Numerical anchors and their strong effects on software development effort estimates

Erik Løhre[a, b] and Magne Jørgensen[a]

[a] Simula Research Laboratory, Oslo

P.O. Box 134, 1325 Lysaker, Norway

[b] Department of Psychology, University of Oslo

P.O. Box 1094 Blindern, 0317 Oslo, Norway

Corresponding author:

Erik Løhre

Simula Research Laboratory

P.O. Box 134, 1325 Lysaker, Norway

Telephone: +47 41 42 01 35

erikloh@simula.no

Second author:

Magne Jørgensen

Simula Research Laboratory

P.O. Box 134, 1325 Lysaker, Norway

magnej@simula.no

Published version: http://dx.doi.org/10.1016/j.jss.2015.03.015

Abstract (152 words): *The anchoring effect may be described as the tendency for an initial piece of information to influence people's subsequent judgement, even when the information is irrelevant. Previous studies suggest that anchoring is an important source of inaccurate software development effort estimates. This article examines how the preciseness and credibility of anchoring information affects effort estimates. Our hypotheses were that anchors with lower numerical precision and anchor sources with lower credibility would have less impact on effort estimates. The results from three software project effort estimation experiments, with 381 software professionals, support previous findings about the relevance of anchoring effects to software effort estimation. However, we found no decrease in the anchoring effect with decreasing anchor precision or source credibility. This suggests that even implausible anchors from low-credibility sources can lead to anchoring effects, and that all kinds of misleading information potentially acting as estimation anchors in project estimation contexts should be avoided.*

Keywords: software projects; cost estimation; human judgement; anchoring effects; numerical preciseness; source credibility

1. Introduction

Imagine that a software developer is in a meeting about a new, large project, and that a client asks whether the developer thinks it will take less than 20 hours to complete the new project. Although this number is absurdly low, will it affect the final estimate that the software developer produces? Previous research, see Section 1.1, suggests that the software developer will actually give a lower estimate of how long it will take to complete the project after getting this irrelevant question than he otherwise would have. *Numerical anchors* of this kind can strongly influence estimates of software development effort, but are there some contexts in which anchors have stronger effects than others? In this paper, we investigate whether the numerical precision of the anchor and the credibility of the source of the anchor can moderate the strength of the anchoring effect in a software development effort estimation context.

1.1 Background

Estimates of software project cost and effort are necessary for several purposes, e.g., planning, budgeting and bidding. The consequences of highly inaccurate estimates can be severe. If an effort or cost estimate is too low, the provider may choose to produce the product with lower than desired quality to avoid financial losses, the delivery may be delayed with the consequence that the client loses market opportunities, or the profitability of the project can become negative, i.e., the client would not have started the project if presented with a realistic estimate. Too high an effort or cost estimate may result in inefficient resource use and lost business opportunities, e.g., providers may lose bidding rounds due to the price being too high.

By far the most common way to estimate the effort and cost of software systems, in spite of years of research on formal software effort estimation models, is to ask developers with experience in the field to give their best judgement of the most likely effort needed to develop the system (Jørgensen, 2004). Unfortunately, human experts are not always as good at estimating as one could hope: estimates of cost and effort in software projects are often inaccurate, with an average overrun of about 30% (Halkjelsvik & Jørgensen, 2012).

According to the influential heuristics and biases approach to judgement and decision making (Kahneman, 2003), our judgements are frequently based on heuristics, i.e., rules of thumb or mental strategies that satisfice rather than optimize. These heuristics "*reduce[s] the complex task of assessing probabilities and predicting values to simpler judgmental operations"* (Tversky & Kahneman, 1974). When there is a good match between the context and the heuristics, the use of heuristics will frequently produce accurate predictions. Sometimes, however, the use of heuristics leads to biased judgement and poor predictions (Kahneman, Slovic, & Tversky, 1982). Hence, although expert judgement-based effort estimates may be reasonably accurate in some contexts, there are also contexts in which reliance on judgemental heuristics leads to highly inaccurate estimates.

The anchoring effect is one of the best-documented findings in the heuristics and biases approach (Klein et al., 2014). Anchoring occurs when judgements are influenced by an initially presented value (the anchor value). An example could be a client or a manager with unrealistically low cost expectations asking a developer whether he/she thinks a task will take more than three days. The developer´s estimate will then tend to be closer to the anchor value than it would have been had the anchor not been presented. In this example, anchoring could therefore lead to too low estimates, with potential negative consequences such as delays or budget overruns. An anchor tends to influence the subsequent judgment even when participants are explicitly informed that the presented value is not relevant to the judgement in question. Researchers have established the anchoring effect using many different kinds of target judgements and many different kinds of anchors (Furnham & Boo, 2011), including important real-world judgements, such as criminal trial judges' sentencing decisions (Englich & Mussweiler, 2001) and real estate agents' estimates of the value of a property (Northcraft & Neale, 1987).

Several studies have documented the relevance of the anchoring effect in effort estimation:

- Jørgensen and Sjøberg (2004) gave computer science students and software professionals information about customer expectations – they told one group that the client believed that 50 hours and another group that the client believed that 1000 hours would be a reasonable estimate for the total cost of a software project. Even though the participants were informed that the client knew very little about the time needed and that they should not consider this information as relevant, the anchors strongly influenced both students and professionals, with estimates in the "1000 hours"-group much higher than the estimates of the "50 hours"-group. Follow-up questions revealed that the participants were not aware of or strongly underestimated this influence (Jørgensen & Sjøberg, 2004).
- Aranda and Easterbrook (2005) investigated anchoring with students and professionals in a software project effort estimation context, in which the anchoring information was given in statements such as "I admit I have no experience with software projects, but I guess this will take about 2 months to finish". The participants who received a high anchor (20 months) gave much higher estimates than the participants who received a low anchor (2 months) or no anchor at all.
- König (2005) gave student participants a low anchor, a high anchor or no anchor, before asking them to estimate the time needed to find the answers to a set of questions about different items in a commercial catalogue (e.g., "How heavy is the iron folding chair?"). He found the predicted anchoring effect, with the estimates being the lowest in the low anchor group and the highest in the high anchor group, with the control group in between (König, 2005).
- Thomas and Handley (2008) asked student participants to estimate the time it would take them to build a miniature plastic castle by following a set of instructions. The researchers presented anchor values as the time a random participant had spent on the same task and in a second experiment as the duration of an unrelated task. They found anchoring effects on the time estimates (estimates of the low anchor group < estimates of the control group < estimates of the high anchor group) in both experiments (Thomas & Handley, 2008).

