A Preliminary Theory of Judgment-based Project Software Effort Predictions

by:
Magne Jørgensen, Professor of Informatics, Simula Research Laboratory

Magne Jørgensen
Simula Research Laboratory
P. O. Box 134
NO-1325 LYSAKER
NORWAY

Tel: +47 924 333 55
Fax: +47 67 82 82 01
E-mail: magnej@simula.no

A Preliminary Theory of Judgment-based Project Software Effort Predictions

by:
Magne Jørgensen, Professor of Informatics, Simula Research Laboratory

Abstract

An improvement of the software industry's software development effort estimation processes may benefit from a better understanding of the mental, partly unconscious, processes involved in estimating effort. This paper proposes and tests a theory potentially explaining essential parts of typical judgment-based effort estimation processes ("expert estimation"). The theory is based on findings from the human judgment research literature and proposes that judgment-based effort estimation is based on: i) an early categorization of the project to be estimated, ii) a resistance towards a change of the chosen category, and, iii) a "regression" of the effort estimate towards a reference value of the chosen category, where the amount of regression depends on the level of uncertainty of the project work. Implications of the theory are tested with results from three software effort estimation experiments. All examined studies confirmed the theory. There is, however, a strong need for more work, independent evidence and clearer description of scope and concepts part of the theory. Finally, we outline a study planned for further testing of essential parts of the theory.

1. Motivation

Planning, bidding and budgeting of software development projects take as input the estimate of the effort most likely required to complete the project. Surveys of software projects suggest that these estimates are quite inaccurate and strongly biased towards over-optimism (Moløkken-Østvold and Jørgensen 2003). Over-optimistic effort estimates may lead to poor planning, low profitability, and, consequently, products with poor quality.

In spite of more than 40 years of research on formal software effort estimation models, most software project estimates are expert judgment-based (Jørgensen 2004). A review of studies comparing software development effort estimates based on models and expert judgment is provided in (Jørgensen 2006). That paper concludes that available evidence does not, in general, support a replacement of expert judgment-based effort estimates with estimation models. An important reason for this is the amount of "contextual information" present in software development situations, i.e., highly specific information about the project, the client, or the project member possessed by the estimator, but not included in the model. Consequently, besides development of better estimation models, we should try to support and improve the processes underlying judgment-based effort estimates. This improvement should be based on a better understanding of the mental, partly unconscious, processes involved. To date, there has not been much research on this topic. We scanned software effort estimation articles in relevant journals, see review in (Jørgensen and Shepperd 2006), and found no papers with in-depth descriptions of the judgment-based software development effort estimation processes.

In other domains, there may be a similar lack of knowledge about the mental processes involved in judgment-based quantification. According to Brown and Siegler (1993) psychological research on real-world quantitative expert estimation *"has not culminated in any theory of estimation, not even in a coherent framework for thinking about the process"*. This lack of theory does, however, not mean that there are no relevant results. There are many studies that provide results relevant to this issue, particularly among the numerous studies on human judgment under uncertainty.

This paper proposes a theory, or, more correctly, elements of a theory, of expert judgment-based effort estimates based on results from the human judgment literature. Our contribution is mainly in the combination and transfer of theories from general human judgment studies to a software effort estimation context. Section 2 describes the proposed theory and some of its testable implications. Section 3 describes studies that test these implications. Section 4 outlines a planned experiment that tests essential parts of the theory. Section 5 provides a few final remarks.

2. The Theory

The basis of our theory is that many real-world problems are too complex to be answered by following an optimal decision process. People therefore use simple process that, in essence, answers simpler questions than the one asked, but where the answer of the simpler question typically, but not always, is a good substitute for the answer of the more complex question (Kahneman 2003). These processes are to a large extent unconscious (intuition-based). A comprehensive descriptions of properties of intuition-based decision processes is provided in (Hogarth 2001). The following, rather well-established, theories from the human judgment

research are describing elements of such simple intuition-based processes (heuristics) and have provided essential input to our theory:

