Anchoring in Project Duration Estimation

Lorko, Matej and Servátka, Maroš and Zhang, Le

MGSM Experimental Economics Laboratory, Macquarie Graduate School of Management, Ekonomická Univerzita v Bratislave

13 August 2018

Anchoring in Project Duration Estimation

Matej Lorko

MGSM Experimental Economics Laboratory, Macquarie Graduate School of Management, Sydney, Australia
matej.lorko@gmail.com

Maroš Servátka

MGSM Experimental Economics Laboratory, Macquarie Graduate School of Management, Sydney, Australia
and
University of Economics in Bratislava, Slovakia
maros.servatka@mgsm.edu.au

Le Zhang

MGSM Experimental Economics Laboratory, Macquarie Graduate School of Management, Sydney, Australia
Lyla.Zhang@mgsm.edu.au

August 13, 2018

Abstract: The success of a business project often relies on the accuracy of its project duration estimates. Inaccurate and overoptimistic project schedules can result in significant project failures. In this paper, we explore whether the presence of anchors, such as relatively uninformed suggestions or expectations of the duration of project tasks, play a role in the project estimating and planning process. We conduct a controlled laboratory experiment to test the effect of anchors on task duration estimates. We find strong anchoring effects and systematic estimation biases that do not vanish even after the task is repeatedly estimated and performed. We also find that such persisting biases can be caused by not only externally provided anchors, but also by the planner's own initial task duration estimate.

Keywords: Project management, project planning, time management, anchors, anchoring effect, task duration, duration estimation, time estimation, anchoring bias

JEL codes: C91, D83, D91, O21, O22

1. Introduction

Effective project planning processes, capable of producing accurate project estimates, are the cornerstone of successful project management practice. Businesses often undertake multiple projects, many of which are run concurrently and/or sequentially. and in which future project tasks depend on the outcome of the current ones. Such dependencies put a strain on company resources that are being reallocated from one project to another as necessary. Accurate task duration estimates thus play a crucial role in the effective allocation and utilization of these resources.

According to a recent global project performance report (Project Management Institute, 2017), approximately 50 percent of business projects fail to be delivered within the original estimated schedule, and many of them are even not completed at all. Such failures are often caused by unrealistically optimistic estimates of the time needed to complete project tasks. While there exist multiple reasons for inaccurate duration estimates (e.g., planning fallacy, optimism bias, strategic misrepresentation, competence signaling or commitment devices), in this paper we focus on anchoring (Tversky & Kahneman, 1974), a potentially prevalent cause of systematic bias in project planning, which is often ignored (or even unrecognized) by companies and/or planners.

In the context of project management, anchors can appear in a variety of forms, for example, as an initial wild guess (e.g., “How long? Three months, maybe?”), a suggestion (e.g., “Do you think two weeks are enough for you to get it done?”), a customer expectation (e.g., “We would really like to introduce the product to the market before the summer season.”) or, perhaps a tentative deadline (e.g., “The CEO expressed the intention to finish the project by the end of the year.”), all of which can influence task duration estimates, and subsequently the project schedule. Customers or managers would prefer to have their projects completed as soon as possible, *ceteris paribus*. Their suggestions or expectations driven by wishful thinking and over-optimism can, however, lead to underestimation of project duration. As a result, overoptimistically planned projects usually require deadline and budget extensions, distorting company resources and hindering customer satisfaction. In other scenarios, such projects are even cancelled before their completion, resulting in sunk costs without actual benefit to the company. One of the most prominent examples of major project planning failures is the construction of Sydney Opera House, which was completed 10 years behind the original schedule, with the total cost soaring to $102 million, in an extreme contrast with its original budget of $7 million. Moreover, the scope of the venue was not even delivered fully as planned.

Furthermore, the high rate of project failures is not only prevalent in companies possessing less experience with projects but also frequently found in companies with extensive history of project

management. Why are planners unable to effectively learn from their past estimating mistakes? We suspect that anchoring might play a role. Indeed, projects are usually complex, carried out by multiple teams and can last for years. People responsible for project planning are not necessarily involved in the execution of project activities and are often allocated to a different project before the current one is completed. Moreover, schedule delays are frequently attributed to causes other than misestimation such as weather, unforeseen risks, etc. Therefore, the feedback loops are often imperfect and imprecise, with the planning teams not receiving enough details regarding the actual duration of project tasks. As a result, project planners can remain relatively unaware of the full extent of their own mistakes. Subsequently, they may dominantly become anchored on either historical or their own estimates and prone to repeating the same estimating mistakes in future projects. Thus, the anchoring effect and resulting inefficiencies can carry over from one project to another.

Based on the above insights, we conjecture that (i) the presence of numerical anchors influences duration estimates (henceforth just "estimates" for simplicity) and that (ii) without feedback on estimation accuracy, the anchoring effect persists over time, biasing future estimates of the same or a similar task. Moreover, in the absence of externally provided anchors, the planner's first own estimate can act as an anchor in itself. We test our conjectures in a controlled laboratory experiment employing a simple real-effort task of evaluating comparisons of pairs of 2-digit numbers. Subjects are asked to estimate how long it will take them to correctly assess 400 such comparisons. Before the subjects provide their estimates, we ask whether it will take them less or more than [the anchor value] minutes to complete the task. Following the estimates, subjects perform the task. In order to parallel project management decisions in business practice, we present anchors to subjects in the context of the task they are about to perform. In contrast with the existing studies of the anchoring effect in which anchors are often irrelevant to the task at hand and anchor values are randomly drawn (see Furnham & Boo, 2011 for a comprehensive review), our experiment can be considered an instance of applied anchoring research. For a clean identification of the anchoring effect, we also conduct a control treatment in which the anchoring question is not asked. Moreover, unlike previous studies of anchoring in task duration or effort estimation, our experiment employs real incentives for estimation accuracy and task performance, which not only mimics the project management environment outside the laboratory but also makes subjects take their choices seriously as their financial earnings in the experiments are determined by their decisions. Our research also extends the standard one-shot anchoring paradigm into the testing the effect of anchors over a longer horizon. In the experiment, the estimation and the actual task are repeated three times, while the anchoring question is only presented prior to the first estimation. The recurrence of estimating is a crucial element of our design, allowing us to test whether people could adjust their estimates away from the anchor in a repeated

setting. Thus, we can observe how the anchoring effect evolves over time and whether the obtained experience (albeit without feedback) can mitigate it.

We provide clear evidence that numerical anchors can bias estimates even in an environment encompassing multiple features known to alleviate estimation biases (see Section 3 for details) and under the incentive structure that motivates individuals to provide unbiased estimates of their own task completion time. We find strong effects of both low and high anchors. Moreover, we show that the bias caused by anchors is not restricted to the first estimate of a given problem and can persist over time and carry-over to following similar estimations if the external corrective action (e.g. estimation feedback) does not take place between the estimations. Although these anchoring effects slightly diminish with time, they remain statistically significant during the entire experiment except for the third estimation in the Low Anchor treatment. In addition, we find that anchors also influence retrospective estimates of how long the task actually took each subject to complete. Finally, the obtained estimates in the Control treatment display a "self-anchoring" effect, meaning that subjects' future estimates are anchored on their own first estimate.

Our study provides three important implications for the project management practice. First, our findings support the argument that project managers should isolate potentially biasing information such as management or customer expectations from planners (Halkjelsvik & Jørgensen, 2012). Second, a possible approach to mitigate the estimation bias during the planning phase of the current project is to consult historical information regarding the past projects. Our experimental data show that a simple measure predicted by the mean of past task duration can outperform planners' own estimates in terms of estimation accuracy, no matter whether in the presence or absence of anchors. Thus, planners can benefit from the use of simple statistics from similar projects in the past, complementary to the more traditional step-by-step planning based on the specification of the current project. Third, relying on the planners' awareness of their mistakes by themselves is not necessarily an effective strategy. To improve future project estimates it seems to be crucial to provide planners with precise feedback on their estimation accuracy.

2. Relationship to the literature

The concept of anchoring was introduced by Tversky & Kahneman (1974) who propose that the use of an initial starting point in estimation, such as that in the form of a suggestion, can lead to systematic biases due to insufficient adjustment of the estimate away from the starting point (which is referred to as an "anchor"). Thus, for the same problem, different starting points (anchors) lead to different

estimates or values. In Tversky & Kahneman (1974), before the estimation of the percentage of African countries in the United Nations, subjects were presented with either a low anchor (10%) or a high anchor (65%), generated by the wheel of fortune, and asked to consider whether the correct answer lies above or below the proposed value. Subjects' final median estimates (25% in the low anchor group and 45% in the high anchor group) of the percentage of African countries in United Nations were clearly affected by anchors.

The anchoring effect has been subsequently documented in other studies of general knowledge judgments, e.g. Jacowitz & Kahneman (1995); Epley & Gilovich (2001); Blankenship, Wegener, Petty, Detweiler-Bedell, & Macy (2008). The anchoring effect has also been observed in other domains such as negotiation (Galinsky & Mussweiler, 2001; Ritov, 1996), purchasing decisions and valuations (Ariely, Loewenstein, & Prelec, 2003; Alevy, Landry, & List, 2015), probability estimates (Plous, 1989), sentencing decisions (Englich, Mussweiler & Strack, 2006), performance forecasts (Critcher & Gilovich, 2008), social judgments (Davis, Hoch, & Ragsdale, 1986), self-efficacy (Cervone & Peake, 1986), and meta-memory monitoring (Yang, Sun & Shanks, 2017).

