Realism in Assessment of Effort Estimation Uncertainty: It Matters How You Ask

Jørgensen, M.

Simula Research Laboratory

P.O. Box 134, NO-1325 Lysaker, Norway

magne.jorgensen@simula.no

Abstract: Traditionally, software professionals are requested to provide minimum-maximum intervals to indicate the uncertainty of their effort estimates. In this paper we claim that the traditional request is not optimal and leads to over-optimistic views about the level of estimation uncertainty. Instead, we propose, a person different from the estimator should identify minimum and maximum values relevant for planning or bidding purposes and request that the software professionals assess how likely it is that these values are exceeded. Not only does this seem to increase realism, but it also leads to more useful uncertainty assessments. Our claims are based on the results of a previously reported experiment and field studies in two companies. The two software companies were instructed to apply the traditional, and our alternative, framing on random samples of their projects. In total, we collected information about 47 projects applying the traditional framing and 23 projects applying the alternative framing.

Keywords: D 2.9.b Cost Estimation, D 2.9.m Risk Management, D 2.0.b Software Psychology.

1. BACKGROUND

Nothing is more certain than uncertainty. (Anonymous)

Assume that you are asked to provide a bid for a software project. You organize a team of experienced developers and project managers. The team believes that the most likely effort to complete the project is about 1.400 work-hours. Is this sufficient information for proper bidding? Probably not. In addition, you should know the level of *uncertainty* present in the effort estimation. Without that information you would not know how likely you are to lose money on the project, given different bids. Similarly, if you were a project manager you should know the level of uncertainty so that you could determine the size of the project's contingency buffer, i.e., the buffer for dealing with unexpected events. A relevant question in these contexts is: How should you frame the request for uncertainty information from the estimators? It should be no surprise that the following framing is not optimal: *"You don't believe it possible that the project will spend more than maximum 1.700 work-hours, do you?"*. This type of framing may induce an "anchoring effect" [1] and increase the social pressure towards over-optimism [2]. Apart from this and other obviously non-optimal framings, does it really matter how you frame the request?

In this paper we provide empirical evidence that supports the view that the framing does matter and that the traditional minimum-maximum interval framing should be replaced. Instead of asking the estimators to provide minimum and maximum effort values based on given confidence levels, e.g. "almost sure " or "90% confident", it seems to be better to ask them to assess the probability of the actual effort being higher or lower than a value set by a person other than the estimator. For example, our results suggest that it is better to ask the

estimator “How likely is it that the project will require more than 2.000 work-hours?” rather than “What is the maximum cost of the project? Base the maximum effort on a 90% confidence level.” The latter framing is, as far as we have observed, in frequent use and the one suggested by the popular project management technique sometimes referred to as the “PERT System of Three Time Estimates” (PERT = Project Evaluation and Review Technique) [3]. The first framing of the question not only leads to greater realism, but also seems to be able to provide greater amounts of useful information, for example, about the probability of making losses given a particular bid.

Our research on effort estimation uncertainty assessments was carried out in several steps:

- First we tried to identify the size of the systematic over-confidence and to understand the reasons for it [4]. We found that the level of over-confidence was high. Typically, when the estimators claimed to be 90% confident, their minimum-maximum intervals included only about 60% of the actual effort values. Results from our investigations into the reasons for the over-confidence are included in Section 4 of this paper.
- Our next step was to develop and evaluate different variants of formal effort estimation uncertainty models that are designed to *replace* the expert judgment [5, 6]. Several of the formal models did remove the over-confidence, but were “less efficient”. To illustrate what we mean by “less efficient”, consider two software developers A and B. Both of them are supposed to estimate the effort and assess the uncertainty of the estimate on the same project. Developer A is inexperienced and does not possess much information pertaining to uncertainty, while Developer B is experienced and has completed a risk analysis of the project, so that he or she knows a great deal about the uncertainty of the project. To compensate for lack of skill and information, Developer A has to provide wider minimum-maximum intervals to include the actual effort with the same confidence as Developer B. Similarly, to achieve the same level of confidence,

formal estimation models seemingly had to provide much wider intervals than uncertainty estimates based on expert judgment. Our conclusion from our own and other researchers' effort uncertainty model-building attempts, e.g., [7], was that current formal models for the uncertainty of effort estimation were not capable of replacing expert judgment, and that a more promising approach was to support expert judgments of uncertainty instead of replacing them with formal models.

