The Impact of Irrelevant and Misleading Information on Software Development Effort Estimates: A Randomized Controlled Field Experiment

Magne Jørgensen[1,2], Stein Grimstad[1]
[1]Simula Research Laboratory & [2]University of Oslo

Abstract: *Several studies have reported that software development effort estimates can be strongly affected by effort-irrelevant and misleading information without the estimators being aware of this effect. These studies were conducted in laboratory (artificial) estimation contexts. To increase our knowledge about the importance of these effects in field settings, we paid 46 outsourcing companies from Eastern European and East Asian countries to estimate the required effort of the same five software development projects. The companies were allocated randomly to either the original requirement specification or a manipulated version of the original requirement specification. The manipulations were as follows: i) reduced length of requirement specification with no change of content, ii) information about the low effort spent on the development of the old system to be replaced, iii) information about the client's unrealistic expectations about low cost, and iv) a restriction of a short development period with start up a few months ahead (which should, rationally speaking, lead to an increase in effort). All manipulations led to decreased median effort estimates, but only manipulation iv) led to a large, statistically significant decrease. A comparison of the effects of similar types of irrelevant and misleading information in laboratory and field settings suggests that the effect of manipulations i), ii) and iii) where much lower in field settings than in laboratory settings, while the effect of manipulation iv) was almost at the same level. We conclude that the tendency towards a smaller effect in field settings means that laboratory studies are frequently only useful for demonstrating the existence of a software engineering phenomenon, or for understanding it better, and that we need field studies to analyze its importance.*

1. Introduction

The use of unconscious processes of judgment is essential in most software development effort estimation, either as part of expert judgment-based estimation, so-called "expert estimation", or as part of providing input to estimation models (Jørgensen 2004). It is known that these unconscious judgmental processes can be affected, not just by relevant information, but also by information that is irrelevant. Software professionals' effort estimates can, among other things, be strongly influenced by:

- the client's expectation of the effort that will be expended, even when informed that this expectation is not based on any knowledge of the required software development effort (Jørgensen and Sjøberg 2001; Jørgensen and Sjøberg 2004; Aranda and Easterbroook 2005),
- variation in the wording of the requirements that is irrelevant to the effort, such as describing a maintenance task as "development of new functionality" as opposed to "an extension" (Jørgensen and Grimstad 2008),
- information that induces wishful thinking, such as information about future opportunities that will be offered, assuming that high productivity is achieved on the project to be estimated (Jørgensen and Grimstad 2008), and,
- increased amount of information, even when this information has nothing to do with the effort required (Grimstad and Jørgensen 2007).

In several of the above studies, independent software professionals established that the information was irrelevant for actual effort usage and/or the estimators were explicitly instructed *not* to use that information in their estimation work. In (Jørgensen and Sjøberg 2001), we asked the software professionals whether or not they had used the client's effort expectations as input to their estimates. The software professionals claimed that they did not use this information at all or that it had only a minor impact on their estimates. However, the measured impact of the information was very large. This supports the claim that essential steps

in judgmental effort estimation are based on unconscious processes. We discuss the unconscious steps of judgment-based effort estimation in (Jørgensen 2005). Similar impacts on judgment from clearly irrelevant information have been found in numerous other professions, e.g., among professionals who are involved in judicial decision making (Englich, Mussweiler et al. 2006) and property pricing decisions (Northcraft and Neale 1987). Software professionals clearly do not differ from other people with respect to the extent to which they are influenced by irrelevant information.

The effect that irrelevant and misleading information has on software professionals can have unfortunate consequences. It can, for example, lead to effort estimates that are too low, and hence to loss-inducing bids, project management problems, and low client satisfaction. Therefore, better knowledge about what type of, when, and how irrelevant and misleading information affects effort estimates is essential if software effort estimation is to become more accurate.

In this paper, we use the term 'estimation-irrelevant information' to denote information that is not causally related to the actual use of effort and consequently should not influence the estimate. This information is frequently relevant for other purposes, e.g., bidding or planning, and may well correlate with the actual use of effort. For example, it is likely that the clients' expectations about cost or the length of a requirement specification correlate with the actual use of effort, although the clients' expectations of low cost or a short requirement specification do not themselves act causally to reduce the actual use of effort. We do not discount the possibility that there are real-world estimation situations in which the information that we term estimation-irrelevant is relevant for the actual use of effort, e.g., situations in which an expectation that the cost will be low implicitly say something about a client's expectations about the quality of the software. Consequently, the actual relevance of information may be difficult to determine without knowledge of the intentions of whoever is providing the information.

Previous research on the effect of irrelevant information on effort estimates does not include field studies. The fact that this is so provided an important motivation for the field study described in this paper. While it may be reasonable to generalize, to a certain extent, from the results of previous studies to some types of real-life estimation situations, e.g., small maintenance tasks and early project effort estimates that are based on limited information, laboratory-based results are not necessarily relevant for field settings of the type in which carefully selected experts spend several work-days estimating a project.. For example, it is possible (a) that the high time pressure on estimation in the laboratory-based studies increased the use of surface indicators that usually correlate with use of effort, e.g., the number of pages in the specification, and (b) that spending more time on the estimation work would lead to the use of a greater number of causal variables. The use of a greater number of causal variables may, in turn, make software professionals better able to resist the influence of irrelevant information. In short, field studies may be required to assess the effect sizes of irrelevant and misleading information in typical effort-estimation field settings.

The remainder of the paper is organized as follows: Section 2 describes the design of the field study. Section 3 describes the results. Section 4 discusses the results and the limitations of the study. Section 5 concludes.

2. Design of the Study

2.1 Research Questions, Motivation, and Hypotheses

Table 1 presents our research questions, the motivations for the research questions, and the hypotheses. The context of all research questions and hypotheses is the estimation of software projects in field settings.

