Characteristics of Software Engineers with Optimistic Predictions

Magne Jørgensen[1,2], magnej@simula.no (*)
Bjørn Faugli[2], bjorn.faugli@osir.hihm.no
Tanja Gruschke[1,3], tanjag@simula.no
[1]Simula Research Laboratory, Norway,
[2]Hedmark University College, Rena, Norway
[3]University of Oslo, Norway

(*) Corresponding author:
Address: PO Box 134, NO-1325 Lysaker, Norway
Email: magnej@simula.no
Telephone: +47 924 333 55
Fax: +47 67 82 82 01

Abstract: *This paper examines the degree to which level of optimism in software engineers' predictions is related to optimism on previous predictions, general level of optimism (explanatory style, life orientation and self-assessed optimism), development skill, confidence in the accuracy of their own predictions, and ability to recall effort used on previous tasks. Results from four experiments suggest that more optimistic software engineers are characterized by more optimistic previous predictions, higher confidence in the accuracy of their own predictions, lower development skills, poorer ability or willingness to recall effort on previous tasks, and higher optimism scores. However, a substantial part of the variation in the level of optimism seems to be random.*

Keywords: Software development effort estimation, expert judgment, predictions of optimism

1. Introduction

Organizations that develop software have, in general, a bad reputation for effort overruns. One way of reducing or eliminating the strong bias towards optimistic effort predictions would be the mechanical use of formal estimation models, which would lead to unbiased models with output unaffected by political issues and wishful thinking. These advantages have motivated a great deal of research on formal cost estimation models. However, empirical studies suggest that the more flexible method "expert estimation" is typically just as accurate (Jørgensen 2006). Possibly for that reason, most software companies rely on expert estimation and seldom use formal cost estimation models (Paynter 1996; Hill, Thomas et al. 2000).

The level of optimism and accuracy in predictions varies a lot among software professionals (Jørgensen 2006). Consequently, the ability to select software engineers who are less likely to provide strongly optimistic predictions of effort and other project parameters is essential to ensure manageable software development projects. Unfortunately, there has been little research on the selection of estimation experts in software engineering. In two prior studies, we found that more accurate and less optimistic effort estimates were, to some extent, related to the amount of relevant experience possessed by the estimators (Jørgensen and Sjøberg 2002; Jørgensen and Carelius 2004). We have been unable to find studies on other characteristics of software engineers who make optimistic predictions. The identification of such characteristics is the goal of this paper. We hope that the acquisition of more knowledge about this topic may be a first step to support software organizations' selection of software engineers who are less likely to provide strongly optimistic predictions.

We examine the following set of characteristics (indicators) that have the potential to predict software engineers' level of optimism when estimating:

1) The level of *optimism on previous prediction tasks*. We hypothesise that software engineers who have been optimistic in their previous predictions are more likely to be optimistic in subsequent predictions. If we should

find that there is no such connection, this may imply that the random component of software engineers' level of optimism dominates and that no other indicators are likely to be able to predict the level of optimism in similar contexts, either. The connection between the previous and the future level of optimism within a particular context may indicate the size of the systematic component of a person's level of optimism. Consequently, it indicates the upper boundary for the other indicators' ability to predict the level of software engineers' optimism in similar contexts.

2) The ASQ test of optimism (Attributional Style Questionnaire (Seligman 1995)) measures *explanatory style*. An optimistic explanatory style is defined as the pattern of explaining positive events by appeal to factors permanent, general and internal to the subject, while explaining negative events by appeal to factors unstable, specific and external to the subject. We hypothesise that those with a more optimistic explanatory style have a stronger tendency towards optimistic predictions than those with a less optimistic explanatory style. The motivation for this hypothesis can be exemplified by the response to the following statement (Statement 1 of the questionnaire): "*The project you are in charge of is a great success.*" The possible responses are "*I kept a close watch over everyone's work.*" and "*Everyone devoted a lot of time and energy to it.*" Selecting the first alternative would mean that the respondent thought that the reason for a project's success would typically be related to their own performance and control, while selection of the second alternative would indicate that the respondents attributed the success to less controllable and external events. Our hypothesis implies that people who select the first alternative would have a stronger tendency towards optimistic predictions of their own performance. The ASQ test is a popular test of optimism and been used in, for example, the recruitment of salesmen. The test is included as Appendix A.
3) The LOT-R test of optimism (Life Orientation Test-Reduced (Scheier and Carver 1985)) measures *life orientation* or life expectation. A positive life orientation is defined as the tendency to believe that one will generally experience good outcomes in life. We hypothesise that those with a more positive life orientation (a higher level of so-called "dispositional optimism") have a stronger tendency towards optimistic predictions than those with a less positive life orientation. We would, for example, believe that the agreement with the statement "*In uncertain times, I usually expect the best.*" (Statement 1 in the LOT-R test) is connected with optimistic predictions. The LOT-R test measures different aspects of optimism than the ASQ test, but the two tests are correlated to a certain extent. Correlations between these two measures of optimism in the range of 0.3 to 0.6 are reported in (Gutkovich, Rosenthal et al. 1999). The LOT-R test is included as Appendix B.
4) The *self-assessed level of optimism* based on the answer to the effort estimation related question: "*How optimistic are you?*" and the alternatives: *a) Much less optimistic than average, b) Somewhat less optimistic than average, c) About average, d) Somewhat more optimistic than average, e) Much more optimistic than average*. We hypothesise that those who perceive themselves as being more optimistic regarding effort estimation also have a stronger tendency towards more optimistic effort estimates than those who perceive themselves as being less optimistic. In one of the studies reported in this paper (Study D) we measured the correlation between ASQ and self-assessed level of optimism to be 0.3. The similarity of question formulations suggests that there may be a correlation between this indicator and the LOT-R scores, as well.
5) *Software development skill* is measured as the effort required to complete a software task with sufficient quality. We hypothesise that those who are better at solving a task will have less optimistic predictions about the tasks, due to a better understanding of what has to be done.
6) *Confidence in the accuracy* of the effort estimate is measured by how probable the estimator believes it is that the actual effort will fall within the effort prediction interval +/-10% of estimated effort. Previous studies, e.g., (Jørgensen, Teigen et al. 2004), suggest that software professionals are strongly overconfident regarding the accuracy of their own predictions. Confidence and optimism may be based on similar elements, e.g., a lack of willingness or ability to revise beliefs in the light of previous experience. Consequently, we hypothesise that greater confidence that the effort estimates will fall within the prediction interval is an indicator of more optimistic effort estimates. It is also possible to argue that greater confidence in the accuracy of one's estimates indicates greater knowledge about how to solve the task and, consequently, less optimistic estimates. However, the low correlation (r=0.26) between estimation accuracy and confidence found in (Jørgensen 2004) suggests that level of confidence is a poor indicator of estimation accuracy. Mainly for that reason, we adopt the first hypothesis.
7) The *ability to remember* the actual use of effort of previous tasks. We hypothesise that those with better recall are less likely to produce optimistic predictions, i.e., that it is easier to stay optimistic when one has a poor memory.