- As a consequence of that conclusion, we evaluated several strategies for judgment support in student experiments. One of these experiments, evaluating the framing variant described in this paper, gave promising results and we replicated the experiment with software professionals [8]. The framing results were robust to the change from students to software professionals. A brief description of that experiment is provided in Section 2 of this paper.
- Experimental results do not always generalize to industrial contexts, even when recruiting software professionals instead of students as participants. For this reason, we evaluated the uncertainty framing in a full-scale industrial context. The design and results of that study are described in Section 3 of this paper.

2. Increased Realism and Faster Learning [8]

In the experiment reported in [8] we evaluated experimentally two framings (formulations) of the request for the assessment of effort estimation uncertainty. Following an estimation of the most likely effort required for a software development task we asked half of the participants to respond to the TRADITIONAL framing:

Assess the minimum and maximum effort so that the probability of including the actual effort is believed to be about 90%.

The other half responded to the ALTERNATIVE framing:

Assess the probability of including the actual effort in the minimum-maximum interval in the interval [50% of estimate; 200% of estimate] (The 50% and 200% values were selected based on the recommendations in NASA: Manager's handbook for software development [9] and were intended to represent high confidence minimum and maximum values.)

We paid 29 experienced software developers and project managers of a Norwegian e-commerce software development organization to participate in the experiment. The participants were divided randomly into two groups. The members of one group applied the TRADITIONAL framing, and those of the other group the ALTERNATIVE framing on all estimation tasks. The experiment itself was divided into two parts:

1) TRAINING: Instructions and training in use of the uncertainty assessment process. As training tasks the effort and the effort uncertainty of 10 real-world software development projects were estimated. The estimation process was based on expert judgment and supported by an "experience database" that included information about five similar projects. We provided feedback, in the form of the actual effort, after each estimate. When the 10 estimates were completed, the software developers were asked to analyse and reflect on their own performance, with particular reference to the correspondence between their confidence level and the actual inaccuracy of the estimates.
2) ESTIMATION: Estimation of 30 software enhancements tasks previously conducted in a large telecommunications company. Task 1 was estimated based on a simple "experience database" that included information about five previously completed

tasks; Task 2 was estimated based on the five tasks and the feedback, i.e., the actual effort, on the Task 1 estimate; Task 3 was estimated based on the five tasks and the feedback on the Tasks 1 and 2 estimates, and so on. All estimates and uncertainty assessments were based on expert judgment, i.e., there was no use of tools other than that of a spreadsheet supporting simple calculations.

The estimation context was, to some extent, different from a realistic context. In particular, the estimators would have had more relevant information about the tasks than we provided, spent more time on the estimation tasks, and had greater incentives for accuracy than in an experiment.

A summary of the experimental results is described in Table 1, where "Hit rate" measures the proportion of minimum-maximum intervals including the actual effort, and "Av.Conf." measures the average confidence level of a group's uncertainty assessments. A "Hit rate" similar to "Av.Conf." indicates good correspondence between the assessed uncertainty and the actual level of accuracy. The measures are further described in Section 3.

Table 1: Results from Experiment

	Training tasks		Task 1-10		Task 11-20		Task 21-30	
Group	Hit rate	Av.Conf.	Hit rate	Av.Conf.	Hit rate	Av.Conf.	Hit rate	Av.Conf.
TRADITIONAL	0.58	0.90	0.64	0.90	0.70	0.90	0.81	0.90
ALTERNATIVE	0.83	0.83	0.67	0.73	0.69	0.71	0.73	0.71

As can be seen from Table 1, the participants who received the TRADITIONAL framing approached, very slowly, an acceptable correspondence between "Hit rate" and average confidence level. The participants who followed the ALTERNATIVE framing, on the other

hand, had a close to perfect correspondence between "Hit rate" and average confidence level on all sequences of tasks.

