How to Improve the Accuracy of Project Schedules? The Effect of Project Specification and Historical Information on Duration Estimates

Lorko, Matej and Servátka, Maroš and Zhang, Le

MGSM Experimental Economics Laboratory, Macquarie Graduate School of Management, Sydney, Australia, MGSM Experimental Economics Laboratory, Macquarie Graduate School of Management, Sydney, Australia and University of Economics in Bratislava, Slovakia, MGSM Experimental Economics Laboratory, Macquarie Graduate School of Management, Sydney, Australia

15 August 2019

How to Improve the Accuracy of Project Schedules? The Effect of Project Specification and Historical Information on Duration Estimates

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 15, 2019

Abstract: The success of a business project often depends on the accuracy of project estimates. Inaccurate, often overoptimistic schedules can lead to significant project failures. In this paper, we experimentally investigate the effectiveness of two interventions designed to mitigate the pervasive underestimation bias and improve the accuracy of project duration estimates: (1) increasing the quantity of available information prior to estimation by providing historical information regarding the average duration of similar projects in the past and (2) increasing the quality of available information prior to estimation by providing a more detailed project specification. In addition, we also test whether it is more effective to provide historical information together with the project specification or only after the initial beliefs regarding the project duration are formed. We find that increasing both the quantity and quality of project relevant information successfully mitigates the underestimation bias. However, only the provision of historical information is also associated with significant improvement in absolute estimation accuracy. The timing at which such information is disclosed to planners does not seem to influence the effectiveness of the intervention. We also find that subjective confidence in the accuracy of duration estimates does not vary across experimental treatments, suggesting that the confidence in estimates is neither a function of the quantity nor the quality of available information prior to estimation.

Keywords: project management, project planning, time management, duration estimation, historical information, project specification

JEL codes: C91, D83, O21, O22

1. Introduction

A commonality of virtually all business projects is the uncertainty regarding the amount of time and resources needed to deliver requested project outcomes. Proficient planning processes capable of producing adequate project plans are essential for executing a cost-benefit analysis and deciding which projects to initiate. Once a project is underway, a realistic project schedule is a crucial determinant of project success, ensuring effective allocation and utilization of company resources. Accurate project duration estimates are especially important when managing a project portfolio, in which individual projects compete for temporary use of scarce resources. A delay in one project can slow down the progress of other projects within a portfolio that run in parallel and/or sequentially, resulting in increased costs and lower efficiency.

The estimation of project duration appears to be a challenging undertaking, as suggested by the recent global project performance report (Project Management Institute, 2019). The report concludes that approximately 50 percent of business projects fail to be delivered within the original estimated schedule. The high failure rate begs a question of how the accuracy of project duration estimates can be increased. In the current paper, we experimentally test the effectiveness of two interventions advocated by project management methodologies, namely providing a more detailed project specification and disclosing historical information regarding the average duration of similar projects in the past. We also examine the effect of timing at which the historical information is disclosed.

Traditionally, a thorough project specification is perceived as a crucial determinant of estimation accuracy (Project Management Institute, 2013). In this regard, project managers often go to great lengths to equip their planners with as detailed as possible descriptions of project tasks. Arguably, no specification is extensive enough to capture every aspect of the requested outcomes, which is especially true at early stages of a project.[1] Nonetheless, project planners intuitively focus only on the project specification at hand, often failing to realize that it might be incomplete. Because of neglecting the unspecified (or unknown) details, project duration estimates may become understated. [2]

[1] Customer requirements are often not yet developed to the full extent early in the project. As an illustrative example, consider a customer of a software development project who approaches the developer with a description of only the core features of the application, without elaborating in detail on smaller supporting functions that integrate the core features and make the application more ergonomic and user-friendly.

[2] Although the current paper focuses on underestimation caused by incomplete project specification, it is important to keep in mind that there are multiple other factors contributing to inaccurate project duration

Kahneman (2011) refers to the phenomenon of paying attention only to the information provided while effectively ignoring the missing links as the "what you see is all there is" rule.

Kahneman & Lovallo (1993) and Kahneman & Tversky (1979) suggest that the accuracy of project duration estimates can be improved by consulting historical information (also referred to as reference class information or distributional information), i.e., the actual duration of similar projects in the past. The main advantage of utilizing such information in the planning process is that it naturally encompasses the impact of incomplete and unforeseen specifications on project execution. Instead of (or complementary to) estimating how long it takes to complete each requested project deliverable based on its specifications, a company can estimate the duration of completing the whole project based on how long it took similar projects to complete in the past. The approach, labeled as the "analogous estimating technique," is also endorsed by project management methodologies (IPMA, 2015; Project Management Institute, 2013). However, the methodologies suggest consulting the duration or costs of previous projects only in the absence of detailed information regarding the current project. Advocating for the use of this technique more broadly, Flyvbjerg (2006) argues that even though historical information may fail to predict extreme project outcomes, its utilization in the planning process commonly induces more accurate estimates in comparison with a more conventional planning based on project specification, which he describes as "the road to inaccuracy".

Although the practicality of historical information is recognized in project management methodologies, to the best our knowledge, the effect of its utilization in the planning process has not yet been tested in a controlled environment and with real incentives. Lorko, Servátka, & Zhang (2019) provide preliminary evidence that planners could benefit from considering past project duration in the planning process. In an environment where subjects estimate how long it will take them to complete a simple real-effort task, the authors show that over two thirds of subjects would be better off in terms of estimation accuracy if the historical average was used for estimation purposes instead of their own estimate. In the current paper, we study the impact of historical information on estimation accuracy directly, and compare its effectiveness with the impact of providing a more detailed task description. Our experimental design controls for confounding factors such as quality of project deliverables, project costs, risks and unforeseen events, all of which may interfere with the project progress and affect the estimation accuracy in business practice. In all treatments (described in detail in sections 3 and 4), subjects read the description of a real-effort task they are about to

estimates, e.g., optimism bias, strategic misrepresentation, competence signaling, using deadlines as commitment devices or anchoring effects.

perform, estimate the time it will take them to complete the task, indicate their subjective confidence in the accuracy of the estimate, and then execute the task. Subjects are financially incentivized for both their estimation accuracy and performance on the task.

In Experiment 1, we test whether providing historical information (operationalized in our design as the average task duration in the past) in the estimation process can mitigate the estimation bias and increase the accuracy of project duration estimates. Additionally, we explore whether disclosing historical information to planners only after the initial estimate has already been made, is more effective than making it available alongside the project specification. Providing historical information after the initial estimation avoids inducing other biasing mechanisms, such as anchoring (Tversky & Kahneman, 1974).

In Experiment 2, we test whether a more detailed project specification (operationalized in our design as additional relevant information in the task description) also mitigates the estimation bias and improves the estimation accuracy. By linking Experiment 1 and Experiment 2 together through a common baseline treatment, we are able to compare the relative effectiveness of providing historical information with providing a more detailed task description in the duration estimation process. In both experiments, we deliberately provide only a single piece of additional information. The design is geared towards eliciting the lower bound of the two effects, allowing for their direct comparison. We conjecture that estimates incorporating historical information outperform, in terms of their accuracy, not only estimates based on crude (incomplete) specifications, but also estimates based on detailed specifications.

In both experiments, we also examine whether the quantity and quality of available information reflects on subjective confidence in estimates. Although intuitively one might expect to find a positive correlation, according to Kahneman's (2011) “what you see is all there is” rule, planners neglect the missing elements in project specifications. As a result, they may not be able to differentiate between various degrees of ambiguity embedded in different specifications of the same project. Planners equipped with less information or less detailed project specifications can thus produce less accurate, but not necessarily less confident duration estimates.

Our results support the conjecture that disclosure of historical information can significantly mitigate the estimation bias and improve the estimation accuracy, regardless of whether the information is provided together with the task description or after the initial duration estimate. We also find that

although a more detailed task description decreases the frequency as well as the extent of duration underestimation, it also induces a larger variance in individual estimates. The estimates are on average unbiased, but the absolute estimation accuracy is not improved. Finally, in line with "what you see is all there is" rule (Kahneman, 2011), we find that subjective confidence in estimates is similar across all treatments and thus, is not a function of the quantity (Experiment 1) or the quality (Experiment 2) of available information. Subjects do not account for the possibility that they might be missing critical details for accurate estimation and exhibit high confidence in their estimates regardless of what they know about the task.

Our study provides the following important implications for project management practice. First, if data from a meaningful reference class of past projects is available, project managers should consider "anchoring" their planners' estimates on the class average. Providing historical information to planners appears to be more effective than equipping them with overly detailed project specification. Second, project managers can expect initial resistance to use historical information, because planners may not realize its usefulness before they actually experience its benefits. Third, confidence in estimates is not a reliable predictor of estimation accuracy and project managers should be cautious when making decisions based on the planner's confidence in the proposed project schedule.

2. Relationship to the literature

Although both underestimated and overestimated project schedules imply negative consequences for project stakeholders, businesses appear to perceive underestimation to be a more serious issue. The overwhelming focus on underestimation might be driven by the asymmetry of consequences. Direct costs stemming from underestimation are more salient than opportunity costs of resource underutilization arising from overestimation. Moreover, if members of a project team identify instances of overestimation in the project, they can strategically "waste" allocated time and utilize other resources anyway, so the estimation error may go unnoticed.[3]

In academic research, underestimation has also attracted more attention than overestimation. Kahneman & Tversky (1979) coin the term "planning fallacy," which describes a tendency to make overoptimistic plans and forecasts that are close to the best-case scenarios. A symptom of the

[3] "Wasting" time on the job has been anecdotally summarized as the Parkinson's Law, stating that "work expands so as to fill the time available for its completion" (Parkinson, 1955).

planning fallacy is ignoring evidence from past projects that took significantly longer to complete. The underestimation of required resources is pervasive in public works (Engerman & Sokoloff, 2006; Flyvbjerg, Holm, & Buhl, 2002) and also in business projects. Recent project management performance statistics show that a large number of projects is not delivered within the planned schedule, or not finished at all even in companies with extensive history of project management practice (Project Management Institute, 2019), signifying the existence of substantial inefficiencies. The propensity to underestimate the duration is, however, not restricted to large initiatives. Planning fallacy can also be tracked at the level of casual activities such as student predictions of tutorial session completion (Buehler, Griffin, & Ross, 1994) or tax file returns completion date (Buehler, Griffin, & MacDonald, 1997).

