Empirical Software Engineering, 10, 7–29, 2005.

Expert Estimation of Web-Development Projects: Are Software Professionals in Technical Roles More Optimistic Than Those in Non-Technical Roles?

KJETIL MOLØKKEN-ØSTVOLD kjetilmo@simula.no
Software Engineering Department, Simula Research Laboratory, 1325 Lysaker, Norway

MAGNE JØRGENSEN magnej@simula.no
Software Engineering Department, Simula Research Laboratory, 1325 Lysaker, Norway

Editor: Michael S. Deutsch

Abstract. Estimating the effort required to complete web-development projects involves input from people in both technical (e.g., programming) and non-technical (e.g., user interaction design) roles. This paper examines how the employees' role and type of competence may affect their estimation strategy and performance. An analysis of actual web-development project data and results from an experiment suggest that people with technical competence provided less realistic project effort estimates than those with less technical competence. This means that more knowledge about how to implement a requirement specification does not always lead to better estimation performance. We discuss, amongst others, two possible reasons for this observation: (1) Technical competence induces a bottom-up, construction-based estimation strategy, while lack of this competence induces a more ''outside'' view of the project, using a top-down estimation strategy. An ''outside'' view may encourage greater use of the history of previous projects and reduce the bias towards over-optimism. (2) Software professionals in technical roles perceive that they are evaluated as more skilled when providing low effort estimates. A consequence of our findings is that the choice of estimation strategy, estimation evaluation criteria and feedback are important aspects to consider when seeking to improve estimation accuracy.

Keywords: Effort estimation, web development, bidding process, individual differences.

1. Introduction

This paper examines the relationship between the accuracy of expert effort estimates of web-development projects and the estimators' type of competence. In spite of a high number of published estimation models, see (Walkerden and Jeffery, 1997) for an overview, software development effort estimates frequently rely on expert judgement (Hughes, 1996; Moløkken-Østvold and Jørgensen, 2003; Jørgensen, 2004a), i.e., an estimation process where one or more competent people estimate project effort with little or no help from formal estimation models. One reason for the frequent use of expert estimation is that we lack substantial empirical evidence in favor of formal estimation models. We found (Jørgensen, 2004a) fourteen studies comparing the estimation accuracy of expert estimates with that of model estimates for software development and maintenance work. Of those studies, five were in favor of expert estimation (Kusters et al., 1990; Vicinanza et al., 1991; Mukhopadhyay et al., 1992; Pengelly, 1995;

Kitchenham et al., 2002), five found no difference (Heemstra and Kusters, 1991; Ohlsson et al., 1998; Walkerden and Jeffery, 1999; Bowden et al., 2000; Lederer and Prasad, 2000), and four were in favor of model-based estimation (Atkinson and Shepperd, 1994; Jørgensen, 1997; Myrtveit and Stensrud, 1999; Jørgensen and Sjøberg, 2002). Another argument in favor of expert estimates is that experts are more flexible regarding the type and format of the estimation information than formal estimation models.

Expert estimates may be based on a variety of estimation strategies, and conducted by people with different types of competence. The present knowledge on how an expert's strategy and competence affect estimation accuracy is, however, limited. The only software study related to this topic, as far as we know, suggests that neither maintenance skills, measured as frequency of unexpected problems, nor length of experience are good indicators of accurate estimates of one's own maintenance work (Jørgensen and Sjøberg, 2002). Similar lack of correspondence between expertise in completing the task and estimation accuracy is reported in other studies on human judgment (Lichtenstein and Fischhoff, 1977).

This paper focuses on effort estimates used as input to a bidding process. Such estimates may easily be affected by a price-to-win strategy (Boehm, 1984) and experts may, therefore, provide effort estimates that are much too low, due to the so-called ''anchoring-effect'' (Whyte and Sebenius, 1997; Jørgensen and Sjøberg, 2001).

If a company is selected as a contractor based on a bid that is too low, they may later try to skip functionality or expand the project by means of change requests to make the project profitable. This is, however, a dangerous strategy. The customer expects the specified functional solution for a price at least in the neighborhood of what was initially agreed upon, and may select another company for the next project. Clearly, realism in the estimates used as a basis for bidding is important for a company's long-term financial success.

Most of the previous research on software effort estimation has focused on traditional software projects, and there has not been much focus on web-development projects (McDonald and Welland, 2001). Although the challenges in the web-development industry are, to a large extent, similar to those in the traditional software development industry, there may be additional important estimation issues related to the increased focus on graphic and user interaction design in web-development projects. Another important difference from traditional software development companies, which often have few, but large projects, is that many web-development companies provide a very high number of project bids on relatively small projects (Wiegers, 1999). For example, the web-development company described in Section 2 of this paper had less than 100 employees, but nevertheless provided about 500 project bids per year. A high amount of project bids on small projects means that it may be difficult to conduct high quality estimation work on each bid. There may, for example, be insufficient estimation resources available to use many employees for several days on each estimate. The selection of competent personnel to conduct fast and sufficiently accurate expert estimates is, therefore, essential.

Web-development company employees have varied backgrounds and roles. There is, however, little literature describing the different roles in web development (McDonald and Welland, 2001), and we have been unsuccessful in discovering any earlier research

about how employees in roles typical for web-development projects differ in their estimation process and performance, which is the topic of this paper.

The remainder of this paper contains a description of the web-development company and software professionals we used as our case organization and experiment participants (Section 2), the hypothesis and motivation (Section 3), the design of the experiment (Section 4), the results from the experiment (Section 5), a discussion of the results (Section 6), and a conclusion and description of further work (Section 7).

2. The Web-Development Company and its Employees

The company that participated in our study is the Norwegian branch of a large international web-development company. At the time of the study, the branch had about 70 employees. For their last completed budget year (2000), the branch studied had a turnover of 119 million NOK (16.5 million USD). The role of the company is that of a contractor (McDonald and Welland, 2001), producing web-solutions for its customers. The projects were mainly web portals, e-commerce solutions or content management systems.