The remainder of the paper is organized as follows: Section 2 briefly describes the designs of the four empirical studies, the measures used and important study limitations. Section 3 reports the results related to each of the potential indicators of optimistic software engineers. Section 5 summarizes and concludes.

2 Design of the Studies

2.1 Study A: Prediction of Examination Marks

Twenty-five software engineering students at the University of Oslo volunteered to participate. The participants were paid for their participation in a study that collected several characteristics about their study expectations, study technique and examination results. For the purpose of the analysis presented in this paper, we collected information about the students' explanatory style (using the ASQ test) and their prediction of examination marks for a software engineering course that they took. The participants' predictions of their examination marks were collected on three occasions: at the beginning of the semester, just before the examination, and a few days after the examination (before they knew the results). Finally, the actual examination marks were collected. The degree of optimism at the three different occasions was measured as the difference between the actual and the predicted examination mark. Predicting a B grade and receiving a D, for example, yields a difference of 2. A positive difference means that a prediction was optimistic, while a negative difference means that the prediction was pessimistic.

2.2 Study B: Prediction of Effort to Complete a Software Development Project

Fourteen senior project managers from the same Norwegian software development company participated in this experiment. Their task was to estimate the most likely effort necessary to complete a specified software project. All participants received the same information (the requirement specification of an actual project completed by their own company, which had just started), and were instructed to base the effort estimate on the assumption that skill of the project team would be roughly equivalent to the average skill level of development teams within the company. None of the project managers had information about the project other than the requirement specification.

We collected information about the project managers' life orientation (using the LOT-R test) and used the estimated effort as an indicator of optimism. This use of estimated effort as indicator of optimism is based on the assumption that a project manager with low effort estimates is more likely to have an optimistic estimate of effort. Since the actual effort spent on the project is the same for all estimates and our analysis only used the relative difference in level of optimism, the use of the estimated effort instead of the deviation between estimated and actual effort (the estimation error) as a measure of optimism makes no difference to the results.

The project managers may have interpreted the specification differently. This means that both the estimated effort and the deviation between estimated and actual effort of the project are imperfect measures of prediction optimism. This may reduce the degree of connection between the indicator (the LOT-R score) and our measure of optimism (the estimated effort), i.e., that we should expect a stronger connection with a better measure of the level of optimism.

2.3 Study C: Bids for a Project

Seventy-six software professionals from different Norwegian companies participated in this experiment. Their task was to provide fixed price bids for the same projects. The requirements described a project with about 1000 paper documents that were to be scanned and stored electronically. We collected the bids for this project and the life orientation (using the LOT-R test) of each software professional. We used differences in bids as an indicator of level of optimism. This is based on the assumption that a lower bid is more likely to be based on more optimistic effort predictions.

Similarly to Study B, the software professionals may have interpreted the specification differently, which may affect the degree of connection between the indicator (the LOT-R score) and our measure of optimism (the bid).

2.4 Study D: Prediction of Effort to Complete Programming Tasks

Twenty software professionals were hired to estimate their own effort and complete the same five programming tasks. The work sequence was as follows: receive specification of Task 1, estimate effort needed to

complete Task 1, complete Task 1, receive specification of Task 2, estimate effort needed to complete Task 2, complete Task 2, etc. The quality of a task solution was tested and had to be accepted before a developer was allowed to proceed to the next task. All participants received the same requirement specifications and all specifications were of high quality regarding completeness and precision. The total effort used by the software developers to complete the five tasks varied from about 30 to 60 work-hours.

There were differences in material to support the development work and differences in type of feedback received after each task, related to research questions other than those addressed in this paper. The indicators we analyze in this paper had almost the same distributions in the groups with different types of material and feedback, i.e., it is not likely that this affected the analyses in this paper.

We collected information about the indicators: Explanatory style (using the ASQ test), self-assessed level of optimism, development skill (measured as effort to complete the programming tasks), confidence in the accuracy of their own estimates, and ability to remember the actual use of effort one year after completion. The level of optimism was measured as the mean relative error (RE) of the estimate, where RE = (Actual Effort – Estimated Effort) / Actual Effort). A higher RE-value indicates a higher level of optimism.

2.5 The Analysis of Prediction Ability of an Indicator (Hit Rates)

The strength of the connection between an indicator and the software engineer's level of optimism is termed the "hit rate" and was calculated as follows:

i) The indicator values of each software engineer were measured, e.g., a software engineer's indicator value for "explanatory style" (ASQ-score) was calculated based on answers to the questions in Appendix 1.

ii) Each software engineers' level of optimism for one or more predictions was measured, e.g., as the mean relative error (RE) of the effort estimates for five programming tasks in Study D.

iii) The set of all unique pairs of software engineers was constructed. With n software engineers we constructed $[n*(n-1)]/2$ pairs.

iv) For each pair we predicted that the software engineer with the more optimistic indicator value would have more optimistic predictions, e.g., that a software engineer with an ASQ-score of 3 would have a higher RE-value (more optimistic prediction) than a software engineer with an ASQ-score of 1. If two software engineers had the same indicator value or the same level of optimism, we removed the pair from the "hit rate" analysis of that indicator.

v) The "hit rate" was calculated, for each indicator and each study, as the proportion of correctly predicted more optimistic software engineers. A hit rate of x% means that the indicator predicted correctly in x% of the pairs. A hit rate of 50% suggests that the indicator has no predictive value, i.e., that it performs no better than a random selection of software engineers. Hit rates close to 50% indicate that the indicator has little predictive value, but not that there a lack of connection. It is, for example, possible that a statistical test of difference in mean or median values would provide significant p-values. In other words, our hit rate analysis focuses on the usefulness of an indicator to select the most optimistic software engineers and not whether there are statistically significant connections with small effect sizes between variables.

