is displayed in Figure 3.

Figure 3: Effort Estimates per Group



Since Study 2 is a replication of Study 1 and we already had hypothesized the direction of the effect, we used a one-tailed instead of a two-tailed Kruskal-Wallis Test of significance. The statistical significance of the difference was, $\chi^2 = .81$, $p = .19$. When including only the participants with self-assessed competence "Acceptable", "Good" or "Very good" in the analysis, we get $\chi^2 = 1.88$, $p = .09$. The decrease in statistical significance of Study 2 may to some extent have been caused by a larger variation in estimates of those with low estimation competence.

Notice that even non-significant results replicating results of previous studies, i.e., studies that report differences in the same direction and with comparable effects, can provide support of previously reported results. In our case, we have almost the same difference in median estimates in both studies. The important role of replications and its relation to tests of statistical significance is discussed in for example (Hubbard and Vetter 1996).

Studies 1 and 2 establish that the alternative format can lead to lower estimates of total effort than the traditional format. The following two studies (Studies 3 and 4) aim at gaining insight into the robustness of this effect, the mechanisms involved, and to discover conditions that could strengthen or reduce the effect.

We believed that the human judgment literature on anchoring effects (e.g. (Kahneman, Slovic et al. 1982) could provide possible explanations for the effect. If the time-frame we provided (30 work-hours) served as a reference for either the initial or the remaining estimates, we would

expect a longer time-frame to produce higher estimates than a shorter time-frame. More of practical relevance, we wanted to see if the effect was controllable by varying the time-frames. At the same time we tested whether the observed effect mainly applies to software development work, or is more of a general effect.

4. Study 3

4.1 Design

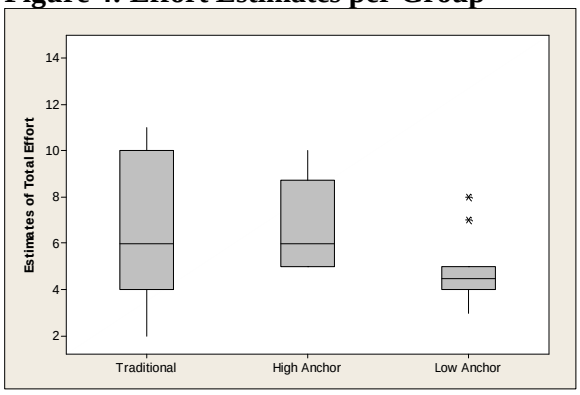
The sample in Study 4 consisted of 45 female and 11 male students taking part in an introductory psychology course at the University of Oslo. Participants were randomly allocated to three different conditions. Two of the participants failed to provide complete effort estimates.

Participants first indicated their reading speed compared to other students on a 7-point scale (1 = "much faster", 7 = "much slower"). Then, they estimated how long, excluding breaks of different kinds, it would take to read and take notes from a particular chapter on the syllabus. The chapter was quite long, with 42 two-column pages. One group received the traditional request asking for total work effort in hours, while the second and third group first received the alternative request, and then estimated the duration for the remaining pages. The second group, which we refer to as the low-anchor group, initially estimated the number of pages they were able to read the first 2 hours. The third group, which we refer to as the high-anchor group, initially estimated the number of pages they would read within the first 5 hours. Total estimates for participants in these two groups were computed by adding the estimate for the remaining pages to the given time-frame of the initial estimate (2 hours/5 hours). For the traditional format we used the total estimate as provided by the participants. All estimates were converted to minutes before analyses. If participants provided intervals, e.g. "from 3 to 4 hours", the middle value was used (210 min).

4.2 Results

A boxplot of the resulting time estimates is displayed in Figure 4. There were no significant differences between the groups on gender or age, but the reading speed seemed to differ slightly between groups, $F = 2.88$, $p = .065$. However, the traditional group, $M = 4.35$, $SD = 1.00$, and the low-anchor group, $M = 4.47$, $SD = 1.01$, did not differ, $t = .34$, $p = .73$, and the high-anchor group actually scored lower (faster reading speed), $M = 3.75$, $SD = .97$, than the other two groups. Since our hypothesis was that a high anchor would lead to higher estimates, which necessarily means a slower reading speed, this difference actually worked against our hypothesis.