- *Understanding by categorization*, i.e., that one important means to understand and judge is by classifying and categorizing the world around us, putting things neatly into boxes (Rakison and Hahn 2004).
- *Importance of first impression*, i.e, the finding that the first impression is based on rather superficial indicators, but nevertheless is long-lasting and difficult to change. It seems as if the main reason for the impact is that new evidence is *assimilated* with the first impression, i.e., that we interpret new evidence differently dependent on our first impression (Gawronski, Alshut et al. 2002). The importance of the first impression is further strengthened by the observation that we are much better in finding confirming than conflicting evidence (Brehmer 1980).
- *Judgments are anchored and adjusted*, i.e., that we tend to base estimates and decisions on 'anchors' or representative values and adjust relative to this starting point. This theory is, among other things, based on the observation that people prefer relative rather than absolute judgments (Kahneman, Slovic et al. 1982).
- *Regression towards the mean*, i.e., the observation that when the uncertainty increases, people tend to predict values closer to the most representative or middle value of the class of relevant objects (Czaczkes and Ganzach 1996). Interestingly, the study reported in (Jou, Leka et al. 2004) suggests that increased experience with a particular type of task leads to higher regression to the reference value of a category, i.e., that the willingness to deviate much from the reference value decreases with more experience.

The theory we propose is based on the above theories and has the following elements:

i) In the beginning of the software development effort estimation work, the estimator, frequently unconsciously, categorizes the project based on an early "first impression".
Comments:
  - The assessment of similarity may typically be based on surface similarity to other, previously completed, projects, i.e., based on indicators that typically correlate with but are not necessarily causally connected with the amount of effort required.
  - The chosen category may in some cased be dominated by one single project, e.g., a project that is representative for that category or is fresh in mind, but may also include a range of projects. It may also be dominated by an artificial "prototypical" project, i.e., a project that is assembled from typical properties of several project of one category.
  - Probably, the creation of a category, the actual categories from which an estimator chooses, and, the strategies in use to choose a strategy differ from individual to individual.
  - The theory does not exclude overlapping project categories.
  - The essential point for our theory is not how the categories look like, but that the first estimate of the approximate size (in effort or another size measures) of the project is established early based on limited information and surface similarity with a category of projects.

ii) In the remaining estimation work, there is a resistance towards change of the chosen category.
Comments:
  - This resistance towards change of category is, amongst others, based on the human tendency to look for supporting rather than conflicting evidence for the initial categorization.

- The information collected after the categorization, i.e., after the first impression of project size has been established, is impacted by the chosen project category. For example, estimating an activity of a project categorized as a "large web-project" would lead to different interpretation of the activity information and, consequently, different effort estimates compared with estimation work based on the *same* information if the project had been categorized as a "medium large web-project".

iii) Estimates regresses towards the reference value of the chosen category, which may be the effort of the most "representative" project. The deviation from the reference value is determined by, among other things, the chosen estimation category, the quality of the project specific information, and, the knowledge about the outcome of similar projects or project activities (the project analogies).
Comment:
   - The theory of "regression towards the mean" describes the theoretical optimal relation between information uncertainty, e.g., as measured by the assumed similarity between the new project and the project analogies. Low level of similarity between the work to be estimated and the project analogies, leads to estimates closer to the reference value of the chosen category. We have described the use of this normative model to improve estimation work and to describe the actual behaviour of software estimators in (Jørgensen, Indahl et al. 2003).

Notice that we do not claim that our theory is an accurate or useful description of the underlying processes for *all* instances of judgment-based effort estimation. An estimator may, for example, decide that he will base the effort estimate on a simple formula using highly objective input. We hypothesize, however, that our theory describes elements of typical *judgment-based* estimation processes in the software industry.

As an illustration on the relation between the theory and actual estimation work, consider the following example.

Example: A software engineer is asked to estimate the effort required to develop software as specified in a textual requirement specification. He starts reading the specification and soon gets a first impression of the project. The first impression implies, among other things, that the project is similar to a medium large database development project completed earlier that year ("Early categorization"). The software engineering starts to break the project work into activities, e.g., project management, design, programming of modules, and testing. When estimating the effort of each activity he mainly recalls previously completed activities of medium large database development projects ("Resistance towards change of category"). Several of the activities were difficult to estimate, due to high implementation uncertainty and lack of information. These activities were estimated to require effort similar to that of typical effort of the chosen project category ("Regression towards reference values"). His total effort estimate is consequently not very far from the effort of the reference project of medium large database development projects.

The proposed theory is at an early stage. The introduced concepts are vague, there are probably missing elements, the strengths of the relationships, and, the scope of the theory is not well described. Nevertheless, there are implications of the theory that is possible to test and use as evidence to support whether the model is worth further work or not. The implications that we will test in Section 3 are the following two:

- When project surface indicators are misleading, the estimate becomes inaccurate.
- Projects of different size belonging to the same size category will be estimated to require similar effort estimates in situations with uncertainty connected to the requirements, i.e., there is a regression towards the reference value of a category.

