Improving the accuracy of project schedules

Lorko, Matej and Servátka, Maroš and Zhang, Le

MGSM Experimental Economics Laboratory, Macquarie Graduate School of Management, University of Economics in Bratislava

7 October 2020

Improving the accuracy of project schedules

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

October 7, 2020

Abstract: How to avoid project failures driven by overoptimistic schedules? Managers often attempt to mitigate the duration underestimation and improve the accuracy of project schedules by providing their planners with excessively detailed project specifications. While this traditional approach may be intuitive, solely providing more detailed information has proven to have a limited effect on eliminating behavioral biases. We experimentally test the effectiveness of providing detailed specification and compare it to an alternative intervention of providing historical information about the average duration of similar projects in the past. We find that both interventions mitigate the underestimation bias. However, since providing detailed project specification results in high variance of estimation errors due to sizable over- and underestimates, only the provision of historical information leads to more accurate project duration estimates. We also test whether it is more effective to anchor planners by providing historical information simultaneously with the project specification or to provide the historical information only after beliefs regarding the project duration are formed, in which case planners can regress their initial estimates towards the historical average. We find that the timing of disclosing information does not play a role as the estimation bias is mitigated and the accuracy is improved in both conditions. Finally, we observe that the subjective confidence in the accuracy of duration estimates does not vary across the interventions, suggesting that the confidence is neither a function of the amount nor the detail of available information.

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

JEL codes: C91, D83, O21, O22

1. Introduction

A common feature of virtually all business projects is the uncertainty regarding the amount of time and resources needed to deliver expected outcomes. Proficient planning processes capable of generating 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 resources in an organization. Accurate project duration estimates are especially important for project tasks or phases that lie on the critical path, with their timely completion being a necessary condition for delivering the entire project on time (Kelley, 1961). Precise schedules are also vital when managing a project portfolio, in which individual projects compete for temporary use of scarce resources. Delays 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 for the entire organization.

The estimation of project duration appears to be a challenging undertaking as approximately 50 percent of business projects are not delivered on time (Project Management Institute, 2019). The high failure rate begs a question of how to improve the accuracy of project duration estimates. 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). Arguably no specification is extensive enough to capture every aspect of the requested deliverables, especially at early stages of a project when they are not yet developed to the full extent. Nevertheless, project managers often go to great lengths to equip their planners with as detailed as possible descriptions of project tasks. Project planners in turn intuitively tend to focus on the project specification at hand, failing to realize that it might be incomplete. By neglecting the unspecified (or unknown) details, project duration estimates may become understated.[1] Kahneman (2011) refers to the phenomenon of paying attention only to the information one is presented with while ignoring the missing links as the "what you see is all there is" rule.

Kahneman & Lovallo (1993) and Kahneman & Tversky (1977) suggest that the accuracy of project duration estimates can be improved by consulting historical information (also referred to as reference class information) regarding the actual duration of similar projects in the past. The main advantage of utilizing historical information in the planning process is that it naturally encompasses the impact of a variety of small obstacles (e.g., omissions in the project specification, misunderstandings of requirements, or unforeseen events) on project execution. The 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 although historical information may fail to predict extreme project outcomes, it commonly induces more accurate estimates compared to 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, its effect has not yet been tested in a controlled environment and with real incentives.

[1] 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 estimates, such as overoptimism, misrepresentation driven by strategic incentives, competence signaling, using deadlines as commitment devices or unintended anchoring effects.

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 to complete a simple real-effort task, more than two thirds of them would be better off in terms of estimation accuracy if the historical average was used for estimation purposes instead of their own estimate. A similar finding can be found in the demand forecasting literature (see Goodwin, Moritz, & Siemsen, 2019, for a comprehensive review).[2]

In the current paper, we study the impact of historical information directly and compare its effectiveness with the impact of providing a more detailed task description. We investigate the effects of our interventions on two outcomes: the estimation bias, which we measure as a relative (signed) estimation error (i.e., estimate – actual task duration), and the estimation accuracy, which we measure as an absolute estimation error (i.e., |estimate – actual duration|). To allow for causal inference, we create a stylized laboratory environment with a project to be undertaken and carefully manipulate the information our subjects have at their disposal. For both interventions, we deliberately provide only a single piece of additional information, eliciting the lower bound of each effect. This conservative design feature is essential for drawing inference about the relative strength of the interventions. Our experimental design controls for confounding factors such as the 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 the experiment, we first test whether anchoring planners on reliable historical information (operationalized as the average task duration in the past) prior to estimation mitigates the estimation

