Expert Judgment in Software Estimation during the Bid Phase of a Project – An Exploratory Survey

Pedro Faria, Eduardo Miranda

University of Coimbra, Carnegie Mellon University

***Abstract*— This article presents the results of an industrial study about the reliability of expert judgment in cost estimation in a medium-sized software company.**

The purpose of the study was to assess current practices within the company, and to use the results as a catalyst for improving the company's cost estimation processes.

The study included the analysis of two aspects: variability and calibration. First, the degree of variability can indicate the absence of a sound and repeatable process for making estimations. Second, measuring the estimators' calibration allowed us to establish their awareness about their lack of specific knowledge when estimating, as well as their ability to properly acknowledge and represent uncertainty.

The results clearly show that, in the context of the organization studied, cost estimation based on unstructured expert judgment is unreliable.

***Index Terms*— software cost estimation, calibration, variability**

I. Introduction

There are many approaches to software estimation. Some focus on using historical data to mathematically derive an estimate. Others try to measure the size of the software to be produced to allow effort estimation based on the size of the task. Others yet use expert opinion, with a focus on trying to compensate the bias and knowledge gaps of individuals betting on group work to achieve better estimates. The focus of this paper is on estimations based on unstructured expert opinion during the bid phase of a project.

Expert opinion can be defined as the judgment of an individual expert or group of experts with respect to a specific subject or unknown measure. In the scenario under study, expert judgment is used to estimate the effort of software projects.

Expert judgment has been considered particularly useful in 1) areas where empirical data is not easily available, and 2) when estimating complex, ill defined or poorly understood problems [1]. Both these dimensions can be seen as underlying reasons why expert judgment is widely used as an approach to software estimation.

The study was conducted in a business division of a medium size organization (2000 employees) focused on the development of bespoken software for domestic and international markets. The company promotes a culture of excellence, quality and commitment, and has adhered to and established a series of good practices and processes to govern and improve its performance. The intention of the organization in supporting this study was to assess its cost estimation capability in the bid phase of a project and, if required, to use the results as a catalyst for change of their estimation practices.

This study shows that for all its virtues, unstructured expert judgment is highly unreliable.

This lack of reliability has direct economic consequences. On one hand, the company may be losing business when the estimates are too high and on the other losing money if these are too low.

The study was conducted as part of the requirements for one course at the Master of Software Engineering Program jointly offered by the University of Coimbra in Portugal and Carnegie Mellon University in the United States.

Before continuing with the analysis it is necessary to establish what we mean by unstructured estimation methods and unreliable estimates.

A. Structured methods for software estimation

By structured methods we refer to any estimation approach consisting of systematic and detailed steps that can be described and replicated. Structured methods exhibit the following properties [2]:

- Correctness. It should be possible to assess whether the method has been applied in accordance with its intent or not, and why.
- Traceability. It should be possible to understand how the outputs were derived from its inputs. No black magic.
- Reusability. The framework should be applicable in different contexts.
- Reduced variability. Other things being equal, the outputs of the method should have a lower dispersion than ad-hoc, anecdotic methods.

By contrast, an unstructured method is any method that does not exhibit the above properties.

B. Reliable estimates

An estimation method produces reliable estimates if its repeated application to the same set of inputs results in estimates that are both close to one another and close to the true value of the quantity estimated. We will refer to the first quality as the precision or variability of the method and to the second as its accuracy. Although the definition of how close is close enough, and what is the true value of a project estimate that has yet to be executed are debatable, it would be easy to accept that if the same method produced for the same project two estimates, one which would result either in the loss of a business opportunity in a competitive environment or in the loss of money should the other be true, is a method that is both inaccurate and imprecise.

Since in this study there was no estimation method being used, in the sense of the structured method defined above, the question about the reliability of the method becomes the question of how reliable the estimators are.