The organization can be viewed as quite immature, since it only had existed in its current form for about two years at the time of the study. It had been incorporated in an international organization after a gradual merging of five different national companies. They were not concerned with either CMM or ISO certification, and were therefore not rated according to these standards.

They had switched between several practices for web engineering during the past few years. The one in use at the time of the study was a company defined work process. It was based on waterfall development, and contained six phases: strategy and concept, specification, development, test, implementation and evaluation. Through a corporate initiative, they were beginning to implement a tailored version of Rational Unified Process (RUP) internationally and at the local branch. This was only in the initial phase at the time of the study.

The employee responsible for a project was usually also responsible for the estimation of that particular project. All estimation was expert based, and no algorithmic models, databases or checklists were in use. However, some of these experts used a predefined estimation process with a project work breakdown structure (Reifer, 2000). Depending on the type and size of the project, the expert responsible could ask other experts (mainly technicians or designers) for input about their areas of expertise. Table 1 shows estimation accuracy information based on a survey of all projects (n = 275) conducted by the company in the period January to October 2001.

Table 1. Company project estimation performance.

Overestimated projects	Projects on target	Underestimated projects
7%	36%	57%

Most projects (57%) were underestimated, and there was a total of 15% write-off (hours underestimated that were non-billable) for all projects. These 275 projects had a total of 94,748 billable and 111,430 actual work hours. This is not uncommon for software development, as observed in several other surveys (Moløkken-Østvold and Jørgensen, 2003).

From the interviews conducted at the company, we found that a typical reason for a project being ''on target'' was that it was paid per work-hour, overestimated, and the ''remaining effort'' was used to improve the solution. For small projects, with a customer-focus on quality of delivery, this may be a reasonable project approach. Other projects that were originally underestimated may also have been completed ''on target'' due to simplifications of the original estimated delivery. For this reason, one should not interpret the project review presented here purely as a measure of estimation skill. A previous paper (Jørgensen and Sjøberg, 2001) contains a more comprehensive discussion on this topic.

Most members of the company are assigned to one of the following four company roles:

- Engagement Manager (EM)—A person in this role is responsible for the customer contact, and usually handles contract negotiations. Typically, EMs have a business, or other non-technical, background and education.

- Project Manager (PM)—The leader of the project. Handles planning, resource allocation and day-to-day supervision of the project. A PM typically has a technical background.

- User Interaction Designer (Design)—A person in this role is responsible for Human Computer Interaction (HCI), graphical design, etc. User Interaction Designers have a large variety of backgrounds, most of them non-technical.

- Technology Developer (Tech)—This role is similar to that of the traditional software developer. Typically, Technology Developers have a technical education and background.

These roles are similar to those described by McDonald and Welland (2001). The only exception is that the company we describe does not have designated domain experts. From what we observed, that role in each project was typically covered by the customers.

The organization divides the project work into four tracks, depending on role: EM-track, PM-track, Design-track and Tech-track. The distribution of workload on the different tracks is similar to that described by McDonald and Welland (2001), with an average proportion of the total project effort of 10% on the EM-track, 20% on the PM-track, 35% on the Design-track, and 35% on the Tech-track. When analyzing the company, we found a clear control hierarchy based on the different roles (see Figure 1).

EMs and PMs have the most contact with customers.

3. Hypothesis

When reviewing data from completed projects more thoroughly, we observed a tendency for projects with heavy Tech-track involvement to be more often underestimated than those with medium or low Tech-track involvement.

A common scenario, when the estimator responsible for a project required assistance, was that people with a technical background estimated the Tech-track work and people with a non-technical background estimated the Design-track work. Consequently, if people with a non-technical background were less prone to underestimate their work, we should find a lower amount of non-billable work (write-off) on projects with a high proportion of non-technical work. To test this relationship, we randomly selected seven projects with a high degree (>40%) of Design-track work. These projects included much less Tech-track work than most of the other web-development projects in the company, with an average distribution of 21%, 54%, 23% and 2% for the different tracks (Tech, Design, PM and EM). The average track distributions of all projects were, as described earlier, 35%, 35%, 20% and 10%. The average write-off of the seven 'high Design-track involvement' projects was only 6%, i.e., a write-off much lower than the average 15%. This finding motivates the hypothesis that software professionals with a technical background are more optimistic than those in non-technical roles. Observational studies of this type, however, can only indicate a relation, since there are other possible explanations for the finding, e.g.:

- There is a greater probability that unforeseen problems will arise in Tech-track work, and hence it is more difficult to estimate. Such problems can be difficult to anticipate in an early stage of the project (Boehm, 1984), i.e., the stage when the bidding estimates are completed.

- When a project is behind schedule, it is easier to cut down on non-technical than technical activities, i.e., the actual use of effort for Tech-track work is less controllable by the developers. Earlier interviews with the employees of the company (Jørgensen and Sjøberg, 2001) indicated a very high ''flexibility'' in user interaction and graphic design activities.

There is, for these reasons, a need for more controlled studies to supplement the observational findings. In particular, there is a need for controlled studies on the level of optimism where comparable tasks are estimated. Motivated by the observations described above, we hypothesized the following relationship between role and estimation performance:

H1: Experts in technical roles provide more optimistic effort estimates than experts in non-technical roles.

An experiment that was designed to test this hypothesis is reported in the following sections.

4. Experiment Design

To test our hypothesis, we randomly selected 20 (five from each company role) software professionals employed by the company described in Section 2. Through payment we were able to select experienced software professionals in all roles. Based on their background and current roles, we categorized the participants in the EM and Design roles as ''Non-technical'' (NT) and the participants in the PM and Tech roles as ''Technical'' (T). All T-group members had a technical background, while the NT-group members had various other non-technical backgrounds (design, economics etc.). All personnel in the technical roles had education as engineers, master of engineering or equivalent. Their previous and current roles, both in the company studied and in earlier places of employment, had either been as technicians, software developers and/or project manager. Personnel in non-technical roles were educated as graphic designers, or held a Master of Arts, an MBA or another, equivalent, qualification. All participants had at least three years of college or university education, and an average of over six years' work experience in the IT-industry.