Misestimation can often be attributed to strategic incentives, for example, gathering political support for the proposed project (Flyvbjerg, 2008). However, a review of psychological studies by Buehler, Griffin, & Peetz (2010) as well as a comprehensive review of empirical duration estimation studies, laboratory and field experiments by Halkjelsvik & Jørgensen (2012) reveal a frequent tendency to underestimate the duration even if there are little to no incentives to manipulate the forecasts. From this perspective, the planning fallacy can be considered an instance of a general optimism bias (Lovallo & Kahneman, 2003).

The current paper explores the effectiveness of mitigating underestimation of project duration by utilizing historical information, a concept originally introduced by Kahneman & Tversky (1979).[4] Kahneman and Tversky suggest that estimation accuracy could be significantly improved by taking into consideration the actual duration of similar projects that have already been completed, and offer a five-step corrective procedure for generating regressive estimates.[5] They propose that planners first select a meaningful reference class for their forecast and then assess the distribution of outcomes, in particular, the average of the reference class. These two steps should be followed by intuitive estimation of the problem at hand and assessment of predictability, i.e., a degree to which the available historical information permits accurate estimation. The final step of the procedure is

[4] Since the concept is based on a statistical regression towards the mean (Nesselroade, Stigler, & Baltes, 1980), it applies to not only underestimation, but also overestimation of necessary project resources, including time.

[5] Flyvbjerg, Skamris Holm, & Buhl (2005) later shorten the procedure to three steps, name it "Reference Class Forecasting" and offer it as an effective tool to mitigate inaccurate demand and cost forecasts in public works. Reference class forecasting was soon endorsed by American Planning Association which encouraged planners to use the procedure in addition to traditional estimating methods (Flyvbjerg, 2008).

correcting the intuitive estimate by adjusting it towards the reference class average. While intuitive, the procedure for producing regressive estimates has not received much empirical attention and testing.

Building on the procedure, Kahneman & Lovallo (1993) distinguish between an "inside view" and an "outside view." The inside view represents an estimation based solely on considerations of project specification and possible risks (a forward-looking strategy) while the outside view represents an estimation based on statistics of similar projects from the past (a looking-back strategy).[6] Kahneman & Lovallo (1993) and Lovallo & Kahneman (2003) advocate for a broader use of the outside view because the history ultimately carries the consequences of a variety of small obstacles (such as omissions in project specification, misunderstandings of requirements, or unforeseen events) on the project performance. Since such obstacles are usually hard to foresee and account for during the project planning process, the estimates produced via inside view are likely to be overly optimistic. If enough historical information is available, the outside view potentially yields more realistic project estimates.

The distinction between the inside and the outside view is, however, not sharply delineated. It is because the inside view, represented by an expert judgment, often facilitates implicit duration standards or experiences from the past, e.g., how long it usually takes to develop a basic software feature, or how many lines of code a developer usually produces within a day. Although planners often declare their estimates as "gut feelings" or "intuition", the judgement is in fact a reflection of their prior knowledge and experience within their expertise (Klein, 1999; Rush & Roy, 2001). Nevertheless, implicit experiences can be susceptible to biases. For example, individuals often remember the duration of previous activities incorrectly (Roy, Christenfeld, & McKenzie, 2005) and without proper feedback on their estimation accuracy, they can become anchored on their own former estimates (Lorko et al., 2019).

In relation to the outside view, Lovallo & Kahneman (2003, p. 61) argue that "the thought of going out and gathering simple statistics about related projects seldom enters a manager's mind." Even when the outside view is more salient as relevant historical information is readily available and easily

[6] For a more comprehensive inside-outside framework incorporating other aspects that influence duration estimates, see Buehler et al. (2010).

accessible, planners typically display a strong tendency to consider the current case as unique and focus only on the details of a specific project at hand (Buehler & Griffin, 2015; Kahneman & Lovallo, 1993; Kahneman & Tversky, 1979). Planners thus intuitively adopt the inside view and effectively neglect historical information when estimating project duration. This proposition is supported by the results of think-aloud procedures (Buehler, Griffin, & Ross, 1994) showing that only a miniscule fraction of participants considered past problems and past successes before making duration estimates of the current task. In addition, Buehler, Griffin, & Ross (1994) find that even when participants are led to consider past experiences, they do not use them to regress their estimates unless they are explicitly instructed to do so.

What factors contribute to planners utilizing historical information? In an applied setting of software development effort estimation, Jørgensen (2010) finds that an increased project ambiguity may drive planners towards paying more attention to historical information. However, the same author claims that planners usually opt to use analogies from the past only if they are "very similar" to the current project (Jørgensen, 2004). The observed reluctance to seek and apply historical information in project planning may be driven by unjustified confidence in intuitive predictions generated by expert judgment. Kahneman & Tversky (1979) claim that intuitive predictions are frequently characterized by overconfidence, which is often caused by putting more weight on the consistency and less weight on the reliability of available information.

Empirical evidence demonstrating the benefits of using historical information for duration estimation is scarce. Two notable studies on the topic include a field experiment focusing on casual daily activities (Roy, Mitten, & Christenfeld, 2008; Experiment 3) and a framed classroom experiment concerning software development effort estimation (Shmueli, Pliskin, & Fink, 2016). Both studies report increased estimation accuracy when the reference class averages are supplied. However, the former one utilizes tasks the duration of which is often beyond the participants' control and the latter one uses only predicted instead of actual accuracy, since the tasks are not performed after the estimation. Also, subjects in neither of the studies are incentivized, possibly resulting in the hypothetical bias (Hertwig & Ortmann, 2001). Moreover, in both studies, historical information is given to participants together with the task description. Under such circumstances, it is impossible to distinguish whether the differences in estimates across treatments are subject to the anchoring effect (König, 2005; Lorko et al., 2019; Thomas & Handley, 2008) or whether the adjustment (regression to the reference class average) of the initial intuitive estimate actually took place.

To summarize, despite the prevalence of inaccurate project estimates in the business world, the research to date has not shed much light on the effectiveness of the correction procedure of regressing the predictions towards the reference class average (Kahneman & Tversky, 1979) and on its adoption by project planners. In this paper, we present the results of a controlled incentivized experiment in which the reference class is a group of subjects from the baseline treatment. We calculate the average actual task duration in our reference class and then provide this average to subjects in the following treatments as historical information. We investigate whether they use this information to improve their estimation accuracy. Unlike the previous studies, the individual estimating the duration of the task is also the one who completes the task. This feature allows us to recreate incentives faced within a project. In addition, by a careful manipulation of the timing when the historical information is provided during the estimation process, we separate the anchoring effect from the regression effect. Furthermore, we examine whether the accuracy can be enhanced by a more traditional and perhaps also more natural approach, which is by providing more detailed project specification. In fact, such conjecture is seemingly so obvious that we are not aware of any study in the area of duration estimation that tests the effect of providing more detailed specification, let alone compares its effectiveness against other interventions.

3. Experiment 1: Historical information

In Experiment 1, we test whether disclosing historical information in the estimation process can induce more accurate and less biased (understated) duration estimates. We measure the estimation (in)accuracy as an absolute value of estimation error (i.e., |estimate – actual duration|), while we measure the estimation bias as a relative (signed) estimation error.

Treatments

Experiment 1 consists of three treatments, implemented in an across-subject design: the Baseline treatment, the Information Before Estimate (henceforth "Info-Before") treatment, and the Information After Estimate (henceforth "Info-After") treatment. In the Baseline treatment, no historical information is provided. Subjects read the instructions with the description of the experimental task and then estimate how long it will take them to complete it. Subsequently they indicate their subjective confidence in the accuracy of the estimate on a Likert scale and execute the task. Upon completion of the task, subjects answer a few questions about the experiment, and complete an incentivized risk attitude assessment (Holt & Laury, 2002) as well as a demographic questionnaire.

The Info-Before treatment and Info-After treatment utilize the same experimental task and follow the same experimental procedure as the Baseline treatment. However, in addition to the task description, subjects in these two treatments also receive information about the average actual task duration recorded in the Baseline treatment. We experimentally manipulate the timing when such information is provided. In the Info-Before treatment, the information is displayed on the screen right after subjects finish reading the instructions. Subjects then provide their estimates, indicate their confidence in the estimate and execute the task, just as in the Baseline treatment. In contrast, in the Info-After treatment, subjects do not receive the historical information until they have provided their estimate and confidence rating. At no stage they are instructed that they will receive such information. Once the historical information is disclosed, subjects in the Info-After treatment are given an opportunity to revise their estimates as well as their confidence rating. To calculate earnings, we use the revised estimate in the Info-After treatment, as explained in the on-screen instructions.

The experimental task

We employ a modified version of individual search task introduced by Mazar, Amir, & Ariely (2008), in which subjects are asked to find two numbers that add up to a target sum of 10 in matrices containing decimal numbers. Each matrix has only one correct answer. Instead of the original twelve numbers within each matrix, we use sixteen numbers, making the task more difficult and taking longer to complete. For the same reason, we make the target sum to be 100 as opposed to 10. A sample matrix is shown in the appendix.

According to Mazar et al., (2008, p. 636), subjects "did not view this task as one that reflected their math ability or intelligence". Thus, the performance in the task is not confounded by prior knowledge, as people are generally skilled in adding numbers and there is little room for learning. Subjects first estimate the total time (in minutes and seconds) it will take them to find correct answers for all 10 matrices together, before they search through the matrices one by one. The instructions describe the task as follows:

You will be shown 10 matrices one by one. Each matrix contains 16 numbers. Two of those numbers add up to exactly 100. You will have to identify those two numbers. You will move on to the next matrix only after you submit the correct answer.