We calculated, but decided not to use, the correlation between the indicator and the measure of optimism in our analyses. The reason was that the correlation and the hit rate provided the same information in most cases. In cases where the hit rate and the correlation gave different results, a more in-depth examination and a graphical display of the data suggested that the hit rate was the measure that had greater validity and was more robust to outliers. In one case, for example, the removal of one single extreme observation changed the correlation from being strongly negative to weakly positive, while the hit rate did not change much.

2.6 Limitations of the Studies

Our results are from different types of study and use different measures of optimism. Consequently, an interstudy comparison of the hit rates of the different indicators is not without its problems. Hence, the comparisons of hit rates should be interpreted carefully, while bearing in mind the differences in study design and measures.

Our analysis of hit rates does not focus on the identification of strongly optimistic software engineers, but instead on the identification of the more optimistic of two software engineers. We need studies in situations that stimulate higher degrees of optimism than those in our studies, e.g., very large software development projects or

projects where the client expects a low price, to study whether the indicators of more optimistic software engineers are similar to those useful for identifying strongly optimistic software engineers.

Optimism and predictions are complex phenomena and include many factors not measured or analysed by our study, e.g., how optimism affects the actual work and the correlation between adapting the work to fit the estimated effort and prediction error. Consequently, it is possible to criticise our study on the basis that we have not really measured optimism, but, for example, how people use predictions to motivate themselves to work harder. This may be so, but to regard it as a criticism of our study is not to the point, because the main purpose of the paper is to identify indicators that can predict the level of prediction optimism among software engineers, i.e., the degree to which they are likely to provide effort estimates that are less than the actual effort. We use the terms "optimistic" or "prediction optimism" mainly because they are useful terms to reflect what is observed, i.e., that the actual outcome is, in some sense, worse than the predicted outcome. Whether the cause of the observed prediction optimism is what we would typically term optimism or not is not essential for our purpose.

Studies A, B and D had relatively few participants (fewer than twenty-five). This means that many of our results depend on the performance of just a few individuals, who may not constitute a representative sample. This is particularly a problem for indicators examined in only one study, but not so much for indicators examined in several studies. The robustness of the results should be interpreted with this in mind.

The use of different subjects in the different studies means that interaction effects between the studied factors are difficult to study. Consequently, we did not include them. It is possible that stronger effects than those found in our studies would be observed if we had studied combinations of factors.

A common objection is that results achieved in artificial settings do not generalize to real-world contexts. However, results from experiments should not be generalized to field settings naively. An important role of experiments in artificial settings is to understand a phenomenon with reduced noise from the environment, as compared to field settings. Generalization to real-world context then occurs by way of a better understanding of basic relationships, i.e., by theory. In the present case, the role of the experiments was to increase understanding of the strength of the connection between the indicators and the level of the optimism. This increased understanding, together with other knowledge, can be used to make testable hypotheses about the size of effects in the real world. This may be just as valid a method of generalization as statistical generalization from sample to population. To illustrate the difference in roles between laboratory experiments and field studies, suppose that we had conducted field studies of the use of the indicators to predict optimism among software engineers. We would then not be able to use the same project for all predictions and there would be more uncontrolled factors. The realism and representativeness may have been higher (but not without problems, either) in a field study, but the added uncontrolled variables would typically make it more difficult to draw conclusions. However, the ultimate test of whether an indicator is useful in the software industry or not should be based on field data.

One may argue that we should have focused on estimation accuracy and not the bias towards optimistic predictions. However, to focus on accuracy would mean that we would have to consider pessimistic predictions as well, while it is the overwhelmingly common presence of optimistic predictions that is the main problem. Further, indicators of inaccurate software engineers may be different from indicators of optimistic ones, and a focus on both accuracy and optimism would have led to a much less focused paper. There is, however, a need for studies on accuracy as well, e.g., studies similar to the study in (Tetlock 2005) observing that "foxes", i.e., those with a more ad hoc strategy for prediction, make more accurate political judgments than "hedgehogs", i.e., those with more focus on using the same prediction strategy on similar tasks.

3 The Results

3.1 Indicator 1: Level of Optimism on Previous Predictions

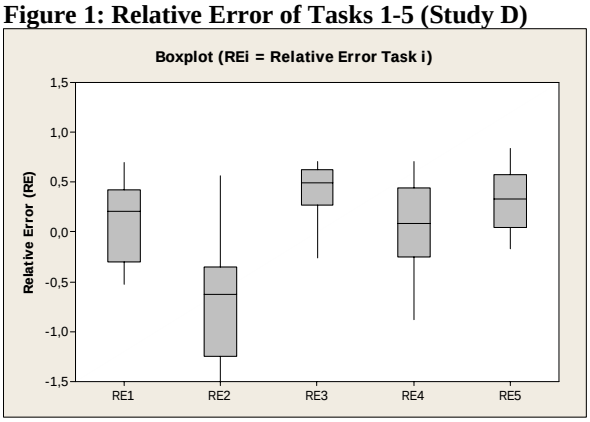
An essential precondition for the meaningfulness of analyzing indicators of optimistic experts in software development effort estimation is that some experts are systematically less optimistic than others. Even in situations with random variance of performance, we easily get the impression that some experts are better. This has, for example, been illustrated in analyses of mutual fund performance. Although the empirical evidence suggests strongly that no mutual fund performs systematically better than others (or better than the stock market index), most investors believe that some mutual funds are better and use patterns of historical performance to support the selection (Lichtenstein, Kaufman et al. 1999).

Study D: We examined the size of the systematic component of the level of prediction optimism, i.e., the degree to which previous prediction optimism could be used to predict future prediction optimism in Study D. All

developers in that study had extensive relevant experience and were quite homogenous with respect to background and skill. It is therefore to be expected that the use of previous optimism as an indicator of future optimism will be better in contexts with larger variance among the software developers, as in Study D.