The influences of anchors have also been found in the domains that are relevant to our research questions, namely the task duration estimation and the effort estimation.[1] For example, estimates of effort required for software development can be anchored on customers' (Jørgensen & Sjøberg, 2004) or managers' expectations (Aranda & Easterbrook, 2005). The anchoring effect can be introduced also by a variation of wording instead of using numerical values, as demonstrated by Jørgensen & Grimstad (2008) who find different work-hours estimates of the same task, labelled "minor extension", "extension" or "new functionality" for different treatment groups.[2] Moreover, Jørgensen & Grimstad (2011) observe an anchoring effect in estimates provided by outsourcing companies in a field setting of software development. In the domain of task duration estimation, König (2005) demonstrates the anchoring effect on estimates of time needed to find answers to questions in a commercial catalogue. Before the estimation and actual task completion, subjects are asked to consider whether they need more or less than 30 (a low anchor) or 90 (a high anchor) minutes to complete the task. Consistent

[1] In project management, the duration of the task is often reported in man-hours (man-days) and is referred to as the effort estimate.

[2] This manipulation can be considered as framing rather than anchoring. However, software development companies relatively frequently use terms such as "extension" or "new functionality" to describe workload requiring a specific number of work-hours. For example, a past employer of one of the authors uses "Minor Enhancement" as a category for every new work that requires approximately 160 work-hours to complete. Thus, the expression is strongly associated with a particular number and serves as a powerful anchor for effort estimation.

with the hypothesis, the estimates in their low anchor treatment are significantly lower than those in the high anchor treatment. The actual time in which subjects complete the task is also measured, however, no significant differences across treatments are found. The author concludes that "estimating the amount of time needed to complete a task is a fragile process that can be influenced by external anchors" (p. 255). Similar results are presented by Thomas & Handley (2008), who also find that significant differences in duration estimates can be caused even by anchors irrelevant to the estimating problem at hand.

Altogether, there exists considerable scientific evidence of anchoring in task duration and effort estimation from laboratory experiments, field experiments, and field questionnaires. However, none of the previous laboratory and classroom experiments incentivized subjects for their estimation accuracy (only flat fees or course credits were used). The lack of real incentives can cause a hypothetical bias (e.g., Hertwig & Ortmann, 2001), and it is therefore questionable whether the anchoring effect is robust when misestimation can cause real losses to the estimator. Indeed, the relatively low magnitudes of anchoring effect found in the above-mentioned field experiment by Jørgensen & Grimstad (2011) can possibly be attributed to the fact that companies producing the estimates were informed that they might be offered additional opportunities for estimation work if their estimates were accurate. In addition, anchoring studies usually employ one-shot tasks, making it impossible to study whether the subjects learn from their previous estimation errors caused by anchors.[3] To the best of our knowledge, the influence of anchors in duration estimation has never been tested in more than one period and we therefore have no knowledge about how the anchoring effect interacts with the planner's experience. In fact, despite the extensive body of anchoring-related research, a relatively little is known about the long-term effect of anchors in general. Ariely et al. (2003) find differences in willingness to accept the payment for listening to annoying sounds between treatments, in which subjects are initially anchored on different amounts of money. The anchored willingness to accept does not converge in repeated estimation, not even in the bidding contest. On the other hand, Alevy et al. (2015) find convergence of prices by the third period in a market setting, where subjects are trading baseball cards. However, since there is no "true" value of the price for

[3] See Smith (1991) for a nice discussion on the importance of interactive experience in social and economic institutions in testing rationality (and implicitly the lack of biases in decision-making). While our experiment does not allow for an interaction (due to the nature of decision-making and the implemented lack of feedback) and is institution-free, it does take a step in the direction proposed by Smith, namely by testing whether the experience itself is sufficient to eliminate the anchoring bias.

listening to annoying sounds or of the price of baseball cards, we lack sufficient evidence from these studies to demonstrate the anchoring bias.

Furthermore, in the domain of duration estimation, many of the earlier studies employed relatively unfamiliar tasks, possibly exacerbating the bias. As suggested by Smith (1991), one might expect that prior task experience will reduce the influence of nuisance factors, such as anchors, in economic decision-making. This claim is supported by empirical evidence directly related to anchoring. For example, Thomas & Handley (2008) show that the subjects who admitted to having performed a similar task in the past are less affected by the anchors in the experimental setting. Similarly, Løhre & Jørgensen (2016) find that subjects with a longer tenure in the profession are less influenced by anchors and thus provide more accurate estimates than less experienced subjects. However, the anchoring effect is still significant even for the most experienced subjects.

While the persistence of the anchoring effect over time and its correlation with subjects' task experience have not been tested, there exists a strand of literature on the effect of experience on the accuracy of non-anchored duration estimates. However, the results are mixed. On the one hand, more experience leads to more accurate estimates in tasks such as reading (Josephs & Hahn, 1995), software development effort (Morgenshtern, Raz, & Dvir, 2007), and running (Tobin & Grondin, 2015). On the other hand, experienced users tend to underestimate the time needed to complete other tasks, such as playing piano pieces (Boltz, Kupperman, & Dunne, 1998), using cell phones and assembling LEGOs (Hinds, 1999) and making origami (Roy & Christenfeld, 2007). Thus, the effectiveness of the experience on estimation accuracy is likely to be task- or context-specific. Possibly, the mixed results can be explained by the fact that focusing on the task duration is more important and salient for some tasks (such as programming or running) than for the others. In a similar vein, Halkjelsvik & Jørgensen (2012) argue that having experience with the task itself does not necessarily imply having experience with its duration estimation. When people do not receive feedback regarding their estimation accuracy (or do not usually estimate the duration of the task in the first place), the increase in their experience with the task can lead to more optimistic and hence less accurate estimates. This proposition is supported by experimental results demonstrating that just prompting for self-generated feedback on the estimation accuracy can reduce future estimation errors (König, Wirz, Thomas, & Weidmann, 2014). Anchors might also affect individual estimation consistency. For example, *ceteris paribus*, the duration estimate by the same person for the same task should be approximately the same. However, Grimstad & Jørgensen (2007) find relatively large variance between the estimates of the same tasks provided by the same experienced software professionals. Can such inconsistency be explained by anchoring effects? Halkjelsvik & Jørgensen (2012, p. 241) note,

the re-examination of the experimental data show, that "the high level of inconsistency is to a certain extent a product of assimilation toward the preceding tasks or judgments."

Overall, the anchoring effect is found to be a pervasive phenomenon in the domain of task duration estimation. However, it is not clear whether anchors persist over time and whether these effects depend on planners' experience. We design an incentivized experiment to fill these gaps and to thoroughly examine the prevalence as well as limitations of the influence of anchors. In companies, the estimates are usually being produced by experienced professionals who are familiar with the task at hand and the estimation is often repeated. We therefore incorporate experience and repetition in our experimental design, together with meaningful incentives for task performance and estimation accuracy. Thus, our study presents a conservative test designed to detect the lower bound of the anchoring effect. One can imagine that if we observe an anchoring effect in our setup, it would be even more prevalent in environments characterized by the absence of these features.

3. Experimental design

We conduct an incentivized laboratory experiment employing an individual real-effort task to test whether numerical anchors influence duration estimates and whether such effects persist over time. Throughout the experiment subjects are prohibited from using their watches, mobile phones and any other devices that have time displaying functions. The laboratory premises also contain no time displaying devices. The clocks on the computer screens are hidden.

The experiment consists of three rounds. In every round, each subject is requested to estimate how long it will take him to complete the upcoming task before the actual task performance starts. In our task, an inequality between two numbers ranging from 10 to 99 is displayed on the computer screen (for sample screenshots, see the Instructions in the appendix) and the subject is asked to answer whether the presented inequality is true or false. Immediately after the answer is submitted, a new, randomly chosen inequality appears. The task finishes once the subject provides 400 correct answers. The advantages of this task are its familiarity (people often compare numbers in everyday life, for example prices before a purchase), and that it has only one correct answer (out of two options), making the estimation process simple. The target number of correct answers (400) was calibrated in a pilot with the goal of finding an average task duration of 600-750 seconds (10 - 12.5) minutes, as the previous research by Roy, Christenfeld, & McKenzie (2005) suggests that tasks exceeding 12.5 minutes are usually characterized by underestimation whereas tasks shorter than that are usually

overestimated. All in all, the design creates a favorable environment for subjects to estimate the duration accurately.

Subjects perform similar two-digit number comparisons in each round. To test whether people are able to overcome the anchoring effect by learning from the experience itself, we provide no feedback regarding the actual duration or estimation accuracy between rounds. Such design captures a common problem of project management present in many companies, namely that project planners do not receive detailed feedback of the actual hours spent by project team members on each task. Even if the actual durations of project activities are evaluated against the project plan, the scheduled delays are often attributed to factors other than inaccurate estimation. Both no feedback and inadequate feedback makes a project planner unlikely to improve his duration estimates.

In the experiment, subjects are financially incentivized for both their estimation accuracy and task performance. The incentive structure is designed to motivate subjects to estimate the task duration accurately, but at the same time to work quickly and avoid mistakes. While the main objective of the experiment is to test the estimation accuracy, our research question requires incentivizing both accuracy and performance. Without incentivizing the task performance, subjects could deliberatively provide high estimates and then adjust their pace in order to maximize their accuracy earnings. Providing incentives for performance creates an environment analogous to duration estimation in project management where the goal is not only to produce an accurate project schedule, but also to deliver project outcomes as soon as possible (holding all other attributes constant). Since there are two dimensions of incentives, there is a concern that subjects might try to create a portfolio of accuracy and performance earnings. While one can control for the portfolio effect by randomly selecting one task for payment (Cox, Sadiraj, & Schmidt, 2015; Holt, 1986), we choose to incentivize subjects for both tasks and minimize the chances of subjects constructing a portfolio by a careful experimental design and selection of procedures. First, subjects are not able to track time throughout the entire experiment. Second, our software is programmed so as to provide neither the count of correct answers nor the total attempts. Both design features make it unlikely for subjects to strategically control their pace and match it with their estimates.[4] We use a linear scoring rule to incentivize both estimation accuracy and task performance earnings. We acknowledge that the linear scoring rule might not be the most incentive compatible one, it is arguably most practical to implement

[4] It is possible that the results could be different if we implemented the pay-one-randomly payoff protocol. We therefore elicit subjects' risk preferences using an incentivized risk attitude assessment (Holt & Laury, 2002) about which they are only informed after the completion of all three rounds. We use these preferences to control for the degree of subjects' risk preferences in a regression analysis.

in an experimental environment than more complex scoring rules (e.g. quadratic or logarithmic) due to ease of explanation to subjects (Woods & Servátka, 2016).

