Empirical Software Engineering, 9, 315–334, 2004.

Group Processes in Software Effort Estimation

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

Editors: Marian Petre, David Budgen and Jean Scholtz

Abstract. The effort required to complete software projects is often estimated, completely or partially, using the judgment of experts, whose assessment may be biased. In general, such bias as there is seems to be towards estimates that are overly optimistic. The degree of bias varies from expert to expert, and seems to depend on both conscious and unconscious processes. One possible approach to reduce this bias towards over-optimism is to combine the judgments of several experts. This paper describes an experiment in which experts with different backgrounds combined their estimates in group discussion. First, 20 software professionals were asked to provide individual estimates of the effort required for a software development project. Subsequently, they formed five estimation groups, each consisting of four experts. Each of these groups agreed on a project effort estimate via the pooling of knowledge in discussion. We found that the groups submitted less optimistic estimates than the individuals. Interestingly, the group discussion-based estimates were closer to the effort expended on the actual project than the average of the individual expert estimates were, i.e., the group discussions led to better estimates than a mechanical averaging of the individual estimates. The groups' ability to identify a greater number of the activities required by the project is among the possible explanations for this reduction of bias.

Keywords: Software development, effort estimation, expert judgment, group processes, expert bias.

1. Introduction

> "When you lie about the future, that's called optimism, and it is considered a virtue. Technically speaking you can't "lie" about the future because no one knows what will happen. When you apply this unique brand of optimism (not lying!) at work, that's called forecasting."
>
> Scott Adams (2002)

Improving the quality of effort estimation is one of the great challenges for software project management. Across a wide range of software projects, from web applications to time-critical financial or medical systems, poor effort estimation is observed, and the problem seems to increase with project size (Gray et al., 1999). To account for this, several formal methods and processes to support estimators, such as COCOMO II (Boehm et al., 2000), ANGEL (Shepperd et al., 1996) and WEBMO (Reifer, 2000) have been proposed. This paper, however, focuses on the estimation of effort by experts. This seems to be the most common estimating method (Jørgensen, 2004), and is employed over a wide range of software projects. One of the main

problems with expert estimation, however, is that it is no better than its participants, and hence is subject to their individual biases (Boehm, 1984; Hughes, 1996) and political pressure brought to bear by the company (Hughes, 1996). A possible method of reducing the risk of unwanted influence on the estimates is to use group discussion as an estimation method. This is not a new idea, but it seems to have been neglected in the empirical research on software engineering. In other fields, however, there is continuous ongoing research on how to best combine the opinions of several experts (Armstrong, 2001).

We believe that new companies in particular, which have limited historical data upon which to draw for experience, may benefit from using groups to improve their effort estimation process. This paper describes a controlled experiment on the performance of estimation groups in a web-development company that uses no formal estimation process or project experience database. A typical web-development project has several characteristics that distinguish it from traditional software development projects (Reifer, 2000). These characteristics include project size, development approach, processes employed and people involved. Some of the differences are due solely to the nature of the technology involved, but others may be related to company size and maturity. Most web-development companies are small compared with traditional software development companies, and smaller companies may face additional estimation challenges, such as access to domain experts and process management (Moses and Clifford, 2000).

In our experiment, 20 individual experts submitted project estimates that ranged from 220 to 2286 hours, with an average of 1088 hours. Estimates from five groups of four experts each ranged from 1100 to 2251 hours. The actual effort expended on the project was 2365 hours, indicating that some, but not all, of the bias towards optimism was eliminated by group discussion. In essence, the results reported in this paper will apply to similar companies and estimation contexts, but the basic idea of using unstructured group discussion to reduce individual bias may also be transferable to more traditional development projects and other estimation contexts. The company studied, our experiences from this and other companies, and the extent to which the study has external validity, will be discussed in later sections.

The remainder of this paper is structured as follows: (2) Background for Expert and Group Estimation, (3) Hypotheses, (4) Research Method, (5) Results, (6) Discussion and (7) Conclusion and Further Research.

2. Background for Expert and Group Estimation

Expert judgment is a commonly employed method for estimating the effort required to complete software projects. As reported in a review of studies on expert estimation by Jørgensen (2004), several independent surveys (Heemstra and Kusters, 1991; Hihn and Habib-Agahi, 1991; Hill et al., 2000; Jørgensen, 1997; Kitchenham et al., 2002a; Paynter, 1996) rate it to be the preferred method among professional software developers. Although expert estimation is commonly used, it is probably not because of its precision that the method is favored. In fact, expert estimation seems to be just

as imprecise as the use of formal estimation models (Jørgensen, 2004). However, expert estimates can be especially useful (and often the only option) for companies that lack documented experience from earlier projects (Höst and Wohlin, 1998) or have limited estimation resources (Moses and Clifford, 2000). This is often the case for young and/or unstable organizations. Both these characteristics apply to most web-development companies, including the one we studied and report on in this paper.