Table 1: Research Questions, Motivations and Hypotheses

Research Questions	Motivation	Hypotheses
RQ1: Does the number of pages in a requirement specification affect the effort estimate of	In a student experiment (*unpublished, contact the authors for the data set and the experimental material*) we found that a decrease in the length of a requirement specification from seven pages to one, brought	**H1:** A reduction in the number of pages in the requirement specification leads to lower effort

software projects?	about by such measures as reducing the font size and the wider margins, without changing the content, resulted in a decrease in the mean effort estimate of about 30%. The students spent less than one work-hour estimating the project, the information about the project was limited, and the students were not experienced estimators. Consequently, it is of interest to study the degree to which spending more time on the estimation work, having more information available, and having more estimation experience, would reduce the influence of the estimation-irrelevant surface indicator "length of specification".	estimates, even when the written content is exactly the same.
RQ2: When and how much do numerical values (anchors) presented early in the requirement specification that are irrelevant to the actual use of effort affect the estimated effort of software projects?	Numerous laboratory experiments have demonstrated that almost any number presented early in the context of human judgment can work as an estimation anchor; see (Englich, Mussweiler et al. 2006) for an overview. However, there is a need to study the effect size and robustness of these findings in field settings, when more time is spent on the estimation work and when the estimators have a higher level of experience than in the previous studies. The two types of irrelevant numerical values examined in our study were i) the effort used to develop the system to be replaced, and ii) the client's unrealistic expectations of very low cost that were not based on knowledge of the required effort. The effect sizes of this type of estimation anchor in laboratory settings vary a great deal. Unrealistic client expectations typically have the largest effect sizes and are reported to decrease the median estimate by 40% or more; see (Jørgensen and Sjøberg 2004; Jørgensen and Grimstad 2008).	**H2a:** Information about the actual, very low effort used to develop the system to be replaced leads to lower effort estimates. **H2b:** Information about the client's unrealistic expectations of very low cost that are not based on knowledge of the required effort leads to lower effort estimates.
RQ3: Does less time available for software development work (shorter development period) result in lower effort estimates?	Previous studies, e.g., (Edwards and Moores 1994; Goodwin 1998), suggest that effort estimation is frequently based on an unfortunate mix of project management concerns (time available) and realistic use of effort (effort estimate). Having less time available should, rationally speaking, seldom lead to less effort spent. The opposite may more frequently be the case. For this reason, estimation models such as COCOMO (Boehm, Abts et al. 2000) and SLIM (Putnam 1978) add effort when there is a schedule compression. In a laboratory experiment (*unpublished, contact the authors for the data set and the experimental material*) we found a 40% decrease in median estimates when the software professionals were informed that the software development should be completed within a two-week period, six months from the time of estimation.	**H3:** The information that the client requires the software to be developed in a short period of time leads to lower effort estimates.

It is also of interest to determine whether the observed effect of irrelevant and misleading information is a result of higher degree of optimism, or a simplification/quality reduction of the planned software product or process. We therefore decided to collect information that would enable this type of determination, e.g., a description of the architecture and data on the effort expended on quality assurance and testing tasks, when we observed that the irrelevant and misleading information had had a large effect.

2.2 Participants

In the second half of 2007, we contacted 240 software companies in Eastern Europe and Asia by email and invited them to estimate five software projects under ordinary payment conditions. The invitation, including the estimation instructions, is provided below as Appendix A. Forty-six companies from various countries agreed to complete the estimation work: Russia (15 companies), Ukraine (5), India (7), Pakistan (5), Bulgaria (4), Romania (3), Belarus (2), Moldovia (1), Poland (1), Serbia (1), Slovakia (1), and Vietnam (1). The companies varied in size from small (fewer than 10 employees) to very large (>1000 employees). All responding companies sent the CVs of the software professionals that would be in charge of the estimation work to document sufficient experience. We accepted only estimators with professional experience from projects similar to those to be estimated, i.e., we allowed only reasonably experienced estimators.

As can be seen from the instructions in Appendix A, we did not inform the companies that their estimates would be used in a research study. This was because we wanted our request to be treated as ordinary estimation work, which it was from the viewpoint of the software companies. We informed the companies that if they did high-quality estimation work for us, we may offer them other opportunities, such as more estimation work. We have already hired for further work several of the companies whose estimation work we assessed be of high quality. The companies were given all the information about the estimation work before they accepted it and were paid a fee that reflected their work effort and that was agreed upon by both parties. None of the individual companies can be identified by the reported data and as far as we can see, the companies could do nothing but benefit from having participated in the study. The companies not only delivered an effort estimate, but also provided descriptions of the proposed architecture, the development platform, essential estimation assumptions, a work break-down of the project, an assessment of the uncertainty of the estimate, and a description of the estimation process they used. The effort spent by the companies on the estimation and estimation-related work varied from about 40 to 100 work-hours.

2.3 Material

The estimation material consisted of five requirement specifications and the estimation instruction in Appendix A. The first requirement specification (S1) was the same for all companies, while the remaining four (S2, S3, S4, and S5) had one original version and one version manipulated with estimation-irrelevant information to test hypotheses H1-H3. S3, S4, and S5 were real-life specifications that had already been implemented by software professionals, while S1 and S2 were specifications developed by students while they were participating in a course on requirements engineering at the San Diego Software Studio. The actual effort when implementing S3 was approximately 800 work-hours. We have no access to the actual effort of S4. The system specified in S5 had been implemented by four different companies. These companies used between approximately 300 and 900 work-hours; see (Jørgensen and Carelius 2004) for more information about the implementations of this specification. A brief description of the requirement specifications and the manipulations that were introduced is included in Table 2. The main criteria used when selecting the specifications were that the estimation work on each specification should not take more than two days, the specifications should describe web applications of a type such that most of the invited companies would have extensive experience, and the specification should be as complete and precise as other specifications that are typical for its type.

