Anchoring in Project Duration Estimation

Lorko, Matej and Servátka, Maroš and Zhang, Le

MGSM Experimental Economics Laboratory, Macquarie Graduate School of Management

12 April 2019

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

March 19, 2019

Abstract: The success of a business project often relies on the accuracy of its schedule. Inaccurate and overoptimistic schedules can lead to 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. Our laboratory experiment contributes to the methodology of investigating the robustness and persistence of the anchoring effect in the following ways: (1) we investigate the anchoring effect by comparing the behavior in low and high anchor treatments with a control treatment where no anchor is present; (2) we provide a more accurate measurement by incentivizing participants to provide their best duration estimates; (3) we test the persistence of the anchoring effect over a longer horizon; (4) we evaluate the anchoring effect also on retrospective estimates. We find strong anchoring effects and systematic estimation biases that do not vanish even after the task is repeatedly estimated and executed. In addition, we find that such persisting biases can be caused not only by externally provided anchors, but also by the planner's own initial estimate.

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

JEL codes: C91, D83, D92, O21, O22

1. Introduction

Effective project planning processes, capable of producing realistic project schedules, 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 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 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 using deadlines as 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 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?"), customer expectations (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."). These mostly unintentional anchors 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 require deadline and budget extensions, distort company resources, and hinder customer satisfaction. In some scenarios, such projects are cancelled before their completion, resulting in sunk costs without bringing any actual benefit to the company. One of the most prominent examples of major project planning failures is the construction of Sydney Opera House. The iconic building 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.

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. If project planners do not receive details regarding the actual duration of project tasks once the project is completed, they may remain unaware of the project being misestimated. Subsequently, they may become anchored on either historical or their own estimates and prone to repeating the same estimating mistakes again. Thus, the anchoring effect and resulting inefficiencies can carry over from one project to similar projects in the future.

Based on the above insights, we conjecture that (i) the presence of numerical anchors influences project 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 similar projects. Moreover, in the absence of externally provided anchors, the planner's first own estimate can act as a future anchor in itself.

We test our conjectures in a controlled laboratory experiment employing a simple real-effort task (representing a stylized project) of evaluating inequalities of two-digit number pairs. Subjects are asked to estimate how long it will take them to correctly assess 400 such inequalities. Before they provide their estimates, we ask whether it will take them less or more than [the anchor time] to complete the task. Following the estimation, subjects work on 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, our experiment can be considered an instance of applied anchoring research.

Our paper contributes to the methodology of investigating the robustness and persistence of the anchoring effect in the following ways. First, for a clean identification of the anchoring effect, we conduct a control treatment in which the anchoring question is not asked. Second, unlike previous studies of anchoring in task duration or effort estimation, our experiment employs real incentives for estimation accuracy and task performance. These incentives mimic the project management environment outside the laboratory and make subjects take their choices seriously as their financial earnings in the experiment are determined by their decisions. Third, our research extends the standard one-shot anchoring paradigm into testing the effect of anchors over a longer horizon. In the experiment, the estimation and the task execution 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 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. Fourth, we evaluate the anchoring effect not only on estimates provided prior to task execution, but also on retrospective estimates.

We provide clear evidence that numerical anchors can bias estimates even in an environment encompassing multiple features known to alleviate estimation biases (i.e., familiarity with the task and task simplicity; see section 3 for details) and under the incentive structure that motivates individuals to provide unbiased estimates of their own performance. 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 estimation but can persist over time and carry over to future estimations if an external corrective action (e.g., estimation feedback) does not take place. Although the observed anchoring effects slightly diminish with time, they remain statistically significant during the entire experiment, except for the third estimation in the Low Anchor treatment. 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 estimates.

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 in the planning phase of the current project is to consult historical information from past projects (see also Lorko, Servátka, & Zhang, 2019). Our experimental data show that a measure as simple as the mean duration of completed tasks 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 utilization of simple statistics from similar past projects, complementary to the more traditional step-by-step planning based on the specification of the current project. Third, relying on planners being aware of their mistakes and correcting them in the future is not necessarily an effective strategy. In order to improve project planning processes in the company, it seems 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 a systematic estimation bias due to insufficient adjustment of the estimate away from the starting point

(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. Subsequently, they were asked to consider whether the correct answer lies above or below the proposed value. The final estimates of the percentage of African countries in United Nations were clearly biased by the respective anchors, yielding a median of 25% in the low anchor group and 45% in the high anchor group.

The anchoring effect has been subsequently observed in several other studies of general knowledge judgments, but also in many other domains (see Furnham & Boo, 2011, for a comprehensive review). The most relevant to our research questions is the anchoring effect evidence in task duration estimation and effort estimation.[1] For example, estimates of effort required for software development can be anchored on managers' (Aranda & Easterbrook, 2005) or customers' expectations, as documented in both laboratory (Jørgensen & Sjøberg, 2004) and field environment (Jørgensen & Grimstad, 2011). The anchoring effect can be introduced also by a variation in wording instead of using numerical values. Jørgensen & Grimstad (2008) find different work-hours estimates of the same task, labelled as "minor extension", "extension" or "new functionality" for different treatment groups.[2]

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. Prior to the estimation and actual task execution, subjects are asked to consider whether they would need more or less than either 30 (a low anchor) or 90 (a high anchor) minutes to complete the task. The estimates in the 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 obtained by Thomas & Handley (2008), who additionally find that significant differences in duration estimates can be caused even by anchors irrelevant to the task or the estimation problem.

[1] In project management, the duration of tasks is often reported in man-hours (man-days) and referred to as "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. The expression is strongly associated with a particular number and serves as a powerful anchor for effort estimation.

The bias induced by anchors thus seems to be a common attribute of task duration and effort estimates. However, most of the previous studies employ relatively unfamiliar tasks, possibly exacerbating the anchoring effect. In addition, the studies are one-shot in their nature, making it impossible to determine whether individuals learn from their previous estimation errors caused by anchors. These features are in contrast with the state of affairs in business environment, where the project duration estimates are usually being produced by experienced professionals who are familiar with the task at hand and where the estimation is often repeated.