A search on the terms ''software effort'' or ''size estimation'' in the leading software engineering journals, IEEE Transactions on Software Engineering, Journal of Systems and Software, Journal of Information and Software Technology and Journal of Empirical Software Engineering, yielded slightly more than 100 relevant papers. Of these, only 16 had expert estimation as a topic, and only one of them had group estimation as a topic. In that paper, Taff and his colleagues present a structured model for expert estimating in groups, called Estimeetings (Taff et al., 1991), but do not compare these estimates with individual or other group based estimates, only with actual effort expended. In other words, papers in the leading journals for software engineering did not contain a single study on individual versus group-based software effort estimation.

Group expert estimates may be categorized according to two different characteristics, though other categorizations are possible.

The first characteristic concerns the involvement of the estimators, and offers two possible approaches: (i) There are designated estimation groups (DeMarco, 1982), whose only objective is to estimate the project, and do not participate in the development process. (ii) Employees who are likely to develop the project are responsible for the estimates. Both approaches have advantages and disadvantages. A separate estimation group may be much less prone to personal or political biases, and more likely to improve their estimation skill over time (DeMarco, 1982). On the other hand, to have such a designated team would require a large organization, and good internal communication. Expert estimators, who are likely to implement the solution themselves, will probably get to know the project better than anyone else, and may have a higher motivation for a thorough project analysis (Hughes, 1996). For smaller organizations, this approach may be the only possibility due to financial and resource allocation restrictions. A recent study by Jørgensen (2003) found that estimation accuracy improved when the estimator participated in project development. This is supported in previous research (Lederer and Prasad, 1993). However, one of the main problems associated with not using independent estimators is that personnel with high stakes in the project, e.g. project managers, may provide unrealistic estimates in order to get their project approved.

The second characteristic is a structured versus an unstructured approach. An elaborate method of reducing problems in groups related to company politics is to employ the Delphi technique (Helmer, 1966), which is often recommended in management papers (Fairley, 2002). The Delphi technique does not involve face-to-face discussion, but anonymous expert interaction through several iterations, supervised by a moderator until an agreed-on majority position is attained. A modification of this technique, which includes more estimation group interaction,

was developed by Barry Boehm and his colleagues, and labeled the Wideband Delphi technique (Boehm, 1981). This technique is a hybrid of unstructured groups and the traditional Delphi method. As in the Delphi technique, there is a moderator (labeled coordinator), that supervises the process and collects estimates. In this approach, however, the experts meet for group discussions both prior to, and during the estimation iterations. This approach has been suggested as an effort estimation method in books and articles on software metrics (Fenton, 1995), software process improvement (Humphrey, 1990), project management (Wiegers, 2000) and effort estimation (Hughes, 1996). The Estimeetings process (Taff et al., 1991) involves a series of group meetings, where at some stage, parts of the requirement specification is handed down to several individual estimators with in-depth knowledge of the problems. It is a complicated process that require several weeks. It was specifically designed for one large project, and we have not found any evidence that it has been used anywhere else.

To the best of our knowledge none of the Estimeetings, the Delphi or the Wideband Delphi techniques has been subject to extensive empirical research in a software engineering context during the last 25 years. We are aware of only one study that describes a company which employed the Delphi approach (Kitchenham et al., 2002a). By contrast, we know from experience that many software organizations apply unstructured group discussions in their work leading to software development effort estimates.

3. Hypotheses

Some of the scepticism towards group-based effort estimation may be attributed to misinterpretation of the research from other research areas, e.g. as presented in some introductory textbooks in psychology. Many of these are quite extensive in their coverage of the possible dangers of group processes (Aronson et al., 1999; Atkinson et al., 1996; Hewstone et al., 1996). Such books may be misleading, if the described results are not understood properly.

The terms group polarization and choice shift refer to two similar, and often confused, concepts in group psychology. These related concepts concern, among other things, how an initially optimistic or risky decision can be rendered even more optimistic or risky by group discussion. Zuber et al. (1992) define choice shift as the difference between the arithmetic average of the individual decisions and the group decision. Group polarization is defined as the difference between individual pre- and post-group discussion responses. Studies conducted on group decision making have found choice shift and group polarization effects in decisions made in a range of different areas, from burglary to management (Bem et al., 1965; Cromwell et al., 1991; Stoner, 1961; Wallach et al., 1964; Zuber et al., 1992). Some of this knowledge may have been oversimplified during the transfer from psychology to other professions, such as project management. The literature in these other professions often states the dangers of group interaction, but not why or in which situations these dangers apply. This may have resulted in an incomplete picture, in which

valuable group work aspects such as motivation (Rowe and Wright, 2001) and information sharing (Fairley, 2002) have been omitted.

We find evidence of this omission in software management textbooks, where both separate estimating groups (DeMarco, 1982) and methods such as the Wideband Delphi technique (Boehm, 1981; Fenton, 1995) are described as countering choice shift and group polarization, with no further explanations of the phenomena. For example, Boehm (1981) states that "... group members may be overly influenced by figures of authority or political considerations."