- Research has demonstrated anchoring effects in software professionals' project effort estimates even when using an extreme anchor (4 hours for a task estimated by the control group to take 160 hours) (Jørgensen & Grimstad, 2008). The same study also found that describing a project as a "minor extension", which may lead the developers to anchor their estimates in smaller tasks, led to a lower effort estimate than when the same project was described as developing "new functionality" (for similar results, see Jørgensen & Grimstad, 2012).
- A study has shown similar tendencies for an anchoring effect in a field context, in which software developers in different outsourcing companies in Eastern Europe and East Asia estimated the same software projects with different estimation anchors (Jørgensen & Grimstad, 2010).
- Jørgensen and Løhre (2012) showed that when an anchoring value is followed by another anchor at the other extreme of the scale (for instance, a low anchor followed by a high anchor), it is the first anchor that exerts the strongest influence on the final effort estimate. Hence, one cannot necessarily neutralize an anchor by presenting an alternative value pointing in the opposite direction. The study also showed that an instruction to forget the anchor does not decrease the anchoring effect – in fact, if anything, the effect on the effort estimate is slightly stronger after such an instruction (Jørgensen & Løhre, 2012).

Overall, previous studies demonstrate that the presentation of an anchor value influences estimates of project or task effort and that the anchoring effect is important to keep in mind for professionals involved in estimation work. This also means that it is important to identify the factors or situations in which the anchoring effect is amplified or attenuated. Software developers could use knowledge of such factors to take extra precautions against specific situations or as a guide to when anchoring effects are less likely to pose a serious risk to the realism of the project's effort estimate.

1.2 Anchoring as communication

In this paper, the attitude change theory of anchoring (Wegener, Petty, Detweiler-Bedell, & Jarvis, 2001) inspired us to view anchoring as a communication process. While traditional studies of anchoring take great care to emphasize to the participants that the anchor is not relevant to the judgement in question (for instance by generating the anchor using a “wheel of fortune”), in a natural communication process numerical values that could influence effort estimates are often introduced as a more or less relevant part of a conversation. In such cases, the participants can allot some informational relevance to the anchor values. For instance, a project manager might ask whether a project will take more than a certain number of hours, just to gain a rough idea of the scope of the project. With this kind of approach, one can hypothesize that subtle changes in different parts of the communication process can influence the anchoring effect. Here, we focus on two somewhat related factors of particular interest regarding anchoring in software development effort estimation, namely how a difference in the numerical preciseness of the anchor and/or the perceived credibility (expertise) of the person providing the anchor affects the strength of the anchoring effect.

Studies within the attitude change approach have found that anchors from highly credible (expert) sources lead to stronger anchoring effects than anchors from less credible (non-expert) sources (Wegener, Blankenship, Petty, & Detweiler-Bedell, 2009; cited in Wegener, Petty, Blankenship, & Detweiler-Bedell, 2010). Another study within the same approach found that more extreme (and hence less credible or plausible) anchors can have less influence on judgements than more moderate anchors (Wegener et al., 2001). A study suggesting that less reliable sources lead to the removal of the framing effect also supports the moderating effect of credibility (Jørgensen, 2013). Together, these studies indicate that both the source of the anchor value and the exact value of the anchor can be important moderators of the anchoring effect.

Relatedly, a recent study in a price negotiation context found that high preciseness of the initial offer indicated more expertise on the anchor provider side and led to stronger anchoring effects than when the initial offer was a round number (Mason, Lee, Wiley, & Ames, 2013). Another similar study (Zhang & Schwarz, 2013), using a more traditional anchoring procedure, also found an increased anchoring effect with more precise numbers, but only when the number was pragmatically relevant. Mason et al. (2013) argue that people see a precise number as indicating a greater level of knowledge and therefore consider it to be more informative of the true value. This explanation again points to a possible influence of the perceived expertise or credibility of the source of the anchor.

In software effort estimation contexts, this could mean that exposure to round anchor values or anchor values from sources without expertise would introduce less bias than exposure to more precise anchors or anchors from competent or relevant sources (provided that the anchors were equally off the mark). Although previous studies of anchoring in effort estimation suggest that informing participants about the low competence of the source of the anchoring value does not eliminate the anchoring effect (Aranda & Easterbrook, 2005; Jørgensen & Sjøberg, 2004), anchor values from competent sources have not been compared directly with anchor values from non-competent sources. Therefore, we do not know whether the effect of anchors on effort estimates is reduced when the anchor stems from a less competent source.

1.3 Research questions and hypotheses

In the current experiments, we hypothesize that we will find a strong effect of initially presented numerical values on software project effort estimates, replicating previous studies. In addition, we examine two related questions not addressed in previous papers on anchoring in the domain of effort estimation:

Q1: If the anchor value is imprecise (round), does it reduce the anchoring effect? For example, does the question of whether the project will take more than 1000 hours influence an effort estimate less than the question of whether it will take more than 998 work-hours?

Q2: If it is clear that the source of the anchor is less credible, does it reduce the anchoring effect? Does it, for example, make a difference whether a clerk without software development experience or the project manager with technical competence asks whether a task will take more than 10 work-hours?

We hypothesize that the answer to both of these questions is yes, based on previous research from other domains showing stronger anchoring effects with more precise anchors (Janiszewski & Uy, 2008; Loschelder, Stuppi, & Trötschel, 2014; Mason et al., 2013; Zhang & Schwarz, 2013) and weaker anchoring effects with less credible sources (Wegener et al., 2009).