Figure 4: Effort Estimates per Group



The difference between the traditional group and the low-anchor, alternative format group was in the expected direction. We can see from Figure 4 that those predicting the amount of work they could complete within 2 hours gave lower total estimates than those receiving the traditional request, Kruskal-Wallis Test, $\chi^2 = 2.06$, $p = .07$ *(one-tailed)*. As another replication of the first two studies this supports the main finding in this paper.

If the phenomenon under investigation is caused by an anchoring effect, we would expect the high anchor to yield higher estimates than the low anchor. Six participants in the high-anchor group predicted that they would finish reading all the 42 pages within the first 5 hours, which was the time-frame given for the initial question in this condition. Since 5 hours was the minimum possible value for the dependent in the high-anchor instructions we could not test differences with rank-based statistics, such as Kruskal-Wallis. Instead we used median tests of the distribution above and below the combined median. The combined median of the low-anchor and the high-anchor conditions was 5 hours. Participants with estimates at the median (5 hours) were grouped with those giving estimates lower than the median. Accordingly, the distributions would not have changed even if we allowed for estimates lower than 5 hours in the high-anchor condition. Table 4 illustrates a significant difference between the low-anchor group and the high-anchor group, $\chi^2 = 10.09$, $p = .001$. Participant that first estimated how much work they could do in 5 hours had significantly higher total estimates than those initially estimating how much they could do in 2 hours.

Table 4: Low Anchor vs High Anchor

	Low Anchor	High Anchor
> Median	3	15
<= Median	13	6

To compare the high-anchor with the traditional format, we performed another median test, see Table 5. The estimates obtained from a high-anchor frame did not differ from the estimates in the traditional format, $\chi^2 = .00$, $p = .97$. Thus, an alternative format with a long initial time-frame did not result in lower total estimates.

Table 5: Traditional vs High Anchor

	Traditional	High anchor
> Median	8	10
<= Median	9	11

A large time-frame of 5 hours in the initial estimation reduced the effects of the alternative format. The current experiment might however have been a too extreme test of our assumption that longer time-frames give higher total estimates. The 5-hour time-frame in the high-anchor condition equals the combined median of the two alternative groups, and the instructions did not state that it was possible to use less than 5 hours on the reading task. For this reason our result might have been affected by demand characteristics, such as participants choosing a value below 42 pages to please the experimenter, or in order to have something left for the next question. Still, 6 participants in the high-anchor condition answered that they were able to complete 42 of 42 pages within the first 5 hours. If the alternative format automatically lead participants to downscale the task, we would expect more responses on this value, or at least more participants with a total estimate of 6 hours or less.

To avoid further speculation we wanted to test if a less extreme anchor, a medium-sized anchor, would lead to higher estimates than a low anchor. The next study (Study 4) used a different task and a third person perspective to further explore the robustness of the phenomena.

5. Study 4

5.1 Design

We initially recruited 54 students in a history of philosophy course at the University of Oslo. Seventeen received an experimental condition irrelevant for the current paper (framing in minutes

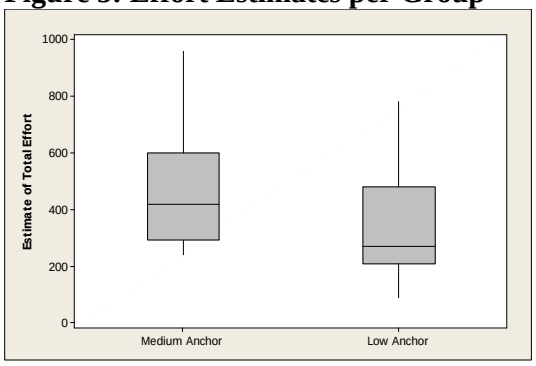
vs. hours), and one participant failed to provide estimates. This left 19 participants in a low-anchor condition and 17 participants in a medium-anchor condition.

The task was to estimate how long it would take for an imagined friend of the participants to write 24 pages of a book by hand. The estimation requests were preceded by a description of an imagined friend who needed a chapter from a book, with no copy machine available. In the low-anchor condition the imaginary friend wanted to know how many pages the participants thought she could finish within the first half hour, and in the medium-anchor condition she asked how many pages she could finish within the first 2 hours. Then both groups estimated how much time it would take to complete the chapter. As in the previous study, the total estimate was the sum of the initial time-frame (half an hour/2 hours) and the remainder estimate.