Fortunately, there are research communities that have investigated the different properties of group decision making. According to Kernaghan and Cooke (1990), the engineering management community has gradually accepted that the output of groups is likely to be superior to its average member. The forecasting community also constantly addresses best practices for combining expert judgments (Armstrong, 2001). Textbooks specialized in group psychology (Brown, 1988; Forsyth, 1999) also present a balanced view on when and how group processes can be beneficial, depending on the task and individuals involved.

A review of the literature on the Delphi technique in forecasting (Rowe and Wright, 2001) suggests that it, on average, outperforms unstructured groups in which group members discuss and interact freely. However, the review has also shown that there are tasks for which unstructured groups are better suited. In some situations, it is possible that extra information and group motivation exist in an unstructured group, and this can facilitate the process and enable it to surpass a Delphi group in performance (Rowe and Wright, 2001). Perhaps typical software estimation processes represent such situations. In an estimation process, there may be several experts who contribute different project experiences and knowledge. Such experiences can more easily be shared in a face-to-face group than through a moderator, as in the Delphi technique.

Earlier reviews of the literature and experiments have concluded that it seems to be less important which combination method, from a set of "meaningful" methods, is used (Fischer, 1981). It may, for example, not matter much whether simple averaging, unstructured groups, the Delphi technique or other combination methods are used. Other factors, such as cost and political issues, should determine which combination method to employ (Fischer, 1981).

Several software engineering textbooks (Boehm, 1981; Kitchenham, 1996) and papers (Boehm, 1984; Fairley, 2002; Hughes, 1996) point out that forgotten tasks are among the major obstacles to successful estimation by experts, especially when employing a bottom-up estimation approach, i.e., the decomposition of a project into activities and the estimation of each activity individually. A group approach to estimation will help to remove this obstacle, because several estimators will identify at least as many activities as the best single estimator alone. This will be especially true if estimates from experts with different company roles and experiences are combined in group discussion.

In sum, previous findings lead us to believe that unstructured groups can be used to reduce individual estimation optimism. The latter belief is evaluated through experimental testing of the following hypotheses:

H1. Group effort estimates are, on average, less optimistic than the average of the experts' individual effort estimates.

H2. Individual effort estimates are, on average, less optimistic after group discussion with other experts than before group discussion.

4. Research Method

The research presented in this paper attempts to follow the guidelines suggested by Kitchenham et al. (2002b), which includes specifying as much information about the organization, the participants and the experiment as possible, in addition to the complete experimental results. The rationale for including so much information is to ensure that the study is easy to replicate, and that the results are appropriately interpreted and transferred.

4.1. The Company Studied

The company studied is a web-development company, and operates as an independent contractor that develops a wide range of complete solutions for its customers. At the time of the study the company had about 70 employees. The employees (excluding administration and support staff) were allocated to the following four business roles: Engagement Manager/Sales and Client Responsibility (EM), Project Manager (PM), User Analyst/Designer (User) and Technical programmer (Tech). These roles are similar to those described by McDonald and Welland (2001). Average participant experience in the IT-business was 6.3 years, in their current role 2.7 years and with effort estimation 1.5 years. Half of the participants were educated to at least Master's level, while the rest had Bachelor's or comparable degrees.

The organization had no formal estimation procedure, their projects had short development cycles and the development processes involved were ad-hoc. These are all typical characteristics of web-development companies (McDonald and Welland, 2001; Reifer, 2000). The company based its estimates on expert judgment. A project estimate was most often provided by the person(s) responsible for the project. This is, to the best of our knowledge, a common situation in web-development companies, and small companies in general (Moses and Clifford, 2000).

4.2. The Estimation Task

Twenty participants were selected at random from the company; five from each of the four company roles (EM, PM, User and Tech). Each participant was required to estimate the most likely effort needed to complete a project, based on a requirement specification including some screenshots from a CD-ROM. The specification, i.e.,

the document describing the software to be developed, was taken from a project currently under development by the company. The complete requirement specification, as presented to the participants, is enclosed as Appendix I.

The project and the customers were anonymous, and none of the participants had any knowledge about them. The project was the development of a "publication solution" for a technical magazine. It was medium to large, compared to the average size of projects developed by the company. The estimation instructions stated that the participants in the experiment should behave as realistically as possible. The participants were also informed that a project crew had not been selected, and that they should base their estimates on average company productivity. These measures were undertaken in order to reduce any political biases (DeMarco, 1982; Thomsett, 1996), e.g. that the estimates would be influenced by a belief about what would be the right price to win the contract. What we wanted to investigate was the group discussion effect on the most likely estimate. During the sessions, the experimenter was present to answer questions and take notes.

The experiment had the following steps:

- The participants developed their individual estimates of the effort needed to produce the software during a 45–60 min estimation session.
- After the individual session, the participants formed groups, with one EM, PM, User and Tech in each group.
- Each group was given about 60 min to agree on an estimate for the same project as in the individual estimation session. These sessions were also videotaped to ensure thorough analysis of the group discussions.
- After the group decision, each participant was asked: "After an individual assignment and following teamwork, what is now your personal opinion about the estimation assignment."

During the experiment each participant completed two questionnaires, in which he provided background information about his experience, education, estimation training and skill level, and gave his comments on the experiment.