The current experiments introduce a new way of varying the precision of an anchor, by comparing traditional single anchor values with interval anchors ("How likely is it that the task will take between 900 and 1100 hours?"). Intervals are highly relevant to software effort estimation, as it is common practice to describe the uncertainty of an estimate by using an interval (Connolly & Dean, 1997; Jørgensen, Teigen, & Moløkken, 2004). In the early phases of a project, when there is a great deal of uncertainty about how the project will turn out, it might be more common to suggest possible ranges for the most likely effort than to suggest single point estimates. It will therefore be interesting to see whether such interval anchors lead to weaker (or stronger) anchoring effects than more precise single anchors.

2. The study design

We invited 423 software professionals from Romania, Ukraine, Argentina and Poland to participate in a set of three experiments. All the participants were required to have good English skills, so that they could properly read and understand specifications written in English. All the participants received a normal hourly wage for their estimation work. Some of the invited participants gave estimates that indicated that they had misunderstood the instructions; for instance, in cases in which the most likely estimate was higher than the maximum estimate or the minimum estimate was higher than the maximum estimate. We excluded 42 participants who gave such erroneous responses to one or more of the software project estimation tasks. This left 381 participants (92.4% male), with a mean age of 29.3

years (*SD* = 5.7 years) and a mean of 6 years of experience in software development (*SD* = 4.7 years).

We distributed the experiments to the participants as an online survey. The participants could complete the experiments at their own pace, but we told them that they should aim to spend approximately an hour on them. The general introduction page informed the participants that they would be asked to estimate the number of work-hours they thought they would need to develop some relatively small software systems. It asked them to assume that they would carry out the development work themselves and that they could use the development technology, e.g., the programming language, development tools and database, that they knew best. It was also explained that the purpose of the survey was to gain a better understanding of how effort estimates are made and how to improve them, and not to evaluate the individual participant's competence, to minimize the problem of socially desirable responses. In all the estimation tasks, we asked the participants to provide an estimate of how much effort the project would take, as well as estimates of the minimum and maximum number of work-hours. The details of the design of each of the estimation tasks will be described separately for each section. In brief, Experiments 1 and 2 investigated different kinds of precision of high (Experiment 1) and low (Experiment 2) anchors, while Experiment 3 compared different degrees of credibility of the source of the anchor.

The distributions of project effort estimates are often right-skewed – there are usually a few estimates that are much higher than the majority of the other estimates. To avoid problems in the interpretation of mean values, we used log-transformed estimates as input in our analyses in all the studies. However, to ease the understanding of the results, we present non-transformed median values for the estimates of the most likely, minimum and maximum number of work-hours alongside median values for the absolute interval width and the relative interval width in the tables. We define the absolute and relative interval width in the following way:

- $AbsoluteIntervalWidth_i = MaxEst_i - MinEst_i$
- $RelativeIntervalWidth_i = (MaxEst_i - MinEst_i) / MLEst_i$

Here, $MaxEst_i$, $MinEst_i$ and $MLEst_i$ are, respectively, the estimated maximum, minimum and most likely effort for project *i*.

3. Experiment 1

3.1 Design

We gave the participants a description of a web-based application for visualizing information about the amount of software development offshoring in a country on a world map (Project A). We assigned the participants randomly to one out of five groups:

- The control group, which we simply asked to estimate the most likely, minimum and maximum number of work-hours needed to develop and test a system meeting the requirements.
- The precise single anchor group, which we asked how likely they thought it was that they would need less than 998 work-hours to complete the software development work on a scale from 1 (very unlikely) to 4 (very likely), and then posed the estimation questions. We intended the anchor likelihood question to be a kind of question that could be raised in a real-world setting, and it should make the participants consider the anchoring value before they gave their estimates.
- The round single anchor group, which we asked to evaluate the likelihood that the development work would take less than 1000 hours before giving their estimates.
- The precise (narrow) interval anchor group, which we asked to evaluate the likelihood that Project A would take between 900 and 1100 hours before answering the estimation questions.
- The imprecise (wide) interval anchor group, which evaluated the likelihood of an interval anchor of 500 to 1500 hours.

In other words, we gave the participants in the anchor groups an anchor value to consider before they gave their estimates of the most likely, minimum and maximum number of work-hours for the project. We varied the precision of the anchor in two ways: by giving a single anchor that was numerically precise (998 hours) or round (1000 hours) and by giving an interval that was precise (900 to 1100 hours) or imprecise (500 to 1500 hours). Using an interval as an anchor is a novel way to investigate how precision influences anchoring. Intervals are arguably less precise than single anchors, so one could expect that intervals should lead to less of an anchoring effect than single anchors. On the other hand, intervals contain two numbers, and if two anchors are stronger than one, one should expect a stronger anchoring effect. We expected a stronger anchoring effect from the precise single anchor (998 hours) than from the round single anchor (1000 hours); similarly, in the interval group, we expected a stronger anchoring effect from the narrow (900 to 1100 hours) than from the wide (500 to 1500 hours) interval. We also calculated the absolute interval width (the maximum number of work-hours minus the minimum number of work-hours) and the relative interval width (the absolute interval width divided by the most likely number of work-hours) for each participant. We expected that the participants in the interval anchor groups, particularly in the imprecise (wide) interval anchor group, might give wider intervals than the other participants.