Table 2: The Five Requirement Specifications

System ID	System domain	Specification length	Specification Manipulation

S1	Dinner reservation.	5 pages	None. All companies received the same specification. This specification was included to enable a control of similarities of groups with different treatments.
S2	Management of appointments at a doctor's office.	12 pages (3 pages)	The text of the two versions of the requirement specification was identical. However, the manipulated requirements specification was reduced to three pages by using smaller fonts, wider margins, and more compressed text. The original version was 12 pages. Test of H1.
S3	Information site for members of an organization.	4 pages	The following text was included early in the manipulated requirement specifications: "*The previous website required about 25 work-hours to build. The estimate of the effort most likely required to build the new website should, however, be independent of this previous use of effort, as the old and the new systems are quite different in size, quality and complexity*." 25 work-hours is a very low value that had nothing to do with the current system. Test of H2a.
S4	An inventory and work order management system for engineering work.	19 pages (5 pages)	The following text was included early in the manipulated requirement specifications: "*The preliminary budget of the new system is $10 000 [corresponding to about 100 work-hours with typical pricing in the country in which it will be built]. The preliminary budget is not built on any knowledge about the actual cost of developing the new system, and will, if needed, be extended to cover the expenses necessary to build a quality system with the desired functionality*." One hundred work-hours is a very low value for this project and the companies were instructed to not use this as input to their effort estimate. The manipulated requirement specification also had fewer pages (five) compared to the original version (19 pages), but the content was identical. The purpose of combining two manipulations was to try to induce a larger effect. However, we have also tried to use the analysis of the estimates of S4 to test H2b.
S5	Database for information about research studied, integrated with an existing website.	11 pages	The following text was included early in the manipulated requirement specifications: "[the client] *expects that the system development starts February 3, 2008 and can be launched on February 23, 2008. This three-week period should include all development and testing.*" A short development period should, as discussed earlier, lead to the use of the same or more effort, not less. Test of H3.

2.4 Data Collection Process

The data was collected as follows:

a) First, we searched the internet for software companies, using free text search on the terms "outsourcing" and "software development. We selected software companies on the basis of the results of this search, inviting only companies whose websites listed relevant experience. The invitation, which we sent by email, included the five requirement specifications and the instructions for performing the estimation work that

are included as Appendix A of this paper. The invited companies were encouraged to ask for more information about the work, if needed. For each of the four requirement specification S2, S3, S4, and S5, a company was randomly allocated to either the original or the manipulated version.

b) The companies that were interested in the estimation work responded by email and sent us the curriculum vitae of the person(s) supposed to be in charge of the estimation work, the required price, and the date by which they would complete the estimation work. We accepted or rejected the offer using this information as the basis for our decision.

c) Then, the companies completed the estimation work for the five projects, in the sequence they wanted, typically within the following two weeks. They sent us questions by email when they needed clarification about the requirements. We tried to respond as similarly as possible to all companies and the responses were made without our knowing whether the company had received the original or the manipulated version of a requirement specification ("blind" responses). The companies sent us their estimation work for approval once they had completed it. Most companies needed one or two revisions before we approved the estimation work. The criteria for approval are described in Appendix A. About 2,000 emails were sent and received during the data collection process.

2.5 Issues of Analysis Considered

The results presented in Section 3 were arrived at in light of the following issues pertaining to the analysis:

- As found in previous studies, e.g., (Jørgensen and Carelius 2004), a high variation in effort estimates and a few unusual values are to be expected. This means that the mean values are not very reliable as the central value of the distributions. Hence, we based the analysis of differences on the median, rather than the mean, effort estimate of a treatment group.
- We included, but did not emphasize, statistical tests of significance of difference of median values in our analyses. The main reason for this is that we believe studies that replicate and extend previous results should be analyzed mainly in the light of previous knowledge and not as isolated observations, as is the case in statistical hypothesis testing based on a data set. For example, even small, non-significant differences in the expected directions provide empirical support for previous findings. In addition, the four manipulations of requirements that we analyzed in our study are all analyses of the impact of estimation-irrelevant information and can be interpreted as a family of experiments. If all effects are in the expected direction, the individual effects found in a family of studies support each other to some extent.

3. Results

3.1 Distributions of Effort Estimates

The effort estimates varied a great deal from company to company, as expected. Table 3 shows the distributions of estimates for the five projects. Q1 is the first quartile. Twenty-five per cent of the values will be lower than or equal to this value. Q3 is the third quartile. Seventy-five per cent of the values will be lower than or equal to this value. Differences in effort estimates are the results of differences in productivity, differences in use of the development environment, differences in properties of the produced software, differences in levels of experience and optimism. We have repeatedly found differences in effort estimates in field settings of the size displayed in Table 3; see, for example (Jørgensen and Carelius 2004). We had intended to use the performance of S1 to determine whether there were substantial differences with respect to performance in the groups that received the original and the manipulated version of S2-S5. Unfortunately, the differences in the sequence in which the companies estimated the projects made this analysis difficult. The reason for this was that when a company estimated S1 after estimating one of the other projects, the estimators were influenced by the estimate that they made of the project that they had estimated immediately before. This means that those that received the manipulated version of the project that they had

estimated immediately before were more likely to provide lower estimates for S1. Consequently, it would not be possible to determine the degree to which observed lower estimates of S1 in one of the groups would point to an actual difference between the groups, or to the effect of the sequence. Analyzing only the estimates of those companies that started with S1 was not an option, because only seven of the companies did this. The effect of the sequence does not constitute any serious threat to the testing of our hypotheses, because the sequence chosen by the companies is not likely to be affected by their allocation to the group with the manipulated or the original specification.