An important issue in expert estimation is overconfidence [3]. The point here is whether experts are able to judge their lack of knowledge when estimating future projects or not. Research has shown that we tend to overstate our own knowledge. To evaluate this aspect of expert estimation we will assess the estimators' calibration by asking them how certain are they about their estimates.

In a nutshell, an estimator is overconfident if the subjective level of confidence he assigns to his estimates is higher than the relative frequency of the events predicted [4]. In other words, if an estimator says s/he is confident in her/his estimations at the 70% level, we would expect to see that over a number of projects, 70% of them come under the estimated cost. If less than 70% of the estimated projects come under the estimated cost we will say that the estimator is overconfident. When an estimator is neither over nor under-confident we say that the estimator is calibrated.

Section II presents the questions we sought to answer with this study. Section III describes the method used to answer them, namely the experiments designed to collect the necessary data. In Section IV we present the results and we analyze the data from the perspective of how the data allows us to answer the research questions and what conclusions we draw from it. Section V discusses some threats to the validity of the results. Finally, Section VI provides some conclusions and recommendations.

II. Research Questions

The study sought to answer the three following questions:

A. Variability

RQ1: Given a common project description, will the estimators arrive to the same estimate?

A negative answer to this question will imply that unstructured expert estimation is unreliable.

RQ2: Are the differences among estimators material?

A high value in this measure will indicate that the company might be losing money in those cases where the bid is low and losing business in those cases where the bid is too high.

B. Calibration

RQ3: Are estimators aware of the assumptions they make and their lack of specific knowledge when estimating future projects? In other words, are the estimators calibrated, over or under-confident in their estimates?

The answer to this question will indicate avenues of improvements to the estimation process, such as providing estimators with more information, checklists or critical thinking training.

III. Survey & Research Method

The following aspects should be highlighted regarding the research method

A. Study demographics

Thirty employees with estimation responsibilities were invited to participate. They all belonged to the sponsoring company. Figure 1 and Figure 2 show a breakdown of the employees' roles and experience. The survey response rate was 47%. The average experience of the respondents is 7.5 years. This allows us to be confident that the results of the study are not due to inexperience or lack of knowledge. All the data was collected using an anonymous online survey with voluntary participation. Participants were divided in two groups (A, B). The assignment of participants to groups was random. The use of two groups and two different projects allowed us to control for project effects, i.e., the observed effects were not due to the choice of project.

Figure 1 – Distribution of roles

Note that percentages exceed 100% because respondents indicated they performed more than one role.

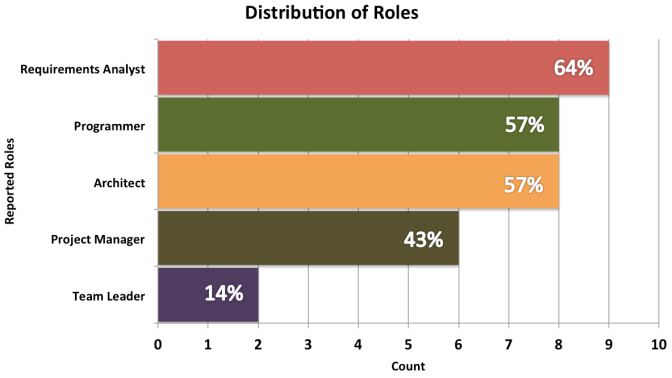
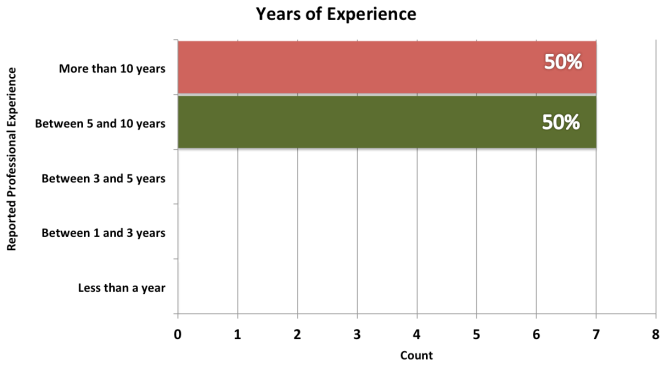


Figure 2 – Distribution of experience Level



B. Variability

To answer questions RQ1 and RQ2 above, a description of one of two past projects executed by the division was provided to each group of participants, requesting them to give their estimation, in staff-days, for the effort needed to execute the project. Each project was described using the typical documentation usually available when applying for a bid (typically the Request-For-Proposal issued by the potential customer). Therefore, the level of detail available for the estimators is representative of a real situation.

