Eliminating Over-Confidence in Software Development Effort Estimates

Magne Jørgensen and Kjetil Moløkken
Simula Research Laboratory,
P.O.Box 134, 1325 Lysaker, Norway
{magne.jorgensen, kjetilmo}@simula.no

Abstract: Minimum-maximum effort intervals are applied in the planning of software development projects in order to, among other things, determine the contingency buffer. Several studies suggest that judgment-based minimum-maximum intervals are based on a systematic over-confidence in the accuracy of the effort estimates. In this paper, we investigate whether the possession by estimators of information about previous estimation error for similar projects reduces this over-confidence. Nineteen realistically composed estimation teams provided minimum-maximum effort intervals for the same software project. Ten of the teams (Group A) received no instructions about the uncertainty assessment process. The remaining nine teams (Group B) were instructed to begin the minimum-maximum effort interval assessment by recalling the distribution of estimation error for similar projects. We found that the recall of the error distribution of the Group B teams did have an impact, but mainly on the assessment of the estimated minimum effort, not on the maximum effort. We discuss reasons for this finding and provide recommendations.

1 Introduction

Assume that you are a project manager and ask one of the software developers to estimate the development effort for a task. The software developer believes that the task requires about 40 work-hours. You want to plan the project task with a low risk of over-run and therefore ask the developer to assess the minimum and maximum effort for the task. You instruct the developer to be "almost certain" that the actual effort will be included in his/her minimum-maximum intervals. The developer assesses the minimum to be about 35 and the maximum to be about 60 work-hours. How confident should you be in the accuracy of that minimum-maximum effort interval? A good guess, supported by the studies presented in *Section 2*, is that the actual probability of including the actual effort in the minimum-maximum interval is only about 60%, i.e., far from the desired confidence level of "almost certain". If you base your contingency buffer on the provided maximum effort there is, therefore, a high risk that your plan is too optimistic.

What should we do about this tendency to provide over-optimistic minimum-maximum effort intervals? This paper describes an attempt to improve the accuracy of the judgment-based minimum-maximum effort intervals through the use of historical uncertainty information. The background for our attempt is described in *Section 3*. The study itself is described in *Section 4*, discussed in *Section 5*, and, concluded in *Section 6*.

2 Studies on Over-Confidence

The accuracy of software development minimum-maximum effort intervals has only been studied by researchers in the last few years (the first study was published, as far as we know, in 1997). Nevertheless, there is strong evidence to suggest that there is a systematic tendency towards too narrow minimum-maximum intervals, e.g.:

- Connolly and Dean [1] report that the actual effort used by student programmers to solve programming tasks fell inside their 98% confidence minimum-maximum effort intervals only in about 60% of the cases, i.e., the intervals were much too narrow to reflect 98% confidence. Explicit attention to, and training in, establishing good minimum and maximum effort values increased the proportion inside the intervals to about 70%, which was still far from the required 98%.
- Jørgensen and Teigen [2] conducted an experiment in which 12 software professional were asked to provide 90% confidence minimum-maximum effort intervals on 30 previously completed maintenance tasks. In total, 360 minimum-maximum effort intervals were provided. The software professionals had access to a small experience database for similar projects and were informed about the actual effort of a task after each minimum-maximum effort interval assessment. Although "90% confident", the professionals included, on average, only 64% of the actual effort values on the first 10 tasks (Task 1-10), 70% on the next 10 task (Task 11-20), and, 81% on the last 10 tasks (Task 21-30). In other words, even after 20 tasks with feedback after each task, there was a systematic bias towards intervals that were too narrow.
- Jørgensen, Teigen and Moløkken [3] studied the software development minimum-maximum effort intervals of 195 student project activities and 49 industry project activities. The minimum-maximum effort intervals of the activities of the student projects were based on a 90% confidence level, and included only 62% of the actual effort values. The effort intervals of the activities of the industrial projects were based on a confidence level of being "almost certain", and included only 35% of the actual effort values, i.e., a strong underestimation of uncertainty.
- Jørgensen [4] studied the effort estimation intervals provided by seven realistically composed estimation teams. The teams assessed the uncertainty of the effort estimate for the same two projects, i.e., in total fourteen project minimum-maximum effort intervals. The actual projects had been completed in the same organization as the software professionals participating in the study. Only 43% of the teams' effort prediction intervals included the actual effort.