3. Evaluating the Framing in an Industrial Context

We replicated, with minor changes, the experiment in a full-scale industrial context, i.e., we compared the uncertainty assessment framings when applied in the software professionals' normal estimation work.

Participants

Two medium-sized Norwegian software development companies, Companies X and Y, participated in the study. Neither of the organizations had implemented any formal process for software effort or uncertainty estimation. The companies' projects and employees were similar, as far as we observed. Over a period of approximately eighteen months we collected information about development projects with an estimated effort of more than ten work-hours and duration of less than approximately eight calendar months. In total, information about 70 projects was collected, with 62 projects from Company X and eight projects from Company Y.

Design of Study

In the projects' estimation phase, the estimators completed a questionnaire that asked for information about the effort estimates. In addition, we instructed the estimators to provide the effort estimation uncertainty assessment with either the TRADITIONAL or the ALTERNATIVE framing. The allocation of uncertainty framing to the projects was based on a *random* selection process controlled by the "Chief Project Manager" in Company X and a person with a similar role in Company Y. In total, 23 of the 70 projects were allocated the

ALTERNATIVE framing, and the remaining 47 projects to the TRADITIONAL framing. The uneven distribution between the two framings, i.e., 23 versus 47 projects, was caused by circumstances not affecting, as far as we are aware, the randomness in the selection process. Immediately after a project was completed the estimator completed a questionnaire that requested information about the actual effort and other information relevant to estimation accuracy.

The TRADITIONAL framing was identical to that in the previous experiment [8], i.e.,

Assess the minimum and maximum effort so that the probability of including the actual effort is believed to be about 90%.

We made a minor change in the ALTERNATIVE framing. We divided the original ALTERNATIVE framing: "Assess the probability of including the actual effort in the minimum-maximum interval in the interval [50% of estimate; 200% of estimate]" into two separate assessments:

a) Assess the probability of the actual effort being more than 50% of the most likely effort (MinConf)*, and*

b) Assess the probability of the actual effort being less than 200% of the most likely effort (MaxConf).

This change in the ALTERNATIVE framing provided two uncertainty assessments per project instead of one, i.e., there were, in total, 46 uncertainty assessments following the ALTERNATIVE framing in our dataset. In addition, the change enabled the estimator to

provide asymmetric assessments. For example, an estimator could assess that she/he was 99% certain that more than 50% of the most likely effort would be expended, but only 70% certain that less than 200% would be required. The change led, in our opinion, to more detailed assessments and better evaluation, while retaining the essence of the ALTERNATIVE framing.

Measures

How to measure the realism of uncertainty assessment is a complex topic, e.g., while estimated effort can be compared with the actual effort, a single uncertainty assessment typically has no obvious reference value. When evaluating a set of uncertainty assessments, however, there are references points related to *bias* and *usefulness* as indicators of estimation accuracy.

To illustrate these reference points, assume that an organization provides minimum-maximum effort intervals for 100 projects, and that each of the intervals is based on 90% confidence. With 90% confidence we expect that 90% of the actual effort values are included in the minimum-maximum intervals. With less than 90% of the actual effort values included, the uncertainty assessments are biased towards "over-confidence" and with more than 90% biased towards "under-confidence". Now, assume that some of the minimum-maximum intervals are wide, e.g., minimum = 100 work-hours and maximum = 800 work-hours, and others narrow, e.g., minimum = 480 work-hours and maximum 520 work-hours. In the long run, we would expect that the estimate of most likely effort would be more accurate in the narrow interval situation, e.g., we would expect wider minimum-maximum intervals to correlate with more inaccurate estimates of most likely effort. For further discussion on other uncertainty assessment evaluation measures, see [4].