All participants were instructed to estimate the effort required to complete a web-development project based on the same requirement specification, employing their preferred estimation strategy. The project to be estimated was a project conducted by the company, selected by their Chief Project Manager. It was described as representative, and meaningful to estimate in about one hour. The actual project had received little publicity, and none of the participants in the study knew about it. The requirement specification given to the participants did not include the actual customer or project name. The complete requirement specification, as presented to the participants, is enclosed as Appendix I.

The project was in its start-up phase when it was selected. The actual effort of the project later turned out to be 2,365 work-hours. The project lasted approximately six months, and was completed several months after the study was conducted. It did not include development of unusually complex software, although it had a rather high proportion of technical activities (56% on Tech-track work vs. an average of 35%), see Table 2.

As estimation input, each participant was given the requirement specification as delivered by the actual customer. The requirement specification was two pages long, which was not atypical for specifications received by the company as a basis for project bids. The participants estimated the most likely effort (ML) measured in work-hours. The participants used 50–60 minutes on the estimation task.

Table 2. Actual project distribution of effort per track.

Track	Proportion of effort
EM	4%
PM	17%
Design	23%
Tech	56%

After this individual estimation, there was a group discussion, the purpose of which was to come to an agreement on a common estimate for the project effort. These discussions were videotaped. In addition to the testing of the hypothesis H1, we wanted to investigate whether discussion in unstructured groups could be a reasonable method for combining the judgment of several experts. The results indicate that it may be beneficial to combine estimates from experts with different backgrounds in order to reduce individual biases. This part is described in a separate paper (Moløkken-Østvold and Jørgensen, 2004).

Two questionnaires were completed by each participant, including information about the individuals' backgrounds and their estimation processes. We apply the information from the questionnaires and videotape in our discussion of the experiment results in Section 6. An overview of the steps in the experiment is presented in Table 3.

Although we did our best to achieve a high level of realism in the experiment, the study had limitations that should be considered when interpreting the results:

- The experimental setting did not enable the participants to contact the customer. However, this situation is similar to the first bidding round of a normal estimation process. If the customer is satisfied with the first bid, further contact is initiated, and more thorough estimation work is conducted.

- The estimators could not discuss estimation issues with colleagues in the study. This restriction on realism was a part of the design of our study on individual estimates, but means that the participants may have made better individual estimations in a more realistic setting.

- The experiment did not allow the participants to use as much time to complete the estimate as they might have wanted. On the other hand, as described earlier, the

Table 3. Experiment steps.

Step	Activity	Description
1	Selection of company	The described company was selected on the basis that it was evaluated as being representative of the web-development community.
2	Selection of task	The task was selected in cooperation with the chief project manager to ensure that the project was representative and meaningful to estimate in about one hour.
3	Selection of subjects	In cooperation with the chief project manager, subjects that had no knowledge of the project, and that could be clearly defined as being either technical or non-technical employees were selected.
4	Individual estimation	The subjects estimated the assignment individually
5	Questionnaire #1	The subjects answered a background questionnaire
6	Group estimation	The subjects formed groups and re-estimated the assignment.
7	Questionnaire #2	The subjects answered a questionnaire related to the assignment and their personal view on the effort required.
8	Debriefing	The subjects were debriefed on the background and purpose of the experiment.

general situation with many bids and small projects implied that participants in all roles were used to provide effort estimates within a short time in real estimation tasks. None of the participants complained about problems with delivering their estimates within the time available.

- The participants knew that this was an experiment and may, therefore, have had a lower motivation for providing accurate estimates. The experiment addressed this threat to validity in three ways. (1) Participants were informed that only serious participation would enable participation in further studies. (2) As a way of motivating serious participation, participants were paid according to the time spent, which was reported as billable hours. (3) Following their individual estimates, the participants met in groups to discuss their estimates. This means that the participants knew that their estimates would be evaluated by other members of their organization.

- It was not common for people in the Design and Tech roles to estimate the total project effort. However, the participants in these roles had participated in several projects and knew approximately how large a part of the total effort their work on different types of projects would be, and could use this knowledge to derive the total effort estimate.

However, since we were studying an eventual difference between technical and non-technical groups, it was most essential to keep the experimental conditions equal for both groups. This was done in the experiment, as all participants received the same treatment. We must, however, be aware of the possibility that the groups might have been affected differently by the experimental conditions. It could affect the results if, e.g., the limited time available were to cause more stress in, and hence lower the performance of, one of the groups. However, we do not believe that this was the case, because participants from both groups regularly worked together on the same assignments, with the same clients and deadlines.

The main argument supporting the internal validity of the experiment is the similarity between the participants' estimates of most likely effort and the actual project's estimate. The project plan describes a planned effort of 1,240 work-hours. The mean estimate of the participants in the experiment was 1,087 work-hours, i.e., only a little less than the originally planned effort. The experimental situation does not seem to bias the results in any significant way.

The external validity of the experiment is more difficult to assess. Only one company and only one project were used. However, our experiences with other companies lead us to believe that the company we studied is similar to many other web-development companies, both in size and (lack of formal) estimation process. Our observations of the company also indicate that they are similar to other small companies who employ an ad-hoc estimation approach, as described by Moses and Clifford (2000).

The time available limited us to use one project in the experiment, but the chief project manager assured us prior to the experiment that it was representative, and we conducted interviews with the employees who participated in the actual project after the project was completed. These interviews revealed no extreme properties of the project, only that their initial estimates had been far too optimistic.

Even if the limitations of the experiment were to have had an unwanted impact on the estimates, this would only be problematic if the impact was much larger on some groups of the participants, e.g., the people from Tech-track. We have no reason to believe that this was the case.

5. Experiment Results

The individual estimates, divided into the NT and T groups are shown in Table 4. All estimates are in work-hours.

The actual effort of the project was 2,365 work-hours, and most participants gave an estimate that was much lower than the actual effort. However, the actual effort of a new project based on the same specification may have required less (or more!) effort than the completed project. The actual effort of the completed project cannot, for this reason, be taken as more than an indication of estimation accuracy in this experiment. Nevertheless, based on the actual effort and the original estimate, we believe that ML-estimates of less than 1,000 work-hours point to strong over-optimism. As can be seen in Table 4, nine out of ten participants in the T group had ML-estimates less than 1,000 work-hours, compared with only two out of ten in the NT group.