In addition to these studies, there are results about uncertainty assessment in other domains. Most studies from other domains seem to report levels of overconfidence similar to that in the software domain; see, for example, the studies described in [5-8]. Lichtenstein and Fischhoff [9] report that the level of over-confidence seems to be unaffected by differences in intelligence and expertise, i.e., we should *not* expect the level of over-confidence to be reduced with more experience. Arkes [10] provides a recent overview of studies on over-confidence that strongly supports the overconfidence claim.

3 Background

The focus of this paper is to investigate whether greater explicit use of the distribution of the estimation error for similar projects improves the minimum-maximum effort intervals. The idea behind of our distributional-based approach can be illustrated by the real-life case described in [11]:

A team that was concerned with the development of a high school curriculum on thinking under uncertainty was conducting a planning session. The question was raised of the time that would be required to complete the first version of a textbook. The participants in the discussion were asked to estimate this value as realistically as possible; the seven estimates ranged from 18 months to 3 years. The team leader then turned to one of the participants, an educator with considerable expertise in the problems of curriculum development, with the following question: 'What has been the experience of other teams that have tried to write a textbook and develop a curriculum in a new area where no previous course of study existed? How long did it take them to complete a textbook, from a stage comparable to the present state of our project?' The chilling implications of the answer appeared to surprise the expert who gave it, as much as they surprised the other participants: 'Most teams I could think of failed and never completed a textbook. For those who succeeded, completion times have ranged from five to nine years, with a median of seven.'

Kahnemann and Tversky [11] comment on this case: *A notable aspect of this anecdote is that the relevant distributional information was not spontaneously used, although it was available to one expert from personal knowledge and could have been estimated quite accurately by several other participants*. Their comment suggests that if we do not remind the estimators to use the historical performance as an indicator of future performance, it may not be used at all.

When viewed in the context of our software estimation uncertainty assessment, the above case suggests that accuracy may be increased when estimators are forced to focus explicitly on the distribution of estimation error of similar projects. Assume, for example, that an estimator is informed that about 50% of the projects similar to the one to be estimated have been subject to effort estimation overruns of more than 30%. This information should, given rational behavior, reduce the probability that the project leader provides strongly over-optimistic maximum values. In other words, it may be harder to remain over-optimistic about the estimation accuracy when possessed of explicit (pessimistic) historical information about previous estimation inaccuracy.

The outcome of explicit use of historical performance is, however, not obvious. There are examples suggesting that people may remain over-optimistic even in the light of historical performance data. The amusing study described in [12] exemplifies resistance to the application of historical data: *Canadians expecting an income-tax refund were asked to predict when they would complete and mail in their tax forms. These respondents had indicated that they typically completed this chore about 2 weeks before the due rate; however, when asked about the current year, they predicted that they would finish, on average, about 1 month in advance of the due date. In fact, only 30% of the respondents were finished by their predicted date - on average they finished, as usual, about 2 weeks before the deadline.* It is possible that the same tendency to neglect the relevance of historical data is present in the context of software development effort estimation uncertainty assessment.

A potential reason for over-optimism, despite the availability of relevant information about historical estimation error performance, is that people tend to apply an "inside view" when planning and estimating work instead of a history-based "outside view" [13]. People typically divide the work into phases and activities (the "inside" of a project) and estimate each of these phases, instead of comparing the current project as a whole (the "outside" of the project) with other projects. As suggested in the high school curriculum case described earlier this section, thinking based on the "inside view" can easily lead to estimates that are too low and views on the uncertainty that are too optimistic, while an "outside view" may increase realism. It is, consequently, not obvious when awareness of previous estimation error on similar projects does improve the realism. A first attempt to study the effect of information about previous estimation errors is described in Section 4.

4 A Study on Use of Estimation Error Distribution

4.1 *The Participants*

Nineteen estimation teams, all from a medium-large Norwegian company that develops, among other things, web solutions, participated in the study. The company manager in charge of all projects ensured that each team was realistically composed for the estimation work to be performed, i.e., that each team possessed the necessary skill and experience. The participants knew they were part of an experiment, but were instructed to behave as similarly as possible to the way in which they would have done in real estimation work. An important incentive for serious estimation work was that all the teams should present their estimation work in the presence of the other teams and the company's management.