We evaluate the **bias**, as in the previous experiment, by a comparison of the average confidence level and the "Hit rate" of each framing approach, where "Hit rate" is defined as:

$$HitRate = \frac{1}{n}\sum_i h_i, \quad h_i = \begin{cases} 1, & \min_i \le Act_i \le \max_i \\ 0, & Act_i > \max_i \vee Act_i < \min_i \end{cases},$$

where min_i and max_i are, respectively, the minimum and maximum values of the prediction interval for the effort estimate of project *i*, Act_i is the actual estimate of project *i* and *n* is the number of estimated projects.

A close correspondence between average confidence level and "Hit rate" does not establish that each *individual* uncertainty assessment is well calibrated. For example, if one assessment is strongly over-confident and another strongly under-confident, the *average* confidence level of these two assessments may still reflect the average level of uncertainty.

As stated earlier, we would expect a correlation between the "wideness" of minimum-maximum intervals and the size of the estimation error, i.e., that not only the average confidence level is appropriate but also that the assessed differences in level of uncertainty among the projects reflects the actual difference in level of estimation accuracy. For this purpose, we introduce a measure of the relative width (RWidth) of a minimum-maximum interval and a measure of the magnitude of relative estimation error (MRE), where

$RWidth_i = (Max_i - Min_i) / Est_i$, where Est_i is the estimated (most likely) effort of project *i*, and

$MRE_i = |Est_i - Act_i) / Act_i$, where Est_i is the estimated (most likely) effort of project *i*.

We evaluate the **usefulness of RWidth as an indicator of estimation error** through the correlation (r-TRAD) between RWidth and MRE for the results from the TRADITIONAL framing. When following the ALTERNATIVE framing in our study, however, the RWidth is constant, i.e., the minimum is always 50% and the maximum always 200% of the most likely effort. For this reason we evaluate the usefulness of *Confidence level* as an indicator of estimation error for the results from the ALTERNATIVE framing. We would expect that a higher confidence in that the actual effort is included in an interval correlates with lower estimation error. We measure this using the correlation (r-ALT) between the average value of MinConf and MaxConf, and MRE. We expect a negative r-ALT, because high confidence should be correlated with low estimation error.

In short, the measure of bias allows us to evaluate the ability to assess the *overall* level of estimation uncertainty. The measure of usefulness as an indicator of estimation accuracy allows us to evaluate the ability to assess the *relative* difference in estimation uncertainty between projects. In the previous experiment, none of the framings were good at assessing the relative difference in estimation uncertainties, i.e., the correlations to MRE were close to zero. We attribute this to the homogeneity of the tasks and the limited task information in the experiment, and expect higher correlations in this study.

Results

Table 2 provides several indicators relevant for comparing the uncertainty assessments of the two processes. We found no systematic uncertainty assessment differences between the two companies (X and Y) and combined all the projects into one dataset.

Table 2: Results from Field Study

	TRADITIONAL framing	**ALTERNATIVE framing**
Hit Rate	74%	87%
Average Confidence Level	90%	88%
Bias (= Hit Rate - Average Confidence Level)	16% over-confidence	1% under-confidence
Indicator of estimation error	0.26 (r-TRAD)	-0.26 (r-ALT)

Notice that the estimators applying the TRADITIONAL framing only had an impact on the "Hit rate", since the confidence level was set at 90% in the uncertainty assessment framing. The estimators applying the ALTERNATIVE framing had no impact on the "Hit rate", only on the average confidence level.

The uncertainty assessments applying the TRADITIONAL framing showed a level of over-confidence similar to that found in the experiment. "90% confident" corresponded, in reality, to "74% accurate". As in the experiment, the correspondence between "Hit Rate" and average confidence level was much better when following the ALTERNATIVE process. The average hit rate of 87% corresponded well with the average confidence level of 88%, i.e., the ALTERNATIVE framing led to more realistic assessment of the *overall* level of uncertainty of the projects.