5.1. Test of Hypothesis 1

The results displayed in Table 4 indicate that the ML-estimates were systematically higher for the NT group than for the T group. We tested the strength of this difference by applying a statistical t-test. To use the t-test properly, the underlying distributions should be approximately normal. An Anderson-Darlington normality test (Christensen, 1998) did not exclude normality ($p > 0.1$ for all relevant distributions), and a visual inspection

Table 4. The Most Likely (ML) effort estimates.

Non-technical (NT) roles		Technical (T) roles	
Role	ML	Role	ML
EM	1200	PM	960
EM	1550	PM	1820
EM	1850	PM	300
EM	547	PM	914
EM	2286	PM	984
Design	1500	Tech	960
Design	1140	Tech	585
Design	1260	Tech	220
Design	620	Tech	660
Design	1500	Tech	900
Mean	**1345**	**Mean**	**830**
Median	**1380**	**Median**	**907**

of the distributions suggested that they were close to normal. There were moderate departures from the normal distribution, but this has generally negligible effects on the validity of both Type I and Type II error calculations (Cohen, 1969). The only exceptions are when there are substantially unequal variances and substantially unequal sample sizes, which was not the case here.

The results from the t-test (one-sided, assuming different variance) are shown in Table 5. We provide the actual p-values, as suggested by Wonnacott and Wonnacott (1990), instead of pre-defining a significance level for rejection. To measure the size of the difference in mean values, we included Cohen's size of effect measure (d) (Cohen, 1969). The size of effect (d) was calculated as: d = (mean value NT group – mean value T group) / pooled standard deviation amongst groups NT and T.

The results in Table 4 show that the p-value was low (0.02). As indicated by the size of effect (d), the difference was considered ''large'', i.e., d was larger than 0.8 (Cohen, 1969). The robustness of the results is illustrated by a non-parametric statistical Kruskal-Wallis (Siegel and Castellan, 1988) test on the median estimates results, which gave a p-value less than 0.03. The analysis, therefore, supports Hypothesis 1.

6. Discussion

The observation that people in technical roles showed a higher degree of optimism in effort estimates does not necessarily imply that the roles or backgrounds were the direct *causes* of the difference. The effects may be due to a third variable. Perhaps difference in, for example, estimation strategy, has a direct effect on the estimates. This section discusses potential reasons for our observations. We investigate both personal differences due to length of experience and gender, and estimation differences such as perceived estimation skill, formal estimation training and estimation strategy. Interesting aspects related to the organizational structure, namely estimation goals and (lack of) feedback are also treated.

6.1. Different Length of Experience

A potential explanation of the observed difference is that the NT-group participants were more experienced. However, as shown in Table 6 there were no large differences between the groups in the measured types of experience.

Table 5. Differences between mean estimates for NT and T groups.

	NT group (mean work-hours)	T group (mean work-hours)	Difference in mean	Pooled standard deviation	t-test (p-value)	Size of effect (d)
Most likely	1345	830	515	486	0.02	1.1

Table 6. Subjects' average experience.

Group	Total experience IT-industry	Experience in current role	Experience in providing effort estimates
NT	5.1	2.0	1.2
T	7.5	3.8	1.7

Even if there were differences in length of experience, they could probably not explain the difference in estimation optimism, as shown by the following analysis. We divided the participants into two equally sized experience-groups, where each member of the ''Long Experience''-group had greater experience than had all members of the ''Short Experience''-group. A Kruskal-Wallis analysis of the difference in median estimates (ML) of the two groups yielded $p = 1.0$ when the groups were divided according to total IT-industry experience, $p = 0.41$ when divided according to experience in current role, and $p = 0.55$ when divided according to estimation experience. This lack of relation between length of experience and estimation performance is similar to the results reported by Jørgensen and Sjøberg (2002). Interestingly, all three analyses exhibited the same trend with respect to the effect of experience: Shorter experience led to higher, i.e., more realistic, estimates. More studies are needed to investigate this relationship.

6.2. Gender Differences

Results reported by Henry (1994) suggest that the level of optimism may depend on gender, i.e., that female estimators provide higher estimates. There were only three female participants in our study (all of them worked in the Design-track), and it is, therefore, unlikely that gender explains all of the variance. Nevertheless, it is interesting to observe that the three female participants had the highest estimates (ML) of those in the Design-track; their median estimate (1,500 work-hours) was much higher than that of the male participants (960 work-hours). The impact of gender on optimism of effort estimates is an interesting topic for further studies.

6.3. Differences in Perceived Estimation Skill

We asked the participants about their opinions on the company's and their own estimation skills. The possible answers ranged from 1 (very poor) to 5 (very good) in questions 1 and 2, and from 1 (low importance) to 5 (high importance) in question 3. Table 7 shows that there was almost no difference in the NT and T-group participants' opinions regarding effort estimates. Interestingly, both the T and NT-group participants viewed their own estimation skill as inferior (!) to that of the company (2.6 vs. 3.0 for the T group and 2.6 vs. 3.1 for the NT group).

Table 7. Subjective opinions on estimation skill and importance.

Subjective opinion on...	NT (mean values)	T (mean values)
1. The company's estimation skill (1–5):	3.1	3.0
2. Own estimation skill (1–5):	2.6	2.6
3. The importance of accurate effort estimates to the company (1–5):	4.4	4.6

The questionnaires were completed after the project was estimated. The participants may therefore have been influenced by the somewhat difficult estimation task of the experiment, and rated their skill lower than they would normally do. Still, this would affect both NT and T-groups to a similar extent. In addition, when we are measuring estimation precision, it is preferable that the estimation task influences the questionnaire answers, and not the other way around.

6.4. Differences in Formal Estimation Training

There were no differences regarding the training in estimation received at universities, or earlier or current employer. In the T group five out of ten, and in the NT group four out of ten, reported that they had received some estimation training.

