First Impressions in Software Development Effort Estimation: Easy to Create and Difficult to Neutralize

Magne Jørgensen & Erik Løhre
Simula Research Laboratory & University of Oslo
magnej@simula.no
erikloh@simula.no

Abstract

The four studies included in this paper examine the strength of the first impression in the estimation of work effort. The studied context is mainly one where the first impressions of software development effort is manipulated (biased) through comparisons with either much too low or too high reference effort values, e.g., by responding to a question from the client whether one believes that the effort will be less than 10 work-hours when the likely usage of effort typically would be in the range of 100-300 work-hours. Then, the software developers are exposed to a subsequent comparison with reference effort values in the opposite direction, e.g., by responding to a question from the client whether one believes that the effort would be less than 800 work-hours. The results from the four studies suggest a strong first impression effect, but also a noticeable effect from the subsequent comparisons. We also observe that the instruction to "forget" the first impression seems to have the opposite effect, i.e., it seems to increase the strength of the first impression. A practical implication of the results is that it is essential that software professionals ensure that their first impression of a project's development effort is based on comparisons with representative reference values and objects. First impressions in software development seem to be easy to manipulate with misleading reference values and difficult to replace.

1. Introduction

"First impressions" are known to be quickly made and resist updates. A possible reason for this may be a strong tendency towards searching for confirming and neglecting non-confirming evidence [1]. The danger of incorrect first impressions was observed by Francis Bacon in his book "Organum scientarum" already in 1620: "*The human understanding when it has once adopted an opinion draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside and rejects, in order that by this great and pernicious predetermination the authority of its former conclusion may remain inviolate*." Human judgment is, however, not always dominated by lack of willingness to update first impressions and there are observations that corresponds to a "last impression" or "recency" effect [2]. A recency effect occurs when the information accessed most recently has the strongest effect on the judgment, possibly due to unconscious mental processes that prefer the use of information that has just been activated.

To what degree it is the "first impression" or the "recency" of activated information that dominates peoples' judgments is the topic of this paper. As far as we are aware of, this is a topic that has not been addressed before in an effort estimation context. If there is a strong "first impression" effect, then software developers should be especially careful about the relevance and representativity of the information they receive very early in the process of estimating the effort of a task. If the "recency" effect is dominating, the relevance of the first information is a lesser concern than that of the relevance of the information exposed to just before estimating the effort.

As indicated in several research papers on software development effort estimation, it seems to be easy to bias the estimation process through inclusion of irrelevant or misleading information [3]. The bias can be substantial and the software professional will typically not be aware of how much she has been affected by the client's request. If the biased first impression based is difficult to update when new information arrives, then the project may end up with over-optimistic plans and delivery problems. The lack of reflections on the information exposed to early in the estimation process may consequently be an important reason for the current strong tendency towards cost overrun and delivery problems in software projects.

The remaining part of this paper is organized as follows: Section 2 presents four studies that examine the strength of the first impression, mainly in comparison with the recency effect. Section 3 briefly discusses the results, the limitations of the studies and the possible practical implication for software development effort estimation.

2. The empirical studies

2.1 Study A

Research question: Is it the first or the last comparison with an unrealistic effort reference threshold that has the strongest influence on the effort estimate?

Participants: Software professionals from six different outsourcing companies in Thailand and Vietnam. In total 108 software professionals took part. All of them had previous experience in estimating the effort of own software development work.

Estimation task: The developers were asked to estimate the effort they most likely would need to develop and test the same software system, i.e., a system used to support the selection of jogging shoes in a shop. The developers should assume that they could use the technology (programming languages, web-technology etc.) that they know best and do all the work themselves.

Study design: The developers were randomly divided into three groups with different treatments. Developers in the control group were asked to first read the specification, then list the activities required to complete the work, and finally estimate the total effort of the development work. The participants in the two other groups (Low-High and High-Low) were instructed, after reading the requirement specification, but before listing activities and estimating the total effort, to complete two comparisons. The first comparisons aimed at creating a first impression of the effort required for the development of the system. The second comparison aimed at an updated impression based on "counter-information". None of the comparisons should, rationally speaking, affect the effort estimates, since they did not contain any information about the actual size or complexity of the project.

The first comparison request was formulated as follows (where X was 10 work-hours for Group Low-High and 400 for Group High-Low): "*The client calls you and asks how likely it is that you will be able to develop and test the system in X work-hours or less. You know that the client has no experience in software development and that you will be allowed to use the effort it takes to develop the system with proper quality. You are nevertheless asked to respond to the client's question*."