There was no difference in the level of correlation between RWidth and MRE (TRADITIONAL framing) and Confidence level and MRE (ALTERNATIVE framing). This suggests that the ALTERNATIVE framing only affects the ability to identify the overall level of uncertainty of similar projects, not the ability to distinguish between high and low uncertainty estimates.

We analyzed whether the differences could have been caused by systematic differences in favour of one of the framings. The analysis included analyses of type of task, project complexity, "know-how" skills of the estimators, estimation skills of the estimator, whether or not the estimators were estimating their own work, the project priorities (cost, time, quality), type of payment (fixed or per hour), organizational role of the estimator, project importance for the customer and the estimators own explanation for high or low estimation accuracy. Although many of these factors may have an impact on the uncertainty of the estimate, (see our previous formal model building attempt on a sub-set of the projects [6]), they could not explain the difference in results as displayed in Table 2. In fact, the only difference we found *went against* the ALTERNATIVE framing. The mean relative estimation error (mean MRE) was substantially higher for the projects following the ALTERNATIVE approach (36% versus 22%), suggesting that these projects may have been more difficult to estimate. As indicated in [10] higher estimation error seems to be connected with higher level of over-confidence.

4. Discussion of the Results

4.1 Potential Explanations

Statistically there is no important difference between the TRADITIONAL and the ALTERNATIVE framings, i.e., the statistical problem described by the TRADITIONAL

framing may easily be transformed into a statistical problem described by the ALTERNATIVE framing. We should, therefore, look at differences in how software professionals *perceive* and *perform* the uncertainty assessment tasks in the two framings. We believe that there are, amongst others, two important differences:

1) There seems to be a better *fit* between the ALTERNATIVE framing and the format of historical project estimation data. For example, assume that a project manager has estimated the most likely effort of a project to be 1000 work-hours. He is then asked about how likely it is that the actual effort will exceed the estimate by more than 50% (ALTERNATIVE framing). He recalls previous projects and believes that about 1 out of 10 previous projects exceeded their estimate by more than 50%. Based on this information, he can easily induce that there is, based on historical information, a 10% probability of exceeding the estimate by more than 50%. Now suppose that the same project manager was asked to provide the maximum effort with 90% confidence (TRADITIONAL framing). The problem is similar, but more complex to transform into a format in which the historical data is useful. The project manager must, for example, follow this process: 1) Find a project such that only 10% of all the projects have a higher relative estimation error, 2) Apply the relative estimation error of that project to calculate the maximum effort of the current project. Assume that project P is the one with only 10% more inaccurate projects and that the relative estimation error of project P is 40%. Then, the 90% confidence maximum effort should be 1000 work-hours * 140% = 1400 work-hours. In our opinion, this process is more complex and requires much more analytical skill than the one following from the ALTERNATIVE framing. Whether software professionals actually use any of these processes in their expert judgments is unclear. Software development uncertainty assessments seem to be highly intuition-based and analyzing the discussions of software professionals when assessing the uncertainty, as we

did in [11], does not provide many clues as to the mental steps involved. It is, nevertheless, possible to *instruct* software professionals to follow a particular process, e.g., a process based on the ALTERNATIVE framing.

2) Software professionals may have goals other than realism in uncertainty assessments. In [4] one of the software developers (applying a TRADITIONAL framing) explained: *"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 we evaluated how project managers actually assessed the skill of software developers and found that they indeed evaluated those software developers as more skilled who provided narrower intervals and exhibited higher confidence. Interestingly, this 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. One benefit of applying the ALTERNATIVE framing may, therefore, be based on the fact that the minimum and maximum values are not provided by the estimator him/herself. The skill evaluation effect may consequently be reduced, since the interval width is not set by the estimator himself and cannot be used as information for skill evaluation. In other words, it may be psychologically more difficult to provide a realistically high maximum effort for a 90% confidence level, risking being evaluated as unskilled, than to claim 90% confidence in the same maximum effort provided by another person.