C. Calibration

To measure the calibration of the estimators (question RQ3 above), the same two groups as in the variability study were provided with brief descriptions (the same for both groups) of ten completed projects, and asked to provide their low and high effort estimations, so they were 90% certain the real effort fell within these two values. Each project was described using two to four paragraphs and the actuals were obtained from accounting records.

An estimator was considered calibrated if the actual values for each of the ten projects fell within the defined ranges nine out of ten times. If the number of correct answers was below that the estimator was considered overconfident.

IV. Results & Discussion

As mentioned before, the response rate was of 47%, not very different from the average response rate seen in other studies [5]. **Table 1** shows the breakdown in terms of response rate. While a fully quantitative analysis has yet to be performed, the results so far seem to confirm that the organization is not different from others studied in the literature [6].

The following sections provide a summarized view of the results with respect to each research topic and question.

Table 1 – Response rate breakdown

Group	Expected Responses	Actual Reponses	Response Rate
Group A	15	7	47%
Group B	15	7	47%
TOTAL	30	14	47%

A. Variability

1) RQ1

Table 2 below shows the respondent's estimates corresponding to the project description assigned to his or her group. The answers obtained exhibit high variability. Even for the estimates that are closer to each other, differences of more than 100 or 200 staff-days are common.

2) RQ2

Assuming a price tag of 500$ per staff-day, this means that depending on who did the estimate for project A the bid could have been at any level between 60,000$ and 1,008,000$. Although the multiple approvals required by the bid process will mediate in the final number, we cannot discount the anchoring effects of the initial estimate. Even if we were to discard the estimates by estimators 3A, 6A & 7A, the bidding range would be 112,500$ to 162,000$, a 44% variation which could easily discriminate between a losing and a winning bid.

It ought to be remembered that winning a bid and making a profit are two different things. It is well documented, that in the presence of schedule commitments, the cost of recovering from an underestimated project tends to be much higher than if the work had been originally planned [7], [8].

Table 2 – Individual estimates in staff-days for each group

Group A		Group B	
Estimator	**Estimate**	**Estimator**	**Estimate**
1	225	1	55
2	320	2	180
3	2016	3	125
4	230	4	220
5	310	5	940
6	120	6	60
7	1010	7	70
Mean	604		236
Median	310		125
Std. Dev.	688		317
Coefficient of Variance	113.89		134.45

B. Calibration

1) RQ3

The estimators' calibration fared no better, see Figure 3. The results show that no estimator was able to get near the 90% target, which would have indicated that their confidence matched the relative frequency of their successes. Only one estimator managed to get the actual effort within his/her range for one project, resulting in a 25% calibration (one out of four).

Another interesting perspective on this data is related to variability. We can use this data to check if our findings are also observed here.

For that purpose, we analyzed the variability of the middle points of the ranges provided by the estimators for all the ten projects. Table 3 below shows the same descriptive statistics used earlier for the variability research topic. As we saw earlier, the dispersion of the estimates is severe and consistent across project.