Previous experience implied that the development work would typically be estimated to require between 50 and 200 work-hours without the above requests, i.e., the values 10 and 400 represented a very low and a high effort usage value. Studies on "anchoring effects" [4] made us expect that the comparisons would lead to a first impression of a relatively small project in Group Low-High and a relatively large project in Group High-Low. This first impression would be present in spite of the information included in the request that the client's request was not based on knowledge about the effort required or the budget for this project.

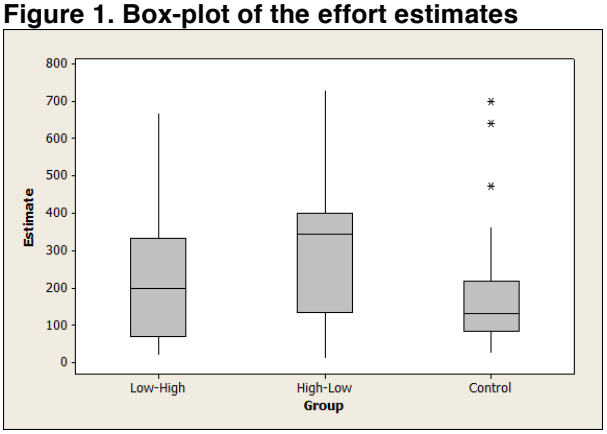
At the next page of the estimation material, so that the developers did not about the request when responding to the first comparison request, the software developers received the second comparison request (where X was 400 work-hours for Group Low-High and 10 work-hours for Group High-Low): "*The client calls you again. This time he asks you about how likely it is that you will use X work-hours or less to develop and test the specified system*."

As can be seen, the participants in Groups Low-High and High-Low were exposed to the same two requests. Only the sequence was different. If the first impression mattered most (least), we should expect that those in Group Low-High had higher (lower) estimates than those in Group High-Low. If the first and second impressions were equally important, we should expect about the same estimates in Groups Low-High and High-Low. If the requests did not affect the effort estimates, after all they contained no estimation relevant information, we should expect that the estimates in Groups Low-High, High-Low and Control were about the same.

Results: The average number of activities identified was the same (about seven) for all three groups. Any differences in work effort between the groups was therefore not likely to be a consequence of different understanding of the activities required to complete the task, but rather how much effort each of them require.

The estimated efforts of the software developers in the three groups suggest that the requests had a strong effect. The median effort of the developers' estimates in Group Low-High (n=37) was 198 work-hours, Group High-Low (n=35) 344 and Group Control (n=36) 132 work-hours. A one-sided Kruskal-Wallis test of difference in median values gives p=0.01. Figure 1 displays a box-plot of the values.

Figure 1. Box-plot of the effort estimates



The results suggest that the first impression was not neutralized with an opposite reference, i.e., the effect of the first impression was stronger than the recency effect. It also suggests that estimation irrelevant information, in this case the client requests based on no knowledge about software development, were effective in creating the first impression of a small (Group Low-High) and a substantially larger project (Group High-Low). The observation that the median estimate in Group Low-High was higher than that of Group Control suggests however that the second request (the 400 work-hour or less question for those in Group Low-High) also had an impact on the final estimate. Otherwise, we would expect the median estimate of Group Low-High to be lower than that of Group Control. The first impression is consequently not impossible to change. To get more insight into the effect of a second comparison on the first impression we designed and completed Study B.

2.2 Study B

Research question: How much is the first impression changed by introducing a second comparison with a much higher or lower effort value?

Participants: The participants in Study B consisted of 114 software developers working in a Norwegian consultancy company.

Estimation task: The participants were asked to estimate the effort they would need to conduct a small copy machine task. The task took, when we tested it, about 140 minutes on a medium fast copying machine, with no major problems occuring.