5. Results

The original individual estimates, the group estimates, and the individual opinions after group discussion are displayed in Table 1. There were no participant dropouts or incomplete responses. All statistical calculations were done with the package MINITAB[1] 13.3 for Windows.

The group estimate was less optimistic than the average expert opinion in four out of five groups. An analysis was performed with a paired t-test (Wonnacott and Wonnacott, 1990), as suggested in similar research on choice shift (Liden et al.,

Table 1. Individual pre-group (before), group and individual post-group (after) estimates.

	Group A		Group B		Group C		Group D		Group E	
Business role	Before	After	Before	After	Before	After	Before	After	Before	After
EM	1200	1000	1550	1500	1850	1500	547	1000	2286	2800
PM	960	1200	1820	1500	300	1550	914	1400	984	2200
User	1500	1200	1140	1500	1260	1500	620	1000	1500	2000
Tech	960	1000	585	1400	220	220	660	1000	900	2000
Average	1155	1100	1273.8	1475	907.5	1192.5	685.3	1100	1417.5	2250
Group	1100		1500		1550		1339		2251	
Actual effort					2365					

1999). Since the hypothesis suggests a direction of effect (groups are less optimistic than average individual experts), the paired *t*-test was one-sided. The result is displayed in Table 2.

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 average values, we included Cohen's size of effect measure (d) (Cohen, 1969), where $d =$ (average value group – average value individuals)/pooled standard deviation amongst groups and individuals. The analysis of possible choice shift on the estimates gave a discernible result ($p = 0.024$), and a large effect ($d = 1.25$).

Group polarization is, as described earlier, the difference between individual responses made before the group session and the responses made after the group session (Zuber et al., 1992). Both median and average individual estimates were less optimistic in the post-group answers than in the initial responses. The median values increased by 478 hours, from 972 to 1450. The average individual estimates increased by 336 hours, from 1088 to 1424. A one sided, paired *t*-test on the before and after values yielded a discernible result ($p = 0.003$), and a medium ($d = 0.62$) effect size (Table 3).

Table 2. Average values for shift between individual and group estimates.

	Group average	Individual average	Difference	Pooled StDev	*p*-value	Size of effect (d)
Estimates	1548	1088	460	368	0.024	1.25

Table 3. Average values, *p*-values and effect size (d) for a change in individual optimism.

	Average individual estimates after group discussion	Average individual estimates before group discussion	Difference	Pooled StDev	*p*-value	Size of effect (d)
Estimates	1424	1088	336	544	0.003	0.62

6. Discussion

The actual effort expended on the project was 2365 work-hours, that is, most participants gave an estimate that was much too low compared with the actual effort expended. However, the actual effort of a new project based on the same specification may have used less (or more!) effort than the completed project, i.e., there are many possible effort usage outcomes of a project based on the same specification. The actual effort expended to complete the project cannot, for this reason, be taken as more than an indication of estimation inaccuracy in this experiment. Nevertheless, based on the actual effort and the original company estimate (1240 hours), we believe that estimates of less than 1000 work-hours point to strong over-optimism. As can be seen in Table 1, 11 out of 20 of the original individual effort estimates indicated a workload of less than 1000 hours. The variation among the individual estimates shows that in reality, opinions regarding the same project may differ by a magnitude of up to ten!

There seems to be a tendency for both the group decisions and the individual post-group discussion decisions to be less optimistic than the original estimates. None of the group decisions, and only 1 out of 20 individual post-group estimates, was less than 1000 hours.

The experiment conducted has several properties that need to be viewed in light of previous research on group processes, applications for researchers and practitioners, and experimental validity.

6.1. Group Processes

When discussing the use of groups in the context of software effort estimation, it is essential to understand what kind of task effort estimation is. Effort estimation is a complex task in which certain properties, such as "quality", are difficult to measure. In his book on group processes, Rupert Brown (1988) differentiates between "group productivity" and "group decision making". In his view, group productivity concerns tasks where there exists a measurable performance, while group decision-making concerns tasks where it is impossible to measure performance.

Regarding group decision making, we have already described some of the previous research on choice shift and group polarization. The literature on group processes (Aronson et al., 1999; Atkinson et al., 1996; Brown, 1988; Forsyth, 1999; Hewstone et al., 1996) and software management (Boehm, 1981; DeMarco, 1982; Fenton, 1995), frequently warns about the possibility that group pressure (e.g. an unspoken "competition" to appear as the most risky or efficient programmer) and political preferences (e.g. a management that demands optimistic estimates), could influence group decisions unreasonably. Observation of post-group opinions that significantly differed from the group responses might indicate that the participants merely complied, for example, with authority. In our experiment, however, the individual post-group estimates were much closer to the group decisions than they were to the

original individual estimates. This, together with responses in the submitted questionnaires, indicates that the participants may have been mostly influenced by the arguments and extra information presented in the group discussion.