4.2 Limitations

There are threats to the validity of the results found in this study and the results cannot be generalized into all types of context. In particular, we believe that the following two issues are important when interpreting the results:

- We evaluated only high confidence uncertainty assessments. The TRADITIONAL framing requested 90% confidence and the ALTERNATIVE framing requested confidence assessments based on rather wide minimum-maximum intervals. The benefits of the ALTERNATIVE framing may be much less, perhaps even disappear, when low or medium confidence uncertainty assessments are compared, for example, when comparing minimum-maximum intervals based on 60-70% confidence (TRADITIONAL framing), and confidence levels of narrow minimum-maximum intervals (ALTERNATIVE framing). There is consequently a need for more studies on the benefits of applying the ALTERNATIVE framing based on lower levels of confidence.
- Providing uncertainty assessments based on instructions in a questionnaire is not the same as being requested to provide an uncertainty assessment by a project or a bidding manager. However, we believe, based on the potential explanations of the increase in realism described in Section 4.1, that a more realistic uncertainty assessment situation would favour the ALTERNATIVE framing, i.e., increase the difference between the TRADITIONAL and the ALTERNATIVE framing. Nevertheless, this reduction in realism should be considered a threat to the validity of the results.

5. Conclusions

The best approach to assessing the uncertainty of effort estimates depends on many factors, e.g., the skill of the estimators, the availability of information about previous projects, the

type of information available about the project to be estimated, etc. The variety of factors means that empirical research can hardly ever be expected to provide general laws that govern the assessment of effort estimation uncertainty. The lack of general laws does, however, not mean that all choices are equally good. Through a series of studies, we have shown that use of the ALTERNATIVE framing is *better supported by empirical evidence* than the use of the TRADITIONAL framing. This means that a software manager asking for information about the uncertainty of an effort estimate of a particular task, who is sensitive to the concept of evidential support, should not apply the traditional PERT-inspired framing: "What is the maximum effort? Be almost sure.". Instead the software manager should, for example, ask "How likely is it that the task requires twice as much effort as estimated?"

Acknowledgement: Thanks professor in psychology at the University of Oslo, Karl Halvor Teigen, and, Mr. Kjetil Moløkken at Simula Research Laboratory for their useful suggestions and interesting discussions.

References

1. Jørgensen, M. and D.I.K. Sjøberg, *The Impact of Customer Expectation on Software Development Effort Estimates.* To appear in: Journal of Project Management, 2004.
2. Epley, N. and T. Gilovich, *Just going along: Nonconscious priming and conformity to social pressure.* Journal of Experimental Social Psychology, 1999. **35**: p. 578-589.
3. Moder, J.J., C.R. Phillips, and E.W. Davis, *Project management with CPM, PERT and precedence diagramming*. 1995, Wisconsin, U.S.A: Blitz Publishing Company.

4. 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.
5. 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.
6. Jørgensen, M., *Regression Models of Software Development Effort Estimation Accuracy and Bias.* To appear in: Journal of Empirical Software Engineering, 2004.
7. Angelis, L., I. Stamelos, and M. Morisio. *Building a software cost estimation model based on categorical data*. in *International Software Metrics Symposium*. 2001. London, England: IEEE Computer Society, Los Alamitos, Calif.: p. 4-15.
8. 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.
9. NASA, *Manager's handbook for software development*. 1990, Goddard Space Flight Center, Greenbelt, MD: NASA Software Engineering Laboratory.
10. Moløkken, K. and M. Jørgensen, *Expert estimation of the effort of web-development projects: Why are software professionals in technical roles more optimistic than those in non-technical roles.* To appear in Journal of Empirical Software Engineering, 2004.
11. Jørgensen, M., *Top-Down and Bottom-Up Expert Estimation of Software Development Effort.* To appear in: Journal of Information and Software Technology, 2004.