These implications will be tested through two experiments conducted previously (Experiments 1 and 2) and one new experiment (Experiment 3). The reader should be aware of that it is likely that we are probably not as good in finding conflicting as supporting evidence for our theory. In addition, Experiments 1 and 2 were partly analyzed before we formulated the first version of the theory. These results may consequently have inspired and impacted the formulation of our theory and be of less value for testing of the theory. Independent studies are consequently needed for better validation of our theory.

3. Testing the Theory

3.1 Early Categorization Based on Misleading Surface Indicators

The following results are derived from an experiment (Experiment 1) described in detail in (Jørgensen and Sjøberg 2004). The experiment is similar to "traditional" anchoring bias experiments, but deviates in that the domain is effort estimation and that the final judgment (the total effort estimate) is based on a decomposition of the problem and judgement about each decomposed part (the project activities). This means that, as opposed to other anchoring studies, the numerical anchor value itself (which is about the total effort) cannot impact the judgments directly (which are about the effort of individual activities), but has to impact the judgments through impact on choice of size category or similar indirect means.

Experiment 1: The experiment examines the work break-down based estimates of 38 computer science students and 12 professional software developers. The participants were asked to estimate how much effort they would need to develop a specified software system. The students and the software professionals were divided into three groups with approximately equal size: The control group (CONTROL), the high customer expectation group (HIGH), and the low customer expectation group (LOW). The CONTROL group received no explicit customer expectation information, while customer expectation in the HIGH and LOW groups was introduced in the respective descriptions of the task, as follows: HIGH group [LOW group]: *"The customer has indicated that he believes that 1000 [50] work-hours (corresponds to about $ 80 000 [$ 4000] in development costs) is a reasonable effort estimate for the specified system. However, the customer knows very little about the implications of his specification on the development effort and you shall not let the customer's expectations impact your estimate. Your task is to provide a realistic effort estimate of a system that meets the requirements specification and has a sufficient quality."*

The resulting bids for the students and the software professionals are displayed in Table 1.

Table 1: Effort Estimates of Experiment 1

Group	***Participants***	***Estimate (median values)***
Low	Students	77 work-hours
Control	Students	224 work-hours
High	Students	404 work-hours
Low	Professionals	77 work-hours
Control	Professionals	176 work-hours
High	Professionals	632 work-hours

The results in Table 1 show that the customer expectations of a project's total cost can have a large impact on human judgment-based effort estimates. We interpret the results as suggesting that the information about the customers' expectation have led the estimators to think of the project as belonging to a particular size category, e.g., "small" or "large" projects. This categorization happened early (i.e., the information was presented early), was impacted by a surface indicator (the customer expectations), and was difficult to change with better understanding and more analysis of the project activities (otherwise the median values of the estimates of the different groups should have been similar). The results are consistent with what our theory predicts, although there may, of course, be alternative explanations.

3.2 Category-Induced Bid and Estimation Anomalies

The use of the chosen category's reference value as starting point for estimation, as predicted by our theory, means that projects that belong to the same project category will be estimated to require similar amount of effort, i.e., they regress towards the same reference value. This similarity can be exploited to create a bidding and estimation anomaly, i.e., the anomaly that the same specification leads to quite different bids and estimates. The two experiments (Experiments 2 and 3) testing this anomaly are both following the same design template:

1) *Preparation Phase*: We create two software development requirement specifications where Specification X is a subset of Specification Y. The difference between Specification X and Y should not be large. A small difference means that it is likely that both Specification X and Y belong to the same size category of projects for most of the bidders or estimators.
2) *Estimation Phase 1*: Let one group of bidders or estimators (Group A) start with an estimate (or bid) based on Specification X, the other group (Group B) with Specification Y.
3) *Estimation Phase 2*: Then, inform Group A that there are added requirements (the difference between Specification X and Y) that need to be estimated or the bids need to be updated. Similarly, inform Group B that there are reduced requirements (the difference between Specification X and Y) and that there is a need for an updated bid or estimate of the reduced specification.

Our theory predicts two, somewhat related, bidding or estimation anomalies in this situation:

i) The estimates or bids produced by Group A for Specification X and Group B for Specification Y will be similar[1] (given a sufficient level of requirement uncertainty), and,
ii) Group A's estimates or bids will be higher than Group B's estimates or bids of Specification Z and Y[2].