The nature of the task involved in our experiment differs significantly from those in most previous studies on group decision making, which typically ask groups of randomly selected people to decide on choice dilemmas, or professionals in an occupation to determine risk associated with different tasks. Those studies suggest that initially risky or optimistic decisions become more extreme after group discussion. Our study, however, yields the opposite result. Possible explanations from a group decision making perspective, is the diverse background of our participants, their shared commitment, and how the group politics were handled.

Our groups were able to identify more activities than did our individuals alone. It was explicitly reported in the questionnaire, by five of the participants in our experiment, that having forgotten activities was the main reason for their estimates being lower than the group estimates. The groups sometimes even identified necessary activities that none of the individuals had reported. In the experiment, none of the participants in each group had the same company role. This may improve the quality of the group (Fairley, 2002), as long as they share terminology and an understanding of the problem (Helmer, 1966). This applies to the participants in our experiment, since they frequently work on the same projects in the same company. Experts with different backgrounds consider the same project from different angles, and are probably able to identify more activities than experts with similar backgrounds. It is possible that groups of experts with similar backgrounds would show less reduction of optimism in their estimates (Jørgensen and Moløkken-Østvold, 2002).

During the video-analysis, we observed how all the groups behaved in a way which allowed the opinions of all participants to be heard and that differing views were discussed openly. They regarded the assignment from a professional point of view, and there was no apparent peer pressure to be either ''risky'' or ''conservative''. All groups resolved the task by a consensus estimate, and there were no incidents of a majority decision, e.g. through a vote. As seen in Table 1, most participants retained the group estimates when they ware asked to re-estimate the tasks individually after group discussion. This can indicate that their personal opinion had had been acknowledged by the group.

In accordance with Brown's productivity aspects, might it not also be possible to measure some kind of productivity in the estimation process? Thus far, we have, in the main, considered the process of estimating effort as a whole. Given the complexity of estimating the effort required to complete software development projects, an analysis of the different parts of the estimation process might yield a better understanding of how far productivity can be measured. Such analysis is, of course, a challenging endeavor and one upon which it would be foolhardy to embark in the present paper, but from a cursory examination it should be clear that some parts of the estimation process are measurable. For example, it would seem to be possible to measure the successful identification of the tasks that need to be performed to complete a software project.

The task of identifying project activities is related to the classic brainstorming process. A brainstorming process often involve several individuals seated together in order to generate input to a specific domain, e.g. a name for a new washing detergent. Brown (1988) reports studies that have found brainstorming procedures to be ineffective, due to both social and coordination problems. On the other hand, if the participants first prepared themselves, productivity increased significantly. Therefore, it is essential to consider not only what projects or part of projects to estimate in groups, but also how this group process is implemented. In our experiment, the individuals prepared themselves during the individual estimation. They first identified tasks they believed were necessary before the group discussion, during the individual estimation session. Since each group consisted of personnel with different roles, they also emphasized different activates. Later, in the group session, they were able to combine their activities, as well as identifying entirely new ones, in order to construct a complete project break-down. This break-down was then used as a basis for the group estimate.

In a more general sense, group productivity can exceed the sum of the individual performances if some kind of group motivation exists (Brown, 1988). In our experiment, and in actual development projects, the participants should be motivated to perform well together with their colleagues.

6.2. Implications for Researchers and Practitioners

It is difficult to discuss our findings in the light of actual estimation practices employed by software professionals. To our knowledge, no surveys have been conducted that analyze the extent to which groups of experts are used in effort estimation. A recent review of all known surveys on software effort estimation (Moløkken-Østvold and Jørgensen, 2003) found that expert judgment is by far the most common method used to estimate software projects. By comparison, formal models, such as COCOMO or FPA-based, are not used to any great extent. The surveys analyzed in the review, however, failed to elicit how expert estimates were made. None of them asked, for example, whether groups of experts were used at any stage of the process.

From all the surveys, experiments and case studies on software estimation that we reviewed, we found only one that identifies a formal group processes actually used in the industry (Kitchenham et al., 2002a). In a case study of 145 projects at an outsourcing company, it was found that three out of 145 projects were estimated by using the Delphi procedure. By comparison, 104 projects were estimated by the project manager alone, and were defined as expert opinion. We found no independent reports of actual use of the Wideband Delphi or Estimeetings procedures discussed earlier.

Even though there are few descriptions in the literature of group-based estimates (formal or informal) this does not necessarily mean that no companies conduct estimates in groups. From our experience as field researchers, and through review of

other studies, we have found that the terms "expert judgment", "expert decision" or "expert estimate" include a wide range of estimation procedures. These expert-based procedures may include varying degrees of group interaction, from none (e.g. a project manager estimates the whole project) to full group interaction (e.g. as in our study). In between these alternatives, there exist other methods to include several experts, e.g. a manager can ask subordinates to each estimate parts of projects, and the manager then adds the estimates for a project total.

In our study, the group-based estimates were less optimistic than the average estimate of individual experts. This may, however, not be the case for other companies or projects. Therefore, it is necessary that the individual companies themselves analyze their projects and decide which projects, and which stages in the estimation process, are suited for group collaboration. As discussed earlier, a typical stage at which the views of several experts may be beneficial is when the requirements are mapped onto different project activities.