In the task description, we intentionally omit the information that numbers in matrices are decimal. Since people do not usually think of decimals when being confronted with the word "number", such

omission in the specification makes the task look easier than it really is, creating a discrepancy between the intuitive estimate and the actual task duration. The discrepancy provides an adequately calibrated environment that is crucial for testing the effectiveness of factors capable of mitigating the estimation bias and improving the accuracy of duration estimates.

Historical information

In the Info-Before and Info-After treatments, we present the historical information to subjects as follows:

Please consider the following additional information. This task was already performed by participants in a previous session. On average it took them X minutes and Y seconds to complete the task.[7]

In line with the procedure proposed by Kahneman & Tversky (1979), we operationalize the historical information as a single data point (the average duration recorded in the Baseline treatment) instead of the whole distribution. The use of a single piece of information allows us to draw clear-cut conclusions regarding the adjustment away from the initial estimate. This would not be the case if the whole distribution was provided because of the inability to attribute the potential change in behavior to a particular information from the distribution. Last but not least, it is arguably easier for subjects to interpret information conveyed in the form of a simple average in comparison with a whole distribution of outcomes.

Incentives

Subjects are financially incentivized for both their estimation accuracy and task performance. The incentive structure is designed to motivate them to estimate task duration accurately, and at the same time to execute the task fast and avoid mistakes. Providing incentives for accuracy as well as 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.

[7] The implemented values were 18 minutes and 13 seconds.

We implement a linear scoring rule to incentivize the estimation accuracy.[8] According to the rule, the estimation accuracy earnings depend on the absolute difference between the actual task duration and the estimate. The maximum earnings from a precise estimate are AUD 18. The accuracy earnings decrease by AUD 2.40 for every minute away (i.e., 4 cents for every second away) from the actual task duration, as shown in Equation (1). We do not allow for negative estimation accuracy earnings. If the difference between the actual and estimated time in either direction exceeds 7.5 minutes (450 seconds), the estimation accuracy earnings are set to zero.[9] This design feature is implemented because of our expectations of a significant estimation mistakes due to the omitted details in the task description that could cause many subjects to end up with negative earnings. Our experimental setting parallels a common practice in the business world 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.

$$\text{Estimation accuracy earnings} = 18 - 0.04 * |\text{actual task duration in seconds} - \text{estimated duration in seconds}| \quad (1)$$

The task performance earnings, presented in Equation (2), depend on the actual task duration and on the number of correct and incorrect answers provided. The shorter the duration and the fewer mistakes, the higher the earnings. We penalize subjects for incorrect answers in order to discourage them from randomly clicking, guessing, or systematically trying all combinations. The experimental incentives are parallel to incentives in business practice where it is not only the speed but also the quality that matters. We expected subjects to complete the task in 15 minutes (900 seconds) on average. Without incorrect answers, such pace would earn them AUD 10 for their task performance, making the task performance earnings comparable with the estimation accuracy earnings.

[8] We acknowledge that the linear scoring rule might not be the most incentive compatible one. However, it is more practical to implement than more complex scoring rules (e.g., quadratic or logarithmic) due to ease of explanation to subjects (Woods & Servátka, 2016).

[9] The 450-second threshold was derived from the average task duration observed in pilot experiments (around 900 seconds). Since the instructions provide only a crude task description, we opted to set the threshold at the level of so-called "Rough Order of Magnitude" estimate, used in the initial project stages when the exact project scope is not yet fully developed. The project management methodology for estimating duration requires the Rough Order of Magnitude estimates to fall within the range of +75%/-25% from the actual duration (Project Management Institute, 2013). Since our estimation accuracy earnings are symmetric for underestimation and overestimation, we implemented a range of +/-50%.

$$\text{Task performance earnings} = \frac{300*(3*\text{number of correct answers} - \text{number of incorrect answers})}{\text{actual task duration in seconds}} \quad (2)$$

Since the experiment recreates two types of incentives faced by planners in a business setting, there is a concern that subjects might try to create a portfolio of accuracy and performance earnings (Cox & Sadiraj, 2018). While it is possible to control for the portfolio effect by randomly selecting one type of incentives for payment (Cox, Sadiraj, & Schmidt, 2015; Holt, 1986), we opt to minimize the chances of subjects constructing a portfolio by a careful experimental design and selection of procedures to preserve the parallelism. Our procedures (described below in detail) ensure that subjects are neither able to keep track of the elapsed time nor are provided with the number of matrices already solved, making it difficult to submit strategic estimates and control their working pace.[10]

Procedures

The experiments were conducted in the MGSM Vernon L. Smith Experimental Economics Laboratory at the Macquarie Graduate School of Management in Sydney. Subjects were recruited using the online database system ORSEE (Greiner, 2015). The experiments were programmed in zTree software (Fischbacher, 2007).

Before the start of the experiment, subjects sitting in individual cubicles were asked to put away their watches, mobile phones and any other devices that show time, to prevent them from measuring the elapsed time. The laboratory premises did not contain any time displaying devices. The clocks on computer screens were hidden. After reading the instructions, subjects were given a few minutes to ask questions regarding the experiment. Once all questions were privately answered by the experimenter, the experiment proceeded with the decision-making part. At the end of the experiment, subjects privately and individually received their experimental earnings in cash in the control room at the back of the laboratory.

[10] The design of the incentive structure is similar to the one used in Lorko et al., (2019), where no evidence of the portfolio effect is found. Moreover, since the task performance earnings are strictly declining with time, the only possibility to create a portfolio is to strategically overestimate (inflate) the time necessary to complete the task. Such behavior would be in sharp contrast with our conjectures, according to which we expect subjects to underestimate the task duration. The experimental data allows us to verify whether overestimation takes place.

Hypotheses

Since we deliberately describe the task in a way that it appears relatively easy to complete, we hypothesize that subjects in the Baseline treatment will exhibit a tendency to underestimate its duration. Due to disclosure of historical information, we hypothesize that the prevalence of underestimation will be lower in the other two treatments. We further conjecture that estimates in the Info-Before treatment will be more accurate than the revised estimates in the Info-After treatment, as in the latter case subjects will need to decide whether to adjust their initial estimates towards the historical average or to persist with their initial estimates. It is conceivable that subjects may be reluctant to fully incorporate the historical information in their estimation process due to cognitive dissonance or the cost of cognitive effort and thus the adjustment of the initial estimate towards the historical information can be insufficient. Therefore, we expect to find unbiased estimates with no systematic tendency to underestimate or overestimate the task duration in the Info-Before treatment but not necessarily in the Info-After treatment.

- *Hypothesis 1*
    - $Estimates_{BASELINE} < Duration_{BASELINE}$
    - $Estimates_{INFO\text{-}AFTER} < Duration_{INFO\text{-}AFTER}$
    - $Estimates_{INFO\text{-}BEFORE} = Duration_{INFO\text{-}BEFORE}$
    - $Estimates_{BASELINE} < Estimates_{INFO\text{-}AFTER} < Estimates_{INFO\text{-}BEFORE}$

Since subjects in all treatments are incentivized for both estimation accuracy and task performance, we hypothesize to find no differences in the actual duration of the task across our treatments, akin to earlier findings (Buehler, Griffin, & MacDonald, 1997; Lorko et al., 2019). In combination with the conjectured differences in estimates, we hypothesize that the Baseline treatment will result in the largest estimation bias and lowest estimation accuracy.

- *Hypothesis 2*
    - $Duration_{BASELINE} = Duration_{INFO\text{-}AFTER} = Duration_{INFO\text{-}BEFORE}$
    - $Accuracy_{BASELINE} < Accuracy_{INFO\text{-}AFTER} < Accuracy_{INFO\text{-}BEFORE}$
    - $Bias_{BASELINE} > Bias_{INFO\text{-}AFTER} > Bias_{INFO\text{-}BEFORE}$

Subjects in the Baseline treatment do not receive any information other than the task description whereas in the other two treatments, subjects receive an additional piece of information that might

aid their estimation and boost their confidence. Thus, one might expect subjects in the Baseline treatment to be less confident in their estimates. However, the "what you see is all there is" rule, predicts relatively high subjective confidence also in the Baseline treatment, as subjects are unaware of what they do not know. We therefore state our hypothesis as not finding any differences in the confidence in estimates across treatments.

- *Hypothesis 3*
    - $Confidence_{BASELINE} = Confidence_{INFO\text{-}AFTER} = Confidence_{INFO\text{-}BEFORE}$

Main Results

A total of 103 subjects, randomly assigned into our three treatments, participated in the experiment. However, 7 of those subjects found the task too difficult and gave up without completing the experiment.[11] We thus analyze only the behavior of the remaining 96 subjects (39 females) with a mean age of 22.8 a standard deviation of 4.5 years. Of these remaining subjects, 38 participated in the Baseline treatment, 29 in the Info-After treatment and 29 in the Info-Before treatment. We opted for a larger sample size in the Baseline treatment, in order to obtain a more robust average task duration. On average, subjects spent 50 minutes in the laboratory and earned AUD 18.60. The summary statistics are presented in Table 1. For the Info-After treatment, we present both the initial estimates elicited before the provision of the historical information, as well as the revised estimates that were elicited after the historical average was disclosed to subjects. Unless specifically stated, we use revised estimates for testing the treatment effects.

[11] Out of the 7 subjects that gave up finishing the task, 5 were in the Baseline treatment, 1 in the Info-After treatment and 1 in the Info-Before treatment.

Table 1: Summary statistics and the test of similarity between the estimates and the actual task duration

Treatments	Baseline (N = 38)	Info-After (N = 29)		Info-Before (N = 29)
		Initial est.	Revised est.	
Mean estimate, SD (seconds)	601 (704)	456 (427)	814 (377)	798 (329)
Mean actual duration, SD (seconds)	1093 (573)	986 (528)		914 (404)
Mean bias, SD (seconds)[a]	-492 (757)		-171 (521)	-115 (365)
Mean absolute error (seconds)	725 (530)		425 (338)	275 (262)
Median estimate (seconds)	270	300	900	900
Median actual duration (seconds)	919	847		818
Median bias (seconds)	-539		-164	-68
Median absolute error (seconds)	682		412	184
Comparison of the estimates and the actual duration (p-values)				
Wilcoxon matched-pairs signed-rank test	<0.001	<0.001	0.09	0.17

Notes: a: The bias is calculated as a relative estimation error (Estimate – Actual duration). SD refers to standard deviation.