The estimation accuracy earnings depend on the absolute difference between the actual task duration and the estimate. In every round, the maximum earnings from the precise estimate are AUD 4.50. The estimation accuracy earnings decrease by AUD 0.05 for every second deviated from the actual task duration, as shown in Equation (1). However, we do not allow for negative estimation accuracy earnings. Thus, if the difference between the actual and the estimated time in either direction exceeds 90 seconds, the estimation accuracy earnings are zero for the given round.[5] This particular design feature is implemented because our expectations of a strong anchoring bias and the related estimation inaccuracy that could cause many subjects to end up with negative (and possibly large negative) earnings. Our setting parallels a common practice in companies where planners are praised or rewarded for their accurate estimates of successful projects but are usually not penalized for inaccurate estimates when a project fails.

$$\text{Estimation earnings } = 4.50 - 0.05 * |\text{actual time in seconds } - \text{estimated time in seconds}| \quad (1)$$

The earnings from task performance, presented in Equation (2), depend on the actual task duration as well as on the number of correct and incorrect answers. The shorter the duration, the higher the earnings. We penalize subjects for incorrect answers in order to discourage them from fast random clicking. Such design is parallel to the business practice where not only the speed but also the quality that matters. We expected subjects to complete the task within 10-12.5 minutes and thus earn between AUD 3.70 and 4.70 per round for their performance, making the task performance earnings comparable with estimation accuracy earnings.

$$\text{Performance earnings } = \frac{7 * (\text{number of correct answers} - \text{number of incorrect answers})}{\text{actual time in seconds}} \quad (2)$$

The experiment consists of three treatments (Low Anchor, High Anchor, and Control) implemented in an across-subjects design, meaning that each subject is randomly assigned to one and only one treatment. In contrast to most of the extant studies on numerical anchoring, we include a Control treatment that allows us to test for a general estimation bias and the possibility of "self-anchoring,"

[5] The 90-second threshold was derived from the observed task duration in pilots (600-750 seconds). The project management methodology for estimating requires the definite task estimates to fall within the range of +/- 10% from the actual duration (Project Management Institute, 2013). We increased this range to 12-15% to make estimation accuracy earnings more attractive.

i.e. whether the first estimate anchors future estimates of the same task. In addition, we use estimates from the Control treatment to calibrate the low and high anchor values. The low anchor value is set at the 7th percentile and the high anchor value at 93rd percentile of the Control treatment estimates, in line with the procedure for measurement of anchoring effect proposed by Jacowitz & Kahneman (1995). The implemented values are 3 and 20 minutes. The Low Anchor and High Anchor treatments are conducted according to the same experimental procedures as the Control treatment. However, before the Round 1 (and only before the Round 1) subjects answer an additional question containing the anchor, in the following form:

Will it take you less or more than [the anchor value] minutes to complete the task?

4. Hypotheses

We hypothesize that anchors influence the estimates in Round 1. Specifically, estimates in the Low Anchor treatment are significantly lower than those in the Control treatment, and estimates in the High Anchor treatment will be significantly higher than those in the Control treatment. Furthermore, since our subjects do not receive feedback on their estimation accuracy, we expect the anchoring effect to carry over to subsequent estimates in Round 2 and Round 3.

- *Hypothesis 1*
    - $Estimate_L^1 < Estimate_C^1 < Estimate_H^1$
    - $Estimate_L^2 < Estimate_C^2 < Estimate_H^2$
    - $Estimate_L^3 < Estimate_C^3 < Estimate_H^3$,

    where the superscript $(1, 2, \text{or } 3)$ refers to Round $1, 2,$ and $3,$ respectively
    and the subscript $(\text{L}, \text{C}, \text{or H})$ refers to the Low Anchor, Control, and High Anchor treatment.

Since the subjects are incentivized not only for their estimation accuracy but also for how quickly they can finish the task, we expect them to work as fast as they can, independently of the treatment. In other words, we hypothesize that anchors do not have any effect on the actual task duration.

- *Hypothesis 2*
    - $Duration_L^1 = Duration_C^1 = Duration_H^1$
    - $Duration_L^2 = Duration_C^2 = Duration_H^2$
    - $Duration_L^3 = Duration_C^3 = Duration_H^3$

By combining Hypotheses 1 and 2, we expect an underestimation of task duration in the Low Anchor treatment but an overestimation in the High Anchor treatment. This is due to the presence of the anchoring effect in estimation but not in task performance. Since subjects are not exposed to an anchor in the Control treatment (and since our design provides favorable conditions for unbiased estimates), we expect to find no systematic bias in task duration estimates in the Control treatment.

- *Hypothesis 3*
  - $Estimate_L^t < Duration_L^t$
  - $Estimate_C^t = Duration_C^t$
  - $Estimate_H^t > Duration_H^t$ $\qquad where\ t = 1, 2, 3$

5. Main results

A total of 93 subjects (45 females, with a mean age of 20.7 and a standard deviation of 4.5 years) participated in the experiment that took place in the Vernon Smith Experimental Economics Laboratory at the Macquarie Graduate School of Management in Sydney.[6] Subjects were recruited via the online subject-pool database system ORSEE (Greiner, 2015). The experiment was programmed in zTree software (Fischbacher, 2007). After completing all three rounds, subjects provided answers to a few questions about the task, completed the risk assessment, and the demographical questionnaire. At the end of the experiment, subjects privately and individually received their experimental earnings in cash. The average subject spent 45 minutes in the laboratory and earned AUD 16.50.

First, we present the results from data aggregated across all three experimental rounds. The distribution of the actual task duration displays a skewed-shape distribution with asymmetric truncation, typical in the domain of task performance (see Figure 1a).[7] The distribution of estimates (see Figure 1b) follows a similar pattern, however, the skewness is less pronounced, mostly because of the inflated estimates in the High Anchor treatment. The Shapiro-Wilk test of normality indicates

[6] One subject was dropped from the sample because of her lack of comprehension. She repeatedly estimated the duration of the entire experimental session (i.e. the sum of all three rounds) instead of each round. The subject was debriefed while getting paid, and we discovered her poor command of English. When asked about the actual duration of the third round only, after the experiment, she made the same mistake again. Removing this data point does not change the treatment effect results anyway.

[7] The distribution of performance is usually skewed to the left because of the lower bound on the possible task duration. In our case there exists a minimum time in which it is possible for a human to provide 400 correct answers.

that the distributions are not normal ($p<0.001$ for both pooled actual duration and estimates); we therefore analyze the treatment effects using non-parametric tests.[8]

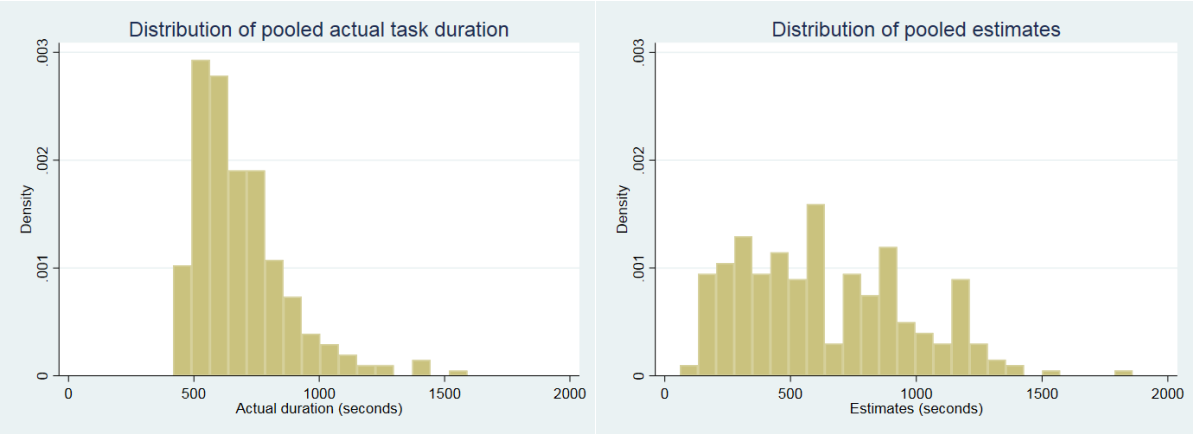


Figure 1a. The distribution of pooled task duration **Figure 1b. The distribution of pooled estimates**

Next, we analyze subjects' behavior in each round (see Table 1 for summary statistics). In line with our Hypothesis 1, the estimates in the Low Anchor treatment are the lowest, whereas those in the High Anchor treatment are the highest across all three experimental rounds. However, the absolute differences diminish from one round to another (see Figure 2a). We analyze the changes using the Wilcoxon matched-pairs signed-rank test and find that the estimates in the Low Anchor treatment rise over time (statistically significantly from Round 1 to Round 2 but insignificantly from Round 2 to Round 3), while the estimates in the High Anchor treatment gradually fall over time, with statistically insignificant decrease between rounds. The estimates in the Control treatment are relatively stable, consistently positioned between the estimates of the Low Anchor and High Anchor treatments and do not change significantly from one round to another. Using the Mann-Whitney test (p-values are listed in Table 2), we find that differences between the estimates in the Low Anchor and High Anchor treatments are statistically significant across all three rounds, supporting our Hypothesis 1 that the anchoring effect persists over time. Even though subjects display some degree of learning from the task experience and move their estimates away from the anchor, the adjustment is insufficient and the anchoring effect diminishes rather slowly.