3.2 Results

Table 1. *Median estimates of the most likely, minimum and maximum number of work-hours needed to complete Project A, along with the median absolute and relative interval width.*

	Control	Precise single anchor	Round single anchor	Precise interval anchor	Imprecise interval anchor
Anchor value	N/A	998	1000	900–1000	500–1500

n	78	75	82	77	69
Most likely effort	40	160	151	160	200
Minimum effort	32	120	116	125	150
Maximum effort	53	300	200	240	250
Absolute interval width	20	80	70	80	100
Relative interval width	.50	.50	.57	.50	.60

The responses to the anchor likelihood question (e.g., “How likely is it that you will need less than 1000 work-hours to complete the software development work?” or “How likely is it that you will need between 500 and 1500 work-hours to complete the software development work?”) showed that the participants in general thought the anchors were very high. In the single anchor groups, 90.7% of the participants in the precise single anchor group and 89% of the participants in the round single anchor group indicated that it was “likely” or “very likely” that the project would take less time than the anchor value. The chi-square test, which is used to determine whether distributions of categorical variables differ from one another, did not show significant differences in the response patterns of the two single anchor groups, χ^2 (3, *N*=157) = 5.124, *p* = .163. This means that participants thought the precise and round single anchor values were equally implausible. For the interval anchor groups, the anchor values were seen as quite unlikely, with 59.7% of the participants in the precise interval anchor group and 62.3% of the participants in the imprecise interval anchor group indicating that it was “very unlikely” or “unlikely” that the project would actually take somewhere between the suggested lower and upper bounds, with no significant differences between the two groups, χ^2 (3, *N*=146) = 3.113, *p* = .374. Note that this means that around 40% of the participants in the interval anchor groups found it “likely” or “very likely” that the project might take around 1000 hours. This shows that even though the anchors in this experiment were very high, not all the participants saw them as entirely implausible.

To establish that an anchor effect was present, we ran a simple ANOVA, which is used to analyse whether the means two or more groups differ from each other. Using the five different groups as a between-subjects factor the analysis showed a clear effect according to the conventional criterion of p < .05, F(4,376) = 20.722, p < .001, η^2_p= .181. Pairwise post hoc comparisons (Tukey) found the control group to be significantly different from all the anchor groups, all p's < .001, while the anchor groups did not differ from each other, all p´s > .47. This demonstrates a clear anchoring effect, which is also evident simply from looking at the most likely estimates in Table 1. The median estimate of the control group is much lower than the estimates of the other groups, which were all exposed to a high anchoring value(s).

Because the control group is so clearly different from the other groups, it was excluded from further analyses to give a better comparison of the anchor effects in the different anchor groups. A separate ANOVA comparing the most likely estimates of the precise and round single anchor groups showed no effect of precision, F(1,155) = .145, p = .704, η^2_p= .001. Similarly, a separate ANOVA for the interval groups did not find any difference between the narrow and the wide interval anchors, F(1,144) = .480, p = .490, η^2_p= .003. Finally, an ANOVA using the type of anchor (single anchors versus interval anchors) as a between-subjects factor showed a slight tendency for a larger anchoring effect with interval anchors, F(1,301) = 2.031, p = .155, η^2_p= .007. The estimates in the interval anchor groups were overall slightly higher (*Mdn* = 190) than those in the single anchor groups (*Mdn* = 160).

Looking at the absolute interval width, there was no effect of precision; F < 1. However, the intervals in the interval anchor groups were slightly wider than those in the single anchor groups, F(1,301) = 3.252, p = .072, η^2_p = .011, although this effect is not statistically significant at p < .05. There were no significant effects of either precision or the type of anchor on the relative interval width; all F's < 1.4.

The relatively high number of participants in Experiment 1 ensured that if there was a small to medium-sized effect of precision on anchoring, the present experiment should have sufficient statistical power to detect such an effect. With this in mind, the results of Experiment 1 indicate that the precision of the anchor is not very important in this project effort estimation context. The observed anchor effects were similar regardless of whether the anchor was 998 hours or 1000 hours. This is in contrast to previous studies that found a difference between precise and round numbers. We also used the novel approach of including interval anchors and expected that these might have less of an effect on the estimates than single anchors due to decreased preciseness. However, we instead found a slight tendency for a stronger anchoring effect with interval anchors. This effect was not statistically significant, so it should be interpreted with caution, but there are several reasons why interval anchors might lead to a stronger anchoring effect. First, there were two numbers in the interval anchor conditions, and two numbers might lead to stronger anchoring than one number. Second, it might appear more plausible or credible to suggest a range of possible outcomes rather than a single number, even though the range might lie outside the range one would normally have suggested. In either case, the experiment demonstrated a strong influence of anchors on project effort estimates, but did not indicate that the precision of the anchor is an important moderating factor on the strength of the anchoring effect.

4. Experiment 2

4.1 Design

The rationale of Experiment 2 was similar to that of the first experiment, and we again investigated the potential effect of numerical precision on anchoring. However, in contrast to Experiment 1, here we used low rather than high anchor values. In addition, we expected the anchor values in this experiment to be less extreme than those employed in the first experiment. In Experiment 1, the anchor values were clustered around 1000 hours, which turned out to be 25 times higher than the median estimate of the control group. We chose not to include single anchor groups in Experiment 2. As Experiment 1 indicated, single anchor groups do not seem to differ much from interval anchor groups.

The specifications given to the participants in Experiment 2 described a shoe shop owner who needed a software system to support customers looking for jogging shoes, into which the customer could enter his/her weight, how the shoes were supposed to be used, etc. The system, on the background of this input, should give a recommendation to the customer. The participants estimated the effort necessary to develop this system (Project B) after we randomly assigned them to one of three groups:

- The control group, which we gave the traditional estimation questions.
- The precise interval anchor group, which we asked to evaluate how likely it was on a scale from 1 (very unlikely) to 4 (very likely) that the software development work would take between 19 and 21 work-hours before answering the estimation questions.
- The imprecise interval anchor group, which we asked to evaluate the likelihood of an interval anchor of 10 to 30 work-hours.

In the precise interval anchor group, we used a quite strong manipulation of precision. The interval is very narrow, and in addition, the numerical values are precise rather than round. In comparison, we used a wider interval and numerically "round" values in the imprecise interval group. If precision does lead to a stronger anchoring effect, one would expect a larger anchoring effect in the precise condition. Since the anchor values were low, the hypothesis was that the estimates in the anchoring groups would be lower than the estimates in the control group.