[2] While demand forecasting differs from project planning in that the actual demand is independent of one's own actions, the prediction of future demand by statistical software still outperforms adjusted expert predictions (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009) most of the time. This result seems to be driven by individuals placing more weight on current observables and neglecting longer horizon outcomes (Kremer, Moritz, & Siemsen, 2011; Massey & Wu, 2005).

bias and improves the estimation accuracy. We then switch off the anchoring effect by disclosing historical information only after the initial estimate has already been made.[3] We test whether individuals regress their estimates towards the historical average and whether this approach is more effective than making the information available alongside the project specification. Finally, we test whether the estimation bias can be mitigated and estimation accuracy improved also by estimating from a more detailed project specification. By linking the interventions together through a common baseline treatment, we are able to directly compare whether providing historical information is more effective than providing a more detailed specification. We conjecture that estimates incorporating historical information outperform, in terms of their accuracy, not only estimates based on a crude (incomplete) specification, but also estimates based on a detailed specification.

After testing our main hypotheses, we explore the mechanism that can encourage planners to seek out and utilize historical information in the estimation process. We elicit the (non-incentivized) willingness to pay for historical information when the benefits have already been experienced by witnessing the improvement in own estimates. We then contrast it with a situation when the information was not provided, and its value is therefore less obvious. For both interventions, we also examine whether the available information reflects on subjective confidence in estimates, measured by a Likert scale. 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 alternative 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.

[3] Previous research shows that anchoring can introduce a bias in estimation if irrelevant information is presented (Lorko et al., 2019; Tversky & Kahneman, 1974). The current study utilizes anchoring as a debiasing technique by nudging the estimates towards the average duration of similar projects in the past.

Our results support the conjecture that disclosing historical information can 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 estimation. We find that while a more detailed task description reduces both the frequency and the extent of duration underestimation, it induces a larger variance in individual estimates. The estimates are on average unbiased, but the estimation accuracy does not improve compared to the baseline accuracy and is significantly worse than the accuracy when historical information is provided. We also find that the willingness to pay for historical information (weakly) increases after witnessing that the information improves the estimates. Finally, in line with "what you see is all there is" rule, we find that subjective confidence in estimates is similar across all treatments and thus does not depend on available information. Subjects do not account for the possibility of missing critical details and exhibit high confidence in their estimates regardless of what they know about the task.

Our study provides the following implications for project management practitioners. First, if information regarding similar projects in the past is available, project managers should consider utilizing the average duration of projects from the reference class as a "helpful anchor" for their planners. Providing historical information is more effective than the more traditional approach of providing detailed project specification. Second, project managers can expect initial resistance of planners to embrace historical information, because planners may not realize its usefulness before experiencing its benefits. Third, confidence in estimates does not correlate with 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

While both underestimated and overestimated project schedules imply negative consequences for project stakeholders, businesses appear to perceive underestimation as 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 underutilized resources arising from overestimation. Moreover, if members of a project team identify instances of overestimation in the project, they can strategically "waste" the allocated time and utilize other resources anyway, so the estimation error may go unnoticed.

In academic research, underestimation has also attracted more attention than overestimation. Kahneman & Tversky (1977) coin the term "planning fallacy," which is a tendency to make overoptimistic plans and forecasts that are close to the best-case scenarios, while 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 (Project Management Institute, 2019). Misestimation can often be attributed to strategic incentives, e.g., 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 or 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). Extant research (see Grushka-Cockayne et al., 2018, for a review) identifies several techniques to mitigate the planning fallacy, such as unpacking/decomposing a project into subtasks (Connolly & Dean, 1997; Forsyth & Burt, 2008; Kruger & Evans, 2004), using predictions by observers instead of self-predictions (Newby-

Clark, Ross, Koehler, Buehler, & Griffin, 2000), or averaging independent estimates from a large group of individuals (Eubanks, Read, & Grushka-Cockayne, 2015).