The group process is probably more valuable than mechanical combination of estimates because then the experts discuss not only their estimates, but also the assumptions they made when calculating them (Winkler, 1989). Still, it is not unlikely that similar results would have been found had the unstructured group discussions been replaced with more structured group processes, e.g., the Wideband Delphi technique. The combination of several experts' opinions seems to be beneficial in most cases. The way in which expert opinions should be combined should be decided relative to estimation context factors, such as type of process and project (Fischer, 1981). Projects with high political stakes may not be suited for unstructured group estimation at all, and more formal procedures may be appropriate. Structured approaches, such as the Wideband Delphi technique, have qualities different from unstructured approaches (Rowe and Wright, 2001), and may, for example, be better suited when the level of personal disagreement is high, or political prestige is involved.

Companies may also develop their own Work Breakdown Structures (WBS) and/or estimation checklists to achieve some of the benefits that we found accrue from group discussion, e.g. individual estimators have a lower risk of forgetting activities when using an extensive WBS or high quality checklists. The use of a group collaboration estimation method can then be reserved for especially challenging projects, for example, projects involving new technology and business areas.

6.3. Experimental Validity

The main threat to validity of the study may be that the experimental setting distorts the realism of the estimation process. The following analysis, however, suggests that at least the outcome of the estimation process was realistic. The average project estimate provided by the experiment participants (1088 hours) did not deviate substantially from the actual effort estimate made by the company (1240 hours). We

must, however, be aware of the large individual deviation, with estimates from 220 to 2286 hours.

There are also validity threats related to the lack of customer contact and limited time available present in our study. For effort estimates applied as input to a bidding process, however, this was a common situation in the company, which developed more than 500 estimates in the year 2001.

The main problem regarding transfer of results to other organizations may be that the participants in the study were from the same company. The participants were so chosen for practical reasons, but 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).

It may be a threat to validity that only one project was used. The project chosen may, for example, have been especially difficult to estimate, and the optimism reduction may not have been of the same magnitude in other "easier" projects.

The impact from the threats to validity means that our findings may be applicable mainly to small companies with lack of formal estimating processes, when estimating more than average "difficult" projects with limited information and strong time restrictions, i.e., estimation situations typical for web-development companies when providing a project bid.

7. Conclusions and Further Research

The findings in our study are that group estimates and individual estimates after group discussion were less optimistic and more realistic than the individual estimates before group discussion. The main sources of this increase in realism seem to be identification of additional activities and an awareness that activities previously identified may be more complicated than was initially thought.

The unstructured group effort estimation approach presented here may be a simple and inexpensive approach for companies to improve the precision of their estimates.

Our contribution to the software industry is not to invent a new procedure, since group-based estimation procedures are probably employed by a wide range of companies all over the world. We seek, instead, to counterbalance the view on group processes presented in many papers and textbooks on software engineering. As mentioned earlier, group-based estimates are often described as dangerous, and often, these descriptions are not accompanied by any deeper explanations as to when and how such dangers applies.

Two topics for further research are comparison of the unstructured approach studied in this paper with more structured group processes, such as the Delphi (Helmer, 1966) or the Wideband Delphi (Boehm, 1981) techniques, and comparison of unstructured estimating approaches with WBS and checklist-supported estimates.

Acknowledgments

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

Appendix I:

User Requirements

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

The Customer

The customer is the producer of an established technical encyclopaedia that numbers 800 editions with 10,000 illustrations and 1600 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.

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.

Desired Functionality

This is the required functionality.

1. *Basis functionality (searching for and displaying information)*
   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.

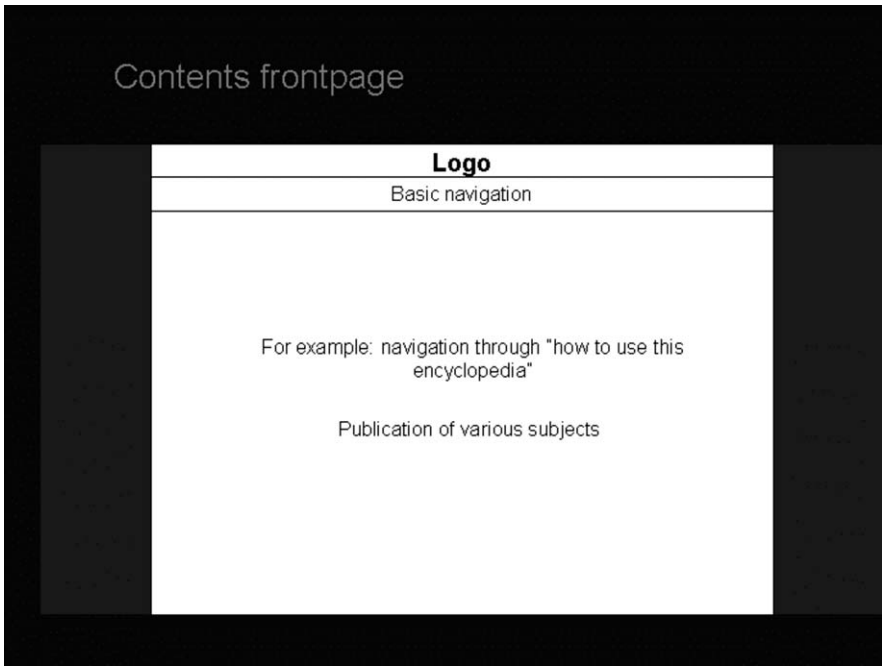

Figure 1. Front page example.