4.2 Results

Table 2. *Median estimates of the most likely, minimum and maximum number of work-hours needed to complete Project B, along with the median absolute and relative interval width.*

	Control group	Precise interval anchor	Imprecise interval anchor
Anchor value	N/A	19–21	10–30

n	126	125	130
Most likely effort	112	46	50
Minimum effort	80	35	40
Maximum effort	150	60	71
Absolute interval width	50	24	30
Relative interval width	.50	.50	.50

An analysis of the participants' responses to the anchor likelihood question showed no significant differences between the precise and the imprecise interval group, χ^2 (3, N=205) = 4.410, p = .220, with 82% of the participants in the precise interval group and 69.5% of the participants in the imprecise interval group indicating that it was "very unlikely" or "unlikely" that the amount of effort for the project would be within the lower and upper bounds of the interval anchors.

A simple ANOVA comparing the most likely estimates of the three groups showed a clear anchoring effect, $F(2,378) = 19.244$, $p < .001$, $\eta^2_p = .092$, with pairwise post hoc comparisons (Tukey) confirming that the control group was significantly different from both the precise anchor group and the imprecise anchor group, p´s < .001. As shown in Table 2, the median estimate in the control group is more than twice as high as that in the two anchor groups. However, a separate ANOVA on the two anchor groups showed no difference between the most likely estimates of these groups, $F < 1$. The anchor groups also did not differ in absolute interval width, $F < 1$, or in relative interval width, $F(1,253) = 1.227$, $p = .269$, $\eta^2_p = .005$.

These results mirror the results of the first experiment: there does not seem to be any influence of anchor precision on the anchoring effect for estimates of software project effort. Interval anchors lead to a strong anchoring effect, but it does not seem to matter much whether the interval is narrow or wide or whether the numbers are precise or round.

5. Experiment 3

5.1 Design

In Experiment 3, we focused on the credibility of the source of the anchor as a potential moderator of the anchoring effect. We described to the participants a web-based library system that displays information about scientific publications that should be stored in an SQL database (Project C). After reading the specifications, we asked the participants in the control group to estimate the most likely, minimum and maximum effort. There were three anchoring groups, all receiving a low anchor of 10 hours, but we varied the source of this anchor:

- We told the low-credibility group to assume the following: "An administrative person in your company, with no background in software development, is responsible for the registration of all work larger than 10 work-hours into a database. Without looking at the requirement specification or having any idea about the complexity and size of the work, he asks you whether you think it will take you less than 10 work-hours to complete the work." This information was supposed to indicate that 10 hours was a number without any relevance to the estimation of the project at hand.
- The high-credibility group read the following instructions: "Your company manager has a background in software development. He takes a look at the requirement specification and asks you whether you think it will take you more than 10 work-hours to complete the work." This information, from a person with technical competence, should give more weight to the anchor value.
- The "neutral" anchor group was simply instructed to consider an anchoring value in the traditional fashion: "Do you think it will take you more than 10 hours to complete the work (development and testing)?"

In this experiment, we did not ask the participants to evaluate the likelihood of the anchoring value, but simply "Do you think it will take you more than 10 hours to complete the work?" as a yes/no question. The hypothesis in the current experiment was that the anchoring effect would be the strongest in the high-credibility group, while it would be weaker in the low-credibility group.

5.2 Results

Table 3. *Median estimates of the most likely, minimum and maximum number of work-hours needed to complete Project C, along with the median absolute and relative interval width.*

	Control group	Low-credibility anchor	High-credibility anchor	Neutral anchor
Anchor value	N/A	10	10	10
n	92	97	98	93
Most likely	60	32	30	32
Minimum	40	24	24	25
Maximum	80	48	40	40
Absolute interval width	30	24	16.5	20
Relative interval width	.50	.57	.58	.50

We performed an ANOVA on the most likely estimates of the four groups, which showed a tendency for an effect, $F(3,376) = 2.176$, $p = .090$, $\eta^2_p = .017$. Pairwise post hoc comparisons (Tukey) comparing the control group with the three other groups showed that the project effort estimates of the neutral anchor group, $p = .144$, the high credibility anchor group, $p = .111$, and the low-credibility anchor group, $p = .287$ were not significantly different from the estimates of the control group. However, as seen in Table 3, the median most likely estimate in all anchor groups was around 30 hours, as compared to 60 hours in the control group. When all the anchor groups are combined, an ANOVA comparing participants who received an anchor with control participants showed a statistically significant difference,

$F(1,378) = 6.296$, $p = .013$, $\eta^2_p = .016$. Together, this indicates that the anchor value of 10 hours led participants to give lower estimates than they otherwise would have.

After excluding the control group, we compared the most likely estimates of the anchor groups with an ANOVA and found no difference between groups, $F < 1$. Most importantly, post hoc pairwise comparisons showed no difference between the low-credibility group and the high-credibility group, $p = .880$. All other comparisons also showed no difference between the three anchor groups, *p*´s > .92. An analysis of the interval width did not show any difference between the anchor groups for either the absolute or the relative interval width, *F*'s < 1.

The results of Experiment 3 do not give support to the hypothesis that the anchor effect would be moderated by the credibility of the source of the anchor. Overall, the estimates of participants in the anchor groups were lower than estimates of participants in the control group, and there were no signs of any difference between the three anchor groups. In other words, software project effort estimates seem to be similarly influenced by anchoring values from both credible and non-credible sources.

6. Discussion

6.1 Summary of findings

The aim of these experiments was to investigate whether the preciseness of the anchor and/or of the perceived credibility of the source of the anchor moderates the anchoring effect in software project effort estimation. Finding moderators for the anchoring effect could be helpful for people involved in estimation by providing them with knowledge of situations in which they should be more or less concerned about biased project effort estimates due to anchoring. In our studies, we varied the precision of the anchor value by using round or precise numbers (Experiments 1 and 2), by using single values or intervals (Experiment 1) and by using wide or narrow intervals (Experiments 1 and 2). We varied credibility by describing the source of the anchor as competent versus non-competent (Experiment 3). We hypothesized that more precise anchor values and a more credible source would lead to a stronger anchoring effect. However, in the three experiments described here, we found no support for our hypotheses. Experiment 1 showed no difference between the anchoring effects for precise and round numbers. In addition, even though we can argue that intervals are less precise than single numbers, using intervals as anchors gave a slight tendency for a stronger anchoring effect than single anchors. Experiment 2 showed no difference between a narrow interval using precise numbers and a wider interval using round numbers. Both types of intervals led to a similarly strong anchoring effect. Finally, Experiment 3 showed no influence of the credibility of the source of the anchor. Even though we made it clear that an anchor value should not be relevant to the judgement at hand, the effect was of a similar magnitude as when the anchor stemmed from a more credible source. However, we emphasize that our findings demonstrated strong anchoring effects for software project effort estimates, even when we gave experienced software professionals a full hour to work on the study materials. This underscores the practical relevance of anchoring effects for people involved in the estimation of software projects and strengthens the confidence in previous findings demonstrating anchoring in effort estimation.