- *Result 1: The anchors influence the task duration estimates and the anchoring effect persists over time.*

Table 1. Descriptive statistics

[8] Parametric tests yield qualitatively similar results, indicating the robustness of our findings. The details are available upon request.

Treatments	Low Anchor (N = 31)			Control (N = 27)			High Anchor (N = 35)		
Rounds	R1	R2	R3	R1	R2	R3	R1	R2	R3
Mean estimate, SD (seconds)	393 (252)	472 (261)	520 (303)	577 (314)	634 (270)	608 (236)	868 (385)	809 (314)	775 (302)
Mean duration, SD (seconds)	751 (242)	646 (173)	631 (164)	741 (200)	655 (143)	612 (132)	791 (220)	672 (157)	659 (172)
Mean bias, SD (seconds)[a]	-358 (323)	-174 (234)	-110 (251)	-164 (378)	-21 (288)	-4 (235)	77 (396)	136 (285)	115 (280)
Mean bias (%)	-45%	-27%	-19%	-17%	0%	0%	15%	21%	20%
Underestimation proportion	84%	74%	61%	66%	52%	59%	49%	40%	34%
Overestimation proportion	16%	26%	39%	33%	48%	41%	51%	60%	66%
Median estimate (seconds)	330	345	420	480	600	550	890	870	780
Median duration (seconds)	683	583	571	705	630	571	728	646	617
Test of similarity between the estimates and the actual duration (p-values)									
Wilcoxon matched-pairs signed-rank test	<0.001	0.001	0.024	0.039	0.755	0.614	0.310	0.014	0.027
Kolmogorov–Smirnov test	<0.001	<0.001	0.001	0.010	0.100	0.187	0.063	0.003	0.033
Test of changes in estimates (trends) between rounds [b]	Significantly rising in R2 (p = 0.015), insignificantly in R3 (p = 0.374)			No significant changes from one round to another (p = 0.225, p = 0.427)			Insignificantly decreasing (p = 0.107, p=0.294)		

a: We calculate the bias as a relative estimation error (Estimate – Actual duration). b: Wilcoxon matched-pairs signed-rank test (p-values). SD refers to standard deviation.

The inclusion of the Control treatment in the design enables us to identify the effects and persistence of both anchors individually, by comparing the differences in estimates between the anchor treatments and the Control treatment (see Table 2). We find significant differences in all comparisons, with one exception – the estimates in Round 3 of the Low Anchor treatment are similar to those in the Control treatment, while the estimates in the High Anchor treatment are significantly higher. This suggests an asymmetry in the persistence of the anchoring effect, namely that the effect of the low anchor is less persistent than the effect of high anchor. However, we are careful in interpreting such result. First, the asymmetry of the anchoring effect may be just a reflection of the natural asymmetry of estimation errors. While it is unreasonable to underestimate the task duration towards zero or negative time, overestimation errors are not limited. As such, there is not much scope for the estimates in the Low Anchor treatment to be extremely far away from the actual task duration. Second, the difference in persistence might be also attributed to the difference between the estimation target (the actual task duration) and the implemented values of the low and high anchors. The mean actual task duration in Round 1 is a little over 12.5 minutes and thus the low anchor (3 minutes) is on average further away from the target value than the high anchor (20 minutes). However, the mean actual task duration is lower in the following rounds (on average approximately

11 minutes in Round 2 and 10.5 minutes in Round 3), which provides less room for adjustment in the Low Anchor treatment in comparison with the High Anchor treatment. Hence, it is more probable for the estimates in the Low Anchor treatment to approach the estimates produced in the unbiased Control treatment.

Table 2. Non-parametric tests of treatment effects

		Mann-Whitney test (p-values)		
	Median (Low Anchor / Control / High Anchor)	Low Anchor vs. High Anchor	Control vs. Low Anchor	Control vs. High Anchor
Estimates				
Round 1	330 / 480 / 890	<0.001	0.010	0.004
Round 2	345 / 600 / 870	<0.001	0.013	0.034
Round 3	420 / 550 / 780	0.001	0.101	0.026
Duration				
Round 1	683 / 705 / 728	0.227	0.779	0.375
Round 2	583 / 630 / 646	0.240	0.543	0.712
Round 3	571 / 571 / 617	0.393	0.785	0.284

Figure 2b displays the actual task duration by treatments and rounds. We find a significant improvement in performance over time ($p < 0.001$ for the Wilcoxon matched-pairs signed-rank test between Round 1 and Round 2 as well as that between Round 2 and Round 3 yields). Nevertheless, there are no significant differences in the actual task duration across treatments (see Table 2) in any round, supporting our Hypothesis 2. This confirms that subjects are working fast in order to maximize their performance earnings and that they are not trying to manipulate their working pace in order to create a portfolio of their experimental earnings.

- *Result 2: The anchors have no effect on task performance.*

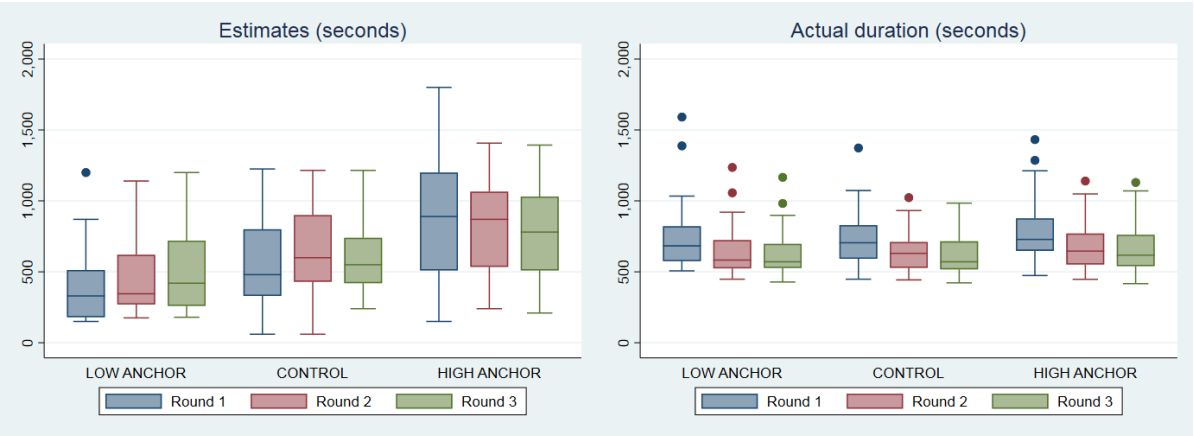


Figure 2a. Estimates by treatment and round **Figure 2b. Actual task duration by treatment and round**

We next analyze the estimation bias, which is measured by relative estimation errors (i.e. estimate – actual duration; see Table 1 and Figure 3 for aggregate data and Figure 5 for data on individual level). We find that the majority of subjects (73% on average) in the Low Anchor treatment underestimate the time it would take them to complete the task, and the average estimate is 30% lower than the actual duration. In a similar vein, overestimation is prevalent in the High Anchor treatment (on average 59% of subjects overestimate, and the average estimate is 18% higher than the actual task duration), consistent with our Hypothesis 3. For the Control treatment, we find prevalence of underestimation in Round 1 where the estimates are on average 17% lower than the actual task duration. There might be multiple reasons for this bias, such as wishful thinking, optimism or providing a shorter estimate as a commitment device. Note, however, that the bias diminishes in the following two rounds, in which the average relative estimation errors are close to 0. Using the Fisher's exact test, we find that the proportions of underestimation and overestimation differ across treatments, which is also robust for all three rounds (p-value is 0.010, 0.018, and 0.058 for Round 1, Round 2, and Round 3, respectively). Since the actual task duration is similar across treatments, the estimation bias is caused by the anchored estimates.

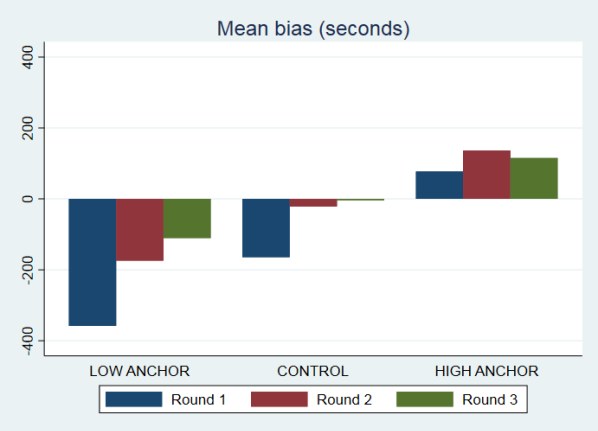


Figure 3. Mean estimation bias by treatment and round

We further compare each individual's estimates with their actual task duration using the Wilcoxon matched-pairs signed-rank test (see Table 1). We find statistically significant differences in the Low Anchor treatment, which are robust for all three rounds. For the High Anchor treatment, we find that the estimates in Round 1 do not differ from the actual task duration; however, the estimates become significantly different from the actual task duration in the following rounds. This is driven by improved

performance without subjects' adequate adjustment of their estimates. In the Control treatment, we find no statistically significant differences between the estimates and the actual task duration in Rounds 2 and 3. Overall, the results show that the estimates in the Control treatment are the most accurate. Furthermore, we test the differences in distributions between estimates and actual duration using Kolmogorov–Smirnov test (the p-values are presented in Table 1 and the cumulative probability functions in Figure 4). Although the Kolmogorov-Smirnov test does not pair the observations, it yields quantitatively similar results to the Wilcoxon matched-pairs signed-rank test, reported also in Table 1.