The current paper explores whether the planning fallacy can be mitigated by estimating from a more detailed project specification and by regressing the estimates towards the average duration of past projects.[4] Kahneman & Tversky (1977) offer a 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. These steps are followed by intuitive estimation of the problem at hand and assessment of predictability, i.e., the degree to which the available historical information permits accurate estimation. In the final step of the procedure, the intuitive estimate is adjusted towards the reference class average.

Interestingly, the procedure for producing regressive estimates has not received much empirical attention and testing. Two notable studies 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 improved estimation accuracy when the reference class averages are supplied. However, Roy et al., (2008) employ tasks the duration of which is often beyond the participants' control while Shmueli et al., (2016) rely only on the predicted accuracy rather than the actual one, as the tasks are not performed after the estimation. Also, subjects in neither study 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

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

[5] Flyvbjerg, Skamris Holm, & Buhl (2005) later shorten the procedure and call it "Reference Class Forecasting". Reference class forecasting was later endorsed by the American Planning Association, which encouraged planners to use it complementary to more traditional estimating methods (Flyvbjerg, 2008).

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 regression of the initial intuitive estimate towards the reference class average actually takes place.

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 subsequent treatments as historical information. We investigate whether they use this information to improve their estimation accuracy. Unlike in previous studies, the individual estimating the duration of the task is also the one who executes the task. This feature allows us to recreate incentives faced within a business project. In addition, by carefully manipulating the timing when the historical information is disclosed, we separate the anchoring effect from the regression effect. Furthermore, we examine whether the accuracy can be improved by a traditional approach of providing more detailed project specification and compare the effectiveness of detailed specification against historical information. While providing detailed project specification is endorsed by project management methodologies (IPMA, 2015; Project Management Institute, 2013), we are not aware of any study in the area of duration estimation that tests the effectiveness of such approach, let alone compares it against other interventions.

3. Experimental design

In our experiment, we test whether (i) disclosing historical information and (ii) providing a more detailed task specification can induce more accurate and less biased duration estimates. The experiment consists of four treatments, implemented in an across-subject design: Baseline, Information Before Estimate (henceforth "Info-Before"), Information After Estimate (henceforth "Info-After"), and Detailed Description.

The experimental task

We employ a modified version of the individual task introduced by Mazar, Amir, & Ariely (2008), in which subjects search for two numbers that add up to 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 provided in the appendix. 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", the 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.

Treatments

In the Baseline treatment, no historical information is provided. Subjects read the instructions with the description of the experimental task and then estimate how much time they would need 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 complete an incentivized risk attitude assessment (Holt & Laury, 2002) as well as a demographics questionnaire.

In the Info-Before and Info-After treatments, subjects also receive information about the average actual task duration recorded in the Baseline treatment (18 minutes and 13 seconds). We operationalize the historical information as a single data point, in order to be able to draw a clear inference regarding the adjustment of the estimate towards the average. This would not be the case if the whole distribution was provided because one could not attribute the effect to a particular information from the distribution. In addition, it is arguably easier to interpret information conveyed as a simple average compared to a whole distribution of outcomes.

In the Info-Before treatment, the historical information is disclosed before the estimation. In contrast, in the Info-After treatment, subjects receive the information only after they have provided their estimate and confidence rating. Once the historical information is disclosed, subjects in the Info-After treatment are given an opportunity to revise their estimate and confidence rating. To calculate the earnings, we use the revised estimate, as explained in the on-screen instructions.

In the Detailed Description treatment, we provide subjects with a more informative task description. In particular, subjects are shown a sample matrix in the instructions 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 the correct answer.

Incentives

We financially incentivize subjects for their estimation accuracy as well as task performance, motivating them to provide accurate task duration estimates, and at the same time to execute the task fast and avoid mistakes. By providing incentives for both accuracy and performance, we create an environment similar to duration estimation in business practice, in which it is not only the project schedule accuracy that counts, but also the speed of project delivery. The earning functions are illustrated in Figure 1.

Estimation accuracy incentives follow a linear scoring rule, in which the earnings depend on the absolute difference between the actual task duration and the estimate. The earnings peak at AUD 18 for a spot-on estimate and decrease by AUD 2.40 for every minute (i.e., 4 cents for every second) the estimate differs from the actual task duration, as shown in Equation (1). An equal penalization in both directions instead of, say, heavier penalty for being late, discourages undesirable strategic behavior, such as inflating estimates to minimize chances of not finishing "on time". If the difference between the estimated and actual time is larger than 7.5 minutes (450 seconds), we set the estimation accuracy earnings to 0 to avoid negative earnings.[6] Note that our experimental setting is similar to what planners often experience in business practice – while accurate estimates leading to successful project completion are commonly rewarded, planners are typically not punished if their estimates are inaccurate.