One might expect that prior experience will reduce the influence of nuisance factors, such as anchors, in economic decision-making.[3] Thomas & Handley (2008) report that subjects who admitted to having performed a similar task in the past are less affected by anchors in their experimental setting. Similarly, Løhre & Jørgensen (2016) find that subjects with a longer tenure in the profession are less influenced by anchors and provide more accurate albeit still biased estimates than those with less experience. However, these findings are rather incidental than obtained in controlled conditions. Furthermore, having experience with the task itself does not necessarily imply having experience with its duration estimation (Halkjelsvik & Jørgensen, 2012). To the best of our knowledge, the influence of anchors in duration estimation has never been tested in more than one period and there are no empirical findings on whether task experience mitigates the anchoring effect. In fact, despite the extensive body of anchoring-related research, relatively little is known about the long-term effect of anchors in general.

In addition to the potential issue of insufficient experience with the task, none of the previous laboratory and classroom experiments incentivized subjects for their estimation accuracy, as only flat fees or course credits were used. The lack of real incentives can result in a hypothetical bias (e.g., Hertwig & Ortmann, 2001) and it is therefore questionable whether the anchoring effect is robust to environments where misestimation can result in losses.

To summarize, there exists a considerable body of empirical literature on anchoring in task duration and effort estimation with data from laboratory experiments, field experiments, and field questionnaires. While the anchoring effect is a pervasive phenomenon, it is not clear whether it persists over time and how it interacts with experience. To thoroughly examine the prevalence as well as limitations of the influence of anchors, we design an experiment incorporating task experience and

[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 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.

repetition, together with meaningful incentives for task performance and estimation accuracy. Our study therefore 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 pronounced in environments characterized by absence of these features.

3. Experimental design

We conduct an incentivized laboratory experiment employing an individual real-effort task to test whether numerical anchors bias duration estimates and whether such effect persists over time. The experiment consists of three rounds. In every round, subjects first estimate how long it will take them to complete the upcoming task. Next, they execute the task.

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, e.g., 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 suggests that tasks exceeding 12.5 minutes are usually characterized by underestimation whereas very short tasks are usually overestimated (Roy, Christenfeld, & McKenzie, 2005). Our design thus creates a favorable environment for subjects to estimate the task duration accurately.

Subjects perform similar two-digit number comparisons in each of the three rounds. To test whether individuals are able to overcome the anchoring effect by learning from the experience itself, we provide no feedback on the actual task duration or the 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 regarding the actual hours spent by project team members on each task. Even if the actual duration of project tasks is evaluated against the project plan, the schedule delays are often attributed to factors other than inaccurate estimation. Both no feedback and inadequate feedback make project planners unlikely to improve the accuracy of their 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. Incentivizing both accuracy and 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).

We implement a linear scoring rule to incentivize the estimation accuracy. According to the rule, the estimation accuracy earnings depend on the absolute difference between the actual task duration and the estimate.[4] In every round, the maximum earnings from a 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. If the difference between the actual and the estimated time in either direction exceeds 90 seconds, the estimation accuracy earnings are set to zero for the given round.[5] This particular design feature is implemented because of our expectations of a strong anchoring bias and the related estimation inaccuracy that could cause many subjects to end up with negative earnings. Our setting parallels a common practice in companies where planners are praised or rewarded for their accurate estimates after a successful project completion but are usually not penalized for inaccurate estimates when a project fails.

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

The task performance earnings, 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 it is 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 the estimation accuracy earnings.

[4] While we acknowledge that the linear scoring rule might not be the most incentive compatible one, it is arguably more practical to implement in an experimental environment than more complex scoring rules (e.g., quadratic or logarithmic) due to ease of explanation to subjects (Woods & Servátka, 2016).

[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.

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

Since there are two type of incentives, there is a concern that subjects might try to create a portfolio of accuracy and performance earnings (Cox & Sadiraj, 2018). While one can control for the portfolio effect by randomly selecting one type of incentives for payment (Cox, Sadiraj, & Schmidt, 2015; Holt, 1986), we choose to incentivize subjects for both types 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 as they 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 and the clocks on the computer screens are hidden. 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 working pace and match it with their estimates.[6]

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," i.e., whether the first estimate anchors future estimates of the same task. We use the estimates obtained in the Control treatment to calibrate the values of the low and the high anchor. The low anchor value is set at the 7th percentile and the high anchor value at the 93rd percentile of the Control treatment estimates, in line with the procedure for measurement of the 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 Round 1 (and only before 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?

Since we intentionally do not use arbitrary values for anchors, the anchoring manipulation in our experiment could possibly establish reference points for subjects to compare their estimates or

[6] 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 aversion in our regression analysis. We find no significant effect of elicited risk preferences on estimates or actual task duration.

performance with.[7] If our anchors were perceived as reference points, subjects could view their experiment performance either positively (mostly in the High Anchor treatment) or negatively (mostly in the Low Anchor treatment). Such a reference point effect could influence their mood and overall behavior. To investigate this possibility, after completing Round 3 but before learning about their payoffs, we asked subjects to evaluate on a 10-point scale how much they enjoyed the experiment. We find no differences in self-reported enjoyment across treatments, suggesting that the implemented anchors were not perceived as reference points.

4. Hypotheses

We hypothesize that the presence of anchors will systematically bias task duration estimates in Round 1. Specifically, we hypothesize that the estimates in the Low Anchor treatment will be significantly lower than those in the Control treatment, and the 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 hypothesize that the anchoring effect will carry over to subsequent estimations 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 $1, 2, \text{and } 3$, respectively
    and the subscript $(\text{L}, \text{C}, \text{or H})$ refers to the Low Anchor, Control, and High Anchor treatment.

We further hypothesize that anchors will have no effect on the actual task duration. This hypothesis is based on the imposed incentive structure that prompts subjects not only to estimate accurately but also to complete the task fast, independently of the treatment, and motivated by the previous findings in the literature on anchoring in task duration estimation. In particular, König (2005) finds that while duration estimates are systematically biased by anchors, the actual task duration is similar across

[7] Although anchors and reference points are sometimes used almost as synonyms, the literature draws a sharp distinction between the two concepts. While anchors are defined as extreme salient values influencing judgments, reference points represent salient neutral points on evaluation scales that alter the slope of the utility function, dividing the scale of values into "gains" and "losses" (Chapman & Johnson, 2002; Kahneman, 1992).

treatments. In Thomas & Handley (2008), the reported averages of actual task duration also seem not to differ across treatments, however, the authors do not report the statistical test.