Table 3: Effort Estimation Distributions

System	Minimum	Q1	Median	Q3	Maximum
S1	45	132	193	339	1320
S2	61	186	330	438	1200
S3	160	316	545	729	2280
S4	240	681	895	1316	3371
S5	18	134	190	347	1160

3.2 Hypothesis 1: Length of Specification

The data gives no or, at best, very weak support for an effect of the length of the specification on the effort estimates of S2 in field settings; see Figure 1. Although the median value (the line inside the grey box) is a little bit lower, the Q3 value is actually higher for the three-page specification group. Table 4 displays the median values for both groups and shows an 11% decrease with shorter specification length, but this median value is not very robust. A one-sided Kruskal-Wallis test of difference in median values gives the p-value 0.5.

Figure 1: Boxplot of Estimates of S2

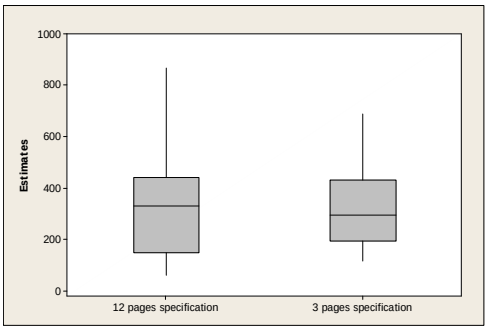

Table 4: Median Estimates of S2

Group	Median Effort Estimate
12-page specification	330 work-hours (n=22)
Three-page specification	295 work-hours (n=24)

3.3 Numerical Anchors

3.3.1 Hypothesis 2a: Information about the System to be Replaced

The data gives only weak support for an effect of information about the effort of the previous system (the low numerical anchor) on the effort estimate of S3; see Figure 2. Table 5 shows the median values for both groups and shows a 15% decrease when exposed to a low anchor. The actual effort when developing the system was about 800 work-hours. It is thus

evident that most estimates seem to be optimistic. A one-sided Kruskal-Wallis test of difference in median values gives the p-value 0.3.

Figure 2: Boxplot of Estimates of S3

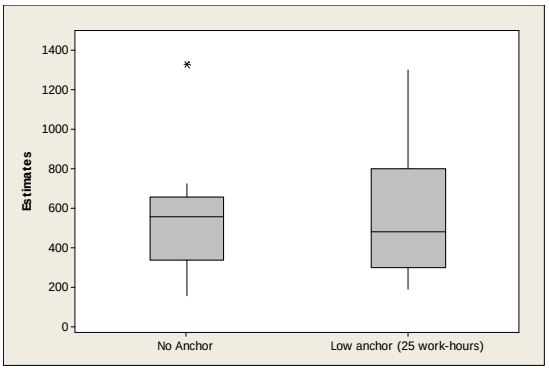

Table 5: Median Estimates of S3

Group	Median Effort Estimate
No anchor	568 work-hours (n=19)
Informed that the previous system required about 25 work-hours to build	481 work-hours (n=27)

3.3.2 Hypothesis 2b: Unrealistic client budget

As described in Section 3.2, we found that a reduced length of number of pages in a specification had, if anything, only a small effect on estimation. We decided therefore to use the manipulation of S4 as a test of the impact that an unrealistic client budget had on the estimates, i.e., we assumed that any observed effect would be due mainly to the client's expectation and not the length of the specification. This expectation on the part of the client was introduced, as described earlier, by providing the information that the client had a preliminary budget representing 100 hours of development work, and that "*The preliminary budget is not built on any knowledge about the actual cost of developing the new system, and will, if needed, be extended to cover the expenses necessary to build a quality system with the desired functionality.*"

The data provide some support for an effect from the above manipulation; see Figure 3. Table 6 shows the median values for both groups and shows that the companies' estimates were 10% lower when they were exposed to the manipulated requirement specification. A one-sided Kruskal-Wallis test of difference in median values gives the p-value 0.1. The information displayed in the boxplot in Figure 3 suggests that the median value does not give the full picture of the impact of the manipulation. When selecting a company on the basis of a bidding round, those with estimates that are most strongly affected by the low numerical anchor are more likely to be chosen as a provider, i.e., clients are likely to choose providers among those with the lowest bids. Figure 3 suggests that many more companies in the group that received the manipulated specification provided very low estimates than in the group that did not receive it.

Figure 3: Boxplot of Estimates of S4

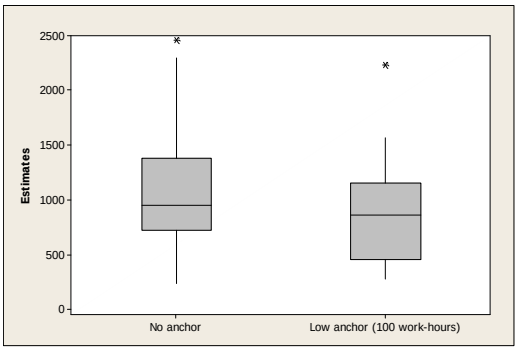

Table 5: Median Estimates of S4

Group	Median Effort Estimate
No anchor	954 work-hours (n=23)
Unrealistic client expectation about low cost	860 work-hours (n=23)

3.4 Hypothesis 3: Short Development Time

The data provide strong support for an effect of the three-week development time with startup several months ahead on the effort estimate of S5; see Figure 4. Table 6 shows the median estimates for both groups. The median estimates was 34% lower for companies that were exposed to the information that the client expects the system to be developed during a three-week period several months ahead. The actual effort of the four companies that previously developed this system was from about 300 to 900 work-hours. Thus, most of the estimates provided by the companies in our study seem to be optimistic. A one-sided Kruskal-Wallis test of difference in median values gives the p-value 0.02.