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

The task performance earnings are calculated based on the actual task duration and on the count of correct (being always equal to 10) and incorrect answers, as shown in Equation (2). The faster the task execution and the fewer mistakes, the higher the earnings. Penalizing subjects for incorrect answers disincentivizes random clicking, guessing, or systematic trying of all combinations. An incentive structure that encourages not only speed but also quality parallels incentives encountered in business situations. We expected subjects to complete the task in 15 minutes (900 seconds) on average.

[6] We derived the 450-second threshold from the median task duration observed in pilot experiments (approximately 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 (Project Management Institute, 2013) for duration estimation requires the Rough Order of Magnitude estimates to fall in the range of +75%/-25% from the actual duration. Since our estimation accuracy earnings are symmetric for underestimation and overestimation, we implemented a range of +/-50%.

Without incorrect answers, the expected pace would earn subjects AUD 10 for their task performance, making the earnings from task performance comparable with the expected estimation accuracy earnings. We reward task performance separately instead of incorporating the performance and estimation accuracy together into one payoff function, because it is easier to understand for subjects and also because it preserves the motivation to continue performing the task even if the subject realizes that his estimate was too low and his estimation accuracy earnings will likely be zero.

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

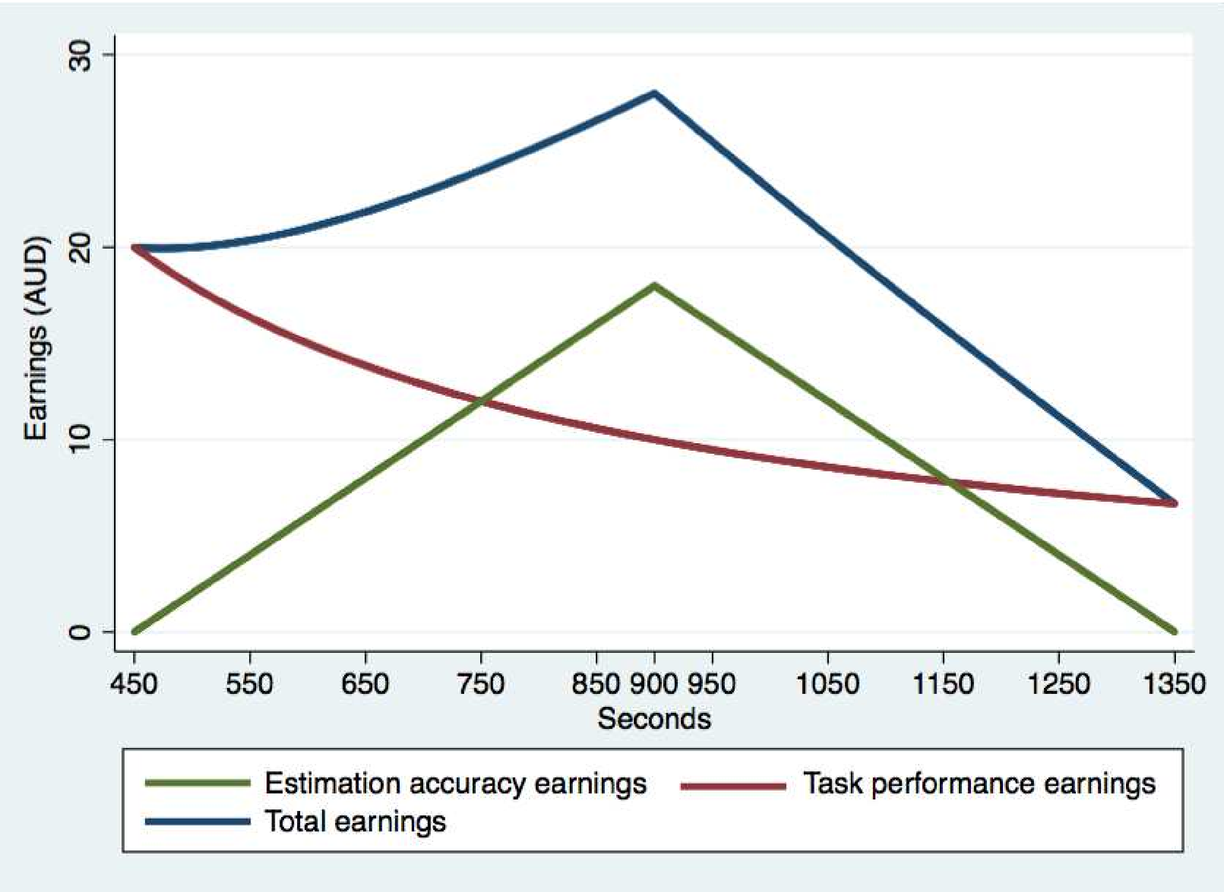

Figure 1: Earning functions

Notes: Figure 1 illustrates the earning functions for the scenario of estimate being 900 seconds, which is what we expected the representative task duration to be. Incidentally, the median actual task duration across all subjects participating in the experiment turned out to be 906 seconds (see Table 1 for the breakdown according to treatments). The highest earnings (resulting from finishing the task in exactly 900 seconds while making no mistakes) in this scenario yield AUD 28, with AUD 18 being for estimation accuracy and AUD 10 for task performance. Note that given the slope of the task performance earnings function, a subject in this scenario could earn more than AUD 28 solely for task performance, by finishing the task in 321 seconds or less; however, the fastest recorded actual task duration in the experiment was 361 seconds.

Since there are two types of incentives, there is a possibility that subjects might construct an earnings portfolio (Cox & Sadiraj, 2018). Although the portfolio effect can be controlled for by randomly selecting only one type of incentives for payment (Cox, Sadiraj, & Schmidt, 2015; Holt, 1986), we opt to preserve the parallelism and minimize the likelihood of portfolio effect by a careful experimental design. 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 (Lorko, Servátka, & Zhang, 2020). 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.