The median relative error (median RE) of all tasks and all developers in Study D was 14%, which means there was a weak tendency towards optimistic effort estimates. Weak optimism, or even pessimism, seems to be typical when estimating the effort of small tasks (Jørgensen and Moløkken-Østvold 2004). The visual display of the data in Figure 1 indicates that a typical pattern may be that the degree of optimism regarding effort estimation among developers is strongly affected by the performance on estimating the most recently completed task, i.e., an effort overrun of the previous task leads to more pessimism for the subsequent task. However, this tendency may also be caused by differences in task characteristics.

Figure 1: Relative Error of Tasks 1-5 (Study D)



If there is a tendency to overreact to the estimation optimism outcome of the previous task, as suggested by Figure 1, this means that an analysis of systematic patterns in a software engineer's level of prediction optimism should integrate the predictions of at least two tasks. For this reason, we decided to analyze how well a software developer's level of optimism (mean RE) for Tasks 1-3 predicted the level of optimism for Tasks 4-5.

The hit rate when using differences in mean RE of Tasks 1-3 to predict the more optimistic software engineer on Tasks 4-5 was 68%. This means that when selecting the less optimistic estimator out of two, based on the previous level of optimism, we would choose correctly in about two thirds of the cases.

Although the hit rate is not impressive, the results nevertheless suggest that the level of optimism was not totally random, i.e., some developers seem to be systematically less optimistic than others. It does, however, also suggest that there is a substantial random component in the measured level of optimism, i.e., that we cannot expect any optimism indicator to be very accurate. In situations like that in Study D, a hit rate of about 68% may be an upper boundary value. If, for example, an indicator has a hit rate of 60% in similar situations, this indicator consequently describes about (60-50)/(68-50) = 56% of the systematic component of the level of optimism.

3.2 Indicator 2: Explanatory Style (The ASQ Test of Optimism)

A software engineer with an optimistic explanatory style would, as described in the introduction, explain positive events by appeal to factors permanent, general and internal to the subject, while explaining negative events by appeal to factors unstable, specific and external to the subject. An optimistic explanatory style may motivate software engineers to work harder and have a stronger belief in their own skill, both attributes that are basically positive for efficient project work. It may, on the other hand, be negative for the ability to learn from estimation experience and, for example, to get a realistic view regarding one's own ability to control projects.

The explanatory style indicator was evaluated in two studies, Study A (prediction of examination mark in a software engineering course) and Study D (prediction of effort to complete programming tasks). The calculation of the ASQ-score is described in (Seligman 1995).

Study A: The level of optimism for the prediction of examination marks, dependent on closeness to the knowledge of the examination results, is illustrated in Table 1.

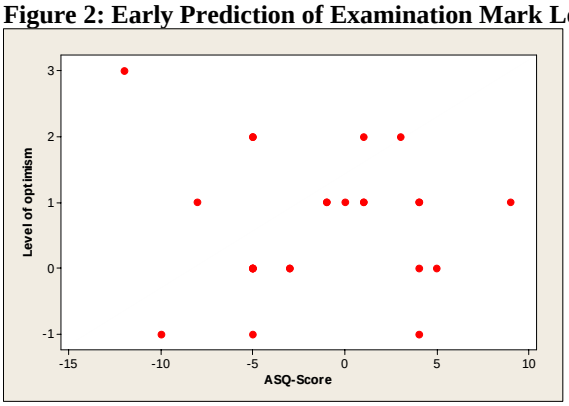
Table 1: Distribution of Predicted and Actual Examination marks

Examination mark	Early Prediction	Just Before Examination Prediction	Just After Examination Prediction	Actual Examination Mark
A	2	1	0	3
B	14	9	5	6
C	9	14	15	9
D	0	1	3	7
E	0	0	2	0
F	0	0	0	0

At the beginning of the semester (Early Prediction), there was a strong bias towards optimistic predictions, e.g., sixteen students believed they would get an A or B, while only nine students actually achieved one of these grades. The students became more realistic as the date of the examination loomed closer (Just Before Examination Prediction) and slightly pessimistic just after the examination (Just After Examination Prediction). One potential reason for this is the, conscious or unconscious, strategic use of optimistic predictions. At the beginning of the semester, optimistic predictions may stimulate harder work. Around the examination period, when there is little or no possibility of affecting the outcome, a less optimistic outlook may be useful to avoid disappointment. This explanation may also be relevant to predictions of software development effort, i.e., optimism may be higher when the time of evaluation is far away. An alternative explanation is, of course, that the students knew more about their own performance when they got closer to the examination. However, the explanation that there is a shift from prediction optimism about examination marks towards more realism, and even pessimism, when getting closer to the examination is supported by findings reported in several studies, e.g., (Manger and Teigen 1988). It seems, therefore, to be a robust finding that the time horizon has an important role to play regarding the level of optimism.

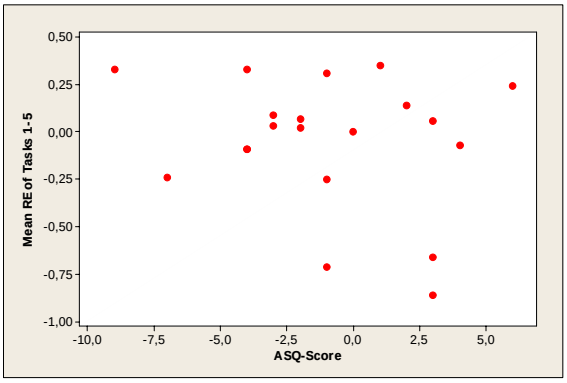
The main analysis of this study is the connection between the ASQ score and the level of optimism of examination mark prediction, i.e., whether a high ASQ score predicts optimistic examination mark predictions or not. The hit rates when using the ASQ score as an indicator of optimism in the early, just-before-examination, and just-after-examination predictions were 54%, 52%, and 45%, respectively, i.e., quite low hit rates. Figure 2 displays the data. Interestingly, the student with the highest level of prediction optimism (predicted A, got a C, i.e., a prediction optimism level of 3), had the most pessimistic explanatory style (ASQ Score of -12). This illustrates that the ASQ score can be quite misleading.