Recently, it has been argued that instead of relying solely on *p*-values, researchers should be more concerned with effect sizes and confidence intervals (CIs; Cumming, 2014). Using CIs makes it easier to evaluate the uncertainty of a research finding, with wide CIs indicating highly uncertain findings, while narrow CIs indicate a lower degree of uncertainty. Figure 1 shows the 95% CIs for the means of the log-transformed most likely estimates of the different groups in Experiment 1. A simple visual inspection confirms that the control group is different from all the other groups in that there is no overlap between the CI of the control group and the CIs of the four anchoring groups. For the anchoring groups, there is a very high degree of overlap, and the 95% CIs are quite narrow, indicating that the results are quite robust. Using the Exploratory Software for Confidence Intervals (ESCI; Cumming, 2013), we find that precision in Experiment 1 leads to a difference in means of 0.01, 95% CI [-0.11, 0.13]. This means that for participants receiving a precise single anchor or a precise interval anchor, the log-transformed mean increases by 0.01, compared with participants receiving a round single anchor or an imprecise interval anchor. In other words, the average effect of precision in this experiment is very close to zero, and the narrow 95% CI indicates that the finding is quite robust. The results from Experiments 2 and 3 show a similar pattern, with clear differences between the control group and the anchoring groups, but with a close overlap between the anchoring groups.

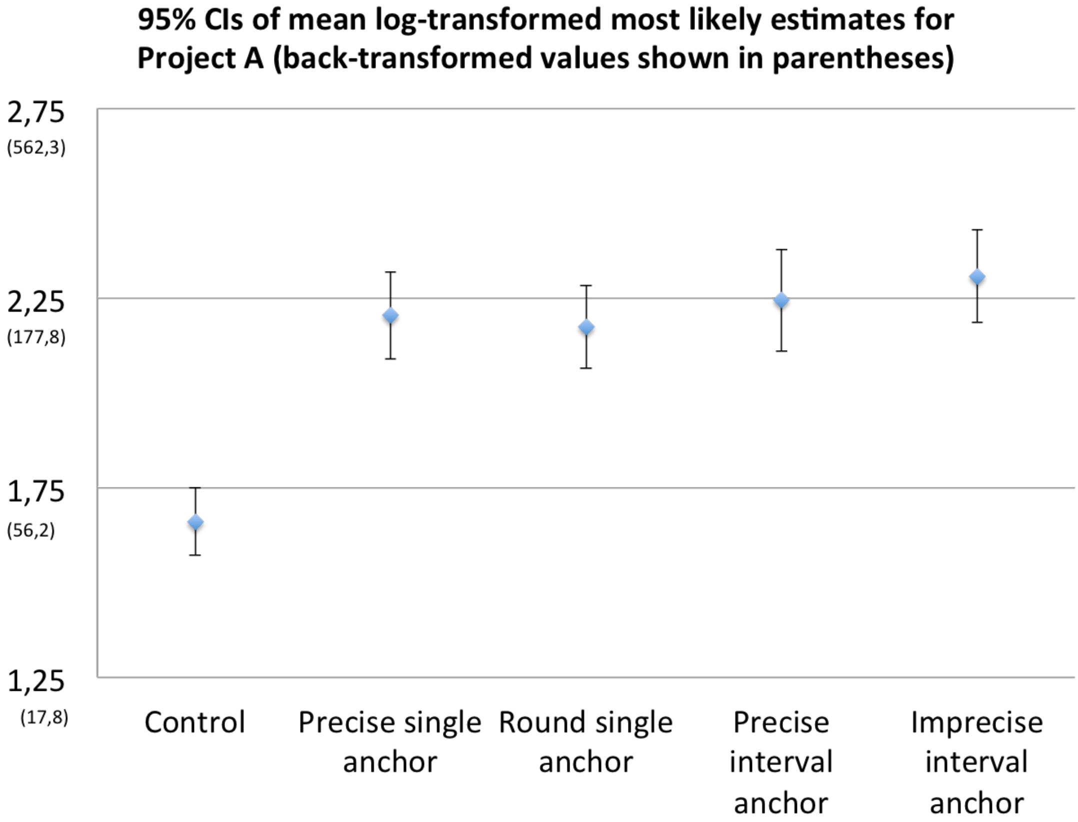

Figure 1. *95% confidence intervals of the mean log-transformed most likely effort estimates of the different groups in Experiment 1, with back-transformed values (in work-hours) shown in parentheses on the y-axis.*

To summarize, our findings indicate on the one hand that anchors have strong effects on estimates of software project effort, and on the other hand, that the anchoring effect is not moderated by numerical preciseness or source credibility. We acknowledge that the realism of studies of this kind can always be criticized. For instance, we cannot say for sure whether source credibility could be more important in situations where participants have personal experience with the source of the anchor. In our Experiment 3, we provided a quite brief description of a person with or without software experience, while in some real-life situations developers will have much more knowledge about the credibility of the speaker. It would be necessary to conduct a large-scale field experiment to be completely certain that our findings also generalize to software effort estimation work in "real life" – however, field experiments also have their own problems, such as a lack of control over extraneous variables. With this in mind, we believe that the current studies were done in a relatively realistic way, and that the results can be informative for people involved in software development estimation work.