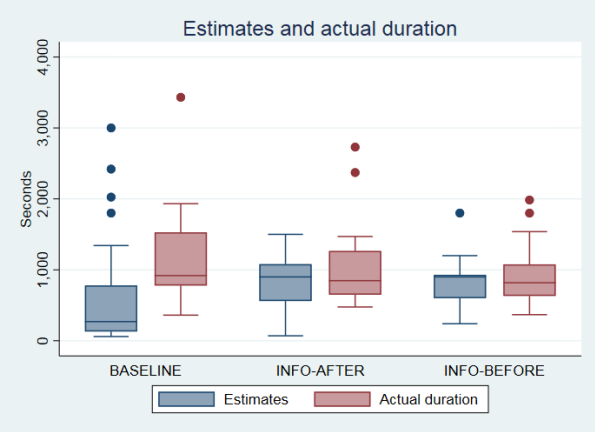
The results of treatment effects are presented in Table 2. Recall that the subjects in the Info-Before treatment received information about the historical average before their initial estimation, while the subjects in the Baseline treatment received no such information. As a result, we find significantly higher estimates in the Info-Before treatment than in the Baseline treatment (Mann-Whitney test, p-value <0.01). On the other hand, the subjects in the Info-After treatment were given identical instructions before their initial estimation as the subjects in the Baseline treatment and were not provided with any historical information at first. Unsurprisingly, they provide similar estimates as the subjects in the Baseline treatment (Mann-Whitney test, p-value = 0.98). However, upon disclosure the historical information, estimates in the Info-After treatment significantly increase, as the subjects adjust their initial beliefs towards the historical average (within-subjects Wilcoxon matched-pairs signed-ranks test, p-value <0.01). These revised estimates are significantly higher than the estimates in the Baseline treatment (Mann-Whitney test, p-value <0.01) and similar to the estimates in the Info-Before treatment (Mann-Whitney test, p-value = 0.73). Regarding the task execution, in line with our Hypothesis 2, we find no differences in the actual task duration across our three treatments (p-values for non-parametric Mann-Whitney tests comparing actual task duration are presented in Table 2, while data are graphically displayed in Figure 1).

Table 2: Treatment effects

		Mann-Whitney test (p-values)		
	Median (Baseline / Info-After / Info-Before)	Baseline vs. Info-After	Baseline vs. Info-Before	Info-After vs. Info-Before
Estimates (seconds)	270 / 900 / 900	<0.01	<0.01	0.73
Actual duration (seconds)	919 / 847 / 818	0.29	0.21	0.71
Bias (seconds)	-539 / -164 / -68	0.02	<0.01	0.76
Absolute error (seconds)	682 / 412 / 184	<0.01	<0.01	0.09
Confidence (Likert)	4 / 4 / 4	0.56	0.84	0.45

Result 1: The estimates in the Baseline treatment are significantly lower than the estimates in both treatments with historical information. The timing when the information is provided (either prior to the estimation or after the initial estimate is made) does not influence the estimates. The actual task duration does not differ across all three treatments.

Figure 1: Estimates and actual task duration

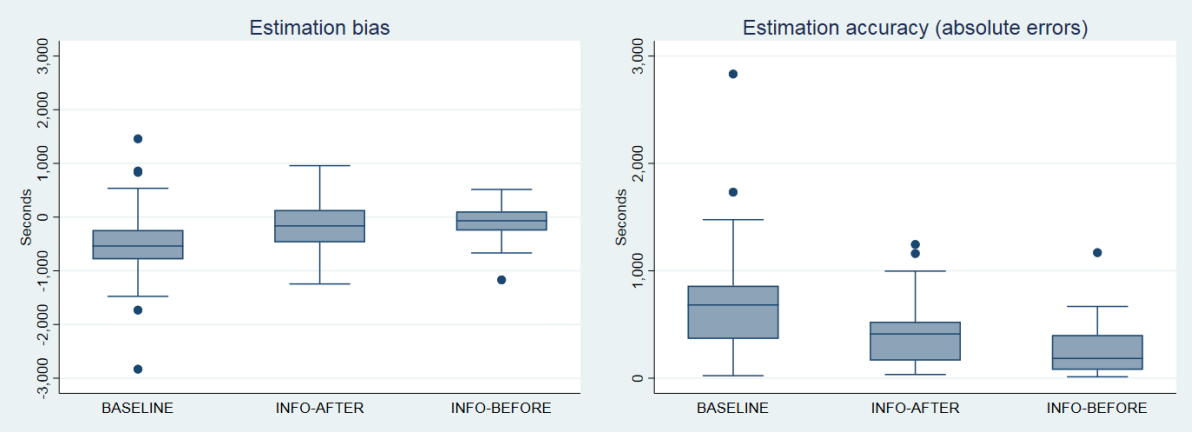


Note: Figure 1 displays box plots of estimates and actual task duration, by treatments.

Our data also provide support for Hypothesis 2, which states that subjects in the Baseline treatment are more likely to underestimate the time necessary to complete the task, resulting in the largest estimation bias and lowest accuracy. As we also predicted, the subjects in the Info-Before treatment exhibit the smallest bias and the highest accuracy. Nevertheless, treatment effects regarding the estimation bias and accuracy parallel the previous results. In particular, we find the bias to be

significantly larger and the accuracy to be significantly lower in the Baseline treatment than in both the Info-Before and Info-After treatments that provide subjects with historical information. We do not find significant differences in the bias and the accuracy between the Info-After treatment and the Info-Before treatment (p-values are presented in Table 2, aggregate data are displayed in Figures 2a and 2b, and individual-level data in Figure 3).

Figure 2: Estimation bias and accuracy

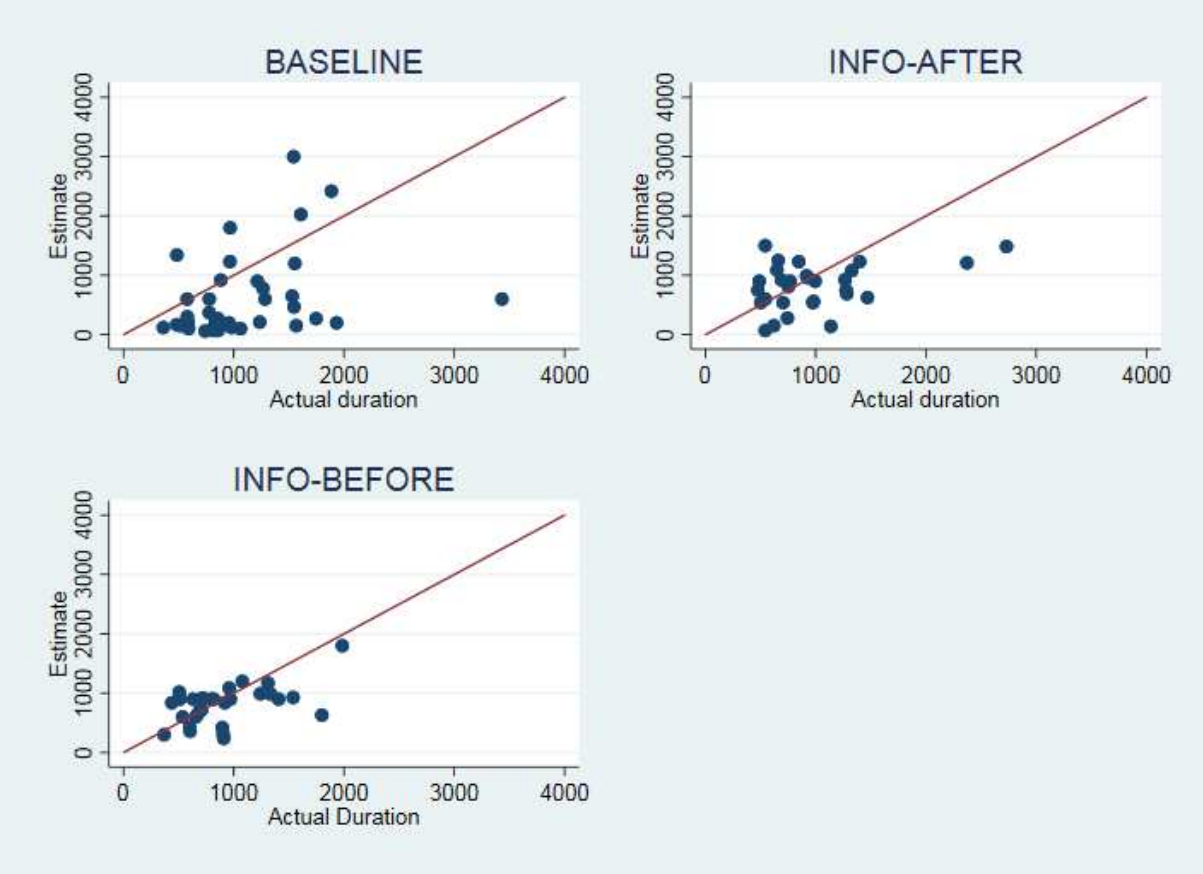


Notes: Figure 2a (left panel) displays box plots of estimation bias (relative estimation error), by treatments. Figure 2b (right panel) displays box plots of estimation accuracy (absolute estimation error), by treatments.

We also examine the estimation bias by conducting a within-subject analysis, comparing the estimates with the actual duration (Wilcoxon matched-pair signed-ranks test; p-vales are presented in Table 1). Although we find prevalence of underestimation in all treatments, the test confirms that the subjects in both the Info-After treatment and Info-Before treatment are significantly less biased than the subjects in the Baseline treatment.