Experiment 2: Our research institute (Simula Research Laboratory) specified a tool to support our web-based bidding experiments (SIMBID). There were two versions of this specification: a reduced version (Specification X) and a full version (Specification Y). Specification Y had all functionality specified in Specification X and additional functionality.

[1] With "similar" we here mean that the average difference is substantially lower than the average difference in bids or estimates produced by *the same company* for Specification X and Y.

[2] This anomaly can be exploited to get lower bids on software projects, e.g., through starting with a large specification, remove functionality step by step and repeatedly ask for updated bids. We strongly warn against applying this strategy. The reason, as discussed in "Jørgensen, M. and S. Grimstad (2005). Over-optimism in Software Development Projects: "The winner's curse". Proceedings of IEEE CONIELECOMP, Puebla, Mexico, IEEE Computer Society: 280–285.", is that projects based on over-optimistic effort estimates and too low price easily lead to low quality software, very low provider flexibility, and other problems for the client.

The length of Specification X was 22 pages, while the length of Specification Y was 27 pages. The mean bids for the reduced (Specification X) and the full (Specification Y) specification are shown in Table 2. Bids from 29 outsourcing companies from different countries were received in the range from $ 750 to $ 16 600 for Specification X and from $ 2000 to $ 37 000 for Specification Y. The full experiment is reported in (Jørgensen 2006).

Table 2: Bidding Results in Experiment 2

Group	***N***	***Mean Bid: Specification X***	***Mean Bid: Specification Y***
A	12	$ 11,195	$ 15,482
B	17	$ 8,841	$ 12,589

As can be seen from Table 2, the implications i) and ii) of our theory are both supported by the results. There is only a small difference between Group A's mean bid for Specification X and Group B's mean bid for Specification Y, i.e., the difference between $ 11,195 and $ 12,589. In addition, Group A's bids based on Specification X and Y were, consistent with our theory, substantially higher than those of Group B.

Experiment 3: This experiment focuses on the effort estimates, and not the bids as in Experiment 2. Forty-two university students following lectures in software engineering participated in the experiments. Similarly to Experiment 2, they were randomly divided into two groups (A and B) and asked to estimate the effort most likely required to complete tasks as described in Specification X and Y. Specification X was a sub-set of Specification Y. The tasks were similar to previously completed programming tasks by the students, i.e., they had sufficient experience to enable meaningful effort estimates. The estimates for Specification X varied from 150 to 750 minutes, while those for Specification Y varied from 180 to 800 minutes.

Table 3: Estimate Results

Group	**N**	**Mean Estimated effort: Specification X**	**Mean Estimated Effort: Specification Y**
A	20	261 minutes	482 minutes
B	22	232 minutes.	311 minutes

As in Experiment 2, the implications i) and ii) of the theory were supported by the results. There is only a 16% difference between Group A's mean estimate based on Specification X (261 minutes) and Group B's mean estimate based on Specification Y (311 minutes). This is a substantially smaller difference than the difference between Group A's and B's mean estimates of Specification X and Y, which was 46% and 25%, respectively. In addition, as predicted, Group A's mean effort estimates for Specification X and Y were substantially higher than that of Group B. An interesting observation is that the mean increase (221 minutes) of Group A was much higher than the mean decrease (79 minutes) of Group B. It seems as if, although the initially chosen size category (261 vs 311 minutes) seems to have been the same, Group A and B have chosen different size categories for the update. This is an effect that was not expected, not found in Experiment 2, and illustrates the need for further work on the theory.

4. Outline of an Experiment

The following experiment is planned for execution later this year. The main purpose of the experiment is to test the implication that the impressions (estimates) formed early in the

process will be similar to those produced later based on much more information, decomposition of the project into activities and increased understanding of the project implications.

Planned Experiment: We plan to invite about 40 software engineers to participate. The participants will be divided, randomly, into three groups (A, B, C). Each participant is asked to estimate the same four projects (P1, P2, P3 and P4). The main difference is in how much time they have available for each estimate. Phase 1 is used to test the similarity of the participants in the different groups on the same project. Table 4 outlines the study design.