Table 3 – Descriptive statistics for intervals' mid points

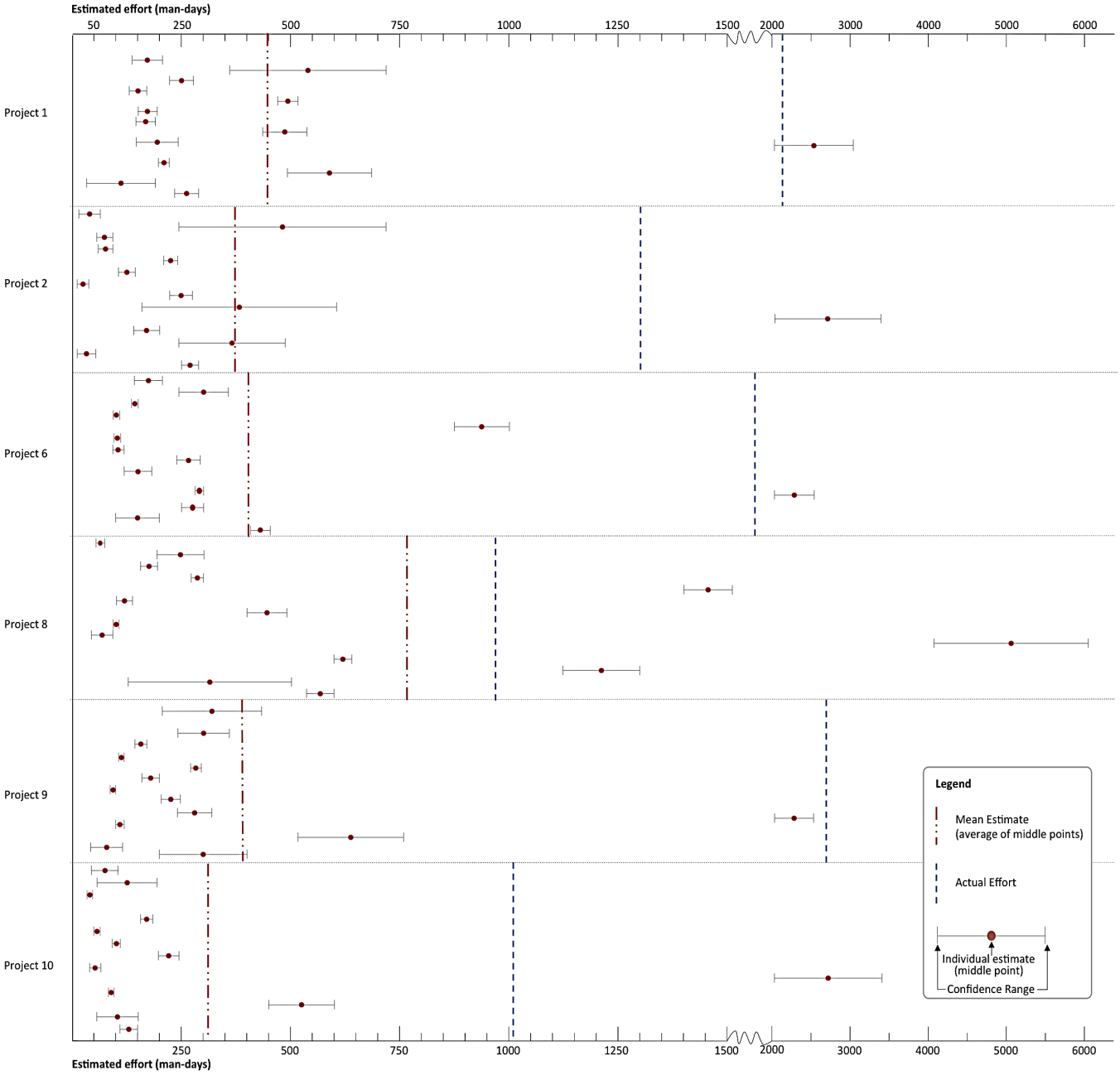
Project	Actual (staff-days)	Mean*	Standard Deviation	Coefficient of Variance
1	2200	449	618	1.37
2	1300	370	682	1.84
3	#	602	819	1.35
4	#	671	1281	1.90
5	#	480	937	1.95
6	1800	401	580	1.44
7	#	434	549	1.26
8	968	763	1302	1.70
9	2600	382	561	1.46
10	1000	319	693	2.17

* Arithmetic mean of the mid-point of the 14 answers received
# It was not possible to obtain the actual effort by the end of this study. With respect to calibration, only the six projects with actuals were considered.