Figure 2: Early Prediction of Examination Mark Level of Optimism vs ASQ Score



Study D: There was a similarly poor connection between ASQ score and level of optimistic predictions in Study D, where software professionals estimated and completed programming tasks. The hit rate was only 53%. Figure 3 displays the data.

Figure 3: Mean Relative Estimation Error Tasks 1-5 vs ASQ Score

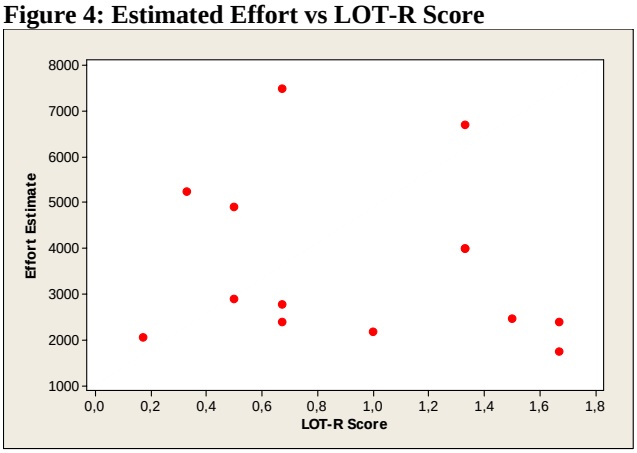


3.3 Indicator 3: Life Orientation (The LOT-R test of optimism)

The software engineers' life orientation ("dispositional optimism") was measured using the LOT-R test in Studies B and C. The LOT-R questionnaire is included as Appendix 2. Notice that statements 2, 5, 6, and 8 of the questionnaire are just filler items and not used for the analysis of optimism. Statements 1, 4, and 10 are positive statements, while statements 3, 7 and 9 are negative statements. To calculate the LOT-R Score we gave the answer alternative A the value 2 to for positive and -2 for negative statements, B the value 1 for positive and -1 for negative statements, C the value 0, D the value -1 for positive and -1 for negative statements, and E the value -2 for positive and 2 for negative statements. The LOT-R score was calculated as the mean of the values of the six relevant questions. The higher is the LOT-R score, the more optimistic is the life orientation.

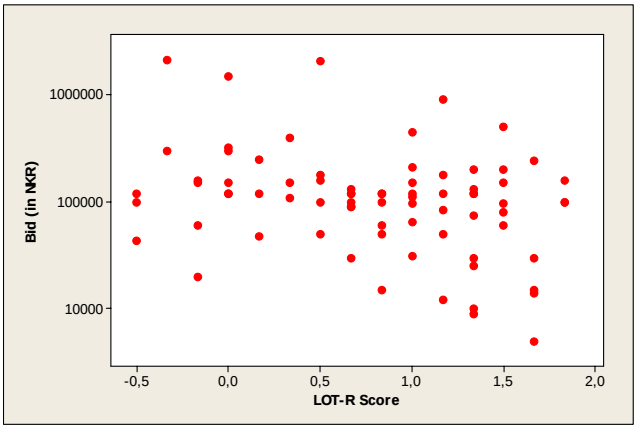
Study B: The senior project managers in Study B had a mean LOT-R score of 0.95. The hit rate when applying the LOT-R score to predict effort estimate was 56%. The data are displayed in Figure 4.

Figure 4: Estimated Effort vs LOT-R Score



Study C: The software professionals in Study C had a mean LOT-R score of 0.80, i.e., the average level of dispositional optimism was similar to that in Study B. The hit rate was 60%. The connection between bid (in Norwegian Kroner) and LOT-R Score is displayed in Figure 5, which uses a logarithmic scale on the y-axis, due to the very high variation in bids.

Figure 5: Bid vs LOT-R Score



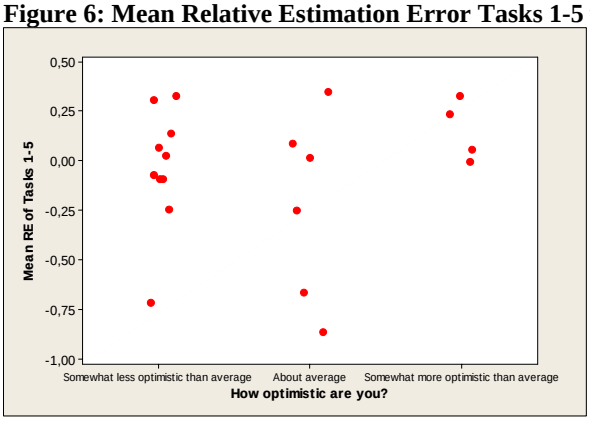
The LOT-R test performed slightly better than the ASQ test. When considering that studies B and C used imperfect measures of prediction optimism, this means that the LOT-R test may, in reality, have a better hit rate than indicated in the analyses in other situations. The prediction tasks of ASQ and LOT-R were, however, different, so we need more studies to examine the robustness of this finding.

3.4 Indicator 4: Self-assessed Level of Estimation Optimism

Study D: In Study D, we asked the software developers to assess how optimistic they were. Self-assessed level of estimation optimism is a more direct and easier way of determining the level of optimism than the ASQ and LOT-R tests. The alternative responses to the estimation-related question "How optimistic are you?" were as follows: a) Much less optimistic than average, b) Somewhat less optimistic than average, c) About average, d) Somewhat more optimistic than average, e) Much more optimistic than average.

None of the developers used the extreme values a) and e). The self-assessed level of estimation optimism yielded a hit rate of 56%. Figure 6 displays the relative estimation error in relation to the self-assessed level of estimation optimism.