4.2 *The Estimation Task*

The estimation teams were instructed to estimate the effort of a real project, which was to implement a web-based system for storing and retrieving research studies at our research laboratory (Simula Research Laboratory). Earlier we had received project bids from other web-development companies based on the same twelve-page requirement specification. The lowest bid reflected a use of effort of about 30 work-hours, the highest about 800 work-hours. The high variation in estimates was partly due to the flexibility in how the requirements could be implemented and partly to differences in development performance. Based on previous project bids we expected a high variation in the most likely estimates among the estimation teams.

4.3 *The Study Design*

The nineteen estimation teams were separated at random into two groups. The Group A teams received the estimation instructions:

a) Estimate the most likely effort of the project, assuming the normal productivity level of your company.
b) Assess the uncertainty of the estimate as % of most likely effort (most likely effort = 100%).

The format of the uncertainty assessment was as follows:
We are "almost sure" that the actual effort is between:
Minimum: _____ % of most likely effort
Maximum: _____ % of most likely effort

The Group B teams received the instructions:

a) Estimate the most likely effort of the project, assuming the normal productivity level of your company.
b) Recall what you believe is the "historical estimation error distribution" of similar projects in your company. Use the table below. *(The table instructed the estimation team to assess the frequency (in %) of projects with: More then 100% effort overrun, 50-100% effort overrun, 25-49% effort overrun, 10-25% effort overrun, +/- 10% effort estimation error, 10-25% too high effort estimates, 26-50% too high effort estimates, and, more than 50% too high effort estimates.)*
c) Assess the uncertainty of the estimate as % of most likely effort (most likely effort = 100%).

The only difference in the instructions was step b) for the Group B teams.

The teams spent between 1 and 1.5 hours on the estimation work, i.e., less than in a normal estimation situation. From a statistical point of view, the effect of spending less than the usual amount of time on estimation should be a widening of the minimum-maximum effort intervals, i.e., that the minimum-maximum intervals would indicate a higher uncertainty in the effort estimates.

The *hypothesis* that we tested in this study is that information about the distribution of estimation error for similar projects leads to minimum-maximum effort intervals that more accurately reflect the actual uncertainty of the effort estimates.

There is no obvious test of the accuracy of individual minimum-maximum effort intervals. "Almost sure" does not imply that *every* minimum-maximum effort interval should include the actual effort. Assume, for example, that the actual effort turns out to be higher than the maximum effort in a project where the estimator was "almost sure" to include the effort in the minimum-maximum effort interval. The estimator may still have provided a proper uncertainty assessment. It may, for example, be the case that the project experienced an unusual amount of bad luck and that the actual effort in 99 out of 100 similar project executions would have fallen inside the interval. In other words, the optimal evaluation of minimum-maximum intervals would be based on a large set of effort minimum-maximum intervals. This type of evaluation is not possible in the current study. Instead, we had to base our evaluation of the minimum-maximum intervals on the assumption that the actual uncertainty of the team's effort estimates is close to or higher[1] than that reflected in the estimation error for previous, similar projects. For example, if among similar projects there had been a 10% frequency of projects with more than 100% effort over-runs, a 90% confidence maximum value should be 200% of the estimated most likely effort. The study [11] described in Section 3, and the results reported in study [14],

[1] Reflecting the limitations regarding time spent on the estimation task.

suggest that the empirical error distribution is, in many cases, a good indicator of future effort estimation uncertainty. Even if the assumption were not true in our experimental context, the results from the study could be useful for analyzing how the uncertainty assessments are impacted by information about previous estimation error, e.g., to analyze differences in how the error distribution information is applied in the minimum and maximum assessment situations.

4.4 *The Results*

Table 1 shows the estimates of most likely effort, minimum effort, and maximum effort (minimum and maximum in % of most likely effort). Table 2 shows the estimation error distributions provided by the estimation teams in Group B.