Table 4: Outline of Design of Planned Experiment 1

Group	***Phase 1: Normal use of time***	***Phase 2: 1 min.***	***Phase 3: 5 min.***	***Phase 4: Normal use of time***
A	P1	P2	P3	P4
B	P1	P3	P4	P2
C	P1	P4	P2	P3

Our theory predicts that:

- The mean effort estimates provided with strongly restricted time, i.e., in Phase 2 and 3, will be similar to those provided with the normal time to produce estimates (Phase 4). This will happen due to the early categorization and resistance towards change of category.
- Less time for estimation (i.e., higher uncertainty) leads to effort estimates closer to the project category's reference effort value. This, in turn, leads to lower variance in estimates. Our theory would consequently predict that the variance of the estimates in Phase 2 will be lower than that of Phase 3, which again will be lower than that in Phase 4.

The theory is weakened if the predictions are false, although there may be alternative explanation for both rejecting and confirming observations.

4. Final Remarks

The work described in this paper is in its early stage and we expect that there is a need for many years of work to establish a useful theory that can explain and predict when and why software engineering produce too pessimistic, realistic and overoptimistic effort estimates. Particularly, we need more insight into the concept of work effort size categorization. We expect that there will be significant changes in the proposed theory resulting from our own and others future experiments and field studies. In particular, we expect that there will be important individual differences, e.g., that the theory is more valid for some people than other. Preliminary evidence, for example, suggests that so-called "mixed-handers" are much more impacted than "strong-handers" by the categorization-induced bias in Experiment 3. "Mixed-handers" are those with no strongly dominating hand. "Mixed-handedness" is hypothesized to be a measure on willingness to update beliefs and judgments (Jasper and Christman 2005).

The goal of proposing the theory is, as stated earlier, not only to explain and understand, but to enable improved effort estimation processes. Possible improvements are already emerging from our theory, e.g., that it is essential that potentially misleading surface indicators creating incorrect first impressions of the project are removed from the requirement specifications of software projects. As soon as the theory is more robust we will increase the focus on how to apply the theory to design better estimation processes.

References

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

Brown, N. R. and R. S. Siegler (1993). "Metrics and mappings: A framework for understanding real-world quantitative estimation." Psychological Review **100**(3): 511 - 534.

Czaczkes, B. and Y. Ganzach (1996). "The natural selection of prediction heuristics: Anchoring and adjustment versus representativeness." Journal of Behavioral Decision Making **9**(2): 125-139.

Gawronski, B., E. Alshut, et al. (2002). "Processes of judging known and unknown persons: How the first impression influences the processing of new information (in German)." Zeitschrift-fur-Sozialpsychologie **33**(1): 25-34.

Hogarth, R. M. (2001). Educating Intution. Chicago, University of Chicago Press.

Jasper, J. D. and S. D. Christman (2005). "A Neuropsychological Dimension for Anchoring Effects." Journal of Behavioral Decision Making **18**: 343–369.

Jou, J., G. Leka, E, et al. (2004). "Contraction bias in memorial quantifying judgment: Does it come from a stable compressed memory representation or a dynamic adaptation process." Americal Journal of Psychology **117**(4): 543-564.

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. (2006). "The Effects of the Format of Software Project Bidding Processes." To appear in International Journal of Project Management.

Jørgensen, M. (2006). "Estimation of software development work effort: Evidence on expert judgment and formal models." Submitted to International Journal of Forecasting.

Jørgensen, M. and S. Grimstad (2005). Over-optimism in Software Development Projects: "The winner's curse". Proceedings of IEEE CONIELECOMP, Puebla, Mexico, IEEE Computer Society: 280–285.

Jørgensen, M., U. Indahl, et al. (2003). "Software effort estimation by analogy and "regression toward the mean"." Journal of Systems and Software **68**(3): 253-262.

Jørgensen, M. and M. Shepperd (2006). "A Systematic Review of Software Development Cost Estimation Studies." IEEE Transactions on software engineering (in press).

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

Kahneman, D. (2003). "A Perspective on Judgment and Choice: Mapping Bounded Rationality (based on his Nobel Prize lecture)." American Psychologist **58**(9): 697-720.

Kahneman, D., P. Slovic, et al. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge, United Kingdom, Cambridge University Press.

Moløkken-Østvold, K. and M. Jørgensen (2003). A review of software surveys on software effort estimation. International Symposium on Empirical Software Engineering, Rome, Italy, Simula Res. Lab. Lysaker Norway: 223-230.

Rakison, D., H and E. Hahn, R (2004). The Mechanisms of Early Categorization and Induction: Smart or Dumb Infants? Advances in child development and behavior. R. Kail, V. San Diego, CA, US, Elsevier Academic Press. **Vol. 32:** 281-322.