Result 2: The estimates in the Baseline treatment exhibit the largest estimation bias and lowest estimation accuracy. Providing historical information in the Info-After and Info-Before treatments aids estimation, which is reflected by a lower bias (less underestimation) and improved accuracy in comparison with the Baseline treatment.

Figure 3: Individual-level estimates and actual task duration



Notes: Figure 3 displays scatter plots of individual-level estimates (vertical axis) and actual duration (horizontal axis), by treatments. 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.

Robustness

We test the robustness of our main results using Jonckheere–Terpstra trend test for ordered alternative hypotheses. In particular, we test our directional hypotheses stating that the estimates and estimation accuracy will be the lowest in the Baseline treatment, higher in the Info-After treatment, and the highest in the Info-Before treatment. We find support for predicted trends in analysis of both the estimates (p-value <0.01) and the estimation accuracy (p-value <0.01). We also find additional support for the hypothesis that all treatments will result in similar actual task duration (Kruskal-Wallis test, p-value = 0.38).

Auxiliary analysis

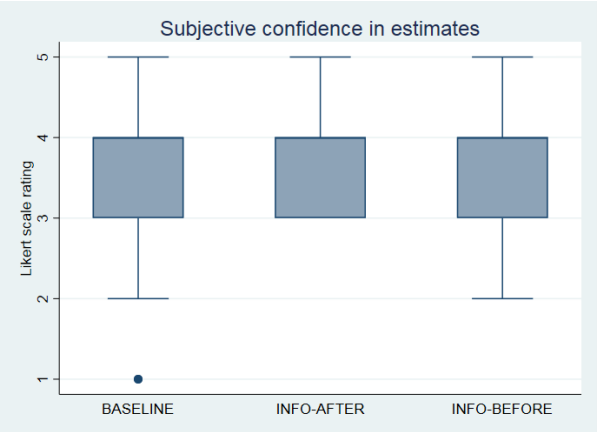
To test whether the behavior of our subjects is consistent with the "what you see is all there is" rule, we analyze the ratings of subjective confidence in estimates (summary statistics are presented in Table 3, data graphically displayed in Figure 4). Immediately after duration estimation, subjects were asked to indicate their subjective confidence in the accuracy of the estimate on a 5-point Likert scale. In particular, subjects filled in the sentence "I am ..... that my estimate will be accurate," with either very confident (in the statistical analysis the assigned value is 5), confident (4), neither confident nor unconfident (3), unconfident (2), or very unconfident (1). Subjects were informed that the answer to this question was not payoff relevant.

Table 3: Summary statistics of the subjective confidence in estimates

Treatments	Baseline (N = 38)	Info-After (N = 29)		Info-Before (N = 29)
		Initial est.	Revised est.	
Mean confidence (SD)	3.7 (0.8)	3.5 (0.7)	3.8 (0.7)	3.7 (0.8)
Median confidence	4	4	4	4

We find support for our Hypothesis 3 as well as the "what you see is all there is" rule. The ratings of subjective confidence in estimates are similar across all treatments (Kruskal-Wallis test, p-value = 0.78; for p-values of pair-wise Mann-Whitney tests, see Table 2). In general, subjects report relatively high confidence, as the median confidence in all treatments is 4 out of the maximum of 5, irrespectively of whether they received historical information prior to the estimation or not.

Figure 4: Subjective confidence in estimates



Note: Figure 4 displays box plots of subjective confidence in estimates, by treatments.

Result 3: Providing historical information (increasing the quantity of task relevant information) does not affect the subjective confidence in estimates.

Finally, to shed light on whether the individuals recognize the importance of historical information, we analyze responses to the non-incentivized willingness-to-pay question asked at the end of the experiment in the Info-After treatment and the Info-Before treatment. The question asked subjects to consider that the information regarding the average actual duration of the task in the past was not given for free and required them to state the maximum they would be willing to pay in order to obtain such information. From the analysis we eliminated 14 subjects who stated that they would be willing to pay more than AUD 18, which was the threshold of the maximum attainable earnings from the estimation accuracy, reducing our sample size to 23 subjects in the Info-After treatment and 21 subjects in the Info-Before treatment. The median willingness-to-pay in these reduced samples is AUD 5.00 in the Info-After treatment and AUD 2.50 in the Info-Before treatment. The difference is statistically significant (Mann-Whitney test, p-value = 0.02). We speculate that the subjects in the Info-After treatment are willing to pay more because they have experienced the benefits of the historical information when updating their original estimates. However, we note that the test results are not significant if we include the eliminated subjects.

4. Experiment 2: Detailed description

Experiment 1 hypotheses assume that the subjects in the Baseline treatment would underestimate the time necessary to complete the task because of the omission in the task description. This omission would lead to subjects expecting to find integer numbers in the matrices, in which case, the task would be arguably easier to complete. Our Baseline treatment data indeed reveal heavy underestimation of time to complete the task. To test whether providing a more detailed task specification can have similar effects to providing historical information, that is whether such intervention can also produce less biased and more accurate duration estimates, we design and conduct Experiment 2.

Treatments

In Experiment 2, we utilize data from the Baseline treatment of Experiment 1 and compare them to the behavior of subjects in the additional "Detailed Description" treatment. In the Detailed Description treatment, we use the same experimental task, incentive structure, and procedures as in the Baseline treatment. However, as the name of the treatment suggests, we provide subjects with a more

informative task description. In particular, subjects in the Detailed Description treatment are shown a sample matrix in the instructions (provided in the appendix) and thus are aware that numbers in matrices are decimal. We explicitly mark the correct answer inside the sample matrix to prevent subjects from practicing the task and learning how much time it takes them to find a correct answer.

Hypotheses

The task description differs across treatments in a way that the task seems easier in the Baseline treatment in comparison with the Detailed Description treatment. We therefore hypothesize to find significantly higher (and hence less understated) estimates in the Detailed Description treatment than in the Baseline treatment.

- *Hypothesis 4*
    - $Estimates_{BASELINE} < Estimates_{DETAILED\ DESCRIPTION}$

Since we also expect no significant differences in the actual task duration across treatments (in parallel to Hypothesis 2 of Experiment 1), we conjecture that subjects in the Detailed Description treatment will provide less biased and more accurate duration estimates.

- *Hypothesis 5*
    - $Duration_{BASELINE} = Duration_{DETAILED\ DESCRIPTION}$
    - $Accuracy_{BASELINE} < Accuracy_{DETAILED\ DESCRIPTION}$
    - $Bias_{BASELINE} > Bias_{DETAILED\ DESCRIPTION}$

Intuitively, estimates based on a less informative task description could be associated with lower confidence. However, our across-subjects design (i.e., only one version of the task description provided to an individual subject) makes it difficult for subjects to realize that some essential information is missing. Hence, in line with the "what you see is all there is" rule, we expect subjects to focus only on what is provided to them in the instructions and display fairly similar confidence in estimates in both treatments.

- *Hypothesis 6*
    - $Confidence_{BASELINE} = Confidence_{DETAILED\ DESCRIPTION}$

Main results

A total of 36 subjects participated in the Detailed Description treatment. Two of them gave up and did not finish the task, leaving us with 34 observations. Thus, in combination with 38 observations from the Baseline treatment, we analyze the behavior of 72 subjects (31 females) with a mean age of 22.8 and a standard deviation of 3.8 years. The average earnings in this experiment (i.e., averaged over the Baseline and Detailed Description treatments) were AUD 14.70.

Table 4: Summary statistics and the test of similarity between estimates and actual task duration

Treatments	Baseline (N = 38)	Detailed Description (N = 34)
Mean estimate, SD (seconds)	601 (704)	1149 (1287)
Mean actual duration, SD (seconds)	1093 (573)	1144 (565)
Mean bias, SD (seconds)[a]	-492 (757)	5 (1369)
Mean absolute error (seconds)	725 (530)	1012 (904)
Median estimate (seconds)	270	525
Median actual duration (seconds)	919	1017
Median bias (seconds)	-539	-211
Median absolute error (seconds)	682	734
Mean confidence (SD)	3.7 (0.8)	3.5 (0.9)
Median confidence	4	4
Comparison of the estimates and the actual duration (p-values)		
Wilcoxon matched-pairs signed-rank test	<0.001	0.76

Notes: a: The bias is calculated as a relative estimation error (Estimate – Actual duration). SD refers to standard deviation.

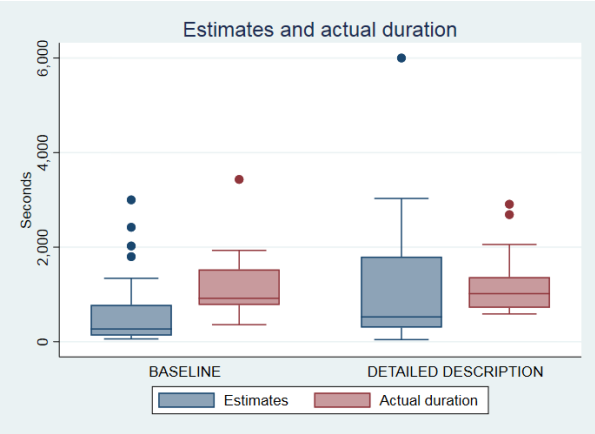
The summary statistics are presented in Table 4, while the treatment effects are presented in Table 5 and box plots in Figure 5. In line with our Hypothesis 4, the estimates in the Detailed Description treatment are on average almost two times higher than in the Baseline treatment, with the difference being statistically significant (Mann-Whitney test, p-value = 0.049). The actual task duration does not differ across treatments (Mann-Whitney test, p-value = 0.66).

Table 5: Treatment effects

	Median Baseline / Detailed Description	Mann-Whitney test (p-values) Baseline vs. Detailed Description
Estimates (seconds)	270 / 525	0.049
Actual duration (seconds)	919 / 1017	0.66
Bias (seconds)	-539 / -211	0.04
Absolute error (seconds)	682 / 734	0.33
Confidence (Likert)	4 / 4	0.28

Result 4: Providing a more detailed task description mitigates underestimation of task duration.