6.5. Differences in Choice of Estimation Strategy

Based on an informal analysis of the estimation material completed by the participants, and of the videotape of the group discussion, it was possible to derive a classification of the estimation strategies applied. One should be cautious when interpreting these results, because we only had access to the participants' mental processes through the discussion, not, for example, through a think-aloud protocol. Dividing the strategies into the broad categories, top-down and bottom-up (Kitchenham, 1996), we categorized the distribution of estimation strategies as shown in Table 8.

In top-down estimation, a project is reviewed as a whole, and the project's effort estimates are derived, for example, through a recall of the effort used on similar projects,

Table 8. Distribution of estimation strategy.

	NT group (# participants)	T group (# participants)
Bottom-up	2	8
Top-down	6	1
Uncertain	1	1
Mixed	1	0

i.e., the effort estimate is based on an ''outside view'' of the project. The project estimate can then be broken down into phases, activities, or tracks. As described in (Boehm, 1984), major advantages of this method are its efficiency, and the fact that it can be combined easily with more formal analogy based estimation strategies (Hughes, 1996).

In bottom-up estimation the project is divided into components or activities, where each component or activity is estimated individually. The total effort is then calculated as the sum of all the component and activity effort values that are identified. The advantage of this method is that each activity is estimated and is available for future project plans (Kitchenham, 1996). The risks are that activities can be easily forgotten, and that the risk budget covering unexpected tasks is not sufficiently large (Boehm, 1984; Kitchenham, 1996). One reason that technicians prefer bottom-up estimation may be that they are familiar with the method, since they will often be required to estimate each activity at later project stages.

The estimation material and the video-analysis of the subjects in the group-discussions indicate that a major reason for the optimistic estimates was forgotten activities. It was explicitly reported in the questionnaires by four T participants and one NT participant that missing activities were among the reasons for their estimates being lower than those of the other participants. Informal interviews with the Chief Project Manager of the company and other company staff also indicated that the main source of over-optimism was forgotten activities.

We conducted a Kruskal-Wallis test on the median values of those participants who followed a clear bottom-up or top-down strategy, to search for an eventual impact of the choice of strategy on the estimates. Table 9 shows that the p-value does not approach a significant level, but that the median value of the top-down based estimates are higher than those based on a bottom-up strategy.

We should not exclude the possibility that the effect of the estimation strategy is important, based on the statistical test alone, because:

- The power of the analysis described in Table 9 is low, due to large variance and few observations.

- The group discussions, the questionnaire answers, results from similar studies (Buehler et al., 1994), and the direction of the measured difference in estimates, all point in the same direction, i.e., all available evidence suggests that the choice of estimation strategy may have an impact on the level of optimism in estimates.

Table 9. Median estimates grouped by strategy.

Strategy	Estimate
Bottom-up (N = 10)	937
Top-down (N = 7)	1260
p-value	0.44

The difference between the NT- and T-group participants' estimates was, however, much larger than the difference between the bottom-up and top-down strategy participants estimates, i.e., it is unlikely that choice of strategy can account for all of the difference between the NT and T groups.

6.6. Difference in Estimation Goals

The way in which a company measures its employees' competence can affect estimation performance. Estimation accuracy is only one out of many possible estimation goals. Other goals for the estimators may be, for example, to signalize personal competence, or to have sufficient time available to deliver a good project result.

As reported in earlier interviews (Jørgensen and Sjøberg, 2001), software professionals in the Design-track were not evaluated for ''speed of delivery''. More important evaluation factors were design quality and usability, which are more difficult to measure (Rosenfeld and Morville, 1998). For designers, their estimation goals may be to receive as much time as possible within the projects limits, to achieve the best possible result. To achieve this goal, they prefer to provide the project managers with less optimistic estimates at an early stage.

The software professionals from the Tech-track, on the other hand, are expected to deliver functional programs, and ''speed of delivery'' is an important evaluation measure of their competence. One way to signal competence is, therefore, to provide low estimates. This is the basic concept of Tesser's *Self-evaluation maintenance theory* (Aronson et al., 1999). If a person feels threatened, e.g., by other technicians potentially delivering low estimates on similar tasks, they will adjust their behavior, e.g., deliver very optimistic estimates, in order to preserve their self-esteem.

For software professionals in the PM and EM roles, the estimation goals may be more complex. A possible explanation for the optimism of PMs is that most of them were programmers, who had been promoted to PMs, i.e., they may still tend to think very much as programmers. In addition, a PM's estimates may be strongly influenced by the expectations from the management, e.g., a PM with low estimates may achieve an immediate positive feedback from the management. On the other hand, PMs may also desire high estimates on their project, as this increases the probability that the project is completed on time. The degree to which these two evaluation considerations affect estimates may depend on how much weight the PMs put on short-term (immediate positive evaluation from low estimates) versus long-term competence evaluation benefit (positive evaluation when the project is completed successfully due to realistic effort estimates).

The videotaped group discussion showed that almost all the EMs were concerned about how much the customer was willing to pay for the project. This should have led to a bias towards very low effort estimates, i.e., the opposite of what we observed. On the other hand, EMs are also responsible for the pay-off of the projects. Similarly to the PMs, the EMs may estimate with a trade-off in mind between short-term positive evaluation (get the contract) and long-term positive evaluation (make the company profitable).

6.7. Differences in Feedback

Lack of feedback prevents learning from experience when estimating effort. As described by Jørgensen and Sjøberg (2001) the estimation feedback to the employees of the company is very limited. The feedback on total project effort may be better than that on individual tasks, i.e., it may be in favor of EMs and PMs. This may have caused the EMs to perform better than the designers and the PMs to perform better than the technicians in the experiment. However, these systematic role-dependent differences cannot explain the difference in optimism between the participants in the T and NT roles.

In the questionnaire on the participants' background, four of the ten T-group participants reported that they usually underestimated effort; four believed that they usually were on target, while two reported that they typically overestimated the effort. The corresponding numbers for the NT-group participants were seven (underestimation), three (on target) and zero (overestimation). This indicates that the NT-group participants were more aware of their biases.