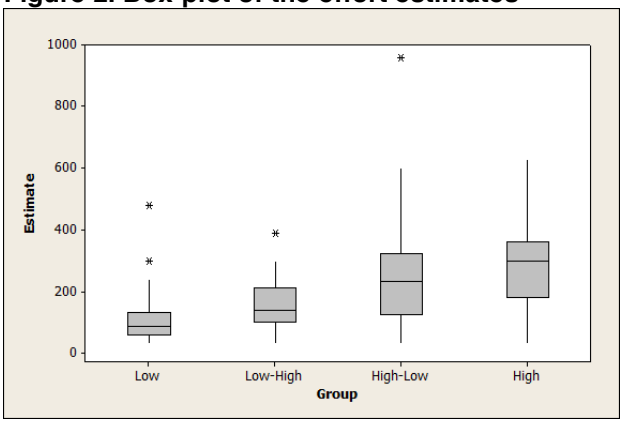
Study design: The participants were randomly divided into four groups with different treatments. All participants started with reading the specification of the copy machine task. In between the reading of the specification and the estimation of the effort the different groups received different comparison request:

- Group Low: "*Do you think you need more than 30 minutes to complete the task?*"
- Group Low-High: i) "*Do you think you need more than 30 minutes to complete the task?*" and ii) "*Do you think you need more than 8 work-hours to complete the task?*"
- Group High-Low: i) "*Do you think you need more than 8 work-hours to complete the task?*" and ii) "*Do you think you need more than 30 minutes to complete the task?*"
- Group High: "*Do you think you need more than 8 work-hours to complete the task?*"

As with Study A, the second comparison request (ii) was presented on the next page so that the participants did not know about that comparison when responding to the first comparison request. We expected that the median values of the effort estimates would be increasingly higher in the sequence: Low < Low-High < High-Low < High. This follows from the previous finding in Study A where the first impression was updated, but not neutralized by a second comparison.

Results: The rank of the median estimates of the groups were as expected. Those in Group Low (n=30) had median estimates of 88 minutes, those in Group Low-High (n=29) 140 minutes, those in Group High-Low (n=28) 233 minutes and those in Group High (n=27) 300 minutes. A Kruskal-Wallis one-sided test of different median values gives $p<0.001$. Figure 2 shows a box-plot of the effort estimates of the different groups.

Figure 2. Box-plot of the effort estimates



The results of Study B imply, similar to Study A, that there is a first impression effect stronger than the recency effect, i.e., those in the Low-High group have on average lower estimates than those in the High-Low group (140 vs 233 minutes). It does, however, also show that there is a substantial effect of the second comparison with a reference effort value. Recalling that a realistic use of effort to solve this task is about 140 minutes, we see that presenting the opposite effort reference to an estimator who is strongly biased

towards very low or high values may "balance" the impression and lead to more realistic effort estimates.

Assume that the mid-point between the median estimates of the Low and the High group ((88 + 300)/2 = 194) is approximately the value that on average would be estimated given that the first impression and the recency effect were about equally strong, i.e., this would be the median effort estimate if the strength of both effects were about 50% in the studied context. We also have that the strength of the first impression is 100% if the average estimate is 88 minutes for those in the Group Low-High and 300 minutes for those in the High-Low group. The median estimate of those in the Low-High group is 140 minutes. This is in the middle of 100% (88 minutes) and 50% (194 minutes) impact from the first impression, which means that the weight of the first impression is about 75%. We also observe that the median estimate of the High-Low group is 233 minutes. This represents a weight of the first impression of 68% (=0.5+ 0.5*(233-194)/(300-194)=0.5+0.18), i.e., slightly lower strength of the first impression than in the Low-High group. The difference in strength of the first impression may consequently vary with the extremity of the manipulation. It is in our case possible that the strength of the first impression based on a suggestion of 8 work-hours is weaker because 8 work-hours are perceived as less believable than 30 minutes.

In Studies A and B we informed that the references used for comparisons were irrelevant for the actual use of effort. This did not remove the strong effect from the comparison. What would happen if we explicitly told the software professionals to *forget* the information? Would this weaken or strengthen the effect of the first impression. This is the topic of Study C.

2.3 Study C

Research question: How much is the first impression weakened or strengthened by the explicit instruction to forget the effort threshold used in the first comparison?

Participants: The participants in Study C consisted of 64 software professionals working with in-house software development in a Norwegian company.

Estimation task: A copy machine task, similar to the one in Study B.