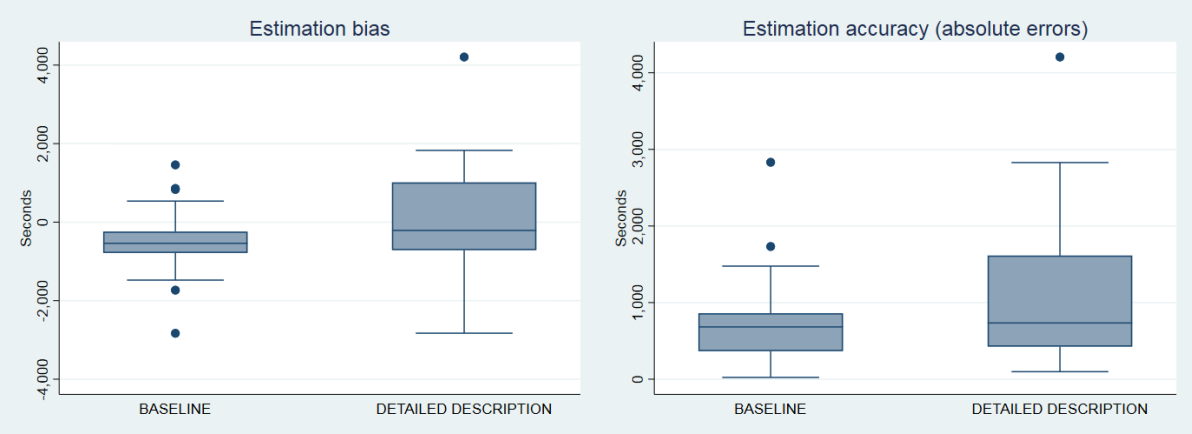
Figure 5: Estimates and actual task duration



Note: Figure 5 displays box plots of estimates and actual task duration, by treatments.

The subjects in the Detailed Description treatment exhibit a mean estimation bias of only 5 seconds. However, the small bias itself does not necessarily imply high estimation accuracy, which depends on the severity of both overestimates and underestimates. We find a large variance in estimates, which range from a couple of minutes to almost 2 hours (individual-level data are displayed in Figure 7) in the Detailed Description treatment. Although the subjects provide unbiased estimates on average, their estimation accuracy is slightly worse (but not statistically significantly) than in the Baseline treatment (see Table 4 for summary statistics and Figure 6 for box plots). Our Hypothesis 5 stating that providing more detailed description leads to less biased and more accurate estimates is supported only partially.

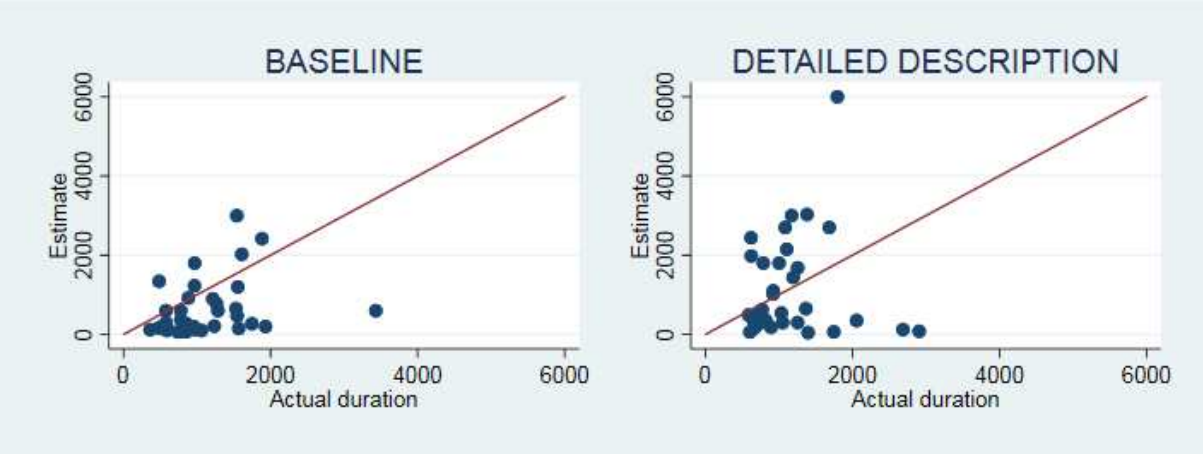
Figure 6: Estimation bias and accuracy



Notes: Figure 6a (left panel) displays box plots of estimation bias (relative estimation error), by treatments. Figure 6b (right panel) displays box plots of estimation accuracy (absolute estimation error), by treatments.

Result 5: Providing a more detailed task description leads to a significantly smaller estimation bias but does not improve the estimation accuracy.

Figure 7: Individual-level estimates and actual task duration

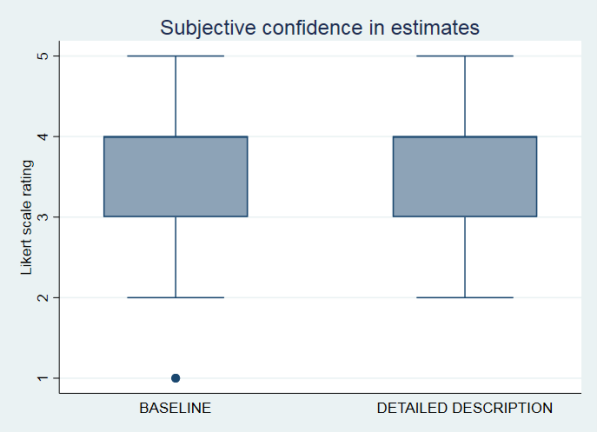


Notes: Figure 7 displays scatter plots of individual-level estimates (vertical axis) and actual duration (horizontal axis), by treatments. Precise estimates are on the red 45-degree line. A dot above the red line indicates overestimation, while a dot below the red line indicates underestimation.

Auxiliary analysis

Similarly to Experiment 1 analysis, we do not find any differences in subjective confidence in estimates between the treatments (Mann-Whitney test, p-value = 0.28; see also Figure 8 for box plots), which is in line with our Hypothesis 6 and provides further support for the "what you see is all there is" rule. Subjects report similar confidence ratings irrespectively of how detailed the task description is.

Figure 8: Subjective confidence in estimates



Note: Figure 8 displays box plots of subjective confidence in estimates, by treatments.

Result 6: Providing a more detailed task description (increasing the quality of task relevant information) does not affect the subjective confidence in estimates.

Finally, we again analyze the willingness to pay for historical information. This time, after the completion of the experimental task, we asked subjects to consider that there was such information available before the estimation and state the maximum amount they would have been willing to pay in order to obtain this information. From the analysis, we have eliminated 17 subjects who stated that they would be willing to pay more than AUD 18, reducing our sample size to 31 subjects in the Baseline treatment and 24 subjects in the Detailed Description treatment.[12] The median willingness to pay is AUD 3.50 in the Baseline treatment and AUD 2.80 in the Detailed Description treatment. The difference is not statistically significant (Mann-Whitney test, p-value = 0.12). Subjects in the Experiment 2 treatments display no difference in willingness to pay for the historical information, providing further support for the "what you see is all there is" rule. Indeed, if people do not account for the possibility that they are missing critical evidence for their judgment and display similar level of

[12] The result is robust to including these subjects.

confidence in their estimates, we would not expect to find differences in their willingness to pay for additional information.[13]

5. What to provide: historical information or more detailed description?

The common Baseline treatment in our two experiments allows us to directly compare the effect of the two implemented interventions. In the earlier analysis, we find that both the provision of historical information (in Experiment 1) and the provision of more detailed task description (in Experiment 2) mitigate the underestimation of the time necessary to complete the task. The estimation bias (in comparison with the Baseline) is significantly reduced in the Info-Before treatment (Mann-Whitney test, p-value <0.01), the Info-After treatment (Mann-Whitney test, p-value = 0.02) as well as in the Detailed Description treatment (Mann-Whitney test, p-value = 0.04).

In contrast, we find a similar estimation bias in all comparisons across the three treatments with an intervention, i.e., providing historical information or a detailed task description (the Mann-Whitney test p-values are 0.76 for the Info-Before vs. Info-After comparison, 0.76 for Info-Before vs. Detailed Description, and 0.84 for Info-After vs. Detailed Description). The results suggest that the effects of both interventions are of similar sizes, making the treatments directly comparable for the following analysis of estimation accuracy.

When analyzing the improvement in estimation accuracy against the Baseline treatment, we find that the intervention implemented in Experiment 1 is effective, while the intervention implemented in Experiment 2 is not. The absolute estimation error is reduced (against the Baseline) in the Info-Before treatment (Mann-Whitney test, p-value <0.01) and the Info-After treatment (Mann-Whitney test, p-value <0.01), but not in the Detailed Description treatment (Mann-Whitney test, p-value = 0.33). Furthermore, we find no statistically significant differences in estimation accuracy between the Info-Before and the Info-After treatments (Mann-Whitney test, p-value = 0.09). We do, however, find that subjects in the Detailed Description treatment are less accurate than subjects in both the Info-Before

[13] The willingness to pay in treatments of Experiment 2 is relatively similar to the willingness to pay found in the Info-Before treatment of Experiment 1. The subjects in the Info-After treatment of Experiment 1 were willing to pay more than the subjects in any other treatment. Again, this may be caused by the fact, that the subjects in the Info-After treatment were actually the ones who used the historical information to update their initial estimates.

treatment (Mann-Whitney test, p-value <0.01) and the Info-After treatment (Mann-Whitney test, p-value <0.01). Thus, in terms of estimation accuracy, the effect of historical information significantly outperforms the effect of more detailed task description.

Result 7: Providing historical information as well as providing detailed task description significantly reduces the estimation bias. However, only the provision of historical information also significantly improves the estimation accuracy.

Robustness

To verify the robustness of our results, we conduct regression analysis. The regression results (presented in Table 6) are consistent with non-parametric tests presented earlier. In particular, we find that both our interventions are associated with higher and thus less biased estimates, but only the provision historical information significantly improves the estimation accuracy. Also, we find no effect of any intervention on the actual task duration.