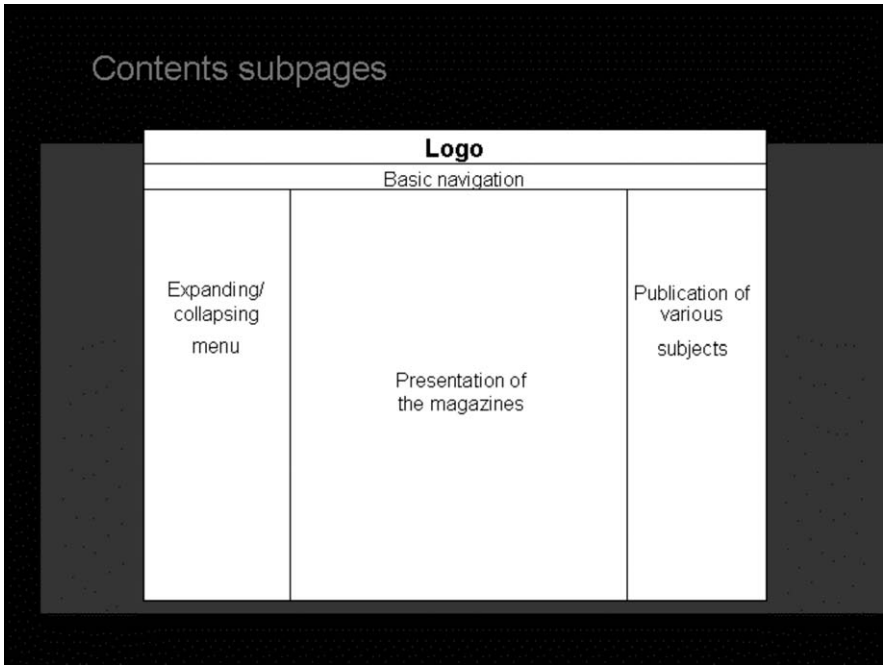

Figure 2. Sub page example.

References

Adams, S. 2002. *Dilbert and the way of the weasel*. HarperCollins Publishers Inc.

Armstrong, J. S. 2001. *Principles of forecasting*. Boston: Kluwer Academic Publishers.

Aronson, E., Wilson, T. D., and Akert, R. M. 1999. *Social Psychology*. Addison-Wesley Educational Publishers Inc.

Atkinson, R. L., Atkinson, R. C., Smith, E. E., Bem, D. J., and Nolen-Hoeksema, S. 1996. *Hilgard's Introduction to Psychology*. Orlando: Harcourt Brace College Publishers.

Bem, D. J., Wallach, M. A., and Kogan, N. 1965. Group decision making under risk of aversive consequences. *Journal of Personality and Social Psychology* 1(5): 453–460.

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

Boehm, B. W. 1981. *Software engineering economics*. New Jersey: Prentice-Hall.

Boehm, B. W. 1984. Software engineering economics, *IEEE Transactions on Software Engineering* 10(1): 4–21.

Brown, R. 1988. *Group Processes*. Cambridge: Blackwell Publishers.

Cohen, J. 1969. *Statistical power analysis for the behavioral sciences*. New York: Academic Press, Inc.

Cromwell, P. F., Marks, A., Olson, J. N., and Avary, D. W. 1991. Group effects on decision making by burglars. *Psychological reports* 69: 579–588.

DeMarco, T. 1982. *Controlling software projects*. New York: Yourdon Press.

Fairley, D. 2002. Making accurate estimates. *IEEE Software* 19(6): 61–63.

Fenton, N. E. 1995. *Software Metrics*. London: Thompson Computer Press.

Fischer, G. W. 1981. When oracles fail—A comparison of four procedures for aggregating subjective probability forecasts. *Organizational Behavior and Human Performance* 28(1): 96–110.

Forsyth, D. R. 1999. *Group Dynamics*. Wadsworth Publishing Company.

Gray, A., MacDonnell, S., and Shepperd, M. 1999. Factors systematically associated with errors in subjective estimates of software development effort: The stability of expert judgment. In *Sixth International Software Metrics Symposium*, pp. 216–227, IEEE Comput. Soc., Los Alamitos, CA, USA.

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

Helmer, O. 1966. *Social Technology*. New York: Basic Books.

Hewstone, M., Stroebe, W., and Stephenson, G. M. 1996. *Introduction to social psychology*. Oxford: Blackwell Publishers Ltd.

Hihn, J., and Habib-Agahi, H. 1991. Cost estimation of software intensive projects: A survey of current practices. In *International Conference on Software Engineering*, pp. 276–287, IEEE Comput. Soc. Press, Los Alamitos, CA, USA.