- *Result 3: The anchors cause systematic under/overestimation of task duration. The estimates are less biased when anchors are not present.*

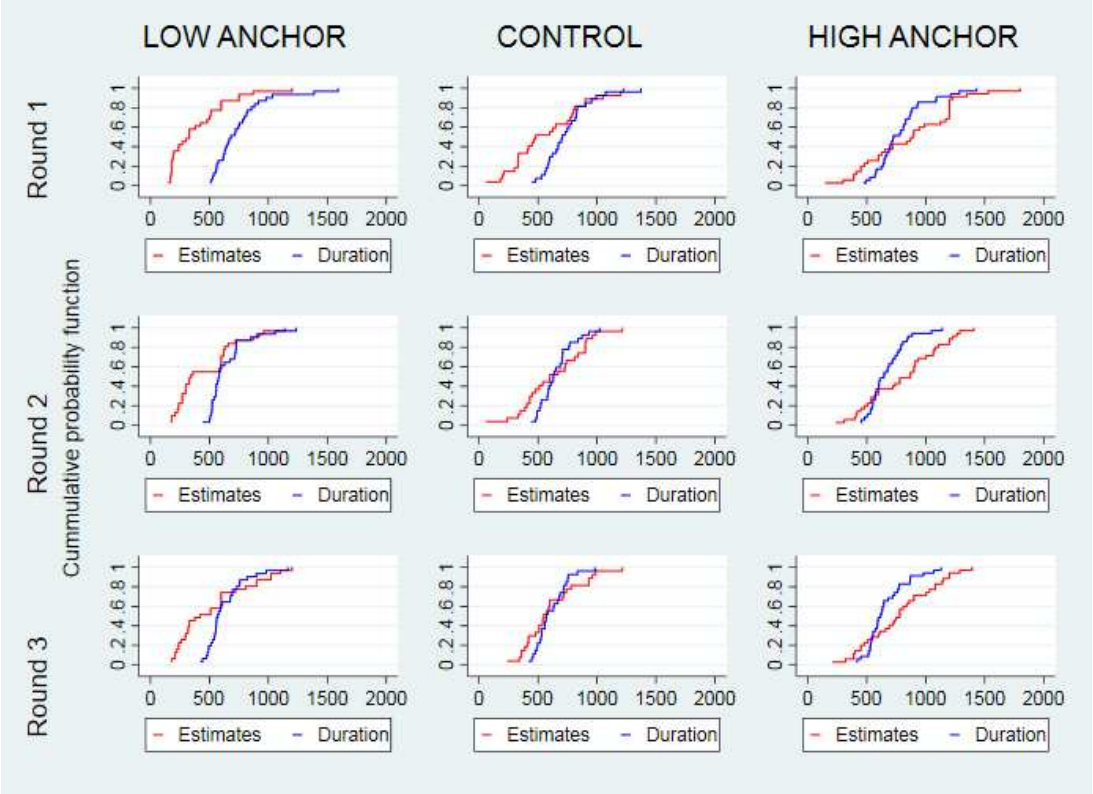


Figure 4. Cumulative probability distributions by treatment and round

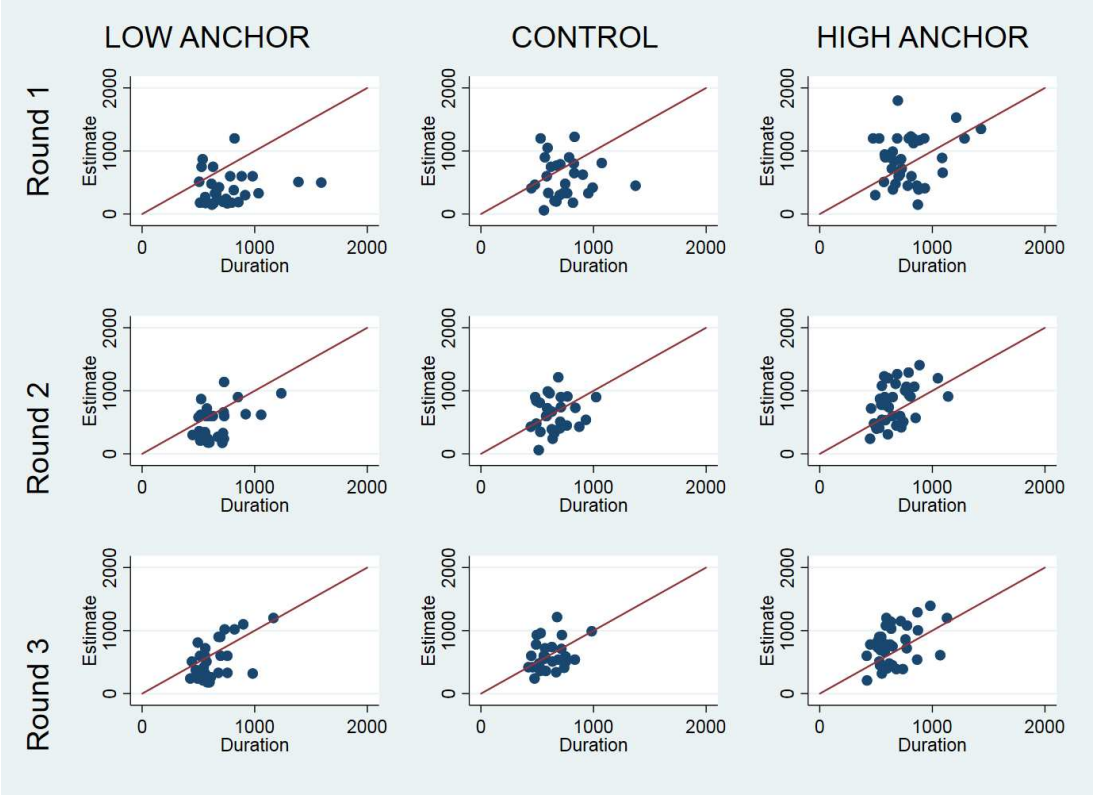


Figure 5. Individual estimates vs. actual task duration by treatment and round

6. Auxiliary results

a. Retrospective estimates

We further investigate the persistence of the anchoring effect by examining subjects' retrospective estimates. We asked subjects to retrospectively estimate the actual duration of Round 3, that is, how long the last task actually took them to complete (see Table 3 and Figure 6). Recall that our experimental task requires a significant cognitive attention, making it relatively hard to keep track of time and precisely construct the actual task duration in retrospect. Thus, we conjectured that without any feedback on the estimation accuracy or actual task duration, the anchors would influence the retrospective task duration estimates in the same direction as the estimates produced before the tasks were performed. In line with this conjecture, we find significant differences in retrospective estimates between the treatments (Mann-Whitney test p-values: <0.001 for Low Anchor vs. High Anchor; 0.003

for Control vs. Low Anchor; and 0.024 for Control vs. High Anchor).[9] Thus, even though subjects in the three experimental treatments perform the same task and it takes them approximately the same time to complete it, their retrospective estimates are significantly different. We conclude that anchors distort not only the estimates before the task, but also the retrospective task duration estimates, which in turn, can influence the future estimates, creating a persistent anchoring effect.

- *Result 4: The anchors influence the retrospective duration estimates in the same direction as the duration estimates produced before the completion of the task.*

Table 3. Retrospective estimates (Round 3)

Treatment	Low Anchor (N = 27)	Control (N = 26)	High Anchor (N = 30)
Actual duration, SD (seconds)	607 (150)	618 (131)	656 (163)
Mean retrospective estimate, SD (seconds)	493 (414)	615 (250)	793 (318)
Mean bias[a], SD (seconds)	-114 (459)	-3 (233)	136 (276)
Mean bias (%)	-17%	0%	22%
Mean absolute error, SD (seconds)	345 (317)	187 (134)	266 (190)
Mean absolute error (%)	50%	30%	41%
Median retrospective estimate (seconds)	360	590	735

a: The bias is calculated as a relative estimation error (Estimate – Actual duration).

[9] We dropped 10 extremely low estimates (from 6 to 30 seconds) from this analysis, reducing the total number of observations in the analysis of retrospective estimates to 83 (27 from the Low Anchor treatment, 26 from the Control treatment, and 30 from the High Anchor treatment). We suspect that these outliers are caused by participants' lack of attention. We asked for the retrospective estimate in the unit of seconds, which might have been overlooked by the particular 10 subjects (the estimates before the task were elicited in both minutes and seconds format). An alternative explanation is that since the accuracy of retrospective estimate was not financially incentivized, the subjects were not paying enough attention. However, since the outliers were almost equally distributed amongst low and high anchor treatments, removing those observations does not alter the result of treatment effect test in any way.