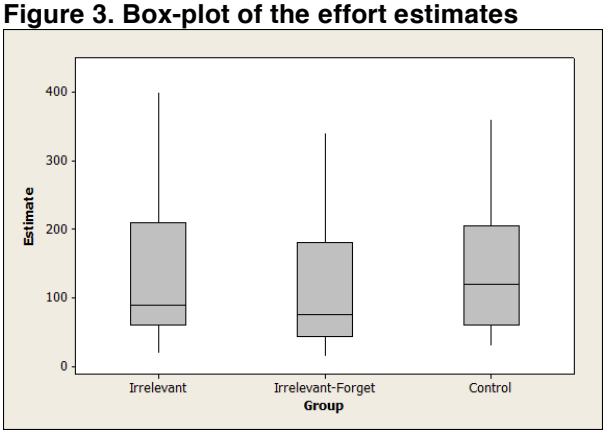
Study design: The participants were randomly divided into three groups with different treatments. All participants started with reading a specification of the copy machine task. Those in the control group were instructed to estimate the effort they would need to complete the task, which included only relevant information about the copy machine task. Those in the two other groups (Groups Irrelevant and Irrelevant-Forget) received in addition to the relevant also some irrelevant task information. The irrelevant information was there to make the developers get a first impression that task was smaller than it actually was. An example of such information was (translated from Norwegian): "*A college at your work place claims that he used approx. 30 minutes on a similar job. Eventually, you understand that the job was not very similar, since the paper feeding in his case was automatic [which was not the case in the specified copy machine task] and there were fewer copies to be made.*"

Those in Group Irrelevant were then, on the next page of the instruction material, given the information that the previous task description had some incorrect and irrelevant information. They were then instructed to use only the new task description when estimating the effort. The new task description included only relevant information, i.e., the same information received by those in the control group. The same treatment was given to those in Group Irrelevant-Forget, with the addition that they were instructed, in bold, large size letters, to: *Try as well as you can to FORGET the first information you received about the copy machine job.*

Results: The box-plots in Figure 3 indicate that those receiving the irrelevant and misleading information (Groups Irrelevant and Irrelevant-Forget) had lower effort estimates, in spite of that they were instructed to use the same information as those in the control group. A one-sided Kruskal-Wallis test of the difference in median estimates between those in Irrelevant-Forget and Control gives for example p=0.12. The median values for those in the control group (n=22) was 120 minutes, for those in Group Anchor (n=21) 90 minutes, and for those in Group Anchor-Forget (n=22) 76 minutes.

The perhaps most interesting result is the difference between the estimates in Groups Anchor and Anchor-Forget. The instruction to try to forget the irrelevant information that affected the first impression seemed to strengthen the effect of that information. A one-sided Kruskal-Wallis test of the difference in median effort values of those in Groups Irrelevant and Irrelevant-Forget groups give p=0.12. While not very strong evidence, this suggests that the strength of the first impression is robust and even increasing with the instruction of "try to forget" the information leading to it.

Figure 3. Box-plot of the effort estimates



In some context there may be more than one element that could lead to updates of the first impression. In Study D, we examine the effect of a sequence of either increasing or decreasing values on the first impression.

2.3 Study D

Participants: Same as in Study A. Five participants did not complete this task, leaving 103 valid responses.

Estimation task: A software system supporting doctors in the management of doctor-patient appointments.

Study design: The participants were randomly divided into three groups. All participants started with reading the requirement specification. Those in the control group then estimated the effort most likely required to develop the system. Those in the other two groups (Groups Increasing and Decreasing) were instructed to assess how likely it was to include the actual use of effort in eight different effort intervals. The only difference between Group Increase and Decrease was that while Group Increase participants did this in an effort increasing sequence (see Table 1), Group Decrease participants did this in an effort decreasing sequence, i.e, the effort intervals in Table 1 were presented in the opposite sequence. The effort estimated by other companies for this project had typically been about 300 work-hours, so the lowest value in Table 1 (10 work-hours) would be unrealistically low, while the highest value (1600 work-hours) would be very high. If the first impression, which would be established by the comparison with 10 work-hours for Group Increasing and 1600 work-hours for Group Decreasing, had a strong and lasting effect, we would expect that those in Group Increasing would believe in less work-effort required than those in Group Decreasing.

Table 1. Estimation intervals of Group Increasing

I think that the probability that I will use less than 10 work-hours is about _____%
I think that the probability that I will use between 10 and 50 work-hours is about: _____%
I think that the probability that I will use between 50 and 100 work-hours is about _____%
I think that the probability that I will use between 100 and 200 work-hours is about _____%
I think that the probability that I will use between 200 and 400 work-hours is about _____%
I think that the probability that I will use between 400 and 800 work-hours is about _____%
I think that the probability that I will use between 800 and 1600 work-hours is about _____%
I think that the probability that I will use more than 1600 work-hours is about _____%

Results: When analyzing the data, we assumed that the effort interval with the highest probability of including the actual effort would represent the software developer's estimate of most likely use of effort for those in Groups Increasing and Decreasing. For those in Group Control, we used the effort category of Table 1 that included their estimate as their estimated category of most likely use of effort. The results are displayed in Table 2. There were only very few observation in the effort intervals < 10 work-hours, 800-1600 work-hours and >= 1600 work-hours. These intervals are consequently joined with the closest effort interval.