Hill, J., Thomas, L. C., and Allen, D. E. 2000. Experts' estimates of task durations in software development projects. *International Journal of Project Management* 18(1): 13–21.

Höst, M., and Wohlin, C. 1998. An experimental study of individual subjective effort estimations and combinations of the estimates. In *International Conference on Software Engineering*, pp. 332–339, IEEE Comput. Soc., Los Alamitos, CA, USA, Kyoto, Japan.

Hughes, R. T. 1996. Expert judgment as an estimating method. *Information and Software Technology* 38(2): 67–75.

Humphrey, W. S. 1990. *Managing the Software Process*. Addison-Wesley Publishing Company, Inc.

Jørgensen, M. 1997. An empirical evaluation of the MkII FPA estimation model. In *Norwegian Informatics Conference*, pp. 7–18, Tapir, Oslo, Voss, Norway.

Jørgensen, M. 2003. An attempt to model software development effort estimation accuracy and bias. In *Proceedings of Conference on Empirical Assessment in Software Engineering—2003 (EASE 2003)*, pp. 117–128, Keele, UK.

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., and Moløkken-Østvold, K. 2002. Combination of software development effort prediction

intervals: Why, when and how? In *Fourteenth IEEE Conference on Software Engineering and Knowledge Engineering (SEKE'02)*, pp. 425–428, Ischia, Italy.

Kernaghan, J. A., and Cooke, R. A. 1990. Teamwork in planning innovative projects: Improving group performance by rational and interpersonal interventions in group process. *IEEE Transactions on Engineering Management* 37(2): 109–116.

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

Kitchenham, B., Pfleeger, S. L., McColl, B., and Eagan, S. 2002a. An empirical study of maintenance and development estimation accuracy. *Journal of systems and software* 64: 55–77.

Kitchenham, B., Pfleeger, S. L., Pickard, L. M., Jones, P. W., Hoaglin, D. C., El Emam, K., and Rosenberg, J. 2002b. Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on software engineering* 28(8): 721–734.

Lederer, A. L., and Prasad, J. 1993. Information systems software cost estimating: A current assessment. *Journal of Information Technology* 8(1): 22–33.

Liden, R. C., Wayne, S. J., Sparrowe, R. T., Kraimer, M. L., Judge, T. A., and Franz, T. M. 1999. Management of poor performance: A comparison of manager, group member, and group disciplinary decisions. *Journal of Applied Psychology* 84(6): 835–850.

McDonald, A., and Welland, R. 2001. Web engineering in practice. In *Proceedings of the Fourth WWW10 Workshop on Web Engineering*, pp. 21–30.

Moløkken-Østvold, K., and Jørgensen, M. 2003. A review of surveys on software effort estimation. In *2003 ACM-IEEE International Symposium on Empirical Software Engineering (ISESE 2003)*, pp. 220–230, Italy: IEEE, Frascati—Monte Porzio Catone (RM).

Moses, J., and Clifford, J. 2000. Learning how to improve effort estimation in small software development companies. In *24th Annual International Computer Software and Applications Conference*, pp. 522–527, IEEE Comput. Soc., Los Alamitos, CA, USA, Taipei, Taiwan.

Paynter, J. 1996. Project estimation using screenflow engineering. In *International Conference on Software Engineering: Education and Practice*, pp. 150–159, IEEE Comput. Soc. Press, Los Alamitos, CA, USA, Dunedin, New Zealand.

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

Rowe, G., and Wright, G. 2001. Expert opinions in forecasting: The role of the Delphi process. In J. S. Armstrong (ed.), *Principles of forecasting: A handbook for researchers and practitioners*, pp. 125–144, Boston: Kluwer Academic Publishers.

Shepperd, M., Shofield, C., and Kitchenham, B. 1996. Effort estimation using analogy. In *International Conference on Software Engineering*, pp. 170–178, IEEE Comput. Soc. Press, Los Alamitos, CA, USA, Berlin, Germany.

Stoner, J. A. F. 1961. A comparison of individual and group decisions involving risks.

Taff, L. M., Borcering, J. W., and Hudgins, W. R. 1991. Estimeetings: Development estimates and a front end process for a large project. *IEEE Transactions on software engineering* 17(8): 839–849.

Thomsett, R. 1996. Double Dummy Spit and other estimating games. *American programmer* 9(6): 16–22.

Wallach, M. A., Kogan, N., and Bem, D. J. 1964. Diffusion of responsibility and level of risk taking in groups. *Journal of abnormal and social psychology* 68(3): 263–274.

Wiegers, K. E. 2000. Stop promising miracles. *Software Development Magazine* (February).

Winkler, R. L. 1989. Combining forecasts: A philosophical basis and some current issues. *International Journal of Forecasting* 5(4): 605–609.

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

Zuber, J. A., Crott, H. W., and Werner, J. 1992. Choice shift and group polarization: An analysis of the status of arguments and social decision schemes. *Journal of Personality and Social Psychology* 62(1): 50–61.

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.

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.