- *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 further hypothesize that subjects will underestimate the task duration in the Low Anchor treatment but overestimate it in the High Anchor treatment. These predicted directions of the anchoring bias are due to the influence of anchors on the estimates but not on the actual task duration. Finally, recall that subjects in the Control treatment are not exposed to an anchor and that our design creates favorable conditions for providing unbiased estimates. We therefore hypothesize that there will be no systematic estimation bias in the Control treatment, parallel to the finding of unbiased task duration estimates under comparable incentive structure in Buehler, Griffin, & MacDonald (1997).

- *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

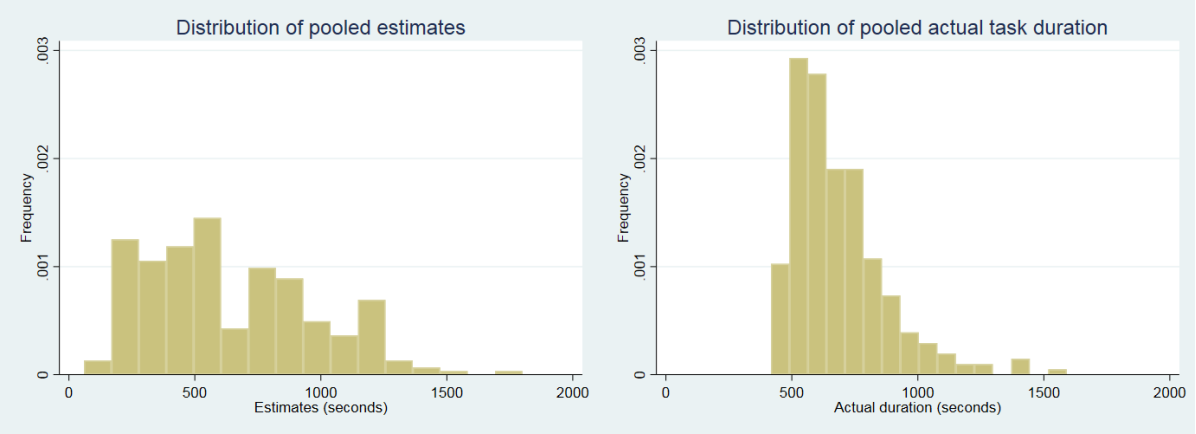
A total of 93 subjects (45 females, with a mean age of 20.7 and a standard deviation of 4.5 years) participated in the experiment that took place in the MGSM Vernon L. Smith Experimental Economics Laboratory at the Macquarie Graduate School of Management in Sydney.[8] Subjects were recruited via the online 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 attitudes assessment, and the demographical questionnaire. At the end of the experiment, subjects privately and individually collected their

[8] 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 upcoming round. The subject was debriefed while getting paid, which is when we also discovered her poor command of English. When verbally asked about the actual duration of Round 3 only, she made the same mistake again. Removing this data point does not change the treatment effect results.

experimental earnings in cash in the control room located at the back of the laboratory. On average, subjects spent 45 minutes in the laboratory and earned AUD 16.50.

First, we analyze the 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 1, right panel).[9] The distribution of estimates (presented also in Figure 1, left panel) follows a similar pattern, however, the skewness is less pronounced, mostly because of the high estimates in the High Anchor treatment. The Shapiro-Wilk test of normality indicates 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.[10]

Figure 1: Pooled estimates and actual task duration



Notes: Figure 1a (left panel) displays a histogram of subjects' task duration estimates in seconds. Figure 1b (right panel) displays a histogram of the actual task duration in seconds. Both figures display data pooled across all three treatments and all three experimental rounds.

Next, we analyze subjects' behavior in each round (see Table 1 for summary statistics). In line with Hypothesis 1, the estimates in the Low Anchor treatment are the lowest, whereas the estimates in the High Anchor treatment are the highest across all three experimental rounds. However, the absolute differences diminish from round to round (see Figure 2a for graphical display). We analyze the within-treatment changes using the Wilcoxon matched-pairs signed-rank test. We 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. The estimates in the High Anchor treatment gradually fall over time, with a statistically insignificant decrease between rounds. The estimates in the Control

[9] 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.

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

treatment are relatively stable, consistently positioned between the estimates of the Low Anchor and High Anchor treatments and do not change significantly from round to round.

Table 1: Summary statistics and treatment comparisons of subjects' estimates and performance

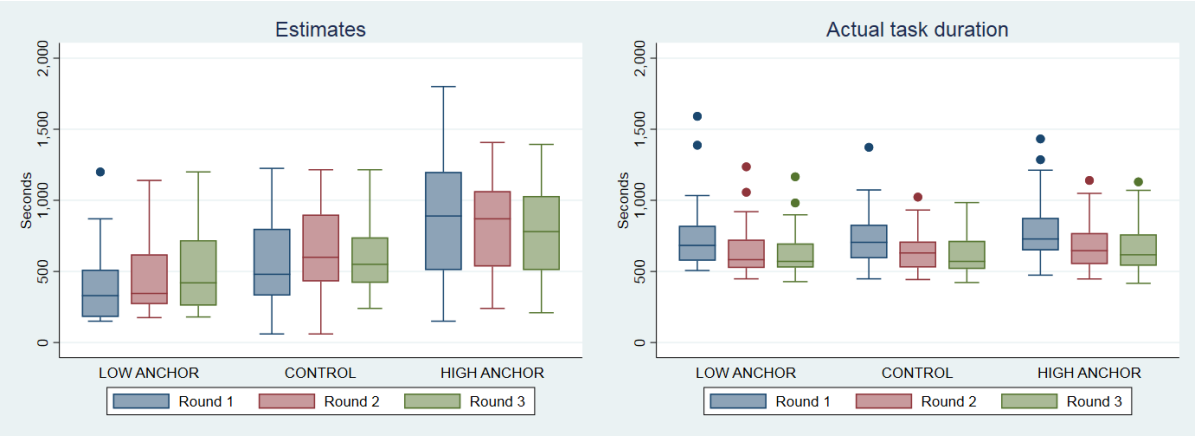
Panel A: Summary statistics									
Treatments	Low Anchor (N = 31)			Control (N = 27)			High Anchor (N = 35)		
Rounds	R1	R2	R3	R1	R2	R3	R1	R2	R3
Mean estimate, SD (seconds)	393 (252)	472 (261)	520 (303)	577 (314)	634 (270)	608 (236)	868 (385)	809 (314)	775 (302)
Mean duration, SD (seconds)	751 (242)	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
Panel B: Tests comparing 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]	Significant increase in R2 (p = 0.015) Insignificant increase in R3 (p = 0.374)			Insignificant increase in R2 (p = 0.225) Insignificant decrease in R3 (p=0.427)			Insignificant decrease in R2 (p = 0.107) Insignificant decrease in R3 (p=0.294)		