In addition, we test the effect of risk attitudes, time spent on estimation, time spent on indicating confidence, subjective confidence in estimate, and demographics (age, gender, education, employment status and self-reported math skill). We find that higher confidence is associated with lower estimates but has no effect on the actual task duration and estimation accuracy. Furthermore, we find a significant positive effect between estimates and the actual task duration as well as between self-reported math skill and the actual task duration.[14]

[14] The observation that self-reported math skill is significantly negatively correlated with the actual task duration is in contrast with the claim that the task does not reflect math skills, made in Mazar et al., (2008). However, we note that our subjects self-reported their math skill after they finished the task, at which point they may have felt how good their performance was. As such, it is not clear which way the causation goes.

Table 6: Linear regression analysis

Dependent variable	(1) Estimate	(2) Actual Duration	(3) Estimation bias	(4) Absolute estimation error
1. Info-After Treatment	323.63**	-94.88	380.79**	-252.81**
	(-140.63)	(-131.26)	(-161.68)	(-118.77)
2. Info-Before Treatment	216.79*	-124.96	316.48**	-413.78***
	(-127.3)	(-124.76)	(-150.28)	(-111.52)
3. Detailed Description Treatment	586.68**	-8.68	526.98*	283.76
	(-266)	(-146.65)	(-267.93)	(-180.34)
4. Age	-5.54	5.46	-10.35	-1.66
	(-23.4)	(-19.59)	(-25.66)	(-20)
5. Female	-84.79	-88.01	13.1	15.81
	(-153.07)	(-96.09)	(-174.69)	(-111.24)
6. Self-reported math skill	27.8	-188.03***	212.58***	-69
	(-79.59)	(-53.23)	(-79.99)	(-50.27)
7. Current degree of study	34.51	1.51	28.97	49.69
	(-79.7)	(-41.54)	(-72.4)	(-53.67)
8. Employment status	52.91	-10.26	57.01	-68.47
	(-68.78)	(-47.06)	(-78.08)	(-52.44)
9. Risk attitudes	45.22	28.89	11.06	-5.62
	(-30.23)	(-26.61)	(-35.89)	(-25.47)
10. Time spent estimating	-2.83	-1.73	-0.77	-1.8
	(-2.37)	(-1.26)	(-2.65)	(-1.53)
11. Time spent indicating confidence in estimate	10.72	11.24	-1.77	7.61
	(-15.37)	(-12.69)	(-19.28)	(-10.99)
12. Subjective confidence in estimate	-243.00**	45.3	-259.98**	-26.8
	(-97.14)	(-77.81)	(-119.6)	(-74.69)
13. Estimate		0.12**		
		(-0.06)		
Constant	1317.07**	206.68	956.86	661.64
	(-569.07)	(-646.97)	(-853.56)	(-568.57)
N	130	130	130	130
R^2	0.13	0.20	0.12	0.24

Note: Standard errors are reported in parentheses. *, **, and *** indicate significance at the 10%, 5%, and 1%-level, respectively.

6. Discussion

An adequate business project schedule is essential for project success and plays a key role in effective allocation and utilization of company resources. In this paper, we investigate the effectiveness of two interventions designed to induce more accurate duration estimates within the project planning process: (1) increasing the quantity of information available before the estimation by providing historical information and (2) increasing the quality of information available before the estimation by providing a more detailed project specification. In Experiment 1, we deliberately omit important information regarding the decimal format of numbers in matrices, making the task appear easier to complete than it really is. This creates a relatively large gap between the intuitive estimate and the time necessary to complete the task. Such gap provides a well-calibrated environment for testing the effect of historical information as a tool for adjusting intuitive estimates towards the average duration of similar tasks in the past. We show that the utilization of historical information in the planning process can significantly mitigate the estimation bias and improve estimation accuracy. We further find that the timing when such information is provided does not matter, at least in the environment encountered by our subjects. We note, however, that the timing might matter in the business practice, where producing initial estimates may be associated with making a commitment towards co-workers or managers. Subsequent adjustment of such initial estimates towards historical average may be seen as poor competence of the planner.

One could object that the task description used in Experiment 1 is too uninformative, not disclosing crucial information regarding the very nature of the task. Although such claim may be true, we note that virtually any project specification is a simplification of the actual project deliverables and companies often have a relatively muddled idea about the precise characteristics of outcomes requested within the project they are about to start. Nevertheless, in order to test whether a more informative task description leads to more accurate estimates, we conduct Experiment 2 in which a sample matrix is added to the task description. We find that a more detailed specification can eliminate the estimation bias (in particular underestimation), which becomes statistically indifferent from zero, resembling the "wisdom of the crowd" phenomenon (Galton, 1907). However, due to the extensive spread of individual estimates, the average estimation accuracy is not improved, akin to the assumption of the "bias-variance trade-off" (Geman, Bienenstock, & Doursat, 1992). The bias-variance trade-off implies that the absence of specific biasing intervention can induce high variance in estimates due to a large number of other environmental factors that can influence them. Hence, letting planners to anchor their estimates on reliable historical information and "biasing" them

towards reference class average appears to be a better strategy than relying on overly detailed project specifications.

Previous literature suggests that planners may not be sensitive to the potential lack of relevant information during the estimation process. In line with this argument we show that subjective confidence in estimates is neither a function of quantity (Experiment 1) nor quality (Experiment 2) of available information and therefore is not a reliable predictor of estimation accuracy. Our subjects provided essentially the same confidence ratings irrespectively of what they knew about the task prior to the estimation. Our results suggest that project managers are better off by not making decisions regarding the adequacy of a project plan based on the confidence displayed by the project planners.

One limitation of our study is that we focus solely on the estimation bias and (in)accuracy stemming from an incomplete project specification. However, misestimation of project duration can also be caused by a complex interplay of multiple other factors, such as risks and unpredictable events. These factors are often hardly foreseeable during the project planning phase, but can induce potentially large schedule delays. Nevertheless, it is likely that the utilization of historical information in estimation can also ameliorate the effect of such factors, a conjecture worthwhile testing in future research. Furthermore, since a project schedule is usually created by more than one professional/project planner, testing the effectiveness of historical information utilization in a group decision-making environment could be another natural extension of the current study.

Another limitation of our study is that in order to maintain control over the data generating process, we only use one task, identical across all subjects, making the selection of the reference class (the Baseline treatment) for extracting historical information straightforward. Since we find no differences in the actual task duration across all treatments, the reference class was selected appropriately, and the historical information calculated from the reference class is a good predictor for individual outcomes of other subjects. Nevertheless, we believe it is worthwhile to investigate the effect of historical information also on complex business projects consisting of multiple tasks that are not so similar to each other.

To consult historical averages in such environment, planners have to first carefully select a meaningful reference class of past projects. Acquiring historical information may be associated with certain costs (e.g., search cost) and if planners do not consider the information valuable, they may be reluctant to seek it. In the current study, we try to elicit the willingness to pay for historical information ex-post,

and tentatively conclude that those who have experienced the benefits of using such information, value it more. A deeper scientific inquiry into the process of reference class selection and a salient elicitation of willingness to pay for historical information are another potentially interesting pathways for future research.

Acknowledgements: This paper is based on Matej Lorko's dissertation chapter written at the Macquarie Graduate School of Management. We thank Barbora Baisa, Michal Ďuriník, Dan Lovallo, the audiences at the 2018 Young Economists' Meeting in Brno, 2018 ESA World Meeting, 2018 Slovak Economic Association Meeting, and 2019 Asia-Pacific 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

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

Buehler, Roger, & Griffin, D. (2015). The planning fallacy: When plans lead to optimistic forecasts. In *The psychology of planning in organizations: Research and applications* (pp. 31–57). https://doi.org/10.4324/9780203105894

Buehler, Roger, 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

Buehler, Roger, Griffin, D., & Peetz, J. (2010). The Planning Fallacy. Cognitive, Motivational, and Social Origins. *Advances in Experimental Social Psychology*, *43*(C), 1–62. https://doi.org/10.1016/S0065-2601(10)43001-4

Buehler, Roger, Griffin, D., & Ross, M. (1994). Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, *67*(3), 366–381. https://doi.org/10.1037/0022-3514.67.3.366

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

Engerman, S., & Sokoloff, K. (2006). Digging the Dirt at Public Expense Governance in the Building of the Erie Canal and Other Public Works. In *Corruption and Reform: Lessons from America's Economic History*.

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

Flyvbjerg, B. (2006). From Nobel Prize to Project Management: Getting Risk Right. *Project Management Journal*, *37*(3), 18–19. https://doi.org/10.1002/smj.476

Flyvbjerg, B. (2008). Curbing Optimism Bias and Strategic Misrepresentation in Planning: Reference Class Forecasting in Practice. *European Planning Studies*, *16*(1), 3–21. https://doi.org/10.1080/09654310701747936

Flyvbjerg, B., Holm, M. S., & Buhl, S. (2002). Underestimating costs in public works projects: Error or lie? *Journal of the American Planning Association*, *68*(3), 279–295. https://doi.org/10.1080/01944360208976273

Flyvbjerg, B., Skamris Holm, M. K., & Buhl, S. L. (2005). How (In)accurate are demand forecasts in

public works projects?: The case of transportation. *Journal of the American Planning Association*, *71*(2), 131–146. https://doi.org/10.1080/01944360508976688

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

IPMA. (2015). *IPMA Competence Baseline (ICB), Version 4.0*. *International Project Management Association*. https://doi.org/10.1002/ejoc.201200111

Jørgensen, M. (2004). Top-down and bottom-up expert estimation of software development effort. *Information and Software Technology*, *46*(1), 3–16. https://doi.org/10.1016/S0950-5849(03)00093-4

Jørgensen, M. (2010). Selection of strategies in judgment-based effort estimation. *Journal of Systems and Software*, *83*(6), 1039–1050. https://doi.org/10.1016/j.jss.2009.12.028

Kahneman, D. (2011). *Thinking , Fast and Slow (Abstract)*. *Book*. https://doi.org/10.1007/s13398-014-0173-7.2

Kahneman, D., & Lovallo, D. (1993). Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking. *Management Science*, *39*(1), 17–31. https://doi.org/10.1287/mnsc.39.1.17

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

Klein, G. (1999). *Sources of Power: How People Make Decisions*. *Personnel Psychology* (Vol. 52). https://doi.org/10.1061/(ASCE)1532-6748(2001)1:1(21)

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

Lorko, M., Servátka, M., & Zhang, L. (2019). Anchoring in project duration estimation. *Journal of Economic Behavior & Organization*, *162*, 49–65.