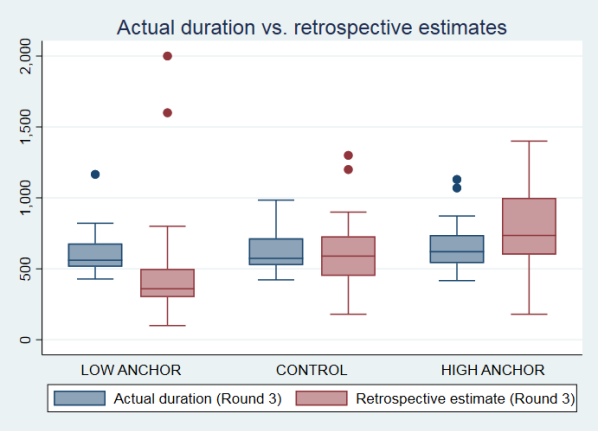


Figure 6. Actual duration vs. retrospective estimates

b. Estimation accuracy

Companies not only strive to have their estimates unbiased (that is, not systematically optimistic or pessimistic), but also strive for the high levels of accuracy (lowest errors) as accurate project duration estimates allow for more effective allocation of company resources. Thus, complementary to the estimation bias, we also analyze the estimation (in)accuracy, which is measured by absolute estimation errors (see Table 4 and Figure 7a) without the sign of the bias being taken into account. We find relatively large average absolute estimation errors, ranging from 40 to 56% in the anchor treatments and from 30 to 45% in the Control treatment.

Table 4. Absolute estimation errors

Treatments	Low Anchor (N = 31)			Control (N = 27)			High Anchor (N = 35)		
Rounds	R1	R2	R3	R1	R2	R3	R1	R2	R3
Subjects: Mean estimate, SD (seconds)	393 (252)	472 (261)	520 (303)	577 (314)	634 (270)	608 (236)	868 (385)	809 (314)	775 (302)
Function: Estimate (seconds)	-	763	659	-	763	659	-	763	659
Subjects: Mean absolute error, SD (seconds)	427 (222)	250 (148)	231 (144)	335 (234)	237 (159)	180 (146)	328 (228)	263 (169)	261 (148)
Function: Mean absolute error, SD (seconds)	-	182 (98)	132 (98)	-	154 (90)	117 (75)	-	151 (99)	132 (107)
Subjects: Mean absolute error (%)	56%	39%	37%	45%	37%	30%	45%	40%	41%
Function: Mean absolute error (%)	-	30%	21%	-	26%	21%	-	25%	20%

Subjects' estimate is better	-	35%	29%	-	33%	33%	-	31%	23%
Function estimate is better	-	65%	71%	-	67%	67%	-	69%	77%

One possible way to improve estimation accuracy is to use historical information as suggested by Kahneman & Tversky (1979). We construct a simple tool which calculates the average actual task duration of all subjects in the last round and uses this calculated average as an individual duration estimate for all subjects in the current round. We compare the accuracy of such a tool with subjects' own estimates, for both the mean absolute estimation errors and the proportion in which the tool outperforms the subjects' estimates (see Table 4 and Figure 7b). Overall, we find that absolute estimation errors of the tool are on average 13.5 percentage points lower than the subjects' estimates and are more accurate on average 69% of the time. In particular, we find that the mean errors of our tool are 162 seconds (with SD of 96 sec.) in Round 2 and 128 seconds (with SD of 95 sec.) in Round 3, compared to subjects' average estimation errors of 251 seconds (with SD of 158 sec.) and 227 seconds (with SD of 148 sec.), respectively. The difference in errors is statistically significant for both Round 2 and 3 (Mann-Whitney test p-values <0.001). Moreover, we find that this holds not only in the anchor treatments, but also in the Control treatment, in which subjects are less biased in their estimation, although the difference in Round 3 is statistically insignificant (Mann-Whitney test p-values 0.048 in Round 2 and 0.149 in Round 3). Our results are consistent with the argument that consulting historical data of average task/project completion time can lead to improvements in estimation accuracy.

- *Result 5: The simple prediction tool based on historical averages outperforms subjects' own estimates in terms of the overall estimation accuracy.*

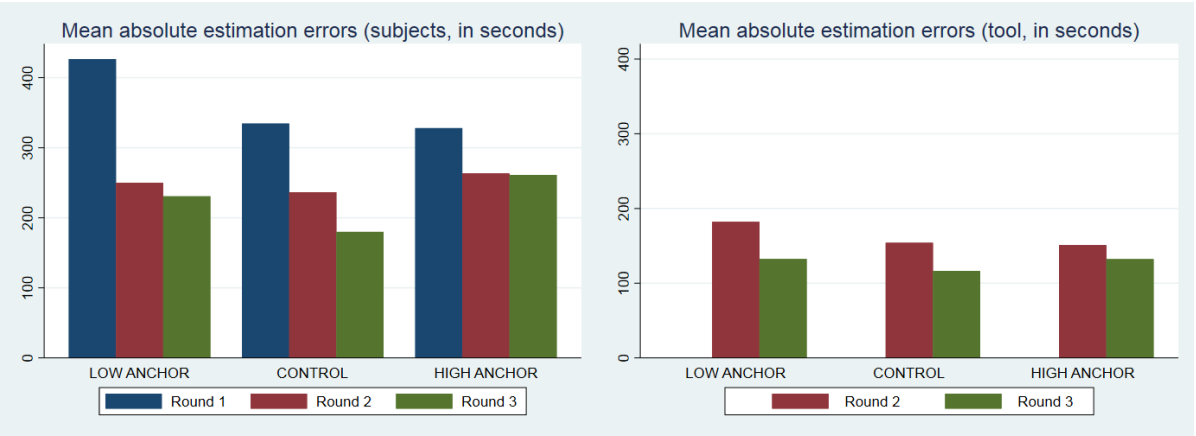


Figure 7a. Mean absolute estimation errors - subjects **Figure 7b. Mean absolute estimation errors – tool**

Note: Since we construct the prediction tool from past actual task duration, we have no prediction for Round 1 (i.e., no blue bars for Figure 7b).

c. Self-anchoring in the Control treatment

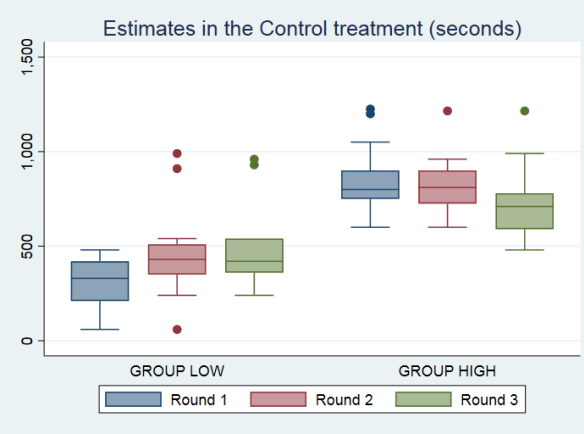
Although the estimates in Round 2 and 3 of the Control treatment are almost unbiased (see Result 3) and best calibrated, the average estimation inaccuracy measured by absolute estimation errors is still relatively high (see Table 4 and Figure 7a). The low bias is driven by a relatively similar number of subjects that underestimated the duration to those who overestimated. On the other hand, the relatively high estimation inaccuracy is caused by an extensive spread of the estimates. These estimates range from 1 minute to over 20 minutes in Round 1 and are often similar in the following rounds at the individual level. We test whether there could exist a "self-anchoring" effect, meaning whether subjects could have become anchored on their own duration estimates produced in Round 1. We divide the subjects who participated in the Control treatment into two subgroups based on the median estimate in Round 1: the "Low group" consisting of 14 subjects with the lowest estimates (from 60 to 480 seconds) and the "High group" consisting of 13 subjects with the highest estimates (from 600 to 1225 seconds). We compare these two subgroups with regards to both estimates and actual task duration (see Table 5, Figures 8a and 8b). We find a strong "self-anchoring" effect, similar to the main anchoring effect described in Results 1 and 2. While there are no significant differences in the actual task duration between the two subgroups, the estimates are significantly different. Subjects, who start with relatively low (high) estimates, generally keep their following estimates low (high). Furthermore, we also find this "self-anchoring" effect in the retrospective estimates.

- *Result 6: Subjects in the Control treatment are anchored on their own initial duration estimates.*

Table 5. Self-anchoring in the Control treatment

Medians (seconds)	Group Low (N = 14)	Group High (N = 13)	Mann-Whitney (p-values)
Estimates			
Round 1	330	800	<0.001
Round 2	430	810	0.001
Round 3	420	710	0.002
Actual duration			
Round 1	715	705	0.846
Round 2	645	610	0.680
Round 3	606	569	0.808
Retrospective estimates			
Round 3	450 (N = 13)[a]	650	0.025

Note: a: one retrospective estimate was dropped, see footnote 9.



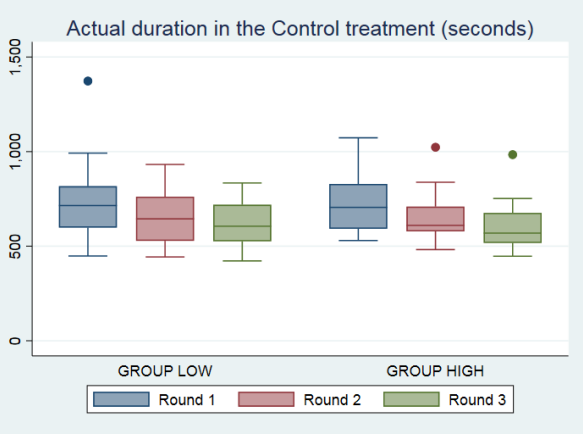


Figure 8a. Estimates in the Control treatment

Figure 8b. Actual duration in the Control treatment

d. Robustness

Finally, we conduct additional analyses to check for possible factors that might have affected our results, such as the number of incorrect answers to comparisons in the experimental task, time spent on the anchoring question, time spent on the estimation, demographics, and risk attitudes. We do not find any significant effects of these factors; the results of the treatment effects are robust. The details are available upon request. One interesting result is that there is a negative correlation between the task enjoyment, elicited in the questionnaire, and the actual task duration as well as a positive correlation between the task enjoyment and estimation accuracy. Thus, to some extent unsurprisingly, those who enjoyed the experiment more also earned more money.