Table 2. Proportion of most likely effort values in the intervals

Effort interval	Group Increasing	Group Decreasing	Group Control
< 50 work-hours	12% (n=4)	9% (n=3)	18% (n=7)
50-100 work-hours	34% (11)	19% (6)	18% (7)
100-200 work-hours	19% (6)	19% (6)	36% (14)
200-400 work-hours	28% (9)	25% (8)	26% (10)
>400 work-hours	6% (2)	29% (9)	3% (1)

A Chi-square analysis of independence suggests that there were significant (p=0.03) differences in the groups' estimates of most likely use of effort. As can be seen, those in Group Increasing were more likely to believe in low effort usage than those in Group Decreasing (and in Group Control). If we, for illustrative purpose, assume that the estimate is the

middle value of the category and 10 work-hours for the lowest effort interval, we find a median effort of 150, 300 and 150 work-hours for Groups Increasing, Decreasing and Control, respectively. A one-sided Kruskal-Wallis test of difference in median estimates gives p=0.03. Our results suggest that the first impression established with the first comparison has a lasting effect, even after multiple comparisons with other reference values. This is the case even when the person completing the estimation work is aware of all effort comparisons he has to do before starting the estimation work.

3. Discussion, implications and conclusion

The common finding in all four studies is that it is easy to create an incorrect first impression of effort required to solve a task, and that this first impression is rather robust towards change. In the contexts we studied it was, however, possible to debias the first impression through counter-information to some extent. Our findings on the debiasing effect of counter-information is similar to the those in [5], where instructions to "consider the opposite" after being exposed to a comparison with a too low or high reference value reduced, but did not neutralize, the estimation bias. Instructing people to try to forget the information leading to the incorrect first impression is, however, not promising as a debiasing technique and seems instead to lead to increase the strength of the first impression.

We have only studied estimation tasks in situations where the estimates are produced with high time pressure and not in real-life estimation situations. There is evidence suggesting that the effects we study tend to be weaker, although still present, in real-life settings [6]. Furthermore, our studies examine only a small sub-set of first impressions and their robustness towards change, i.e., only for the situations where the first impression is affected by a comparison with a reference effort value intentionally set too high or too low to reflect most likely use of effort and then a second comparison in the opposite direction. There are of course many other types of information that could affect first impressions and how much they are updated with new information. There is consequently a need for more studies to increase our understanding of first impressions in effort estimation contexts. It is, for example, possible that we through studies of varying contexts identify contexts where the recency effect is stronger than the first impression effect in effort estimation.

Software practice should however be based on best evidence, even when the evidence is not as strong and general as we would prefer. An implication of our findings is consequently that software developers need to be especially concerned about not being exposed to information leading to incorrect first impressions about the project they are supposed to estimate. A biased first impression, e.g., caused by an early request from the client or manager, could be very hard to debias when relevant information becomes available. Even when the information leading to the first impression is known to be irrelevant, it can create a lasting, incorrect impression of the effort required to develop software.

References:

1. Rabon, M. and J.L. Schrag, *First impressions matter: A model of confirmatory bias.* The Quarterly Journal of Economics, 1999. 114(1): p. 37-82.

2. Tubbs, R.M., W.F.J. Messier, and W.R. Knechel, *Recency effects in the auditor's belief-revision process.* The Accounting Review, 1990. 65(2): p. 452-460.

3. Jørgensen, M. and S. Grimstad, *Avoiding irrelevant and misleading information when estimating development effort.* IEEE Software, 2008. 25(3): p. 78-83.

4. Aranda, J. and S. Easterbrook, *Anchoring and adjustment in software estimation.* Software Engineering Notes, 2005. 30(5): p. 346-355.

5. Mussweiler, T., F. Strack, and T. Pfeiffer, *Overcoming the Inevitable Anchoring Effect: Considering the Opposite Compensates for Selective Accessibility.* Personality and social pscychology bulletin, 2000. 26(9): p. 1142-1150.

6. Jørgensen, M. and S. Grimstad, *The Impact of Irrelevant and Misleading Information on Software Development Effort Estimates: A Randomized Controlled Field Experimen.* IEEE Transactions of Software Engineering, 2011. To appear.