An important consequence of lack of feedback may be that goals other than estimation accuracy become more important. For example, when software developers know that they will not get any feedback on, or evaluation of, an effort estimate, greater weight may be placed on the short-term goal of ''appearing skilled''.

7. Conclusions and Further Work

The experimental results suggest that software professionals in technical roles and with a technical background were more optimistic in their estimates than those in non-technical roles and with a non-technical background. The reasons for this difference are more difficult to pinpoint, but there are probably several contributing factors. As discussed earlier, explanations such as length of experience, perceived estimation skill and formal estimation training seem to have no significant impact. Differences between gender and estimation strategy were difficult to investigate because of the small number of participants, but these variables may have an impact. Although it is impossible to measure and test statistically, evidence from other research and study of the company might indicate that estimation goal and (lack) of feedback could cause individual differences between the groups. In our opinion, the three most likely explanations for the observed difference are:

- *Estimation strategy*. Software professionals with technical competence typically use a bottom-up estimation strategy, while people in non-technical roles must base their estimates on ''outside'' properties of the project, and employ a top-down strategy. Our study indicates that this bottom-up strategy leads to estimation optimism, mainly because of forgotten activities. The indication is, however, not strong and we intend to conduct further studies where the difference in estimation strategy is the main topic. Observations that the recall of very similar, previously completed, projects is a pre-requisite for accurate top-down based effort estimation has already been made (Jørgensen, 2004b).

- *Estimation goals*. The evaluation of competence depends on the software professionals' role in the organization. It seems that programmers are perceived as more competent by others (and themselves) when they provide low estimates. The evaluation criteria of the non-technical and PM roles may be more context-dependent and complex and we need more studies to examine whether, and how, different estimation contexts stimulate optimism, realism and pessimism in expert estimates.

- *Lack of Feedback*. Lack of estimation feedback seems to be typical for software development companies, and implies that other, conflicting goals become more important than estimation accuracy. In particular, we believe that the desire to appear skilled gains stronger weight when there is no feedback. Probably, a combination of these factors influences the estimates. It would, for this reason, be interesting to study the estimation behavior of software professionals in situations where there is more feedback on estimates.

To summarize, there are reasons to believe that the company roles of software professionals affects estimation strategy and goals, which in turn affects software estimation optimism. We believe that company role, in combination with poor project feedback, explains most of the observed difference between software professionals with technical and non-technical backgrounds in our study.

It seems as if the choice of estimation strategy, estimation evaluation criteria, and feedback all seem to affect estimation accuracy. We encourage managers and experts to consider these elements in their estimation process to avoid underestimation. For example, if the estimators are not evaluated on estimation accuracy and apply a bottom-up estimation strategy, there may be a high risk of the estimates being too low.

Acknowledgments

Thanks to Simula Research Laboratory for funding the experiment, Dag Sjøberg for valuable comments, and to all participants and organizers at the company that participated in the study. Kjetil Moløkken-Østvold was funded by the Research Council of Norway under the project INCO.

References

Aronson, E., Wilson, T. D., et al. 1999. *Social Psychology*. Addison-Wesley Educational Publishers Inc.

Atkinson, K., and Shepperd, M. 1994. Using function points to find cost analogies. *European Software Cost Modeling Meeting*. Ivrea, Italy.

Boehm, B. 1984. Software engineering economics. *IEEE Trans. Softw. Eng.* 10(1): 4–21.

Bowden, P., Hargreaves, M., et al. 2000. Estimation support by lexical analysis of requirement documents. *J. Syst. Softw.* 51: 87–98.

Buehler, R., Griffin, D., et al. 1994. Exploring the ''Planning fallacy'': Why people underestimate their task completion times. *J. Pers. Soc. Psychol.* 67(3): 366–381.

Christensen, R. 1998. *Analysis of Variance, Design and Regression. Applied Statistical Methods*. Chapman & Hall/Crc.

Cohen, J. 1969. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press, Inc.

Heemstra, F. J., and Kusters, R. J. 1991. Function point analysis: Evaluation of a software cost estimation model. *Eur. J. Inf. Syst.* 1(4): 223–237.

Henry, R. 1994. The effects of choice and incentives on the overestimation of future performance. *Org. Behav. Human Decis. Process.* 57: 210–225.

Hughes, R. T. 1996. Expert judgment as an estimating method. *Inf. Softw. Technol.* (38): 67–75.

Jørgensen, M. 1997. An empirical evaluation of the MkII FPA estimation model. *Norwegian Informatics Conference*. Voss, Norway, Tapir, Oslo.

Jørgensen, M. 2004a. A review of studies on expert estimation of software development effort. *J. Syst. Softw.* 70(1–2): 37–60.

Jørgensen, M. 2004b. Top-down and bottom-up expert estimation of software development effort.'' *J. Inf. Softw. Technol.* 46(1): 3–16.

Jørgensen, M., and Sjøberg, D. I. K. 2001. Impact of software effort estimation on software work. *J. Inf. Softw. Technol.* 43: 939–948.

Jørgensen, M., and Sjøberg, D. I. K. 2002. Impact of experience on maintenance skills. *J. Softw. Maint. Evolut.: Res. Pract.* 14: 1–24.

Kitchenham, B. 1996. *Software Metrics: Measurement for Software Process Improvement*. Oxford: NCC Blackwell.

Kitchenham, B., Pfleeger, S. L., et al. 2002. An empirical study of maintenance and development estimation accuracy. *J. Syst. Softw.* 64: 57–77.

Kusters, R. J., Genuchten, M. J. I. M., et al. 1990. Are software cost-estimation models accurate? *Inf. Softw. Technol.* 32(3): 187–190.

Lederer, A. L., and Prasad, J. 2000. Software management and cost estimation error. *J. Syst. Softw.* 50: 33–42.

Lichtenstein, S., and Fischhoff, B. 1977. Do those who know more also know more about how much they know? *Org. Behav. Human Perform.* 20: 159–183.

McDonald, A., and Welland, R. 2001. Web engineering in practice. *Proceedings of the Fourth WWW10 Workshop on Web Engineering*.

Moløkken-Østvold, K., and Jørgensen, M. 2003. A review of surveys on software effort estimation. Proceedings of ISESE 2003, IEEE.