Notes: Panel A: Summary statistics of duration estimates, actual duration, estimation bias, and the proportion of overestimates and underestimates by treatments and rounds. Panel B: Statistical tests comparing the estimates and the actual duration; and the tests of trends in estimation. a: The bias is calculated as a relative estimation error (Estimate – Actual duration). b: Wilcoxon matched-pairs signed-rank test (p-values). SD refers to standard deviation.

Using the Mann-Whitney test (p-values are presented 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 with experience subjects move their estimates away from the anchor, the adjustment is insufficient and the anchoring effect diminishes only slightly.

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

Figure 2: Estimates and actual task duration



Notes: Figure 2a (left panel) displays box plots of task duration estimates. Figure 2b (right panel) displays box plots of the actual task duration Both figures display data by treatments and by experimental rounds.

Including the Control treatment in the design enables us to identify the effects and the persistence of both anchors individually, by comparing the differences in estimates between the anchor treatments and the Control treatment. 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 (see Table 2 for p-values). 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 the high anchor. However, such interpretation warrants some caution.

First, the asymmetry of the anchoring effect may be 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 constrained from above. As such, there is not much scope for the estimates in the Low Anchor treatment to be far away from the actual task duration. Second, the difference in persistence might also be 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 slightly over 12.5 minutes and thus the low anchor of 3 minutes is on average further away from the target value than the high anchor of 20 minutes. However, the mean actual task duration is lower in the following rounds (on average being 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 in the unbiased Control treatment.

Table 2: Statistical tests comparing the estimates and the actual task duration between treatments, by experimental rounds

		Mann-Whitney test (p-values)		
	Median in seconds (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. Using pooled data from all three treatments, 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 between Round 2 and Round 3). Nevertheless, there are no significant differences in the actual task duration across treatments (see Table 2 for p-values) in any round, supporting Hypothesis 2.[11] The lack of differences in the actual task duration across treatments is an evidence of subjects not manipulating their working pace, ruling out the portfolio effect concerns. It also supports the interpretation that subjects are working fast in order to maximize their performance earnings.

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

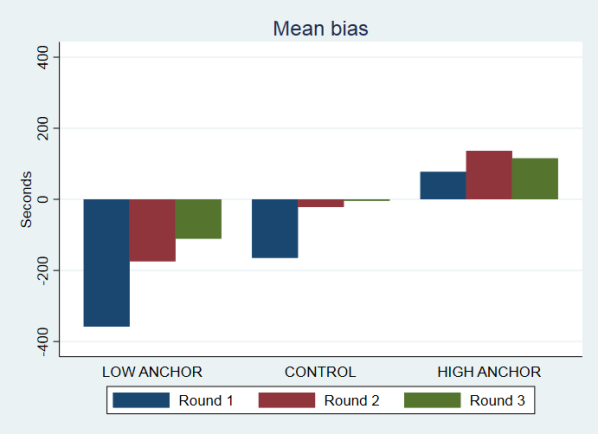
We next analyze the estimation bias, measured as a relative estimation error (i.e., estimate – actual duration; see Table 1 and Figure 3 for aggregate data and Figure 4 for individual-level data). We find that on average 73% of subjects in the Low Anchor treatment underestimate the time to complete the task and the mean estimate is 30% lower than the actual task duration. On the other hand, on average 59% of subjects in the High Anchor treatment overestimate the time to complete the task and the mean estimate is 18% higher than the actual task duration. The different directions of the bias are consistent with our Hypothesis 3.

In the Control treatment, we find prevalent underestimation in Round 1, with the estimates being on average 17% lower than the actual task duration. Reasons for the presence of underestimation in this treatment that does not feature anchors include wishful thinking, optimism, and providing a lower

[11] There are also no significant differences in the number of incorrect answers provided across treatments in any round. For data pooled across all treatments, the median number of incorrect answers is 9 in Round 1, and 7 in both Rounds 2 and 3.

estimate as a commitment device. Note, however, that the underestimation diminishes in the following two rounds, in which the mean relative estimation errors are close to zero. Using the Fisher's exact test, we find that the proportions of underestimation and overestimation differ across our three treatments, with the differences being present in 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 anchored estimates.