Figure 6: Mean Relative Estimation Error Tasks 1-5 vs Self-Assessed Level of Optimism

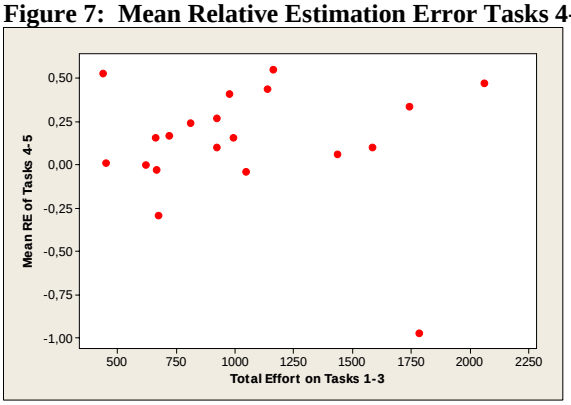


3.5 Indicator 5: Software Development Skill

Study D: The development skill of software professionals in Study D was measured as the total effort spent on completing Tasks 1-3. The use of effort is mainly an indication of the speed with which programming tasks are solved with acceptable quality and is not intended to cover all aspects of development skill. The level of optimism was measured as the mean relative error (RE) on Tasks 4-5. We separated the tasks for measuring skill (Tasks 1-3) and level of optimism (Tasks 4-5), because we could not rule out the possibility that the same nonstudied factor caused both a low use of effort and a higher level of optimism on the same task. Our hypothesis was, as described earlier, that those more skilled (lower use of effort) would be less optimistic.

We found a hit rate of 56%. The data are displayed in Figure 7.

Figure 7: Mean Relative Estimation Error Tasks 4-5 vs Programming Skill on Task 1-3

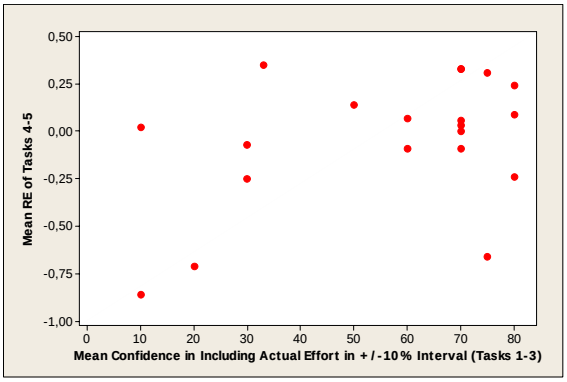


3.6 Indicator 6: Confidence in the Accuracy of One's Own Estimates

Study D: Confidence in the accuracy of the software professionals' own estimates was measured as their assessed probability that the actual effort would fall within the interval +/-10% of their estimated effort. For example, if a developer had estimated that he would use 10 work-hours on a task, he was asked to assess how likely it was that the actual effort would fall within the interval 9 - 11 work-hours. Using an argument similar to that in Section 3.5, we decided to use the mean confidence level of Tasks 1-3 to predict the level of optimism on Tasks 4-5. The analysis of how well the confidence level predicted the more optimistic estimator resulted in a hit rate of 61%.

The data displayed in Figure 8 shows that most developers believed that it was at least 60% probable that the actual effort would fall within the +/-10% interval. However, the overall frequency with which the actual effort fell within the interval +/-10% of the estimated effort was only 15%, i.e., the general level of overconfidence was high.

Figure 8: Mean Relative Estimation Error Tasks 4-5 vs Confidence in Estimation Accuracy of Tasks 1-3



3.7 Indicator 7: Memory

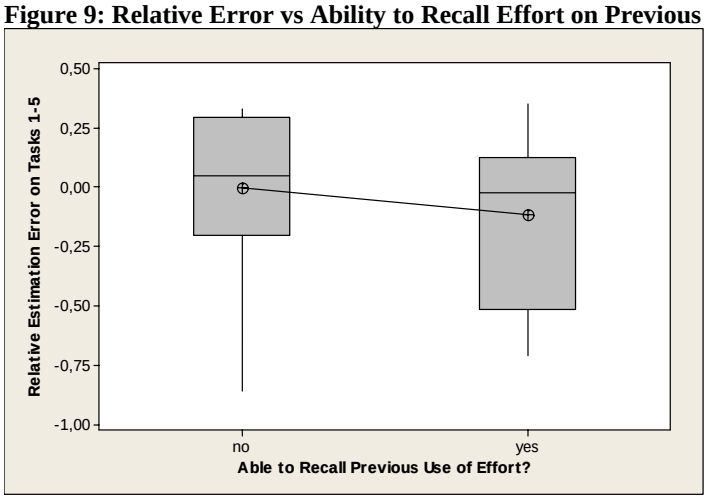
Study D: As argued earlier, a poorer ability to recall previous use of effort may make it easier to stay optimistic. Approximately one year after the five tasks in Study D were estimated and completed, we asked the same software developers to try to recall the effort they had used on the five tasks. If they could not remember it, they were asked to re-estimate the task. Eighteen of the 20 software engineers responded to our request.

Only eight of the 18 programmers based their answers on memory of the effort they used; the other 10 chose to re-estimate. The recall of those who remembered was not very accurate and they had, somewhat surprisingly, a strong tendency to pessimistic recall of their use of effort. On average, the actual effort was only about 50% of the recalled effort, both for those who recalled and those who re-estimated! The reason for this is not

clear, but it may be that they included activities that were not part of the original estimate in the recalled effort or the re-estimate. Even when adjusting for that, there seems to be an interesting tendency towards pessimism when looking back at actual use of effort. The hypothesis of optimistic recall as a reason for optimistic predictions, as suggested in (Roy, Christenfeld et al. 2005), is consequently not supported by our data. Our data is more in line with the pessimistic recall reported in (Buehler, Griffin et al. 1994). It is possible that, whether recall of effort is pessimistic, realistic or optimistic is context-dependent.

The mean RE of the original effort estimates of the eight developers who claimed to remember the actual effort was 0.0 (no bias towards optimism), and 0.12 (weak bias towards optimism) for those re-estimating. The hit rate, predicting more optimism when not remembering the previous use of effort, was 56%. A boxplot of the data is displayed in Figure 9.