7. Discussion

Project management methodology textbooks widely acknowledge the planning phase of the project as the most important phase of the whole project lifecycle because the success of the project ultimately relies on the accuracy of estimates in the project plan. Arguably, it is also the most difficult phase, since the crucial estimates are often requested without having enough information about the exact project scope. Moreover, it is quite common that project managers do not necessarily have the project team assembled and available at this point. As a result, the planning is executed under a high level of both uncertainty and ambiguity, and the estimates are vulnerable to anchoring in a variety of forms. In this paper, we test the effect of anchors that can occur as relatively uninformed suggestions

or expectations from project stakeholders, on the duration of project tasks. The presence of such anchors in the project planning phase often brings negative consequences for the company. While very low estimates almost inevitably lead to project failure, very high estimates contribute to inefficient allocation and utilization of company resources. We hypothesized that anchors could cause a large and predictable estimation bias and that this bias could potentially carry over from one project to another.

Our experimental results support all our hypotheses. We find that the introduction of anchors causes a systematic estimation bias that persists over time, even though we use a relatively familiar task and create favorable conditions for estimating the task duration accurately. The anchoring effect is not eliminated in the environment where subjects are incentivized (akin to Epley & Gilovich, 2005; Wilson, Houston, Etling, & Brekke, 1996) and have experience with the estimated task. Previous research has shown that judgmental anchoring can have a long-lasting effect in terms of overcoming a time gap (at least of one week) between the introduction of anchor and the actual estimation as found by Mussweiler (2001). In addition, it seems that although raising awareness regarding cognitive biases may reduce the anchoring effect, it does not eliminate it (Shepperd, Mair, & Jørgensen, 2018). These findings together with our strong and robust anchoring effects (observed in an environment designed to mitigate the bias), suggest that anchors will play a major role in more complex processes, such as project planning, where project teams often deal with novel and ambiguous tasks.

We believe that the power of anchors in our experiment stems mainly from two factors. The first one is their perceived plausibility. We intentionally deviate from the procedures often used in psychological research where anchors are arbitrary and thus purposely uninformative. In contrast, we determine the values of anchors from the actual estimates provided by our subjects in the Control treatment. Although we do not reveal that anchors actually carry values of other subjects' estimates, due to concerns regarding potential social comparison confounds, we do not use any mechanisms that would discredit the anchors. The reason we present the anchors in a relatively non-arbitrary fashion, is to mimic the project management environment. Our research question focuses on anchors that are generated by project stakeholders and thus are not random. From this perspective our experiment is better thought of as an applied anchoring research and does not (and is not meant to) discriminate between various proposed psychological theories regarding the mechanism driving the anchoring effect e.g. the original anchoring-and-adjustment theory (Tversky & Kahneman, 1974; Epley & Gilovich, 2001) or relatively novel selective accessibility theory (Chapman & Johnson, 1999; Mussweiler & Strack, 1999).

We suppose that the magnitudes of anchors in our experiment are, to some extent, taken as plausible estimate suggestions and perceived as informative, or at least not taken as misleading. This conjecture is supported by subjects' comments collected through open-ended questions at the end of the experiment. None of the subjects mentioned the anchor or the anchoring question when we asked them to provide their thoughts about the experiment. These comments, however, offer some insights regarding how difficult it is to learn from the task experience and adjust the estimates away from the anchor under a relatively high cognitive load. For example, one subject mentions that "*The more I answered questions, the more I kept questioning myself. So even though the tasks were simple, as the rounds went on I found it more difficult to trust my first instinct and would take longer to answer and make more errors".* This statement is consistent with subject's progressively increasing estimates (150 seconds in Round 1, 180 seconds in Round 2, 180 seconds in Round 3, and 200 seconds as the retrospective estimate for Round 3). However, the actual duration of the task for the subject was approximately 600 seconds on average and the subject was in each round roughly 20 seconds faster than in the previous one, with a similar amount of incorrect answers. This shows that in the estimation process the subject placed a disproportionally larger weight on the anchor value (3 minutes) than on own task experience (remembered duration). Interestingly, the subject stayed on the screen at which the anchoring question was presented for 25 seconds, which is two times longer than the average.

The second factor that causes insufficient adjustment of estimates away from the anchors, as subjects gain more experience in estimation, is the absence of feedback regarding the actual task duration or the estimation accuracy. From the perspective of repeated estimation without feedback, our experiment is related to Bayesian updating. A subject's "prior" has the value of the anchor and can be updated by using the signal of relatively noisy remembered task duration. We find that, even though subjects are generally able to detect the direction of their mistakes, they usually underestimate their magnitude, causing a relatively slow convergence of estimates towards the actual task duration. Thus, the insufficient adjustment (updating) causes the anchoring effect to persist over the three experimental rounds. Moreover, if we consider the retrospective estimates produced after Round 3 as candidates for future estimates, the anchoring effect would probably persist at least to fictitious Round 4. Thus, our results suggest that having experienced planners is not sufficient enough to eliminate the anchoring effect in estimating or planning. Although our experiment does not test the effect of feedback, it seems reasonable to conjecture that a proper feedback regarding past results is crucial when the goal is to improve the estimation accuracy and to stop echoing the planning mistakes from the past.[10] In a similar vein, we could expect the anchoring effect in our experiment to disappear

[10] A parallel argument and evidence can be found in Roy, Mitten, & Christenfeld (2008).

if feedback on the actual duration is supplied after each round. We did not implement the feedback feature because we believe that having subjects learn how long it actually took them to complete the task would likely result in a design from which we would not be able to learn about the persistence of the anchoring effect. Nevertheless, we admit that while this conjecture appears to be intuitive, it still should be properly tested and as such constitutes a natural extension of the current study.

Given the robustness of the anchoring effect, another interesting possibility for future research is to focus more on the scenarios and institutions in which anchors might aid the planners to provide better estimates and decision-makers to make better decisions. Recall that our results from the Control treatment show a noticeable sign of the optimism bias in Round 1 estimates. However, after gaining initial experience with the task, the estimates in the following rounds are on average compellingly unbiased, pointing out the independence of estimation errors and relatively equal distribution of under and overestimates. This observation resembles a phenomenon known as the "wisdom of the crowd" (Galton, 1907). Nevertheless, despite no bias at the aggregate level, we find a relatively high estimation inaccuracy, which is parallel to the assumption of "bias-variance dilemma" (Geman, Bienenstock, & Doursat, 1992). The estimates produced without a specific biasing intervention in place (*tabula rasa*) often suffer from a high variance due to a large number of parameters that can influence them. Hence, a small anchoring bias may be beneficial if the goal is to reduce the variance in individual estimates and improve the overall estimation accuracy. Our data reveal that the prediction based on the history of actual task duration data outperforms individual subjects' estimates. Hence, it might be useful to verify whether the use of "good anchors" such as historical averages yields significant improvements in the project estimation process.

Finally, we find evidence of the "self-anchoring effect" in the Control treatment. A group of subjects that produce lower/upper end estimates in Round 1 never make up their estimation differences. Hence, if no anchor is given before the first estimation, in the absence of feedback, future estimates are prone to be anchored on the planner's first own duration estimate. We consider the "self-anchoring effect" to be yet another promising area for future research.

Acknowledgements: This paper is based on Matej Lorko's dissertation chapter written at the Macquarie Graduate School of Management. We wish to thank the audiences of 2017 ESA European Meeting, 2017 Slovak Economic Association Meeting, 2017 ANZWEE, 2017 Behavioral Economics: Foundations & Applied Research Workshop at the University of Sydney and 2018 ESA Asia Pacific Meeting who provided helpful comments and suggestions. Financial support was provided by Macquarie Graduate School of Management.

8. References

Alevy, J. E., Landry, C. E., & List, J. A. (2015). Field experiments on the anchoring of economic valuations. *Economic Inquiry*, *53*(3), 1522–1538. https://doi.org/10.1111/ecin.12201

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

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent Arbitrariness": Stable Demand Curves Without Stable Preferences. *The Quarterly Journal of Economics*, *118*(1), 73–106. https://doi.org/10.1162/00335530360535153

Blankenship, K. L., Wegener, D. T., Petty, R. E., Detweiler-Bedell, B., & Macy, C. L. (2008). Elaboration and consequences of anchored estimates: An attitudinal perspective on numerical anchoring. *Journal of Experimental Social Psychology*, *44*(6), 1465–1476. https://doi.org/10.1016/j.jesp.2008.07.005

Boltz, M. G., Kupperman, C., & Dunne, J. (1998). The role of learning in remembered duration. *Memory & Cognition*, *26*(5), 903–921. https://doi.org/10.3758/BF03201172

Cervone, D., & Peake, P. K. (1986). Anchoring, efficacy, and action: The influence of judgmental heuristics on self-efficacy judgments and behavior. *Journal of Personality and Social Psychology*, *50*(3), 492–501. https://doi.org/10.1037/0022-3514.50.3.492

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, Activation, and the Construction of Values. *Organizational Behavior and Human Decision Processes*, *79*(2), 115–153. https://doi.org/10.1006/obhd.1999.2841

Cox, J. C., Sadiraj, V., & Schmidt, U. (2015). Paradoxes and mechanisms for choice under risk. *Experimental Economics*, *18*(2), 215–250. https://doi.org/10.1007/s10683-014-9398-8

Critcher, C. R., & Gilovich, T. (2008). Incidental environmental anchors. *Journal of Behavioral Decision Making*, *21*(3), 241–251. https://doi.org/10.1002/bdm.586