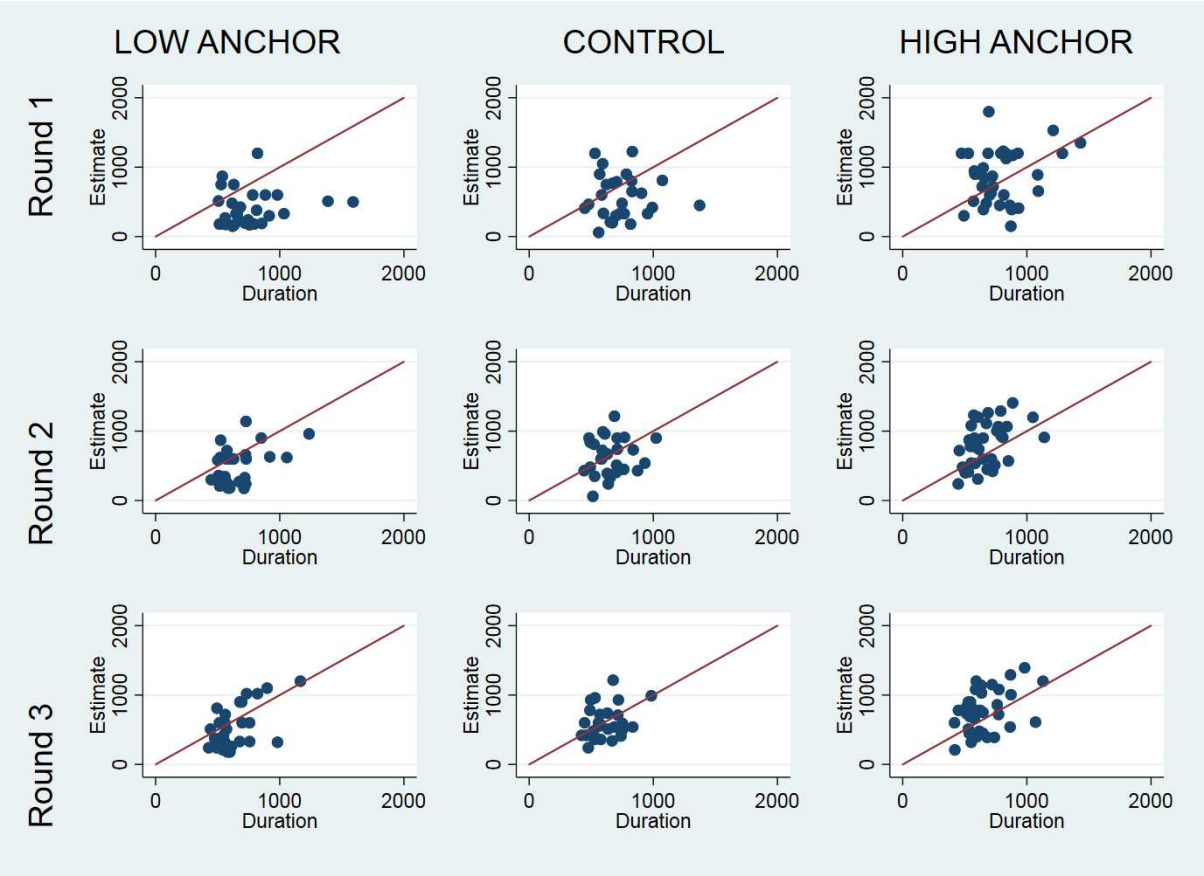
Figure 3: Mean anchoring bias across treatments



Note: Figure 3 displays bar charts of the mean estimation bias by treatments and rounds. Zero on the vertical axis indicates unbiased estimation. Downward sloping bars display underestimation, upward sloping bars overestimation.

We analyze the estimation bias also by comparing each individual's estimates with the actual duration using the Wilcoxon matched-pairs signed-rank test (see Table 1 for p-values). In the Low Anchor treatment, we find the estimates to be significantly lower than the actual duration in all three rounds. In the High Anchor treatment, we find that the estimates in Round 1 do not differ from the actual duration; however, the estimates become significantly higher than the actual duration in the following rounds. The divergence between the estimates and actual task duration is driven by improved performance between rounds, which is not accompanied by adequate adjustment of estimates. In the Control treatment, we find no differences between the estimates and the actual task duration in Rounds 2 and 3.

Figure 4: Within-treatment and within-round comparison of estimates and actual duration

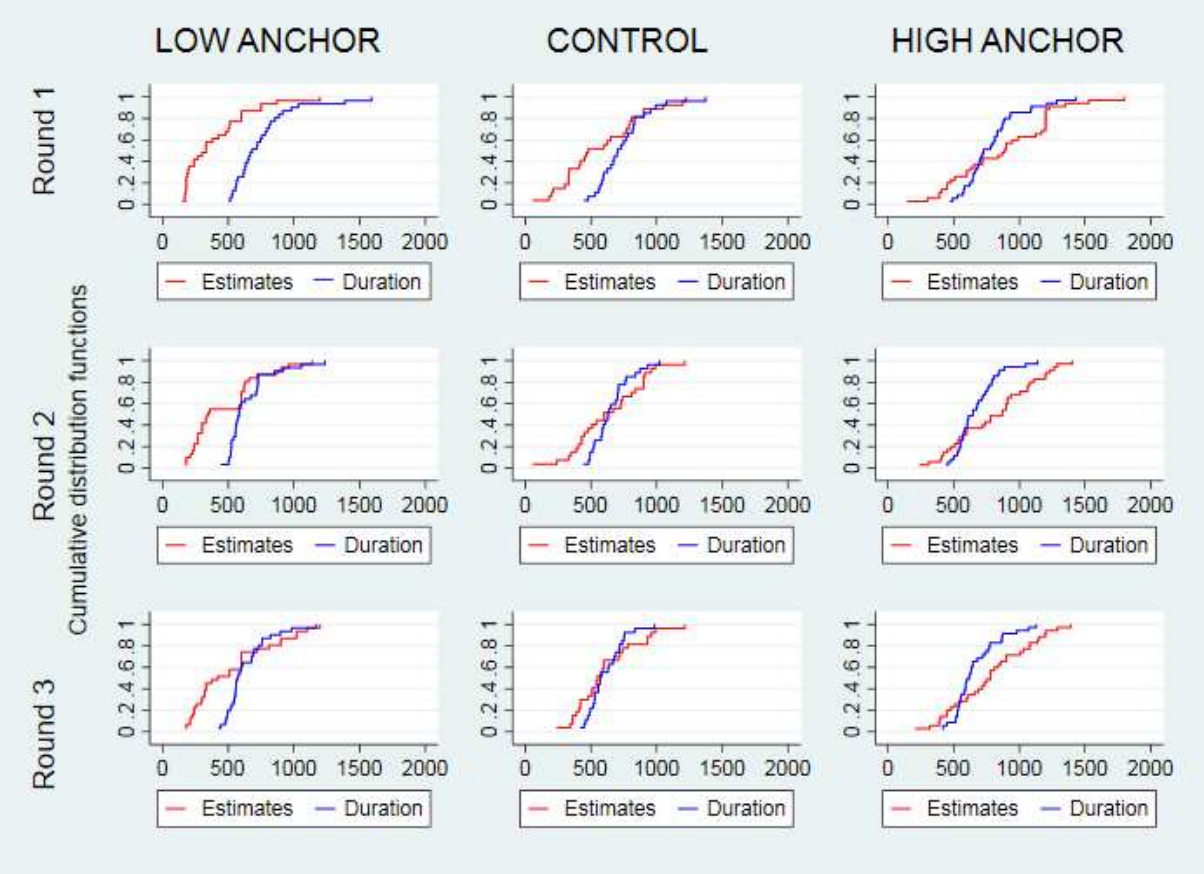


Note: Figure 4 displays scatter plots of individual-level estimates (vertical axis) and actual duration (horizontal axis) by treatment and round. Precise estimates are on the red 45-degree line. A dot above the red line indicates overestimation, while a dot below the red line indicates underestimation.