5.2 Results

Only one participant believed it would take less than two hours to complete the task. In the medium-anchor group the lowest estimate was 240 minutes, i.e. the double of the minimum possible value. Thus, we succeeded in creating a medium-sized anchor, and based on the distribution we could defend the use of rank-based tests. A boxplot of the resulting effort estimates is displayed in Figure 5.

Figure 5: Effort Estimates per Group



The median of the medium-anchor group was 420 minutes, significantly higher than the low-anchor group, which only reached a median estimate of 270 minutes, Kruskal-Wallis Test, $\chi^2 = 3.80$, $p = .03$ *(one-tailed)*. The result confirmed our hypothesis that an initial medium time-frame (2 hours) would lead to higher total estimates than a short time-frame (half an hour).

6. General Discussion

In four studies, an initial estimation request in the format: "How much work can you do in X time?" followed by a request for the remaining work effort (or total work effort) produced lower estimates than the traditional format: "How much time does Y work take?" The shorter the initial time-frames, the larger effects we found. When the time-frame was close to the median of the traditional format, i.e., probably closer to the time needed to complete the whole task, the effect was not present. This section discusses possible explanations for these findings.

6.1 Anchoring Effects

The "anchoring and adjustment" strategy, described in (Kahneman, Slovic et al. 1982), predicts that an estimator will start with a convenient reference value (the anchor) and then adjust the estimate up or down until the estimate feels right. This mental strategy may work acceptably when the reference value is reasonable, e.g., when the anchor is the typical effort of tasks of the same category or the effort of the closest analogy. However, when the anchor is misleading, e.g., consists in a clients' unrealistic budget, the anchoring and adjustment heuristic may lead to quite inaccurate effort estimates (Jørgensen 2007; Jørgensen and Grimstad 2008).

An anchoring effect may in our studies be based on a downward adjustment of the total work effort in the whole project. When asked to judge how much work to be delivered within a specified time-frame, estimators may start with the whole project as a reference (e.g. 8 user stories). Then this anchor is insufficiently adjusted downwards (e.g. towards 2 or 3 user stories). If this initial estimate serves as a reference or anchor for the judgment of the remainder of the project, then the total estimate would be biased. This means that there could be two successive processes involved. First, the initial estimate is distorted by an insufficient downward adjustment from the total amount of work in the project, and then the estimate of the remaining work is based on the efficiency of work covered by the initial estimate. The latter estimate could also be caused by an insufficient upward adjustment of the time-frame specified in the initial estimate (e.g. 30 hours).

The original explanation of the anchoring effect has been challenged. In (Mussweiler and Strack 2001) it is suggested that the anchoring effect is mainly caused by activation of semantic knowledge and confirmatory hypothesis testing. In our case this implies that the effect of the alternative format may have been triggered by the low numerical value (e.g., 30 work-hours) introduced. This value may have acted as an anchor that activated knowledge about trouble-less software projects, e.g., previously completed projects where it was possible to complete a large proportion of the functionality the first week. Perhaps the estimator also engages in automatic confirmatory hypothesis testing, e.g., "Can I deliver 8 user stories in 30 hours?" See (Mussweiler 2003) for more on this.

We have proposed a two-stage model in which the initial estimate is biased by some form of anchoring, and then this estimate itself operates as an anchor, or establishes a standard of efficiency, in which the remainder estimate is based on. We do have other suggestions which are not mutually exclusive or incompatible with our preliminary anchoring account.

6.2 Motivational Bias

Asking how much someone can accomplish within a certain time-frame might have induced a focus on how to achieve the goal, or what is known as an "implemental mind-set" (Gollwitzer 1998). An implemental mind-set, i.e. considering the question "*how* will I do it?", leads to lower perceived vulnerability to both controllable and uncontrollable risks, and enhances self-perception of qualities and skills (Taylor and Gollwitzer 1995). In other words, this way of thinking is found to boost positive illusions. In our studies, we used initial time-frames of different durations and found larger effects for relatively smaller time-frames. If the motivational bias we refer to were the sole explanation of the current finding, it would mean that shorter time-frames lead to a more "how-to" mode than longer time-frames. Although this could be the case, as imagining smaller parts of a task possibly increases vividness and the perception of details, we still suspect that this effect is not a crucial one, as the actors differed from study to study. In Study 1, an average employee in a company was the imagined actor. In Study 2, the participants were told to assume that they were to carry out the software development themselves. And, in Study 4, the participants pretended to give advice to a friend. If parts of the effects stem from a motivational bias, it is not a bias just attached to participant's own effort and abilities.