Figure 4: Boxplot of Estimates of S5

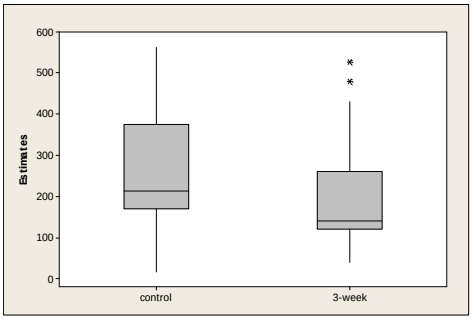

Table 6: The Effect of Information about Short Development Time

Group	Median Effort Estimate
No information about development period (control)	209 work-hours (n=21)
Informed that the client expected the system to be developed during a three-week period	142 work-hours (n=24)

Above, we argued that it is rational to think that a short development period should lead to more, rather than less, use of effort. Yet it may be argued that the information about the schedule is relevant to effort in the sense that a company may want to simplify the solution to make it possible to develop within a short time frame. However, contrary to this, a follow-up analysis of the proposed solutions and the estimation work suggest that it is not likely that the reduced estimates were caused by a simplification of the solutions:

- The specification of S5 was reasonably precise and did not allow much simplification of functionality. (The specification will be sent to interested readers upon request to one of the authors.)
- The average complexity of each of the proposed solutions was assessed by a person with long professional experience as senior architect (the second author of this paper) to be about the same in the two groups. This analysis was "blind", i.e., not based on any knowledge about the group membership.

The companies in the three-week groups may have decided to reduce the quality of the proposed system. An analysis of the proportion of effort that it was estimated would be spent on testing and quality assurance suggests that this could, to some extent, have been the case. The median proportion of effort that it was estimated would be spent on testing and quality assurance was 13% in the three-week group and 21% in the control group. The proportion of effort that it was estimated would be spent on other types of activity, e.g., project management and design, did not vary much. However, the reduced emphasis on testing and quality assurance is not sufficient to explain more than a small part of the large difference in total effort estimates. Adjusting for the difference in testing effort, the median effort estimates of those companies that received the manipulated version of the specification would only be about 20 work-hours higher.

The increased optimism that is found in response to being informed that the development period will be short may be due to the fact that it is easy for companies to think of a 3-week development time as indicating a "small" project and hence to select reference activities (similar activities that had been completed previously) that correspond with this in use of effort. This bias in the selection of reference activities is in accordance with the "selective accessibility" model described in (Mussweiler 2003) and discussed in Section 4. Another possible explanation for the increased optimism is that those in the three-week group were told that the development work should start several months ahead, while those in the control group did not receive this information and may have assumed that the development work was supposed to start very soon. It has been found that people may be more optimistic and confident when they are thinking about activities far ahead, rather than about activities that are supposed to be started very soon (Gilovich, Kerr et al. 1993). We intend to conduct more studies to determine what the explanation is for the increased optimism in this case.

4. Discussion of Results

4.1 Comparison of Laboratory and Field Effect Sizes

Table 7 compares the effect sizes in our study with those of corresponding laboratory studies. The comparison suggests that the effect sizes are typically smaller in field settings than in laboratory studies, sometimes much smaller. Note that this comparison should be interpreted with great care and only as a weak indication. There may, for example, be important differences between the laboratory and field studies that may affect the interpretation of the results related to differences in the size of the estimation tasks and the differences in the population of developers studied.

Table 7: Comparison of Effect sizes of Similar Studies in Laboratory and Field Settings

Treatment	Laboratory settings	Field settings
Shorter specification with same content	30% decrease in median estimates of student's effort estimates (see Table 1)	11% decrease in median estimates of S2 (see Section 3.2)
Low numerical anchors related to effort of system to be replaced or clients cost expectations.	40% or greater decrease in median estimates in experiments using software professionals that have anchors based on unrealistic client expectations (see Table 1).	15% and 10% decrease in median estimates of S3 and S4, respectively (see Section 3.3)
Short development time, combined with start up several months ahead.	40% decrease in median effort (see Table 1).	34% decrease in median estimates of S5 (see Section 3.4)

4.2 A Model of Judgment-based Effort Estimation

It may be useful for the improvement of judgment-based effort estimation methods if we could acquire a better understanding of when and why we should expect irrelevant and misleading information in field settings to affect effort estimation. For this purpose, we explain and apply the model of human judgment presented in (Mussweiler 2003).

A basic assumption of the model of human judgment presented by Mussweiler in (Mussweiler 2003) is that human judgment operates by making comparisons. In our context, this means that effort estimation is based on comparisons of the project or task to be estimated (the target) with one or more references, e.g., similar projects or tasks. Adapting the model presented by Mussweiler to software effort estimation contexts, we believe that the essential steps involved in this comparison-based judgment include the following:

1) Selection of a reference for comparison. In software development effort estimation, the references involved in the estimation of a software project can be of different types, depending on the estimation process. When the estimation method is bottom-up, references at the level of activities may dominate the estimation work, while in top-down estimation, the reference may be one or more similar projects. When providing judgment-based input to an estimation model, e.g., when assessing the experience level of the developers or the complexity of the product, a whole range of references may be accessed. Guiding principles for the selection of references are (a) similarity between the reference and target, and (b) the ease with which the reference can be accessed mentally.
2) Comparison of the target and reference tasks by searching for information that supports similarity or dissimilarity between the target and reference tasks (similarity/dissimilarity testing). An important, well-documented, assumption of the model is that once it has been decided what hypothesis to test, people typically focus on confirming rather than disconfirming evidence (Brehmer 1980; Koehler 1991; Trope and Liberman 1996). Similarity testing, rather than dissimilarity testing, may be the dominant mode in most judgmental contexts (Mussweiler 2003).
3) Synthesis of the collected information about the target and reference tasks in order to estimate the effort that will be required. In (Jørgensen 2005), we present a model for the synthesis of this information in the context of effort estimation that is based on closeness of analogies and level of uncertainty in the closeness assessment. We also provide empirical evidence in support of the model.