Overall, our results show that the estimates exhibit the smallest bias in the Control treatment. As a robustness check, we test the differences in distributions between the estimates and the actual duration using the Kolmogorov–Smirnov test (the p-values are presented in Table 1 and the cumulative distribution functions in Figure 5). 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 low anchor causes systematic underestimation, while the high anchor causes systematic overestimation of the task duration. The estimates are less biased when anchors are not present.

Figure 5: Cumulative distribution functions of estimates and actual duration by treatment and round



Note: Figure 5 displays cumulative distribution functions of estimates (in red) and actual duration (in blue) by treatments and rounds. A red line above a blue line indicates underestimation, while a red line below a blue line indicates overestimation. An overlap represents unbiased estimation.

6. Auxiliary results

a. Retrospective estimates

We further investigate the persistence of the anchoring effect by examining subjects' retrospective estimates. Upon completion of Round 3, we asked subjects to retrospectively estimate how long the last round took them to complete. The summary statistics of retrospective estimates are presented in Table 3 while Figure 6 displays the data graphically.

Recall that our experimental task requires a significant cognitive attention, making it relatively hard to mentally keep track of time and precisely construct the task duration in retrospect. We conjectured that without any feedback on the estimation accuracy or the actual task duration, the anchors would

influence the retrospective duration estimates in the same direction as the estimates produced before the tasks were performed.

We find significant differences in the retrospective estimates between the treatments (the Mann-Whitney test p-value is <0.001 for Low Anchor vs. High Anchor; 0.003 for Control vs. Low Anchor; and 0.024 for Control vs. High Anchor) that provide support for this conjecture.[12] 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 differ significantly. We conclude that anchors distort not only the estimation before the task, but also the retrospective duration estimation, which in turn can influence future estimation, creating a persistent anchoring effect.

Result 4: The anchors influence the retrospective duration estimates in the same direction as the duration estimates produced prior to the task.

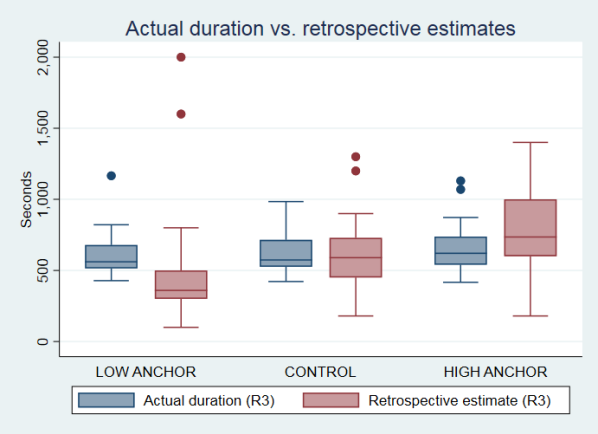
Table 3: Summary statistics of retrospective estimates in Round 3

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

Note: Table 3 presents summary statistics of the task duration in Round 3, retrospective estimates of that duration, and the corresponding bias and absolute errors, by treatments. a: The bias is calculated as a relative estimation error (Estimate – Actual duration).

[12] We dropped 10 extremely low estimates (from 6 to 30 seconds) from the analysis of retrospective estimates, reducing the total number of observations to 83 (27 in the Low Anchor treatment, 26 in the Control treatment, and 30 in the High Anchor treatment). We suspect that these outliers (4 in the Low Anchor treatment, 1 in the Control treatment, and 5 in the High Anchor treatment) were caused by the lack of attention. We asked for the retrospective estimate in the seconds format, which might have been overlooked by these 10 subjects (the estimates prior to the task were elicited in the minutes and seconds format). Also, the accuracy of retrospective estimates was not financially incentivized. However, since the outliers were almost equally distributed amongst the Low Anchor and the High Anchor treatments, removing the 10 observations does not alter the treatment comparison result.

Figure 6: The actual task duration and its retrospective estimates in Round 3



Note: Figure 6 displays box plots of the actual task duration in Round 3 and the retrospective estimates of that duration, by treatments.

b. Estimation accuracy

To ensure an effective allocation of resources, companies require not only that the project duration estimates be unbiased (not systematically optimistic or pessimistic), but also accurate. Therefore, complementary to the estimation bias, we also analyze the estimation (in)accuracy, measured by absolute estimation errors, without the sign of the error being taken into account. The data pertinent to estimation (in)accuracy are summarized in Table 4 and graphically displayed in Figure 7a. We find relatively large mean estimation errors, ranging from 40% to 56% in the anchor treatments and from 30% to 45% in the Control treatment.

One possible way to improve estimation accuracy is to utilize historical information in the planning process, as suggested by Kahneman & Tversky (1979). To test the efficiency of such approach, we construct a simple tool that calculates the mean actual task duration of all subjects in the last round and uses this calculated average as a predicted duration estimate for all subjects in the following round. We compare the accuracy of such tool with the accuracy of subjects' own estimates, for both the mean absolute estimation errors and the proportion in which the tool outperforms the subjects' estimates. The data are summarized in Table 4 and graphically displayed in Figure 7b.

Table 4: Estimation accuracy

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)
Tool: 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%
Tool: Mean absolute error (%)	-	30%	21%	-	26%	21%	-	25%	20%
Subjects' estimate is better	-	35%	29%	-	33%	33%	-	31%	23%
Tool's estimate is better	-	65%	71%	-	67%	67%	-	69%	77%