Table 1: Estimated, Minimum and Maximum Effort

Team	Group	Estimate (work-hours)	Minimum (% of Estimate)	Maximum (% of Estimate)
1	A	800	90	150
2	A	413	75	130
3	A	750	50	200
4	A	5200	50	120
5	A	136	66	150
6	A	1543	65	130
7	A	192	90	115
8	A	149	90	150
9	A	310	60	200
10	A	152	70	150
Mean values – A		**965**	**71**	**150**
11	B	484	70	120
12	B	2280	90	160
13	B	2080	100	150
14	B	752	90	200
15	B	456	75	150
16	B	2640	110[2]	150
17	B	800	90	125
18	B	88	90	140
19	B	820	90	120
Mean values – B		**1160**	**88**	**146**

[2] Statistically, a minimum of 110% of most likely effort is meaningless and should lead to an increase of most likely effort. Estimation team 16 defended the minimum value by pointing to the fact that they never had observed any similar project that used less than 110% of the estimated most likely effort. This viewpoint is, to some extent, valid when not separating estimates of most likely effort and planned effort. See the discussion in Jørgensen, M. and D.I.K. Sjøberg, *Impact of effort estimates on software project work.* Information and Software Technology, 2001, **43**(15), p. 939-948.

Table 2: Distribution of Estimation Error of Similar Projects

Teams (Group B only)										
Estimation Error Category	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**Mean value**
>100% overrun	45	18	10	10	10	5	10	0	18	14
50-100% overrun	20	40	35	20	10	5	20	5	25	20
25-49% overrun	15	22	25	30	30	35	40	20	30	27
10-24% overrun	10	15	25	20	30	45	20	40	15	24
+/- 10% of error	7	4	0	5	10	10	10	20	12	10
10-25% too high estimates	3	1	0	10	5	0	0	10	0	3
24-50% too high estimates	0	0	0	0	5	0	0	5	0	1
>50% too high estimates	0	0	0	0	0	0	0	0	0	0

The large difference in estimates indicates that the estimation teams interpreted, and intended to provide solutions for, the requirements quite differently. This high variation in interpretation of the requirement specification also means that we should analyze the estimates carefully. In principle, we should treat the data as the estimation of 19 different projects, i.e., we should only compare the minimum-maximum effort values of one team with the same team's distribution of error for similar projects. However, since the variation of effort estimates is similar in Groups A and B, we assume that the "average interpretation" in Group A and Group B is similar, i.e., that it is meaningful to compare the mean minimum and maximum values of the Group A and Group B estimation teams.

We instructed the estimators to be "almost certain" to include the actual effort in their minimum-maximum intervals. It is not clear how to translate the concept of being "almost certain" into a statement of probability. However, we believe it is safe to state that being "almost certain" should correspond to there being at least an 80% probability that the actual effort is included in a minimum-maximum effort interval. We further assume a symmetric error distribution and base our interpretation of "almost certain" on a 10% probability of actual effort higher than the maximum effort and a 10% probability of actual effort lower than the minimum effort. These assumptions are open to discussion, but it makes no great difference to the results when changing the "almost certain" interpretation to, for example, a 5% probability of actual effort higher than the maximum and lower than minimum effort. Table 3 applies the above interpretation of "almost certain" and shows the minimum and maximum effort values we would expect if the estimation teams had based their uncertainty assessments on the historical distribution of estimation errors for similar projects. An example to illustrate the calculations is this. Estimation team 13 assessed that 10% of similar projects had more than 100% effort overrun. (see Table 2.) We therefore calculated the maximum effort to be 200% of the most likely effort.

The wide uncertainty categories of the estimation error distribution, e.g., +/- 10% error, means that it is not always obvious which value to choose as a history-based minimum and maximum. Minor deviations between our interpretation of the

history-based and the actual minimum and maximum effort values do not, therefore, necessarily reflect little use of the estimation error distributions.

Table 3: History-Based and Actual Minimum and Maximum

Teams (Group B only)									
Teams	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
History-Based Minimum	90%	100%	100%	90%	90%	100%	90%	90%	90%
Estimated Minimum	70%	90%	100%	90%	75%	110%	90%	90%	90%
Correspondence minimum?	Close	Close	OK	OK	Close	OK	OK	OK	OK
History-Based Maximum	200%	200%	200%	200%	200%	175%	200%	140%	200%
Estimated Maximum	120%	160%	150%	200%	150%	150%	125%	140%	120%
Correspondence maximum?	No	No	No	OK	No	Close	No	OK	No