Procedures

The experiment was programmed in z-Tree software (Fischbacher, 2007) and conducted in the MGSM Vernon L. Smith Experimental Economics Laboratory at the Macquarie Graduate School of Management in Sydney. Subjects, consisting mostly of undergraduate business major students and MBAs, were recruited using ORSEE (Greiner, 2015). Before the start of the experiment, subjects sitting in individual cubicles were asked to put away their watches, mobile phones and any other devices displaying time, to prevent them from measuring the elapsed time. The laboratory premises did not contain any devices that show time. The clocks on computer monitors 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 to the decision-making part. At the end of the experiment, subjects privately received their experimental earnings in cash in the control room at the back of the laboratory.

4. Hypotheses

Historical information

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 underestimate its duration. We hypothesize that the underestimation will be mitigated by the historical information in the Info-Before and Info-After treatments. We further conjecture that estimates in the Info-Before treatment will be more accurate than the revised estimates in the Info-After treatment. It is conceivable that in the Info-After treatment, subjects may be reluctant to fully incorporate the historical information in their estimation due to cognitive dissonance or the cost of cognitive effort. The adjustment of the initial estimate towards the historical information might thus be insufficient. Therefore, we expect to find unbiased estimates that are similar to the actual task duration and 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 we incentivize subjects in all treatments for their estimation accuracy as well as task performance, we hypothesize that there will be no differences in the distributions of the actual duration across our treatments, akin to earlier findings (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 the lowest estimation accuracy.

- *Hypothesis 2*
    - $Duration_{BASELINE} = Duration_{INFO\text{-}AFTER} = Duration_{INFO\text{-}BEFORE}$

- $Accuracy_{BASELINE} < Accuracy_{INFO\text{-}AFTER} < Accuracy_{INFO\text{-}BEFORE}$
- $Bias_{BASELINE} > Bias_{INFO\text{-}AFTER} > Bias_{INFO\text{-}BEFORE}$

Detailed description

Due to the omission in the task description that leads subjects to expect integer numbers in the matrices, the task seems easier in the Baseline treatment compared to 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 3*
  - $Estimates_{BASELINE} < Estimates_{DETAILED\ DESCRIPTION}$

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

- *Hypothesis 4*
  - $Duration_{BASELINE} = Duration_{DETAILED\ DESCRIPTION}$
  - $Accuracy_{BASELINE} < Accuracy_{DETAILED\ DESCRIPTION}$
  - $Bias_{BASELINE} > Bias_{DETAILED\ DESCRIPTION}$