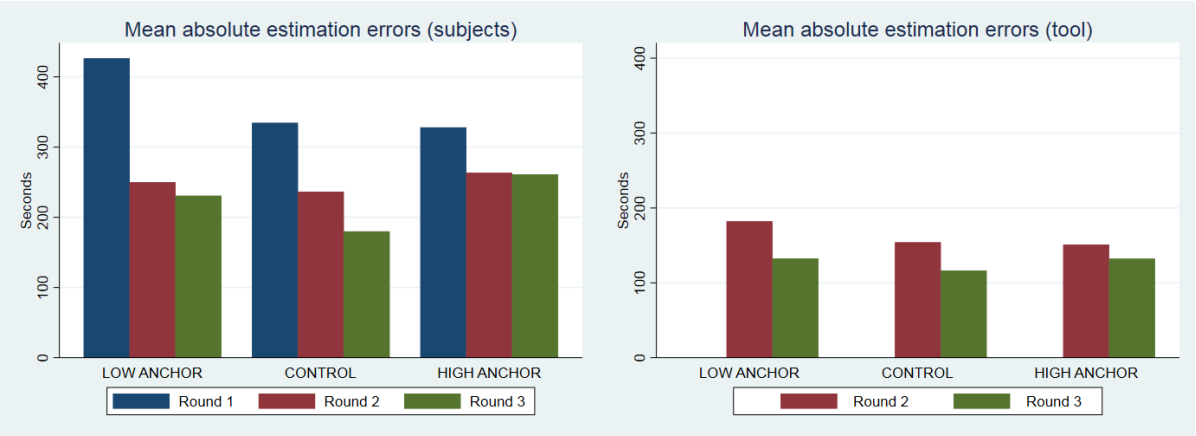
Note: Summary statistics of duration estimates of both subjects and our prediction tool; and the associated absolute estimation errors, by treatments and experimental rounds. The table lists also the proprotions of cases in which subjects are more accurate than the tool and vice versa.

The absolute estimation errors resulting from the use of the tool are on average 13.5 percentage points lower than the subjects' errors; and the tool is 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' mean 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 (both Mann-Whitney test p-values are <0.001).

We find that our tool outperforms subjects' judgment not only in the anchoring 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 (the Mann-Whitney test p-value is 0.048 in Round 2 and 0.149 in Round 3). Our results are consistent with evidence found in Lorko et al., (2019) that consulting historical information regarding the average task/project completion time can lead to improvements in estimation accuracy.

Result 5: A simple prediction tool based on historical averages outperforms subjects' own estimates in terms of estimation accuracy.

Figure 7: Absolute estimation errors



Note: Figure 7a (left panel) displays bar charts of the mean absolute estimation errors for subjects. Figure 7b (right panel) displays bar charts of the mean absolute estimation errors for our prediction tool. Both figures display the data by treatments and by experimental rounds, measured in seconds. 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

The estimates in Round 2 and Round 3 of the Control treatment do not exhibit any systematic bias, the mean relative estimation error is almost zero. However, the mean estimation inaccuracy measured by absolute estimation errors (presented in Table 4 and displayed graphically in Figure 7a) in the Control treatment is still relatively high. What is causing the co-occurrence of a low bias with high estimation inaccuracy?

The low bias is driven by a relatively similar number of subjects who underestimated the duration compared to those who overestimated it. The 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 at the individual level are often similar in the following rounds. We therefore test whether there exists a "self-anchoring" effect, meaning whether subjects in the Control treatment could have become anchored on their own duration estimates produced in Round 1.

For this analysis, we split the subjects who participated in the Control treatment into two subgroups: those who in Round 1 provided an estimate lower than the median and those who provided an estimate higher than the median. The "Low group" consists of 14 subjects with estimates from 60 to 480 seconds while the "High group" consists of 13 subjects with estimates from 600 to 1225 seconds. We then compare these two subgroups with regard to both estimates and actual task duration (see Table 5 for summary statistics and Figures 8a and 8b for graphical display of the data).

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 evidence of the "self-anchoring" effect in the retrospective estimates.

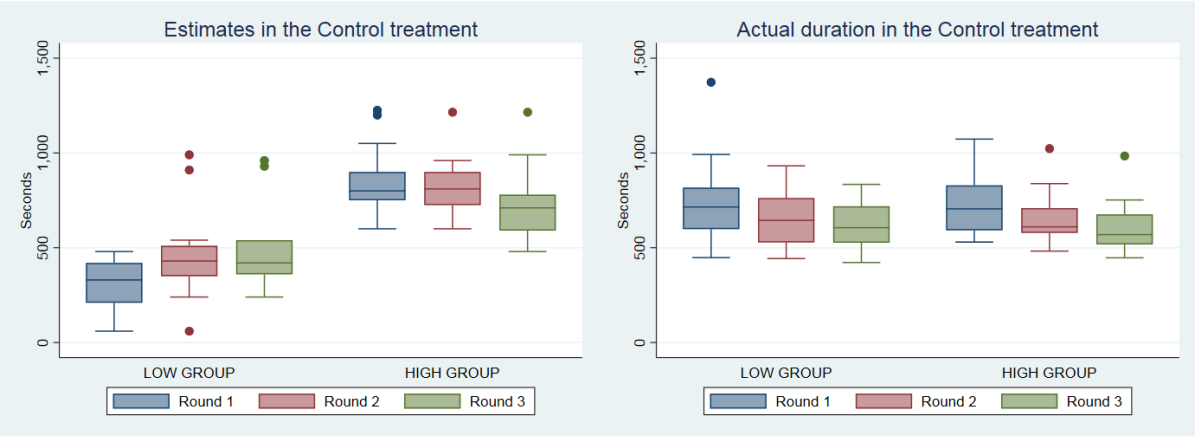
Result 6: Subjects in the Control treatment are anchored on their own initial duration estimates.

Table 5: **Self-anchoring in the Control treatment.**

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

Notes: The table presents the median estimates and the actual duration obtained from the two subgroups in the Control treatment, along with the Mann-Whitney test p-values, by experimental round. The subgroups are divided by the median split of the first estimate. a: one retrospective estimate was dropped; see footnote 12.