Table 3 suggests that all the Group B teams achieved a correspondence between the history-based and their actually estimated *minimum* effort. Comparing the mean minimum effort values of the two groups of estimation teams, see Table 1, shows that the teams in Group B have a minimum that is higher (mean value 88% of most likely effort) than that of the teams in Group A (mean value 71% of most likely effort). The difference in mean values, combined with the results in Table 3, suggests that the minimum values of the estimation teams in Group B were influenced by the error distribution for similar projects. For example, being aware that no similar project ever had been over-estimated by more than 10% (teams 16 and 17) seems to have made it difficult to argue that the minimum effort should be lower than 90% of most likely effort. We interpret this as meaning that the minimum values of the Group B teams were more accurate than those of the Group A teams.

Surprisingly, the same impact from the historical error distribution was not present when estimating the *maximum* effort. Only three of the Group B teams estimated maximum effort values close to the history-based maximum. Comparing the mean maximum effort the two groups of estimation teams shows that the teams in Group B had a mean maximum value similar (mean value of 146%) to that of the teams in Group A (mean value of 150%). In other words, while the estimation teams seemed to have applied the distribution of estimation of similar projects when estimating the minimum effort, only a few of them applied the same information when estimating the maximum effort. We discuss reasons for, and threats to the validity of, these results in Section 5.

5 Discussion of the Results

There are several potential reasons explaining the problems of applying historical error data when providing realistic minimum-maximum effort intervals:

- *Conflicting goals*: A potentially important reason for resistance towards providing sufficiently wide minimum-maximum effort intervals is reported in [3]. One of the software professionals participating in that study stated: *"I feel that if I estimate very wide effort minimum-maximum intervals, this indicates a total lack of competence and has no informative value for the project manager. I'd rather have fewer actual values inside the minimum-maximum interval, than providing meaningless, wide effort intervals"*. In the same study it was evaluated how project managers actually assessed the skill of software developers and it was found that they did, indeed, evaluate those software developers as more skilled who provided narrower intervals and exhibited higher confidence. Interestingly, the evaluation of skill based on the width of the interval persisted even in situations when the managers received the information that the assessments were strongly over-confident. In other words, there seems to be an immediate and assured reward for over-confidence. It follows, therefore, that over-confidence does not necessarily lead to punishment (one may be lucky or the management may choose not evaluate the minimum-maximum intervals). In addition, potential punishment is in most cased delayed until the project is completed. The use of historical data to increasing the accuracy of uncertainty assessments may therefore be hindered by the goal of appearing skilled.
- *The "better-than-average"-bias*: Assume that the estimation teams accepted that previously completed similar projects had an error distribution that suggested a higher maximum value than the one they provided. The estimation teams may nevertheless believe that the estimation error history is not relevant for the uncertainty assessment of their effort estimate. In some cases this may be a valid belief. The organization may, for example, have learned from previous experience how to improve the estimates or reduce the project uncertainty. Another possibility is, however, that the estimation teams are subject to the well-known "I-am-better-than-average"-bias [16], i.e., that most teams believe that their effort estimates are better than those of other teams.
- *The lack of statistical skill:* The minimum-maximum assessment instructions given to the Group B estimation teams did not explicitly state how to apply the distribution of previous estimation error. It is, therefore, possible that some of the teams had problems understanding how to link the estimation error distribution to minimum and maximum effort values. A lack of statistical skill does, however, not explain the difference in the impact of the historical information about error distribution on the assessments of minimum and maximum effort and can only explain a minor portion of the problem.

In this study we have used software professionals, realistically composed estimation teams, and a real-life requirement specification. However, there are at least two important limitations to the validity of the results, both related to the artificiality of the estimation context:

- Unrealistic *time restrictions* mean that we cannot automatically generalize the findings into contexts in which the teams spend much more effort on the estimation tasks. Based on the results in [4], however, we would expect greater knowledge about how to solve a task to lead to more, not less, over-confidence.

- The experimental context may have led to a *greater focus on appearing skilled* than in a realistic minimum-maximum assessment context. The estimation teams knew that they would not have to implement the project and may therefore have focused more on appearing skilled when presenting the results to the other teams and the management. This limitation does, however, not explain the difference in use of the historical distribution of estimation error when deriving the minimum and the maximum effort.