One question is whether the alternative format actually could induce an implemental mindset; another question is to what extent the resulting positive illusion could influence the effort estimates. To our knowledge, no studies have directly explored the latter issue, but some findings regarding planning of a task and "future focus" might point in this direction (See section 6.3).

6.3 Qualitatively Different Information

Does the change in format of the estimation request lead to judgments based on qualitatively different information? Kahneman and Lovallo (1993) made a distinction between the *inside view* and the *outside view* in prediction contexts. An inside view implies that a software developer tries to foresee the necessary steps, break down the task into activities, and think about how to complete each activity successfully, i.e., a strategy based on "looking forward". An outside view relies on the performance of previously completed projects in the estimation process, i.e., a strategy based on

"looking back". Most software estimation processes may be a combination of these two strategies, but the weighting of them may differ based on the situation at hand. Overoptimism of time and effort is reported to increase with more use of a "looking forward" type instead of a "looking back" type of strategy (Buehler, Griffin et al. 1994; Buehler and Griffin 2003). Paradoxically, this suggests that careful planning can bias estimates even more towards overoptimism.

The initial time-frame in the alternative format is more or less arbitrary and does not necessarily follow the natural composition of a task. After one week (the increment in Studies 1 and 2) the software developer might be in the middle of developing a user story. Hence, it may be difficult to utilize knowledge about the effort spent of previously completed subtasks. The use of the alternative framing may therefore lead to less use of historical data. Less use of historical data means that estimates will be based on a "looking forward" type of strategy.

It may also be possible that the alternative format leads to more use of scenario-based strategies through this format's emphasize on how much it is possible to complete in X time. To answer this kind of question, the intuitive response may be to try to foresee what one would do the first day or the first hour, i.e., to apply a "looking forward" strategy. This does however not explain the difference we observed between the lower and higher anchors. But as mentioned in Section 5.2, a shorter time-frame might have brought the estimator even closer to details and a scenario-based representation of the task, thus outweighing historical, "looking back"-based information as the time-frame shortened.

6.4 Subadditivity

Unpacking a judgment into subcomponents can sometimes lead to what is labelled *subadditivity*, i.e., that the whole of something is less than the sum of its part. Tversky and Koehler (1994) found that the judged probability that a patient would die from natural causes was smaller than the sum of the probabilities that the patient would die from heart disease, cancer or other natural causes, which in turn was smaller than the summarized probability of an even finer subdivision (respiratory cancer, digestive cancer etc.). At first glance this does not look like a promising theoretical explanation for our finding, since dividing the task into an initial estimate and a remaining estimate actually caused lower estimates. Put differently, the sum of the parts was less than the whole. But in our case the first estimation request asked participants to predict the amount of work, not time. And the amount of work considered to be completed within the given time-frame was clearly larger than what we can infer from the estimates of the traditional format group.

In Study 4, the number of pages persons assumed they could write in half an hour (Md = 4) was not in proportion with the number of pages participants thought they would complete in two hours (Md = 8). An interval four times as big only doubled the number of pages the participants thought they could read. This resembles the subadditivity found in probability judgments. A small part of a total judgment can be disproportionate to a larger part. Subadditivity in probability judgments is explained in terms of "support" to a claim, i.e. to what extent a person can imagine the occurrence of a proposition. Things that are easy to find support for are rated as more probable. This would clearly not hold up as an explanation for our results, although picturing the work perhaps affect the estimates, as explained earlier.

The subadditivity in our study can perhaps be explained by the function in the Weber-Fechner law. Weber found that the threshold of discriminating changes in stimuli increased linearly as the stimuli increased in intensity. In other words, we are able to discriminate smaller changes in stimuli at lower intensities. It looks like there is a similar law at work in the more abstract domain of numbers, see for example (Dehaene, Dehaene-Lambertz et al. 1998). Neurophysiological research on on number neurons has also shown that the internal representation of numbers in primates follows a logarithmic function (Nieder and Miller 2003). An increase from 1 to 2 is larger than an increase from 4 to 5. Applied to our studies, the first two hours feels quite long, but four hours does not feel like the double. Likewise, the first 30 work-hours feel longer than the next 30 work-hours.