Figure 8: Estimates and actual task duration in the Control treatment based on the median split



Notes: Figure 8a (left panel) displays box plots of estimates for the High and Low groups in the Control Treatment. Figure 8b (right panel) displays box plots of the actual duration for the High and Low groups in the Control Treatment. The subgroups are divided by the median split of the first estimate. Both figures display the data by experimental rounds, measured in seconds.

d. Robustness

Finally, we conduct a regression analysis to check for possible factors that might have affected our results, such as the number of incorrect answers in the experimental task, time spent on the anchoring question, time spent on the estimation, demographics (gender, education, ethnicity, grade average and employment status), and risk attitudes. We do not find significant effects of any of these factors. Also, the results of the treatment effects are robust to using parametric instead of non-parametric tests. The details are available upon request. A couple of interesting findings from these robustness analyses is that there is a negative correlation between task enjoyment, elicited in the questionnaire, and the actual task duration as well as a positive correlation between 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 to be the most important part of the project lifecycle. It is because the success of the project ultimately relies on the accuracy of estimates in the project plan. Arguably, the planning phase can also be the most difficult one, as crucial estimates are often requested without supplying enough information about the exact project scope. In addition, it is quite common that project managers do not have the entire project team assembled and available yet. As a result, the planning is executed under a high level of both uncertainty and ambiguity, and the estimates are vulnerable to biasing effects in a variety of forms. In this paper, we test one such biasing mechanism – the effect of anchors that can occur as a consequence of uninformed suggestions or expectations of the project duration coming from the project stakeholders. We conjectured that anchors could cause a large and predictable estimation bias. The resulting biased estimates can be harmful. While very low estimates almost inevitably lead to project failure, very high estimates contribute to inefficient allocation and utilization of company resources.

Projects are often complex, carried out by multiple teams and can last for years. Individuals responsible for project planning are not necessarily involved in the execution of project activities and are often allocated to a different project before the ongoing one is completed. Moreover, schedule delays are frequently attributed to causes other than misestimation such as external circumstances (e.g., weather) or unforeseen risks. The feedback loops are therefore often imperfect and imprecise,

with the planning teams not receiving enough details regarding the actual duration of project tasks. We thus also conjectured that the effect of being exposed to an anchor can persist over time and potentially carry over to a similar estimation in the future, if the planner is not relevantly informed about his error. Our experimental findings support our conjectures. Despite employing a familiar task and creating favorable conditions for accurate duration estimation, including financial incentives for estimation accuracy and repeated experience with estimation and task performance, we observe that the introduction of anchors causes a systematic estimation bias that persists over time.

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 being exposed to an anchor and the estimation (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 earlier findings, together with our evidence of strong and robust anchoring effects observed in duration estimation of a familiar and repeated task, 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 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 related to potential social comparison or reference point confounds, we do not use any mechanisms that would discredit the anchors.

The reason we present the anchors in a non-arbitrary fashion, is to mimic the project management environment. Our research question focuses on anchors that are generated by project stakeholders and therefore, are not random. From this perspective, our experiment is better thought of as an applied anchoring research. It 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 the selective accessibility theory (Chapman & Johnson, 1999; Mussweiler & Strack, 1999).

We suppose that the values 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 the analysis of 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 itself and adjust the estimates away from the anchor, under a high cognitive load.

For example, one subject mentions that "*[t]he 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."* The statement is consistent with the 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 subject's actual performance time was approximately 600 seconds on average, and in each round the subject was roughly 20 seconds faster than in the previous one, with a similar number of incorrect answers. This shows that the subject placed a disproportionally larger weight on the anchor value (3 minutes) than on own experience (remembered duration) in the estimation process. 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" is the value of the anchor (or his own initial estimate) and can be updated by using the remembered task duration that tends to be noisy. We find that even though subjects are usually able to detect the direction of their mistakes, they underestimate the magnitude of those mistakes, which causes a relatively slow convergence of estimates towards the actual task duration. Such insufficient adjustment enables the anchoring effect to persist over 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. In a nutshell, our results suggest that the task experience alone is not a sufficient remedy to eliminate the anchoring effect in estimating or planning.

How to mitigate the anchoring bias in repeated estimation and stop echoing planning mistakes from the past? It seems reasonable to conjecture that a proper feedback regarding past results is crucial. We did not implement feedback in the current experiment because we believed that having subjects learn how long it actually took them to complete the task would likely cause the anchoring effect to disappear. As a result, such design would not allow us to learn about the persistence of the anchoring

effect. Nevertheless, we admit that while the conjecture of using feedback as a remedy appears to be intuitive, it needs to be properly tested.

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 planners to provide better estimates and decision-makers to make better decisions. Recall that our estimates from the Control treatment show a noticeable sign of an optimism bias in Round 1. After subjects gain experience with the task, the estimates become unbiased, which is an artefact of the independence of estimation errors and approximately equal number of underestimates and overestimates. The average relative estimation error of almost zero resembles a phenomenon known as the "wisdom of the crowd" (Galton, 1907). Despite no bias at the aggregate level, we observe high estimation inaccuracy in our data, which is akin to the assumption of the "bias-variance trade-off" (Geman, Bienenstock, & Doursat, 1992). The bias-variance trade-off principle states that 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 mean task duration from the past significantly outperforms individual estimates of our subjects. The future research could thus verify whether the use of "helpful 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 who produce lower estimates in Round 1 keeps the estimates significantly lower also in following rounds in comparison with the group of subjects who produce higher estimates in Round 1, although the actual performance does not differ between the two groups. 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 avenue 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 editor Daniela Puzzello, an anonymous reviewer, Alex Alekseev, Cary Deck, Michal Ďuriník, Romain Gauriot, Ian Krajbich, 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,

2018 Asia-Pacific ESA Meeting, and 2018 North-American ESA Meeting who provided helpful comments and suggestions. Financial support was provided by Macquarie Graduate School of Management. Maroš Servátka thanks University of Alaska – Anchorage for their kind hospitality while working on this paper.

References

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

Buehler, R., Griffin, D., & MacDonald, H. (1997). The Role of Motivated Reasoning in Optimistic Time Predictions. *Personality and Social Psychology Bulletin*, *23*(3), 238–247. https://doi.org/10.1177/0146167297233003

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

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the Irrelevant: Anchors in Judgments of Belief and Value. In *Heuristics and Biases*. https://doi.org/10.1017/cbo9780511808098.008

Cox, J. C., & Sadiraj, V. (2018). Incentives. In A. Schram & A. Ule (Eds.), *Handbook of Research Methods and Applications in Experimental Economics*. Edward Elgar Publishing Ltd.

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

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), 391–396. https://doi.org/10.1111/1467-9280.00372

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

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

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

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

Kahneman, D. (1992). Reference points, anchors, norms, and mixed feelings. *Organizational Behavior and Human Decision Processes*. https://doi.org/10.1016/0749-5978(92)90015-Y

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

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

Lorko, M., Servátka, M., & Zhang, L. (2019). *The Effect of Project Specification and Historical Information on Duration Estimates*. Working paper.

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

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*.

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

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

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

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

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.