More elements are needed if the model is to be used to explain the influence of irrelevant information on the effort estimates. The following well-documented model elements, mainly belonging to step 1), may be particularly useful for understanding the effects of irrelevant information that we observed in our study:

- *Conversational inference*: A typical, unconscious human response is that information provided in conversational contexts, such as the communication from a client to the provider by means of requirement specifications, is considered to be relevant simply because "*it's there*". Consequently, all types of information that is presented may influence the selection of references, even when it is not relevant for the judgmental task (Grice 1975; Schwarz 1994; Cowley 2006). For example, when the client describes the effort of the previous system in the requirement specification, the software professionals may immediately think that the information is relevant because it has been included in a relevant conversation. Even when they discover that the information is irrelevant for this system, it may not be possible for the professionals to recover an unaffected state of mind. We have documented this problem of returning to an unaffected state in (Jørgensen and Grimstad 2008).
- *Memory accessibility*: Elements that have nothing to do with the use of software development effort may make different references more accessible. For example, in the study described in (Mussweiler and Strack 2000), a high price anchor, which was induced by asking the participants whether the average price for a new car is higher or lower than 40,000 German marks (where 40,000 German marks at

that time was significantly higher than the actual average price), made it easier to recall expensive cars than less expensive cars. Similarly, estimation-irrelevant information, such as a low, preliminary budget may make it easier for the software professionals to recall and use for comparison easy and inexpensive development tasks or projects.

- *Wishful thinking*: The selection of reference projects or tasks may also be influenced by wishful thinking. If the client has indicated that his budget is low, it is not very comforting to suggest an effort estimate that deviates very much from that budget, even when the budget is unrealistic and the software professionals have been informed that it will be extended if necessary. A high estimate may for example make the client believe that the reason for this high estimate is low competence. In order to preserve a good image of one's own abilities, a software professional may select reference projects and tasks in accordance with an unconscious intention to avoid negative responses to the estimate on the part of the client. For example, the selection of small and simple reference projects in the S5 project may be a result of the wish to complete the project within the three-week period using only one developer.
- *Categorical references and surface features:* In estimation settings in which little information is available, the estimators have little experience, and/or there is high time pressure (as is the case in many laboratory studies), categorical references and surface features may increase in relevance for the estimation work. This means that information that typically correlates with, but is not causally related to, use of effort is more likely into influence the effort estimate in laboratory settings. For example, if the time that is allocated for the estimation work does not allow the project to be broken down into activities and the estimator do not recall the completion effort of similar projects, it is understandable that the estimators may be influenced by their knowledge that, for example, medium-sized projects typically have requirement specifications that are the same length the project that they are estimating. It is understandable in light of both the availability of the feature in both the reference projects and the project that they are estimating, and the ease with which the feature similarity can be evaluated.

The above model elements explain why irrelevant information may affect the effort estimates. The model also explains why the effect of irrelevant information may be "diluted" in field contexts, in which there is less time pressure, greater expertise, and a greater amount of relevant information about the task compared with laboratory contexts. Less time pressure, greater expertise, and a greater amount of relevant information about the task means that a single piece of irrelevant, misleading information most likely receives less attention and is, for example, less likely to lead to greater accessibility of conversationally induced references. The model also explains why more time, greater expertise, and a greater amount of relevant information do not guarantee a small effect from irrelevant or misleading information. For example, if information about an expectation of low cost on the part of the client has led to the selection of small reference projects, the use of more time on the estimation work may lead to higher confidence instead of more realistic estimates due to search for confirming rather than disconfirming evidence.

4.3 Limitations of the Study

The high number of estimation experiments and experiments in other domains makes it safe to claim that it is easy to create situations in which irrelevant information affects human judgment strongly. That being so, the issue that we address is not whether or not estimation-irrelevant information is able to influence effort estimates, but the degree to which this is likely to occur in field settings where a software company allocates experienced software professionals to estimation work that lasts for several days. The results of our study indicate that the effects may still be present, although typically not as large as in estimation situations in which time pressure is greater, there is less information about the project to be develped, and it is less important to deliver an accurate estimate, i.e., typical laboratory study settings. However, our study has a number of limitations that mean that the results should be interpreted with care.

Pricing or estimating? We repeatedly instructed the organizations to estimate the most likely use of effort to the client, and not the price. In order to emphasize further that we did not

wish to receive a price, we told them that they would not be chosen to develop the projects that they estimated. This was done to avoid a pricing process in which a preliminary budget may be defended as relevant input. However, we observed that despite our instructions, six of the companies were poor at separating these two concerns. This is consistent with the previous observations of a conflation of concerns about pricing, planning, and estimation among software companies that is reported in (Jørgensen and Sjøberg 2001). It is possible to defend the use of some of the effort irrelevant information if a company conflates concerns about estimation and pricing in its estimation work. For example, a low preliminary budget may be interpreted by the estimator as an indication that low cost is a high priority for client. While we, as clients, intended no such thing, and while we instructed the companies to ignore such information, we cannot fully exclude the possibility that more of the companies conflated pricing and estimation. We conducted the analyses without the six companies that seemed to have provided a price rather than estimated the effort and found that for all systems except S5 (the schedule compression manipulation) this resulted in the manipulations having slightly less effect. This weakly indicates that a pricing process may be even more vulnerable to the influence of irrelevant information than processes that emphasize the estimation of most likely effort.