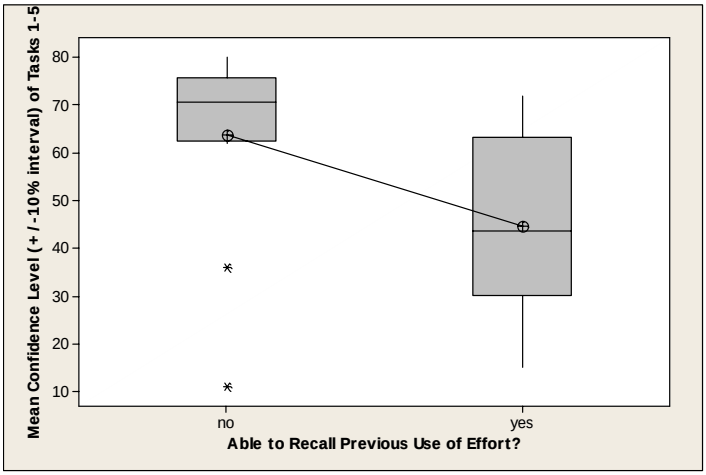
Figure 9: Relative Error vs Ability to Recall Effort on Previous Tasks



The analysis of hit rate is based on the assumption that those who recalled the use of effort had a better memory. However, it is possible that, in reality, we measured the belief that the developers had in their own memory, or their willingness to try to remember, rather than the actual ability to recall previous use of effort. Consequently, more studies should be conducted to examine the relationship between differences in memory and level of optimism. It is possible, for example, that the willingness to try to recall, together with a general tendency towards pessimistic recall, is the main reason for the level of optimism observed, and not differences in memory.

We also compared the level of overconfidence for the +/-10% interval (overconfidence = mean confidence level– proportion of actual effort included in the +/-10% interval), for the two recall groups. Here, the difference was even larger between those who recalled and those who re-estimated; see Figure 10. The mean level of overconfidence of those recalling the effort was 30%, while that of those who did not recall the effort was as high as 49%. This suggests that recall characteristics may be even more strongly related to overconfidence than to optimism.

Figure 10: Confidence Level vs Ability to Recall Effort on Previous Tasks


Mean Confidence Level (+/-10% interval) of Tasks 1-5
80
70
60
50
40
30
20
10
no
yes
Able to Recall Previous Use of Effort?

4 Summary and Conclusion

The hit rates for each indicator are summarized in Table 2.

Table 2: Summary of Hit Rates

Indicator	Study and Hit Rate
Level of optimism (RE) on previous predictions	Study D: 68%
Explanatory style (ASQ)	Study A: 54% (early prediction), 52% (just before examination prediction), 45% (just after examination prediction) Study D: 53%
Life orientation (LOT-R)	Study B: 56% Study C: 60%
Self-assessed estimation optimism	Study D: 56%
Software development skill	Study D: 56%
Confidence in accuracy of estimate	Study D: 61%
Ability to recall previous use of effort	Study D: 56%

As expected, nearly all hit rates were between 50% (no connection) and 68% (a possible hit rate boundary in situations similar to that in Study D). The results suggest that there may be connections between all of the studied indicators and the level of prediction optimism. The most important results from our study are, we believe, the following:

- Optimism on previous predictions was the best indicator of a software engineer's optimism on future predictions.
- Although the variation in levels of optimism among software engineers was not totally random, the level of optimism seemed to have a substantial random component. Otherwise, we should have expected even higher hit rates when applying the previous level of optimism as a predictor of the future level of optimism. As a consequence, we should not expect any indicator or model to be able to select the software engineer with the least optimistic predictions with high precision.
- If one intends to use the general level of optimism as indicator of optimistic predictions, simply ask the software engineer whether he assesses himself to be more or less than averagely optimistic for the particular type of predictions requested. The more complex measures of optimism (ASQ and LOT-R) did not add much predictive value and were sometimes quite misleading.
- Software engineers who were strongly confident in the accuracy of their predictions typically had more optimistic predictions than those who were less confident. This result may be particularly important, since more confidence may easily be interpreted as an indicator of more accurate predictions.
- There was a relation between ability (or willingness) to recall effort on previous tasks and level of optimism in predictions and an even stronger relation between recall ability and overconfidence in predictions.

The question of whether some, or a set of, the examined indicators can or should be used to select less optimistic software engineers is not straightforward to answer and depends on, among other factor, the homogeneity of the sample of software engineers and the importance of decreasing the level of optimism of the predictions. A selection based on our findings would emphasize the identification of software engineers with the following characteristics (in prioritized sequence):

- More realistic predictions on previous tasks
- Lower confidence in accuracy of own predictions
- Higher development skill
- Better ability to recall effort on previous tasks
- Self-perception as being less than averagely optimistic

In addition, as shown in prior research, the selection should emphasize whether or not a software engineer has experience with similar tasks, because a greater amount of highly relevant experience seems to be an indicator of more realistic predictions.

References

Buehler, R., Griffin, D. and Ross, M. 1994. Exploring the "Planning fallacy": Why people underestimate their task completion times. Journal of Personality and Social Psychology **67**(3): 366-381.

Gutkovich, Z., Rosenthal, R. N., Galynker, I., Muran, C., Batchelder, S. and Itskhoki, E. 1999. Depression and Demoralization Among Russian-Jewish Immigrants in Primary Care. Psychosomatics **40**: 117-125.

Hill, J., Thomas, L. C. and Allen, D. E. 2000. Experts' estimates of task durations in software development projects. International Journal of Project Management **18**(1): 13-21.

Jørgensen, M. 2004. Realism in assessment of effort estimation uncertainty: It matters how you ask. IEEE Transactions on Software Engineering **30**(4): 209-217.

Jørgensen, M. 2006. Estimation of software development work effort: Evidence on expert judgment and formal models. To appear in International Journal of Forecasting.

Jørgensen, M. and Carelius, G. 2004. An Empirical Study of Software Project Bidding. IEEE Transactions of Software Engineering **30**(12): 953-969.

Jørgensen, M. and Moløkken-Østvold, K. J. 2004. Reasons for Software Effort Estimation Error: Impact of Respondent Role, Information Collection Approach, and Data Analysis Method. IEEE Transactions on Software Engineering **30**(12): 993-1007.

Jørgensen, M. and Sjøberg, D. I. K. 2002. Impact of experience on maintenance skills. Journal of Software Maintenance and Evolution: Research and practice **14**(2): 123-146.