Davis, H. L., Hoch, S. J., & Ragsdale, E. K. E. (1986). An Anchoring and Adjustment Model of Spousal Predictions. *Journal of Consumer Research*, *13*(1), 25–37. https://doi.org/10.1086/209045

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making. *Personality and Social Psychology*

Bulletin, *32*(2), 188–200. https://doi.org/10.1177/0146167205282152

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, *12*(5), 391–6. https://doi.org/10.1111/1467-9280.00372

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making*. https://doi.org/10.1002/bdm.495

Fischbacher, U. (2007). Z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178. https://doi.org/10.1007/s10683-006-9159-4

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

Galinsky, A. D., & Mussweiler, T. (2001). First offers as anchors: The role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology*, *81*(4), 657–669. https://doi.org/10.1037/0022-3514.81.4.657

Galton, F. (1907). Vox populi (The wisdom of crowds). *Nature*, *75*(7), 450–451.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, *4*(1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125. https://doi.org/10.1007/s40881-015-0004-4

Grimstad, S., & Jørgensen, M. (2007). Inconsistency of expert judgment-based estimates of software development effort. *Journal of Systems and Software*, *80*(11), 1770–1777. https://doi.org/10.1016/j.jss.2007.03.001

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. https://doi.org/10.1037/a0025996

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: a methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383-403; discussion 403-451. https://doi.org/10.1037/e683322011-032

Hinds, P. (1999). The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *Journal of Experimental Psychology: Applied*, *5*(2), 205–221.

Holt, C. A. (1986). Preference reversals and the independence axiom. *The American Economic Review*, *76*(3), 508–515.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655. https://doi.org/10.1257/000282802762024700

Jacowitz, K. E. ., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161–1166. https://doi.org/10.1177/01461672952111004

Joørgensen, 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. https://doi.org/10.1016/S0263-7863(03)00085-1

Jørgensen, M., & Grimstad, S. (2008). Avoiding Irrelevant and Misleading Information When Estimating Development Effort. *IEEE Software*, *25*(3), 78–83.

Jørgensen, M., & Grimstad, S. (2011). The impact of irrelevant and misleading information on software development effort estimates: A randomized controlled field experiment. *IEEE Transactions on Software Engineering*, *37*(5), 695–707. https://doi.org/10.1109/TSE.2010.78

Josephs, R. A., & Hahn, E. D. (1995). Bias and Accuracy in Estimates of Task Duration. *Organizational Behavior and Human Decision Processes*. https://doi.org/10.1006/obhd.1995.1016

Kahneman, D., & Tversky, A. (1979). INTUITIVE PREDICTION: BIASES AND CORRECTIVE PROCEDURES. *TIMS Studies in the Management Sciences*, *12*, 313–327. https://doi.org/citeulike-article-id:3614496

König, C. J. (2005). Anchors distort estimates of expected duration. *Psychological Reports*, *96*(2), 253–256. https://doi.org/10.2466/PR0.96.2.253-256

König, C. J., Wirz, A., Thomas, K. E., & Weidmann, R.-Z. (2014). The Effects of Previous Misestimation of Task Duration on Estimating Future Task Duration. *Current Psychology*, *34*(1), 1–13. https://doi.org/10.1007/s12144-014-9236-3

Løhre, E., & Jørgensen, M. (2016). Numerical anchors and their strong effects on software development effort estimates. *Journal of Systems and Software*, *116*, 49–56. https://doi.org/10.1016/j.jss.2015.03.015

Morgenshtern, O., Raz, T., & Dvir, D. (2007). Factors affecting duration and effort estimation errors in software development projects. *Information and Software Technology*, *49*(8), 827–837. https://doi.org/10.1016/j.infsof.2006.09.006

Mussweiler, T. (2001). The durability of anchoring effects. *European Journal of Social Psychology*, *31*, 431–442.

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, *35*(2), 136–164. https://doi.org/10.1006/jesp.1998.1364

Plous, S. (1989). Thinking the Unthinkable: The Effects of Anchoring on Likelihood Estimates of Nuclear War. *Journal of Applied Social Psychology*, *19*(1), 67–91. https://doi.org/10.1111/j.1559-1816.1989.tb01221.x

Project Management Institute. (2013). *A guide to the project management body of knowledge (PMBOK ® guide)*. *Project Management Institute*. https://doi.org/10.1002/pmj.20125

Project Management Institute. (2017). *PMI's Pulse of the Profession 2017*.

Ritov, I. (1996). Anchoring in Simulated Competitive Market Negotiation. *Organizational Behavior and Human Decision Processes*, *67*(1), 16–25. https://doi.org/10.1006/obhd.1996.0062

Roy, M. M., & Christenfeld, N. J. S. (2007). Bias in memory predicts bias in estimation of future task duration. *Memory & Cognition*, *35*(3), 557–564. https://doi.org/10.3758/BF03193294

Roy, M. M., Christenfeld, N. J. S., & McKenzie, C. R. M. (2005). Underestimating the Duration of Future Events: Memory Incorrectly Used or Memory Bias? *Psychological Bulletin*, *131*(5), 738–756. https://doi.org/10.1037/0033-2909.131.5.738

Roy, M. M., Mitten, S. T., & Christenfeld, N. J. S. (2008). Correcting memory improves accuracy of predicted task duration. *Journal of Experimental Psychology. Applied*, *14*(3), 266–275. https://doi.org/10.1037/1076-898X.14.3.266

Shepperd, M., Mair, C., & Jørgensen, M. (2018). An Experimental Evaluation of a De-biasing Intervention for Professional Software Developers.

Smith, V. L. (1991). Rational Choice: The Contrast between Economics and Psychology. *Journal of Political Economy*, *99*(4), 877–897. https://doi.org/10.2307/2069710

Thomas, K. E., & Handley, S. J. (2008). Anchoring in time estimation. *Acta Psychologica*, *127*(1), 24–

29. https://doi.org/10.1016/j.actpsy.2006.12.004

Tobin, S., & Grondin, S. (2015). Prior task experience affects temporal prediction and estimation. *Frontiers in Psychology*, *6*(July), 1–7. https://doi.org/10.3389/fpsyg.2015.00916

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

Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, *125*(4), 387–402. https://doi.org/10.1037/0096-3445.125.4.387

Woods, D., & Servátka, M. (2016). Testing psychological forward induction and the updating of beliefs in the lost wallet game. *Journal of Economic Psychology*, *56*, 116–125. https://doi.org/10.1016/j.joep.2016.06.006

Yang, C., Sun, B., & Shanks, D. R. (2017). The anchoring effect in metamemory monitoring. *Memory & Cognition*. https://doi.org/https://doi.org/10.3758/s13421-017-0772-6

Instructions

Thank you for coming. Please, put away your watches, mobile phones, and any other devices that show time. The experimenter will check the cubicles for the presence of time showing devices before the start of the experiment.

Also, please note, that from now, until the end of the experiment, no talking or any other unauthorized communication is allowed. If you violate any of the above rules, you will be excluded from the experiment and from all payments. If you have any questions after you finish reading the instructions, please raise your hand. The experimenter will approach you and answer your questions in private.

Please, read the following instructions carefully. The instructions will provide you with information on how to earn money in this experiment.

The experimenters will keep track of your decisions and earnings by your cubicle number. The information about your decisions and earnings will not be revealed to other participants.

Three rounds of the same two tasks

The experiment consists of three rounds. In each round, you will perform two tasks – the comparison task and the estimation task.

The comparison task

The screen will show an inequality between two numbers ranging from 10 to 99. You will evaluate whether the presented inequality is true or false. Immediately after you submit your answer, a new inequality will show up. This task finishes after you have provided 400 correct answers.

Examples:

The estimation task

At the beginning of each round, you will be asked to estimate how long it will take you to complete the comparison task, that is, how long it will take you to provide 400 correct answers.

The earnings structure

Your total earnings (in AUD) from the experiment will be the sum of your comparison task earnings and estimation task earnings for all three rounds.

The comparison task earnings (CTE)

In each round, your comparison task earnings (in AUD) will be calculated as follows:

$$\text{Comparison task earnings } = \frac{7 * \text{ (number of correct answers} - \text{number of incorrect answers)}}{\text{actual time in seconds}}$$

Your comparison task earnings will depend on the actual time in which you complete the task and on the number of correct and incorrect answers you provide. Notice that the faster you complete the task (i.e., provide 400 correct answers), the more money you earn. However, note also that your earnings will be reduced for every incorrect answer that you provide.

The estimation task earnings (ETE)

In each round, your estimation task earnings (in AUD) will be calculated as follows:

$$\text{Estimation task earnings } = 4.5 - 0.05 * |\text{actual time in seconds } - \text{estimated time in seconds}|^{\times}$$

[×] If the difference between your actual and estimated time is more than 90 seconds (in either direction) your estimation task earnings will be set to 0 for the given round.

Notice that the estimation task earnings will depend on the accuracy of your estimate. The calculation is based on the absolute difference between the actual time in which you complete the comparison task and your estimated time.

Your total earnings

Total earnings = $(CTE_1 + ETE_1) + (CTE_2 + ETE_2) + (CTE_3 + ETE_3)$

Notice, that your earnings will be higher:

- The faster you complete the comparison tasks
- The fewer incorrect answers you provide
- The more accurate your estimates are

You need to complete the entire experiment in order to get paid.

When you finish

After the last round you will be asked to answer a few questions about the experiment. The final screen will display the summary of your earnings. When you finish the experiment, please stay quietly seated in your cubicle until the experimenter calls your cubicle number. You will then go to the room at the back of the laboratory to privately collect your experimental earnings in cash.

If you have any questions, please raise your hand.