6.5 Limitations

Studies 1 and 2 include software professionals and realistic estimation work. The context is, however, artificial and the time for the estimation work shorter than in most realistic estimation situations. Although the estimates derived in the traditional format were similar to those produced by ordinary estimation processes, see Section 2.1.1, it may be that more time on the estimation work and higher importance of accurate estimates would have reduced the effect of the alternative format. If anchoring is an essential factor in the observed effect, we believe the results would generalize to field settings with more time spent on estimation work. The main evidence in support of this are the results presented in our recently completed study (Jørgensen and Grimstad 2008). In that study, we found that anchoring effects were present in field settings, although to a lesser extent than typically found in experiments. In more *ad hoc* estimation situations, e.g., when receiving a request from a client or project managers who want a quick answer on the deliveries next week, the differences in setting to our studies may be smaller. Follow-up studies in different types of field settings are however necessary to get more robust knowledge about the practical implications of our findings. There may, for example, be techniques applied by software professionals that reduce the effect of the estimation request format in field settings.

Our primary claim, that an alternative format of the estimation request tend to produce lower and presumably more overoptimistic predictions than a traditional request, is based on data from software professionals. Our second claim, that shorter time-frames produce more optimistic estimates, is however only tested in an estimation context different from software development effort estimation. We can not be sure that this would generalize to software development work. The similarity of the effect of the alternative format in different domains, however, suggests that the results could be valid for a range of effort or time estimation contexts.

7. Conclusion

We found that an initial estimation of output for a specified time (30 work-hours) produced lower estimates of the total effort than the traditional question of how many work hours that was needed to complete a software development project. Within a non-developmental context we found that estimating work output for shorter intervals resulted in lower estimates than estimating the work output for longer intervals. We provided several explanations for the finding, with a particular emphasis on explanations related to anchoring effects.

Regardless of the underlying mechanisms responsible for this effect, our results suggest that the alternative format should be avoided, especially when time-frames are short and in situations where there already is a tendency towards overoptimistic effort or time estimates. Instead of asking: "How much can you complete this week?" a better question will be "How much effort will it take to complete task X?" Our findings do, of course, not exclude that there are even better ways of formulating the estimation request. It may, for example, be even better to ask: "How much effort has similar tasks required previously?" or "How many user stories/use cases are this task?" with the follow-up "How many user stories/use cases do we normally produce per week?" This will stimulate a "looking back" rather than a "looking forward" estimation strategy, which is likely to lead to more realistic effort estimates

References:

Buehler, R. and D. Griffin (2003). "Planning, personality, and prediction: The role of future focus in optimistic time predictions." Organizational Behavior and Human Decision Processes 92: 80-90.

Buehler, R., D. Griffin and M. Ross (1994). "Exploring the "Planning fallacy": Why people underestimate their task completion times." Journal of Personality and Social Psychology 67(3): 366-381.

Dehaene, S., G. Dehaene-Lambertz and L. Cohen (1998). "Abstract representations of numbers in the animal and human brain." Trends in Neurosciences 21(8): 355-361.

Dolado, J. J. (2001). "On the problem of the software cost function." Information and Software Technology 43(1): 61-72.

Farr, L. (1964). Factors that affect the cost of computer programming v.1. L.G. Hanscom Field, Bedford, Massachusetts, united states air force Electronic Systems Division.

Gollwitzer, P. M. (1998). Action phases and mind-sets. Handbook of motivation and cognition: Foundation of social behavior. E. T. Higgins and R. M. Sorrentino. New York, Guilford. 2: 53-92.

Goodwin, P. (2000). "Improving the Voluntary Integration of Statistical Forecasts and Judgment." International Journal of Forecasting 16(1): 85-99.

Heemstra, F. J. and R. J. Kusters (1991). "Function point analysis: Evaluation of a software cost estimation model." European Journal of Information Systems 1(4): 223-237.

Hihn, J. and H. Habib-Agahi (1991). Cost estimation of software intensive projects: A survey of current practices. International Conference on Software Engineering, Austin, TX , USA, IEEE Comput. Soc. Press, Los Alamitos, CA, USA.

Hubbard, R. and D. E. Vetter (1996). "An empirical comparison of published replication research in accounting, economics, finance, management, and marketing." Journal of Business Research 35(2): 153-164.

Jorgensen, M., K. H. Teigen and K. Molokken (2004). "Better sure than safe? Over-confidence in judgement based software development effort prediction intervals." Journal of Systems and Software 70(1-2): 79-93.