Moløkken-Østvold, K., and Jørgensen, M. 2004. Group processes in software effort estimation. *Empirical Software Engineering.* 9(4): 315–334.

Moses, J., and Clifford, J. 2000. Learning how to improve effort estimation in small software development companies. *24th Annual International Computer Software and Applications Conference*. Taipei, Taiwan, IEEE Comput. Soc, Los Alamitos, CA, USA.

Mukhopadhyay, T., Vicinanza, S. S., et al. 1992. Examining the feasibility of a case-based reasoning model for software effort estimation. *MIS Quarterly* June, 155–171.

Myrtveit, I., and Stensrud, E. 1999. A controlled experiment to assess the benefits of estimating with analogy and regression models. *IEEE Trans. Softw. Eng.* 25: 510–525.

Ohlsson, N., Wohlin, C., et al. 1998. A project effort estimation study. *Inf. Softw. Technol.* 40: 831–839.

Pengelly, A. 1995. Performance of effort estimating techniques in current development environments. *Softw. Eng. J.* September, 162–170.

Reifer, D. J. 2000. Web development: Estimating quick-to-market software. *IEEE Softw.* 17(6): 57–64.

Rosenfeld, L., and Morville, P. 1998. *Information Architecture for the World Wide Web*. Sabastopol, O'Reilly and Associates, Inc.

Siegel, S., and Castellan, N. J. 1988. *Non-Parametric Statistics for the Behavioral Sciences*. McGraw Hill College Div.

Vicinanza, S. S., Mukhopadhyay, T., et al. 1991. Software effort estimation: An exploratory study of expert performance. *Inf. Syst. Res.* 2(4): 243–262.

Walkerden, F., and Jeffery, R. 1997. Software cost estimation: A review of models, process, and practice. *Adv. Comput.* 44: 59–125.

Walkerden, F., and Jeffery, R. 1999. An empirical study of analogy-based software effort estimation. *J. Empir. Soft. Eng.* 4: 135–158.

Whyte, G., and Sebenius, J. K. 1997. The effect of multiple anchors on anchoring in individual and group judgment. *Org. Behav. Human Decis. Mak.* 69(1): 75–85.

Wiegers, K. E. 1999. Software process improvement in web time. *IEEE Softw.* 16(4): 78–86.

Wonnacott, T.H., and Wonnacott, R.J. 1990. *Introductory Statistics*. John Wiley & Sons.

Appendix I

1. User Requirements

This is the requirement specification for the project, as delivered by the customer.

2. The Customer

The customer is the producer of an established technical encyclopaedia that numbers 800 editions with 10,000 illustrations and 1,600 tables. They have 20,000 subscribers. The publication frequency is low, with two shipments each year, each containing several magazines.

The magazines exist both on paper and CD-ROM. The CD-ROM contains some extra features, and there are plans to add other sources of information.

There is also a simple intranet version of the CD-ROM that is used internally and by a handful of existing customers.

All documents are created in MS-word, and approved and converted to HTML by a central unit. They have no need for a very complicated CMS-system.

3. Status

The starting point for a web version is the existing CD-ROM and intranet based system. The primary goal is to build on the functionality of these systems, but also to offer the encyclopaedia commercially over the Internet, both to companies and individuals. It is natural to use this opportunity to look at the possibilities for a new medium, as well as revising existing production and administration routines.

All documents that will be used exist on the CD-ROM in HTML versions, and can be copied directly to the website. No changes are needed. Design is of less importance.

All the users of the site are expected to be technical competent people, with experience and knowledge of the CD-ROM and/or paper versions.

4. Desired Functionality

This is the required functionality.

1. Basis functionality (searching for and displaying information), see figures 2 and 3

Display documents in HTML-format (document including local table of contents (TOC) in magazine)
Navigation through an expanding TOC

Navigation through an index
Free search (Expansion of the existing version)

2. Downloading of documents and pictures

The system must allow the downloading of documents in other formats.
PDF versions of documents for better prints where such exist
Figures in high-resolution bitmap (TIFF)
Figures in vector format (DWG)
Displaying of video-clips
The system must handle the fact that not all documents and figures exist in all formats.

3. Extranet functionality

Only paying subscribers shall have access to the service. For companies this can be implemented by access-limitation on a net-level (IP). Company customers can then skip logon procedures with usernames and passwords. An alternative method is to use personal subscriptions. The system must be able to handle different types of subscriptions that grant different degrees of access to the system.

4. Trade solution

Possibility of subscribing via the web
Possibility of buying a single magazine for downloading or delivery by mail
Possibility of paying online by credit card or other forms of payment

5. Demo-/sales-version

The system shall have an open part, granting access to some functionality, such as navigation and search, and limited content. This functionality should be combined with a function that allows the purchase of single magazines for downloading.

6. Reply service

This contains an overview of the FAQ. The answers should have links to magazine editions with extensive information. The user is checked, and receives the option to log in, buy a single magazine, or buy a subscription.

7. User adjustment and information

Possibility of having personal settings for each user
Possibility for users to register own comments to the magazine
Possibility for the editors to publish comments to the magazines
Possibility for users to give feedback directly from a magazine, reporting errors etc.
Discussion forum tied to magazines

8. Administration of users

The administration system stores information on all customers, both subscribers and buyers of single issues. The administration system must monitor the use of the system. One must be able to extract different types of statistics, as well as blocking users who abuse the system or fail to pay their subscription.

9. Integration with existing administrative systems

Subscriber- and logistics system, Agresso
Customer and support, Superoffice
Degree of integration must be based on cost/value aspects.

10. Adjustment of the production system

The production system is based on a personally developed database, containing a parser that translates documents from MS Word format to HTML. The system must be adapted to a new medium and a new system. Other possible changes to consider:
Parsing of word documents to XML instead of HTML, which will add to the system's flexibility.
Expansion of the database to support the administration of manuscripts.
Adapt the base for the production of additional documents used in the printed issue (TOC, index list, overview of new, changed and expired magazines etc.)