Taking into consideration the potential explanations and major limitations of our study, we believe that our most robust finding is the difference in the use of historical information when providing minimum and maximum effort, i.e., that it seems to be more difficult to accept the relevance of historical estimation performance when assessing maximum (worst case) compared with minimum (best case) use of effort. A potential consequence of that finding is described in Section 6.

6 Conclusion and Further Work

The experiment described in this paper is a first step towards better processes for effort estimation uncertainty. Our results indicate that in the attempt to ensure history-based maximum effort values, it is not sufficient to be aware of the estimation error distribution of similar projects. Based on the identification of potential reasons for not applying the estimation error distribution we recommend that the uncertainty assessment process be extended with the following two elements:

- More detailed instructions on how to apply the distribution of previous estimation error to determine minimum and maximum effort for a given confidence level.
- The presence of an uncertainty assessment process facilitator who ensures that the distribution of previous estimation error is properly applied. The facilitator should be statistically trained in uncertainty distributions and require that deviations from the history-based minimum and maximum are based on sound argumentation.

We intend to introduce and evaluate the extended history-based uncertainty assessment process in a software organization.

References

1. Connolly, T. and D. Dean, *Decomposed versus holistic estimates of effort required for software writing tasks.* Management Science, 1997. **43**(7): p. 1029-1045.
2. Jørgensen, M. and K.H. Teigen. *Uncertainty Intervals versus Interval Uncertainty: An Alternative Method for Eliciting Effort Prediction Intervals in Software Development Projects*. In *International conference on Project Management (ProMAC)*. 2002. Singapore: p. 343-352.
3. Jørgensen, M., K.H. Teigen, and K. Moløkken, *Better Sure than Safe? Overconfidence in Judgment Based Software Development Effort Prediction Intervals.* To appear in: Journal of System and Software, 2004.

4. Jørgensen, M., *Top-Down and Bottom-Up Expert Estimation of Software Development Effort.* To appear in: Journal of Information and Software Technology, 2004.
5. Alpert, M. and H. Raiffa, *A progress report on the training of probability assessors*, in *Judgment under uncertainty: Heuristics and biases*, A. Tversky, Editor. 1982, Cambridge University Press: Cambridge. p. 294-305.
6. Kahnemann, D., P. Slovic, and A. Tversky, *Judgement under uncertainty: Heuristics and biases*. 1982: Cambridge University Press.
7. Tversky, A. and D. Kahneman, *Judgment under uncertainty: Heuristics and biases.* Science, 1974. **185**: p. 1124-1130.
8. Yaniv, I. and D.P. Foster, *Precision and accuracy of judgmental estimation.* Journal of behavioral decision making, 1997. **10**: p. 21-32.
9. Lichtenstein, S. and B. Fischhoff, *Do those who know more also know more about how much they know?* Organizational Behaviour and Human Decision Processes., 1977. **20**(2): p. 159-183.
10. Arkes, H.R., *Overconfidence in judgmental forecasting*, in *Principles of forecasting: A handbook for researchers and practitioners*, J.S. Armstrong, Editor. 2001, Kluwer Academic Publishers: Boston. p. 495-515.
11. Kahneman, D. and A. Tversky, *Variants of uncertainty*, in *Judgment under uncertainty: Heuristics and biases*, D. Kahneman, P. Slovic, and A. Tversky, Editors. 1982, Cambridge University Press: Cambridge, United Kingdom. p. 509-520.
12. Griffin, D. and R. Buehler, *Frequency, probability, and prediction: Easy solutions to cognitive illusions?* Cognitive Psychology, 1999. **38**(1): p. 48-78.
13. Kahneman, D. and D. Lovallo, *Timid choices and bold forecasts: A cognitive perspective on risk taking.* Management Science, 1993. **39**(1): p. 17-31.
14. Jørgensen, M. and D.I.K. Sjøberg, *An effort prediction interval approach based on the empirical distribution of previous estimation accuracy.* Journal of Information and Software Technology, 2003. **45**(3): p. 123-136.
15. Jørgensen, M. and D.I.K. Sjøberg, *Impact of effort estimates on software project work.* Information and Software Technology, 2001. **43**(15): p. 939-948.
16. Klein, W.M. and Z. Kunda, *Exaggerated self-assessments and the preference for controllable risks.* Organizational behavior and human decision processes, 1994. **59**(3): p. 410-427.