Jørgensen, M. (2004). "Realism in assessment of effort estimation uncertainty: it matters how you ask." Software Engineering, IEEE Transactions on 30(4): 209-217.

Jørgensen, M. (2006). "The Effects of the Format of Software Project Bidding Processes." International Journal of Project Management 24(6): 522-528.

Jørgensen, M. (2007). "Estimation of Software Development Work Effort: Evidence on Expert Judgment and Formal Models." International Journal of Forecasting 23(3): 449-462.

Jørgensen, M. (2007). Individual Differences in How Much People are Affected by Irrelevant and Misleading Information. Second European Conference on Cognitive Science, Delphi, Greece, Hellenic Cognitive Science Society.

Jørgensen, M. and S. Grimstad (2008). "Avoiding Irrelevant and Misleading Information When Estimating Development Effort." IEEE Software May/June: 78-83.

Jørgensen, M. and S. Grimstad (2008). "The Impact of Irrelevant and Misleading Information on Software Development Effort Estimates: A Randomized Controlled Field Experiment." submitted.

Kahneman, D. and D. Lovallo (1993). "Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking." Management Science 39(1): 17-31.

Kahneman, D., P. Slovic and A. Tversky (1982). Judgment under uncertainty: Heuristics and biases. Cambridge, United Kingdom, Cambridge University Press.

Moløkken, K. and M. Jørgensen (2003). A review of software surveys on software effort estimation. International Symposium on Empirical Software Engineering, Rome, Italy, Simula Res. Lab. Lysaker Norway.

Moløkken, K. and M. Jørgensen (2005). "Expert Estimation of Web-Development Projects: Are Software Professionals in Technical Roles More Optimistic Than Those in Non-Technical Roles?" Empirical Software Engineering 10(1): 7-30.

Mussweiler, T. (2003). "Comparison Processes in Social Judgment: Mechanisms and Consequences." Psychological Review 110(3): 472-489.

Mussweiler, T. and F. Strack (2001). "The semantics of anchoring." Organizational Behaviour and Human Decision Processes 86(2): 234-255.

Nieder, A. and E. K. Miller (2003). "Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex." Neuron 37: 149-157.

Sanders, D. E. and L. P. Ritzman (1991). "On knowing when to switch from quantitative to judgemental forecasts." International Journal of Forecasting 11(6): 27 - 37.

Taylor, S. E. and P. M. Gollwitzer (1995). "The effects of mindset on positive illusions." Journal of Personality and Social Psychology 69: 213-226.

Tversky, A. and D. J. Koehler (1994). "Support Theory: A Nonextensional Representation of Subjective Probability." Psychological Review 101(4): 547-567.

Webby, R. G. and M. J. O'Connor (1996). "Judgemental and Statistical Time Series Forecasting: A Review of the Literature." International Journal of Forecasting 12(1): 91-118.

Appendix A: Requirements of the application RDinner

Monday through Sunday evenings, [company name removed] serves dinners to employees who work late. To ensure that the proper amount of food is ordered, the dinner administrators take reservations.

The current reservation system was constructed when dinner was served in only one building and few people were served. Since its creation, dinner service has expanded tremendously, number of locations has increased, and the current system is considered to lack functionality, flexibility and user friendliness.

A new system is needed. This system should be able to access through internet browsers, e.g., Internet Explorer, Opera, Firefox and Netscape. The system should enable reservations from at least 20 locations and at least 200 dinner reservations per location.

There are four distinct types of users for this system: diners, proxies, assistants and administrators:

- **Diners** are employees who make dinner reservations for themselves.
- **Proxies** are people who make dinner reservations on behalf of other diners, e.g., a secretary of a manager.
- **Assistants** manage the day-to-day operations of serving dinners, usually at a single location, including ordering food and keeping records of expenses.
- **Administrators** are responsible for managing dinner-related operations at all of the dinner locations, e.g., the fields to be used when reserving dinners.

The user stories (requirements) are:

U1: **Diners** should be able to reserve dinners.

U2: **Diners** should be able to manage their reservations.

U3: **Proxies** should be able to reserve on behalf of other employees.

U4: **Proxies** should be able to manage reservations on behalf of other employees.

U5: **Assistants** should be able to read lists of reservations.

U6: **Assistants** should be able to print out lists of reservations.

U7: **Administrators** should be able to read new dinner serving parameters.

U8: **Administrators** should be able to update dinner serving parameters.