Unusual estimation situation? The typical estimation situation of software development companies is to estimate the effort of projects for bidding, budgeting, or planning purposes. In our study, they provided only the estimate of most likely effort, but they had no opportunity to develop the project. On the one hand, this may have led to less "wishful thinking" (and consequently the irrelevant information may have had less effect!), but on the other hand, may have resulted in the use of estimation processes different from those that are usually followed. Fortunately, an analysis of the estimation processes used by the companies suggests that the processes were no different from those reported to be in typical use in the software industry; see, e.g., (Heemstra and Kusters 1991; Hihn and Habib-Agahi 1991). As an illustration, 41 companies applied a bottom-up, expert judgment-based effort estimation process and five used formal models for effort estimation. There were some variations in the use of judgment-based estimation processes. For example, two companies combined expert judgment with simple rules of thumbs and five applied the PERT method of minimum-maximum effort estimates. All companies produced proper structures for activities, which was a minimum requirement on our part for accepting the estimation work. As stated earlier, we informed the companies that they would be evaluated, and would be more likely to get more paid work from us, on the basis of the quality of the estimation work, not on the effort estimate itself. Note that our emphasis on the quality of the estimation work in the instruction, and not for example the price, is likely to result in irrelevant information having less impact than in bidding processes, due to greater use of deep-structure indicators than surface indicators as input to the estimation process.

Biased selection of companies? Outsourcing companies in low-cost countries, such as India and Russia, were selected for this study because they were less expensive than, for example, Norwegian or North-American companies. Within the segment of low-cost outsourcing companies, it is possible that those that accepted the work and payment conditions (about 20% of the invited companies) were not representative of companies in those countries or of companies in other regions. They may, for example, spend less time on quality assurance and have less expertise and, as a consequence, be affected by the irrelevant information more easily. Unfortunately, there are no data available to show whether or not there is a bias in one direction or the other and we cannot exclude this possibility. However, due to the fact that we accepted only estimators who had relevant experience and emphasized that we would evaluate the quality of the estimation process ,it is also possible that the estimation processes and skills were better than average. Without better knowledge about the world-wide population of software developers and companies, it is difficult to know to what degree a non-random sample of companies represents the underlying population. So far, we have seen no survey of software effort estimation that selected software companies by random sampling, other than in national contexts.

Low commitment? The software professionals were paid for their work and they knew that high-quality estimation work might result in their receiving more work from us as clients. However, it is nevertheless possible that some of the companies were not committed strongly to delivering proper estimation work, because they knew that they would not have the opportunity to develop the projects that they estimated. However, our observations led us to interpret only

one company as not aiming at serious estimation work and we excluded that company from the analysis. The quality of the estimation work was increased by our own quality assurance and the instructions to describe estimation assumptions, technical architecture, development environment, a work break-down structure, and the estimation process that they used.

5. Conclusion

Previous studies on the effect of estimation-irrelevant information on the effort estimates have systematically shown large effects. Our study is however the first to investigate this topic in a controlled field setting. We find that the field setting typically led to irrelevant information having smaller effects than in more artificial experimental settings, which typically have greater time pressure to estimate and have less information available. However, we also show that the effects in field setting can nevertheless be substantial. In particular, we found that a group that received the information that the system had to be developed in a three-week period starting a specified date several months ahead in time, produced effort estimates that were much lower than those produced by a control group that did not receive this information. In addition, the information about an absurdly low preliminary budget seems to have influenced the effort estimates, despite the fact that we told the companies that the budget would be increased if necessary.

In (Jørgensen and Grimstad 2008), we report that the only robust way to avoid effects generated by irrelevant and misleading information seems to be to remove it from the requirement specification or neutralize it before the requirement specification is handed over to the software professionals completing the estimation work. It is very difficult, perhaps impossible, to return to an unaffected state once one has been exposed to irrelevant, potentially misleading information.

Acknowledgement: Thanks to all people who contributed to this study, in particular, Wenche Angel, Hans Christian Benestad, Eric Busboom, Mike Cohn, Tanja Gruschke, Eric Busboom, Hege Johnsrud, and Dag Sjøberg.

References:

Aranda, J. and S. M. Easterbroook (2005). Anchoring and Adjustment in Software Estimation. European Software Engineering Conference / ACM SIGSOFT Symposium on the Foundations of Software Engineering (ESEC/FSE'05), Lisbon, Portugal.

Boehm, B., C. Abts, A. W. Brown, S. Chulani, B. K. Clark, E. Horowitz, R. Madachy, D. Reifer and B. Steece (2000). Software cost estimation with Cocomo II. New Jersey, Prentice-Hall.

Brehmer, B. (1980). "In one word: Not from experience." Acta Psychologica **45**: 223-241.

Cowley, E. (2006). "Processing exaggerated advertising claims." Journal of Business Research **59**: 728-734.

Edwards, J. S. and T. T. Moores (1994). "A conflict between the use of estimating and planning tools in the management of information systems." European Journal of Information Systems **3**(2): 139-147.

Englich, B., T. Mussweiler and F. Strack (2006). "Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making." Personality and Social Psychology Bulletin **32**(2): 188-200.

Gilovich, T., M. Kerr and M. V. Medvec (1993). "Effect of temporal perspective on subjective confidence." Journal of Personality and Social Psychology **64**: 552-560.

Goodwin, P. (1998). Enhancing judgmental sales forecasting: The role of laboratory research. Forecasting with judgment. G. Wright and P. Goodwin. New York, John Wiley & Sons**:** 91-112.

Grice, H. P. (1975). Logic and conversation. Syntax and semantics 3: Speech acts. P. Cole and J. L. Morgan. New York, Academic Press**:** 41-58.