Jørgensen, M., Teigen, K. H. and Moløkken-Østvold, K. J. 2004. Better sure than safe? Over-confidence in judgement based software development effort prediction intervals. Journal of Systems and Software **70**(1-2): 79-93.

Lichtenstein, D., Kaufman, P. J. and Bhagat, B. 1999. Why Consumers Choose Managed Mutual Funds Over Index Funds: Hypotheses from Consumer Behavior. The journal of consumer affairs **33**(1): 187-205.

Manger, T. and Teigen, K. H. 1988. Time Horizon in Students' Predictions of Grades. Scandinavian Journal of Educational Research **32**(2): 77-91.

Paynter, J. (1996). Project estimation using screenflow engineering. International Conference on Software Engineering: Education and Practice, Dunedin, New Zealand, IEEE Comput. Soc. Press, Los Alamitos, CA, USA: 150-159.

Roy, M. M., Christenfeld, J. S. and McKenzie, C. R. M. 2005. Underestimating the Duration of Future Events: Memory Incorrectly Used or Memory Bias? Psychological Bulletin **131**(5): 738-756.

Scheier, M. F. and Carver, C. S. 1985. Optimism, coping and health: Assessment and implications of generalized outcome expectancies. Health Psychology **4**: 219-247.

Seligman, M. E. P. 1995. Learned optimism. New York, Knopf.

Tetlock, P. E. 2005. Expert political judgment: how good is it? how can we know?, Princeton University Press.

Appendix 1: ASQ Questionnaire

Read the description of each situation and vividly imagine it happening to you. Then choose the cause that is more likely to apply to you. You may not like the way some of the responses sound, but don't choose what you think you should say or what would sound right to other people; choose the response that's most like you.

1. The project you are in charge of is a great success.
I kept a close watch over everyone's work.
Everyone devoted a lot of time and energy to it.

2. You and your spouse / partner / boyfriend / girlfriend make up after a fight.
I forgave him/her.
I'm usually forgiving.

3. You get lost driving to a friend's house.
I missed a turn.
My friend gave me bad directions.

4. Your spouse / partner / boyfriend / girlfriend surprises you with a gift.
He/she just got a raise at work.
I took him/her out to a special dinner the night before.

5. You forget your spouse's / partner's / boyfriend's / girlfriend's birthday.
I'm not good at remembering birthdays.
I was preoccupied with other things.

6. You get a flower from an admirer.
I am attractive to him/her.
I am a popular person.

7. You run for a community office position and you win.
I devote a lot of time and energy to campaigning.
I work very hard at everything I do.

8. You miss an important engagement.
Sometimes my memory fails me.
I sometimes forget to check my appointment book.

9. You run for a community office and you lose.
I didn't campaign hard enough.
The person who won knew more people.

10. You host a successful dinner.
I was particularly charming that night.
I am a good host.

11. You stop a crime by calling the police.
A strange noise caught my attention.
I was alert that day.

12. You buy your spouse/partner/boyfriend/girlfriend a gift and he/she doesn't like it.
I don't put enough thought into things like that.
He/she has very picky tastes.

13. You gain weight over the holidays and can't lose it.
Diets don't work in the long run.
The diet I tried didn't work.

14. Your stocks make you a lot of money.
My broker decided to take on something new.
My broker is a top notch investor.

15. You win an athletic contest.
I was feeling unbeatable.
I train hard.

16. Your fail an important examination.
I wasn't as smart as the other people taking the exam.
I didn't prepare for it well.

17. Your boss gives you too little time to finish a project, but you get it finished anyway.
I am good at my job.
I am an efficient person.

18. You lose a sporting event for which you have been training for a long time.
I'm not very athletic.
I'm not good at that sport.

19. Your car runs out of gas on a dark street late at night.
I didn't check to see how much gas was in the tank.
The gas gauge was broken.

20. You lose your temper with a friend.
He/she is always nagging me.
He/she was in a hostile mood.

21. You are penalized for not returning your income tax forms on time.
I always put off doing my taxes.
I was lazy about getting my taxes done this year.

22. You ask a person out on a date and he/she days "no."
I was a wreck that day.
I got tongue-tied when I asked him/her to the dance.

23. A game show host picks you out of the audience to participate in the show.
I was sitting in the right seat.
I looked the most enthusiastic.

24. You save a person from choking to death.
I know a technique to stop someone from choking.
I know what to do in a crisis situation.

Appendix 2: LOT R

Please be as honest and accurate as you can throughout. Try not to let your response to one statement influence your responses to other statements. There are no "correct" or "incorrect" answers. Answer according to your own feelings, rather than how you think "most people" would answer.

A = I agree a lot
B = I agree a little
C = I neither agree nor disagree
D = I DISagree a little
E = I DISagree a lot

	Answer
1. In uncertain times, I usually expect the best.	____
2. It's easy for me to relax.	____
3. If something can go wrong for me, it will.	____
4. I'm always optimistic about my future.	____
5. I enjoy my friends a lot.	____
6. It's important for me to keep busy.	____
7. I hardly ever expect things to go my way.	____
8. I don't get upset too easily.	____
9. I rarely count on good things happening to me.	____
10. Overall, I expect more good things to happen to me than bad.	____

Biographies:

Magne Jørgensen received the Diplom Ingeneur degree in Wirtschaftswissenschaften from the University of Karlsruhe, Germany, in 1988 and the Dr. Scient. degree in informatics from the University of Oslo, Norway in 1994. He has several years of industry experience as software developer, project leader and manager. He is now professor in software engineering at University of Hedmark, Rena and the leader of the software cost estimation research group at Simula Research Laboratory. His main research interest is judgment-based software development effort estimation.

Bjørn Faugli has a degree in electronic engineering from the University of Birmingham, UK, in 1975 and is now an assoc. professor in informatics at Hedmark University College, Rena and at the University of Oslo. He has experience as software developer, project leader and manager of a research institution. His main research interests is systems development in combination with ICT supported learning processes.

Tanja M. Gruschke received the Cand. Scient. degree in Software Engineering from the University of Oslo, Norway, in 2004. She is currently working on her Ph.D. degree in informatics at the software cost estimation research group at Simula Research Laboratory. Her Ph.D. research is on how to train professional software developers to better assess cost estimation uncertainty.