Lovallo, D., & Kahneman, D. (2003). Delusions of Success: How Optimism Undermines Executives' Decisions. *Harvard Business Review*. https://doi.org/10.1225/R0307D

Mazar, N., Amir, O., & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*, *45*(6), 633–644. https://doi.org/10.1509/jmkr.45.6.633

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.88.3.622

Parkinson, C. N. (1955). Parkinson's Law. *The Economist*, 1–5.

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. (2019). *PMI's Pulse of the Profession 2019*.

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

Rush, C., & Roy, R. (2001). Expert Judgement in Cost Estimating: Modelling the Reasoning Process. *Concurrent Engineering*, *9*(4), 271–284. https://doi.org/10.1177/1063293X0100900404

Shmueli, O., Pliskin, N., & Fink, L. (2016). Can the outside-view approach improve planning decisions in software development projects? *Information Systems Journal*, *26*(4), 395–418. https://doi.org/10.1111/isj.12091

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

Appendix: **Instructions**

(Note: used for Baseline, Info-After and Info-Before treatments)

Thank you for coming. Please note that the use of watches, mobile phones, any other devices that show time and calculators is not allowed during this experiment. The experimenter will check the cubicles for the presence of time showing devices and calculators before the start of the experiment.

Also, please note that from now on, until the end of the experiment, no talking or any other unauthorized communication is allowed. If you violate any of these rules, we will have to exclude you 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 explain how you can earn money in this experiment. Your decisions and earnings will not be revealed to other participants.

Task

You will be shown 10 matrices one by one. Each matrix contains 16 numbers. Two of those numbers add up to exactly 100. You will have to identify those two numbers. You will move on to the next matrix only after you submit the correct answer.

Before you start working on the task, you will be asked to estimate how long it will take you to complete it. That is, how long it will take you to provide correct answers for all 10 matrices.

Earnings

In this experiment, you can earn money based on the accuracy of your estimate and on your task performance.

Estimation accuracy earnings

Your estimation accuracy earnings (in AUD) will be calculated as follows:

$$\text{Estimation accuracy earnings} = 18 - 0.04 * |\text{actual time in seconds} - \text{estimated time in seconds}|$$ [×]

[×] If the formula returns a negative number, your estimation accuracy earnings will be set to 0.

Your estimation accuracy earnings depend on the absolute difference between the actual time it takes you to complete the task and your estimated time. Notice that the more accurate your estimate is, the more money you earn

Task performance earnings

Your task performance earnings (in AUD) will be calculated as follows:

$$\text{Task performance earnings } = \frac{300 * (3 * \text{number of correct answers} - \text{number of incorrect answers})}{\text{actual time in seconds}}$$

Your task performance earnings depend on the actual time it takes you to 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 correct answers for all 10 matrices), the more money you earn. Also note that your earnings will be reduced for every incorrect answer you provide.

Your total earnings

Your total earnings from the experiment will be the sum of your estimation accuracy earnings and your task performance earnings.

Notice that:

- the more accurate your estimate is;
- the faster you complete the task;
- the fewer incorrect answers you provide;

the more money you earn.

<u>When you finish</u>

After you complete the task, 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. You need to complete the entire experiment in order to get paid.

If you have any questions, please raise your hand.

Instructions

(note: used for Detailed Description treatment)

Thank you for coming. Please note that the use of watches, mobile phones, any other devices that show time and calculators is not allowed during this experiment. The experimenter will check the cubicles for the presence of time showing devices and calculators before the start of the experiment.

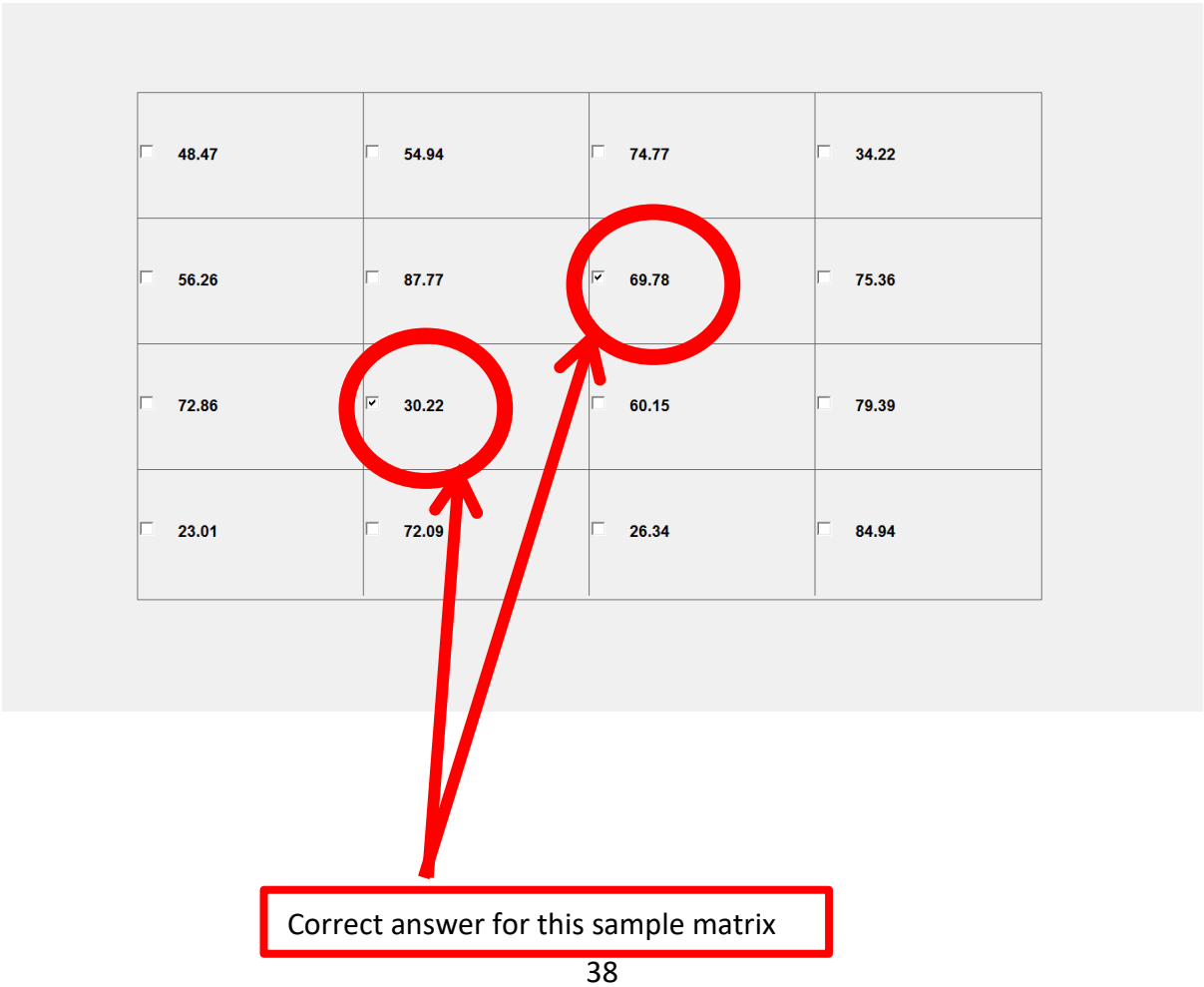
Also, please note that from now on, until the end of the experiment, no talking or any other unauthorized communication is allowed. If you violate any of these rules, we will have to exclude you from the experiment and from all payments. If you have any questions after you finish reading the instructions,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

Please read the following instructions carefully. The instructions will explain how you can earn money in this experiment. Your decisions and earnings will not be revealed to other participants.

Task

You will be shown 10 matrices one by one. Each matrix contains 16 numbers. Two of those numbers add up to exactly 100. You will have to identify those two numbers. You will move on to the next matrix only after you submit the correct answer.

Before you start working on the task, you will be asked to estimate how long it will take you to complete it. That is, how long it will take you to provide correct answers for all 10 matrices.



Earnings

In this experiment, you can earn money based on the accuracy of your estimate and on your task performance.

Estimation accuracy earnings

Your estimation accuracy earnings (in AUD) will be calculated as follows:

$$\text{Estimation accuracy earnings } = 18 - 0.04 * |\text{actual time in seconds } - \text{estimated time in seconds}|$$ ˣ

ˣ If the formula returns a negative number, your estimation accuracy earnings will be set to 0.

Your estimation accuracy earnings depend on the absolute difference between the actual time it takes you to complete the task and your estimated time. Notice that the more accurate your estimate is, the more money you earn.

Task performance earnings

Your task performance earnings (in AUD) will be calculated as follows:

$$\text{Task performance earnings } = \frac{300 * (3 * \text{number of correct answers} - \text{number of incorrect answers})}{\text{actual time in seconds}}$$

Your task performance earnings depend on the actual time it takes you to 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 correct answers for all 10 matrices), the more money you earn. Also note that your earnings will be reduced for every incorrect answer you provide.

Your total earnings

Your total earnings from the experiment will be the sum of your estimation accuracy earnings and your task performance earnings.

Notice that:

- the more accurate your estimate is;
- the faster you complete the task;
- the fewer incorrect answers you provide;

the more money you earn.

When you finish

After you complete the task, 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. You need to complete the entire experiment in order to get paid.

If you have any questions, please raise your hand.