6.2 The role of experience

Increased experience and background knowledge of the participants may decrease the sensitivity to an effect of precision or credibility. To investigate this possibility, we split the participants in our experiments into two groups according to their length of experience. We used a median split, which gave a low-experience group (n = 175) with up to 4 years of experience and a high-experience group (n = 206) with 5 years of experience or more. If length of experience diminishes the effect of precision in anchoring, one could expect that low-experience participants would show more of a tendency for an effect of precision or credibility than high-experience participants. Reanalysis of all three experiments with experience as an additional between-subjects factor did not show any interactions between experience and precision or credibility. This indicates that low-experience developers are neither more nor less sensitive to an effect of precision or credibility on anchoring than high-experience developers. However, for all three experiments, the high-experience participants seemed to be less influenced by the anchor itself. As an example, an ANOVA conducted on the most likely estimates of the anchoring groups in Experiment 3 with the type of anchor (high credibility, low credibility, neutral) and experience (low experience, high experience) as between-subjects factors showed no effect of the type of anchor and no interaction between the two factors, both *F*'s < 1, but a significant effect of experience, $F(1{,}282) = 9.350$, $p =$

.002, η^2_p = .032. The effect of experience is due to the estimates of the high-experience group (*Mdn* = 40) being higher than the estimates of the low-experience group (*Mdn* = 29). Considering that we gave a low anchor of 10 hours to the participants in Experiment 3, this indicates that the anchoring effect was stronger for the low-experience participants. Similarly, the estimates of the high-experience participants were less influenced by the high anchoring value in Experiment 1 and by the low anchoring value in Experiment 2. This finding is important as it shows that expertise (defined as length of experience) can indeed be an important moderator of the anchoring effect, but that the experience of the receiver of an anchor is more important than the indicated expertise of the source of an anchor. We should also highlight here that even though the anchoring effect seemed to be stronger in the low-experience group, it is strongly present even for participants with many years of experience as software developers. Hence, even extensive experience does not fully protect against the negative effect of introducing a numerical anchor.

6.3 Theoretical considerations

The attitude change approach to anchoring inspired the hypotheses investigated in the current experiments (Wegener et al., 2001); this approach predicts weaker anchoring effects when the anchor stems from a less credible source. However, the findings were not in line with what one would expect from this theoretical approach, and as such, our studies do not strengthen this particular explanation of the anchoring effect. As pointed out above, there might be differences in the populations (or alternatively in the domains) studied that could explain the discrepancy between this and previous studies, but it nevertheless seems safe to say that the current studies show that the attitude change approach is not sufficient to explain all anchoring effects. The anchoring effect in our experiments is remarkably robust, even though the participants generally believed that the anchors they considered were much too high or much too low for the tasks in question. Our findings appear to be in line with the selective accessibility model (Mussweiler & Strack, 2001). According to this view, the participants, after receiving an anchor, test the hypothesis that the anchor actually is the correct answer to the particular question they are given. A process of confirmatory hypothesis testing leads them to retrieve evidence that is in agreement with this hypothesis. When the final judgement is made, the heightened accessibility of evidence consistent with the anchor leads to a judgement that is closer to the anchoring value than it otherwise would have been. In particular, the fact that we found clear anchoring effects for interval anchors as well supports the view that the numerical values lead participants to test an initial hypothesis that the project is very large (Experiment 1) or very small (Experiments 2 and 3). And even though they see the anchoring values as much too high or much too low, the end result even after extensive estimation work is a final estimate that is biased by the initial value.

6.4 Practical implications

Rather than demonstrating how anchoring effects can be weakened by simple manipulations of numerical precision or source credibility, as we originally expected, the current studies serve to demonstrate the strength of the anchoring effect in software development effort estimation contexts. Even though our participants were professional software developers with a mean experience of six years, we found, in all three experiments, clear effects of an irrelevant anchor value on software project effort estimates. In Experiment 1, the anchor groups gave estimates that were about four to five times higher than those of the control group, while in Experiments 2 and 3, the estimates of the control groups were at least twice the estimates of the anchor groups.

While we do not know how much effort the developers actually would have spent to complete the software projects there are reasons to believe that the control group estimates, i.e., estimates of participants that did not receive anchoring information, are the most realistic ones. As an illustration, the median effort estimated by those in the control group of Experiment 3 was 60 work-hours, while those in the anchoring groups had median estimates of about 30 work-hours. The requirement specification used in Experiment 3 included about 50% of the functionality of software previously developed by six different software companies, see (Jørgensen, to appear in IEEE Software, 2015) for details about the projects and the companies. The average actual effort of those six projects was 260 work-hours. Assuming that 50% of the functionality would take 50% of the effort, the average effort is 130 work-hours. This suggest that although those in the control group, on average, had over-optimistic effort estimates, they were much less over-optimistic than those in the anchoring groups.

The current studies are in line with previous research showing how hard it can be to remove the effect of an anchor. The fact that even an unrealistically low value from a source that obviously has no relevant information about the project can influence an estimate (Experiment 3) indicates that it will not necessarily be very helpful to try to discredit the source or the validity of the anchor once the participants have encountered it. Similarly, it has previously been shown that even though introducing a second anchor pointing in the opposite direction of the first one can make the final estimate less extreme, the effect of the first anchor is still the strongest (Jørgensen & Løhre, 2012). In other words, people involved in estimation work should take the anchoring effect seriously and should not assume that they and their co-workers are so experienced and professional that they are able to disregard irrelevant information.

Why is it so hard to correct for the unwanted influence of an anchor? In their article on what they call “mental contamination” (i.e., having an unwanted judgement, emotion or behaviour due to unconscious or uncontrollable mental processing), Wilson and Brekke (1994) outline the process that is necessary to correct for a bias in judgement. First, the individual performing the judgement needs to be aware that a certain bias has been introduced. Second, the person needs to be motivated to correct this bias. Third, the person needs to be aware of the direction and the magnitude of the bias. Fourth, the person needs to be able to adjust his or her response (to have mental control over the response). These conditions might be difficult to satisfy, the authors argue: people are often unaware of their cognitive processes, it is often hard to observe that a judgement is biased and people also have limited control over their cognitive processes (Wilson & Brekke, 1994). We can make the same argument for anchoring: people are not necessarily aware that an anchor leads to a biased estimate, and even if they are motivated to correct the bias, they might not know how much or even in which direction they should correct their estimate.