Grimstad, S. and M. Jørgensen (2007). The Impact of Irrelevant Information on Estimates of Software Development Effort. The Australian Software Engineering Conference, Melbourne, IEEE Computer Society.

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.

Jørgensen, M. (2004). "A review of studies on expert estimation of software development effort." Journal of Systems and Software **70**(1-2): 37-60.

Jørgensen, M. (2005). The "Magic Step" of Judgment-Based Software Effort Estimation. International Conference on Cognitive Economics Sofia, Bulgaria.

Jørgensen, M. (2005). The "Magic Step" of Judgment-Based Software Effort Estimation. International Conference on Cognitive Economics, Sofia, Bulgaria, New Bulgarian University.

Jørgensen, M. and G. Carelius (2004). "An empirical study of software project bidding." IEEE Transactions on Software Engineering **30**(12): 953-969.

Jørgensen, M. and S. Grimstad (2008). "How to Avoid Impact from Irrelevant and Misleading Information When Estimating Software Development Effort." IEEE Software **May/June**: 78-83.

Jørgensen, M. and S. Grimstad (2008). Judgment-Updating among Software Professionals. The 2nd international conference on software knowledge information management and applications (SKIMA), Kathmandu, Nepal.

Jørgensen, M. and D. I. K. Sjøberg (2001). "Impact of effort estimates on software project work." Information and Software Technology **43**(15): 939-948.

Jørgensen, M. and D. I. K. Sjøberg (2004). "The impact of customer expectation on software development effort estimates." International Journal of Project Management **22**: 317-325.

Koehler, D. J. (1991). "Explanation, imagination, and confidence in judgment." Psychological Bulletin **110**(3): 499-519.

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

Mussweiler, T. and F. Strack (2000). "The use of category and exemplar knowledge in the solution of anchoring tasks." Journal of Personality and Social Psychology **78**: 1038-1052.

Northcraft, G. B. and M. A. Neale (1987). "Expert, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions." Organizational Behavior and Human Decision Processes **39**: 228-241.

Putnam, L. H. (1978). "A general empirical solution to the macro software sizing and estimating problem." IEEE Transactions on Software Engineering **SE - 4**(4): 345-360.

Schwarz, N. (1994). Judgment in social context: Biases, shortcomings, and the logic of conversation. Advances in experimental social psychology. M. P. Zanna. San Diego, Academic Press**:** 125-162.

Trope, Y. and A. Liberman (1996). Social hypothesis testing: Cognitive and motivational factors. Social psychology: Handbook of basic principles. E. T. Higgins and A. W. Kurglanski. New York, Guilford Press**:** 239-270.

APPENDIX A: Request for estimates of software development effort

Simula Research Laboratory is responsible for the collection of independent software development effort estimates on five software projects. The independent effort estimates will be applied to evaluate the realism of other software providers' effort estimates of the projects. The five project requirement specifications are enclosed this mail. The specifications are the clients' original ones, with the exception that we have removed information that identifies the software clients' names. Any requests for more information and clarifications should be directed per email to Dr. Magne Jørgensen at Simula Research Laboratory (magnej@simula.no).

The work we would like you to complete is to estimate the effort (in work-hours) your company, with normal quality of the development work, **most likely** would need to complete each of the projects, including testing. An acceptable interpretation of estimated effort (most likely use of work-hours) is that it is just as likely that the actual effort overruns as underruns the estimated effort.

Your company's role is consequently **not** to provide a bid for the project, but to help our clients to assess the realism of the effort estimates from other software providers. To ensure that your effort estimates are not impacted by estimation behaviour impacted by the wish to win a bidding round (which has repeatedly been found to lead to over-optimistic effort estimates) your company will **not** be considered for the development of the software specified in the specifications.

Your estimation work should be based on the assumption that a **tailor-made system** has to be developed, i.e., **you should not base the estimates on solutions where you adapt existing software packages or where you base the solution on large pre-made components**. All important technology choices you make to estimate the effort should of course be documented. An important reason for this avoidance of pre-made packages is the clients' need for control and ownership of source code. Notice also that it is essential that the projects are estimated independently from each others. Estimation work of high quality may also give you the opportunity to conduct similar and other types of software work for Simula Research Laboratory.

For each project we require that:

- The estimation work is completed by software professionals with documented, relevant competence.
- Your document the choice of technology and architecture.
- You document essential assumptions you have based the estimates on. This is in particular important when the specification is incomplete or difficult to interpret.
- The effort estimates are documented through a work break-down structure, a set of user stories, or a set of use cases. Each element, e.g., each activity, should be estimated.
- The uncertainty of the total effort estimate is documented as specified by us in *<xxx>*.
- You describe the estimation process that you have used.

If this estimation work is of interest to you, please respond to this mail with your price for the estimation work, the CV of the person(s) that will conduct the estimation work, and information about when you will be able to complete the assignment. The person(s) CV should emphasize documentation on estimation experience of projects similar to those included in this mail and be sent to us as soon as possible. The remaining steps are then as follows:

1) We will evaluate whether we choose to use your estimation competence or not. This will be based on your price for the estimation work, the CV and information about when you will be able to complete the assignment.
2) We will notify you about our decision within one week after reception of the CV.
3) The person(s) described in the CV should then provide the estimates of the most likely use of effort of each of the five projects, as specified in this mail. This should be completed as soon as possible, latest at the date specified by you in the initial mail.
4) When the estimation work is accepted, we pay you for the work (Or, if you prefer, we pay you 50% of the price after the first accepted estimate and the rest when all five are accepted.)

Please consider the requirement specification and the estimates you produce as confidential information, i.e., it is essential that this information is not distributed to people outside your company.