11. Interface with other systems

By implementing the magazine in a web setting, integration with other systems should be considered. An interface (API) for communication with other systems (Using SOAP, XML etc.) should be defined. It must be possible to link to documents by URL.

12. Choice of technology

The goal is to use already known technology. This is to handle development and changes with internal resources.
OS: Windows 2000
Web server: Internet Information Server w/ASP
Database: Microsoft SQL Server
Languages: Visual Basic, NET technology

13. Hosting

The system is to be installed on the client's servers.

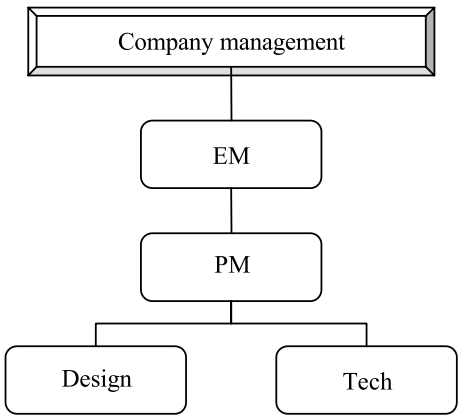

Figure 1. Role hierarchy.

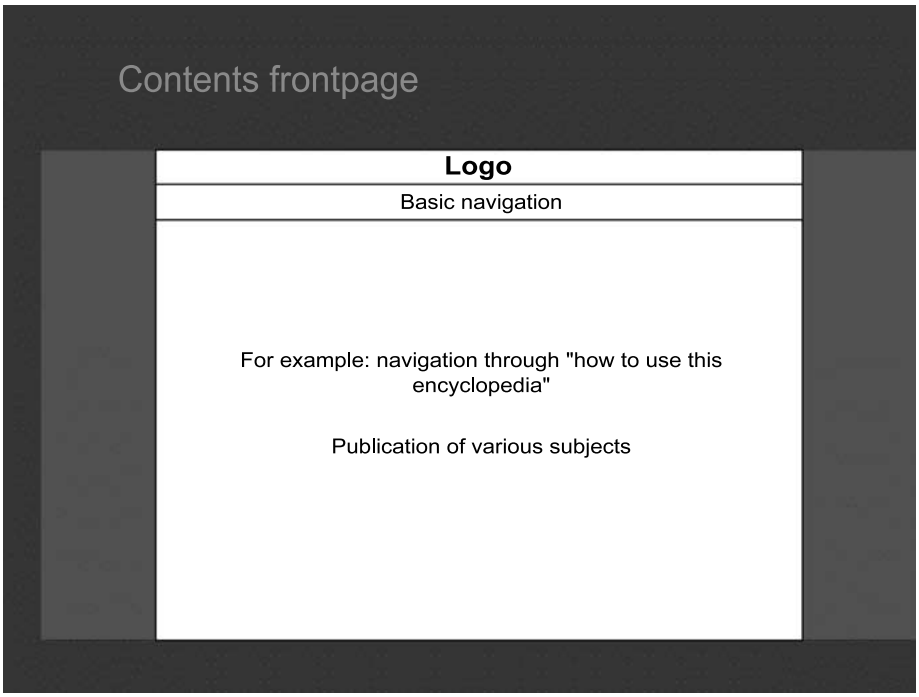

Figure 2. Front page example.

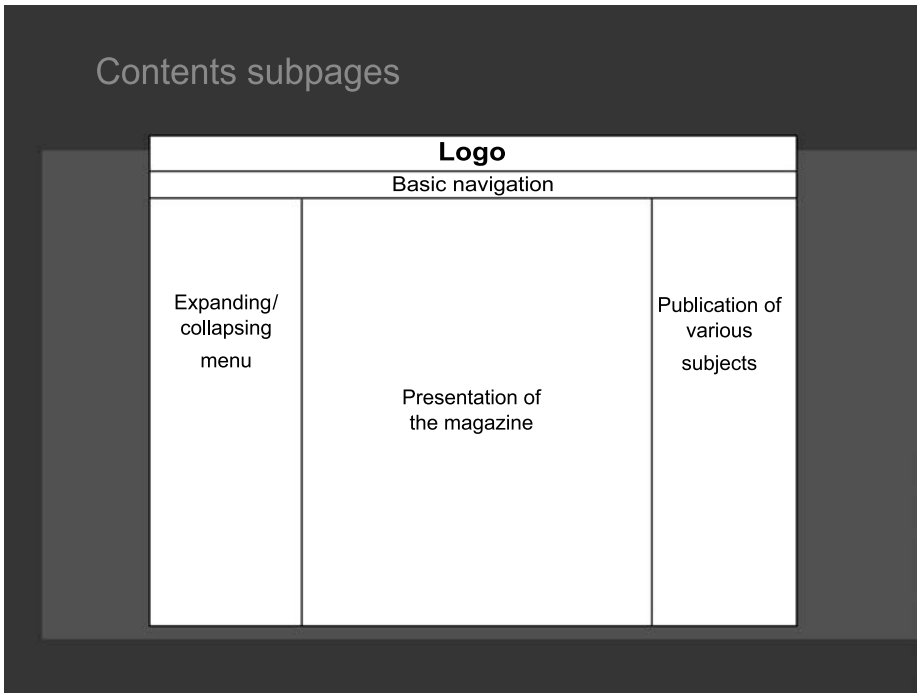

Figure 3. Sub page example.

Kjetil Moløkken-Østvold is a PhD student at the Software Engineering research group at Simula Research Laboratory in Norway. He received his MSc degree in computer science from the University of Oslo, Norway, in 2002. He joined Simula Research Laboratory in October 2002. His main research interests are software effort estimation, group processes and software process improvement.

Magne Jørgensen received the Diplom Ingeneur degree in Wirtschaftswissenschaften from the University of Karlsruhe, Germany, in 1988 and the Dr. Scient. degree in informatics from the University of Oslo, Norway in 1994. He has 10 years industry experience as consultant and manager. He is now professor in software engineering at University of Oslo and member of the software engineering research group of Simula Research Laboratory in Oslo, Norway. His research interests include software cost estimation, uncertainty assessments in software projects, expert judgment processes, and, learning from experience.