V. VALIDITY OF THE RESULTS

The experiments performed suffer from the usual class of problems found in this kind of study: limited sample size and emailed questionnaires, which do not guaranty a totally random sampling. The results, however, seem to be consistent with other studies addressing the same issues in software and other disciplines [4], [9], [10].

Figure 3 – Confidence Intervals for estimates in the calibration study



VI. Conclusion

The purpose of this study was to explore whether or not variability issues and overconfidence, reported concerning unstructured expert estimation, were present in the estimation processes of the software development organization surveyed. In affirmative case, the results obtained should start a discussion about the business implications and possible remediation. Without such diagnosis, any claims about the need to invest in this area would be solely supported by subjective opinion and, therefore, with little probability of being acted upon.

The study clearly confirms that judgment inconsistencies and overconfidence are an issue. This does not mean that expert judgment should not be used as an estimation method but rather points to the need to supplement the judgment processes that account for the biases observed. Methods such as Wideband-Delphi [11], [12], paired comparisons [13] and decision markets [14] could be used to compensate for

variability while self-awareness and calibration training can be used to deal with overconfidence issues.

ACKNOWLEDGMENTS

We would like to thank Sofia Esteves and Pedro Abreu. Their efforts to make this study possible exceed our best expectations. Moreover, a special thanks for those who devoted their time – in some cases personal time – to participate in this study.

REFERENCES

[1] S.-W. Lin and V. M. Bier, "A study of expert overconfidence," *Reliability Engineering & System Safety*, vol. 93, no. 5, pp. 711-721, May 2008.

[2] R. Stegers and A. T. Teije, "From Natural Language to Formal Proof Goal," *Managing Knowledge in a World of Networks*, no. March, 2006.

[3] H. Arkes, "Overconfidence in judgmental forecasting," in *J. S. Armstrong, Principles of Forecasting*, Kluwer Academic Publishers, 2001, pp. 495-515.

[4] D. W. Hubbard, *How to measure anything: finding the value of intangibles in business*. Wiley, 2010.

[5] K. Molkken, "A review of surveys on software effort estimation," *Symposium on Empirical Software*, no. 1325, 2003.

[6] M. Jørgensen, K. H. Teigen, and K. Moløkken, "Better sure than safe? Over-confidence in judgement based software development effort prediction intervals," *Journal of Systems and Software*, vol. 70, no. 1–2, pp. 79-93, Feb. 2004.

[7] F. Freiman, "The Fast Cost Estimating Models," in *Transactions of the 27th Annual Meeting of the American Association of Cost Engineers*, 1983, pp. 26-29.

[8] E. Miranda and A. Abran, "Protecting software development projects against underestimation," *Project Management Journal*, vol. 39, no. 3, pp. 75–85, 2008.

[9] J. J. Christensen-Szalanski and J. B. Bushyhead, "Physicians' use of probabilistic information in a real clinical setting.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 7, no. 4, pp. 928-935, 1981.

[10] A. H. Murphy and R. L. Winkler, "Reliability of Subjective Probability Forecasts of Precipitation and Temperature," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 26, no. 1, pp. 41-47, 1977.

[11] A. Stellman and J. Greene, *Applied Software Project Management*. O'Reilly Media, 2005.

[12] K. Wiegers, "Stop Promising Miracles," *Software Development*, vol. 8, no. 2, pp. 49–53, 2000.

[13] E. Miranda, "Improving subjective estimates using paired comparisons," *Software, IEEE*, no. February, pp. 87-91, 2001.

[14] M. Hearst and R. Hunson, "Building intelligent systems one e-citizen at a time," *IEEE Intelligent Systems*, 1999.