The main take-home message from the present studies is that the best way to safeguard against the anchoring effect in effort estimation is to make sure that no information that can act as a misleading anchor, such as cost expectations of the client, is present. The effect of such anchors is not much reduced with less anchor preciseness or lower credibility of the source. The effort estimates of more experienced software professionals are less affected, but even these estimates are to a large degree affected by the anchoring effect.

Notes

The authors would like to thank Karl Halvor Teigen for helpful comments on an earlier draft of this paper.

References

Aranda, J., & Easterbrook, S. (2005). Anchoring and adjustment in software estimation. *Software Engineering Notes, 30*(5), 346-355. doi: 10.1145/1095430.1081761

Connolly, T., & Dean, D. (1997). Decomposed versus holistic estimates of effort required for software writing tasks. *Management Science, 43*(7), 1029-1045. doi: 10.1287/Mnsc.43.7.1029

Cumming, G. (2013). The new statistics: Estimation for better research. Retrieved from http://www.thenewstatistics.com

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29. doi: 10.1177/0956797613504966

Englich, B., & Mussweiler, T. (2001). Sentencing under uncertainty: Anchoring effects in the courtroom. *Journal of Applied Social Psychology, 31*(7), 1535-1551.

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *Journal of Socio-Economics, 40*(1), 35-42. doi: 10.1016/j.socec.2010.10.008

Halkjelsvik, T., & Jørgensen, M. (2012). From origami to software development: A review of studies on judgment-based predictions of performance time. *Psychological Bulletin, 138*(2), 238-271. doi: 10.1037/a0025996

Janiszewski, C., & Uy, D. (2008). Precision of the anchor influences the amount of adjustment. *Psychological Science, 19*(2), 121-127. doi: 10.1111/J.1467-9280.2008.02057.X

Jørgensen, M. (2004). A review of studies on expert estimation of software development effort. *Journal of Systems and Software, 70*(1-2), 37-60. doi: 10.1016/S0164-1212(02)00156-5

Jørgensen, M. (2013). Relative estimation of software development effort: It matters with what and how you compare. *Ieee Software, 30*(2), 74-79.

Jørgensen, M. (2015). Better selection of software providers through trialsourcing. To appear in *Ieee Software*.

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

Jørgensen, M., & Grimstad, S. (2010). The impact of irrelevant and misleading information on software development effort estimates: A randomized controlled field experiment. *Ieee Transactions on Software Engineering*.

Jørgensen, M., & Grimstad, S. (2012). Software development estimation biases: The role of interdependence. *Ieee Transactions on Software Engineering, 38*(3), 677-693. doi: 10.1109/Tse.2011.40

Jørgensen, M., & Løhre, E. (2012). *First impressions in software development effort estimation: Easy to create and difficult to neutralize.* Paper presented at the 16th International Conference on Evaluation & Assessment in Software Engineering.

Jørgensen, M., & Sjøberg, D. I. K. (2004). The impact of customer expectation on software development effort estimates. *International Journal of Project Management, 22*(4), 317-325. doi: 10.1016/s0263-7863(03)00085-1

Jørgensen, M., Teigen, K. H., & Moløkken, K. (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. doi: 10.1016/S0164-1212(02)00160-7

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *The American psychologist, 58*(9), 697-720. doi: 10.1037/0003-066X.58.9.697

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142-152. doi: 10.1027/1864-9335/a000178

König, C. J. (2005). Anchors distort estimates of expected duration. *Psychological Reports, 96*(2), 253-256.

Loschelder, D. D., Stuppi, J., & Trötschel, R. (2014). "€14,875?!": Precision boosts the anchoring potency of first offers. *Social Psychological and Personality Science, 5*(4), 491-499. doi: 10.1177/1948550613499942

Mason, M. F., Lee, A. J., Wiley, E. A., & Ames, D. R. (2013). Precise offers are potent anchors: Conciliatory counteroffers and attributions of knowledge in negotiations. *Journal of Experimental Social Psychology, 49*(4), 759-763. doi: 10.1016/J.Jesp.2013.02.012

Mussweiler, T., & Strack, F. (2001). The semantics of anchoring. *Organizational Behavior and Human Decision Processes, 86*(2), 234-255. doi: 10.1006/Obhd.2001.2954

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes, 39*(1), 84-97. doi: 10.1016/0749-5978(87)90046-X

Thomas, K. E., & Handley, S. J. (2008). Anchoring in time estimation. *Acta Psychologica, 127*(1), 24-29. doi: 10.1016/j.actpsy.2006.12.004

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. doi: 10.1126/science.185.4157.1124

Wegener, D. T., Blankenship, K. L., Petty, R. E., & Detweiler-Bedell, B. (2009). *Source credibility and numerical anchoring. Raw data*. Purdue University. West Lafayette, IN.

Wegener, D. T., Petty, R. E., Blankenship, K. L., & Detweiler-Bedell, B. (2010). Elaboration and numerical anchoring: Implications of attitude theories for consumer judgment and decision making. *Journal of Consumer Psychology, 20*(1), 5-16. doi: 10.1016/J.Jcps.2009.12.003

Wegener, D. T., Petty, R. E., Detweiler-Bedell, B. T., & Jarvis, W. B. G. (2001). Implications of attitude change theories for numerical anchoring: Anchor plausibility and the limits of anchor effectiveness. *Journal of Experimental Social Psychology, 37*(1), 62-69. doi: 10.1006/Jesp.2000.1431

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin, 116*(1), 117-142.

Zhang, Y. C., & Schwarz, N. (2013). The power of precise numbers: A conversational logic analysis. *Journal of Experimental Social Psychology, 49*(5), 944-946. doi: 10.1016/J.Jesp.2013.04.002