5. Main results

A total of 139 subjects, randomly assigned into our four treatments, participated in the experiment. However, 9 of those subjects (5 in Baseline, 1 in Info-After, 1 in Info-Before, and 2 in Detailed Description) found the task too difficult and gave up before completing the experiment. We thus analyze only the behavior of the remaining 130 subjects (59 females) with a mean age of 22.7 a standard deviation of 4.2 years. Of these subjects, 38 participated in the Baseline treatment, 29 in the

Info-After treatment, 29 in the Info-Before treatment and 34 in the Detailed Description 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 17.20. 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 the revised estimates for testing the treatment effects. The results of treatment effects are presented in Table 2, while the individual-level data are graphically displayed in Figure 2.

Table 1: Summary statistics (data in seconds)

	Treatments	Baseline (N = 38)	Info-After (N = 29)		Info-Before (N = 29)	Detailed Desc. (N = 34)
			Initial est.	Revised est.		
Estimates	Means (SD)	601 (704)	456 (427)	814 (377)	798 (329)	1149 (1287)
	Medians	270	300	900	900	525
Actual duration	Means (SD)	1093 (573)	986 (528)		914 (404)	1144 (565)
	Medians	919	847		818	1017
Bias	Means (SD)	-492 (757)		-171 (521)	-115 (365)	5 (1369)
	Medians	-539		-164	-68	-211
Accuracy (Absolute error)	Means (SD)	725 (530)		425 (338)	275 (262)	1012 (904)
	Medians	682		412	184	734

Notes: SD refers to standard deviation. The bias is calculated as a relative estimation error (= Estimate – Actual duration) averaged across subjects, while the (in)accuracy is measured as the absolute value of the estimation error (= |Estimate – Actual duration|) averaged across subjects.

Historical information

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, henceforth “M-W”, $p < 0.01$). 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, subjects in the Info-After treatment provide similar estimates as the subjects in the Baseline treatment (M-W, p = 0.98), with the median estimate being 270 seconds in Baseline and 300 seconds in Info-After. However, upon disclosing the historical information, estimates in the Info-After treatment significantly increase (with the median being 900 seconds), as the subjects adjust their initial beliefs towards the historical average (Wilcoxon matched-pairs signed-ranks test, p <0.01). These revised estimates are significantly higher than the estimates in the Baseline treatment (M-W, p <0.01) and similar to the estimates in the Info-Before treatment (M-W, p = 0.73). As for the task execution, we do not find differences in the actual task duration or in the number of incorrect answers across the three treatments, resulting in similar task performance earnings.

Table 2: Treatment effects (p-values of the Mann-Whitney tests)

	Baseline vs. Info-After	Baseline vs. Info-Before	Info-After vs. Info-Before	Baseline vs. Detailed Description
Estimates	<0.01	<0.01	0.73	0.049
Actual duration	0.29	0.21	0.71	0.66
Bias	0.02	<0.01	0.76	0.04
Absolute error	<0.01	<0.01	0.09	0.33

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 does not influence the estimates. The actual task duration does not differ across the three 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 the lowest accuracy. As 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 our previous results with the bias being significantly larger and the accuracy significantly lower in the Baseline treatment than in both the Info-Before and Info-After treatments. We do not find significant differences in bias and accuracy between the Info-After treatment and the Info-Before treatment.

Out of 58 subjects pooled from the two treatments with historical information, 49 subjects (84%) provided an estimate lower than the historical average, and 9 subjects (16%) provided a higher estimate. While this might resemble overconfidence, we note that the distribution of virtually any task duration is typically skewed to the right, meaning that over 50% of outcomes are lower than the average (in our Baseline treatment, it is 63%). In the two treatments with historical information, we find a significant positive correlation between the estimates and the actual task duration (Pearson correlation, $r = 0.43$, $p < 0.01$ for pooled data), suggesting that subjects have relatively well calibrated expectations of their own performance when the historical average is available. We do not find a significant correlation between the estimates and the actual task duration in the Baseline treatment (nor in the Detailed Description treatment).

Result 2: The estimates in the Baseline treatment exhibit the largest estimation bias and the lowest estimation accuracy. Providing historical information in the Info-After and Info-Before treatments decreases the bias (resulting in less underestimation) and improves the accuracy.

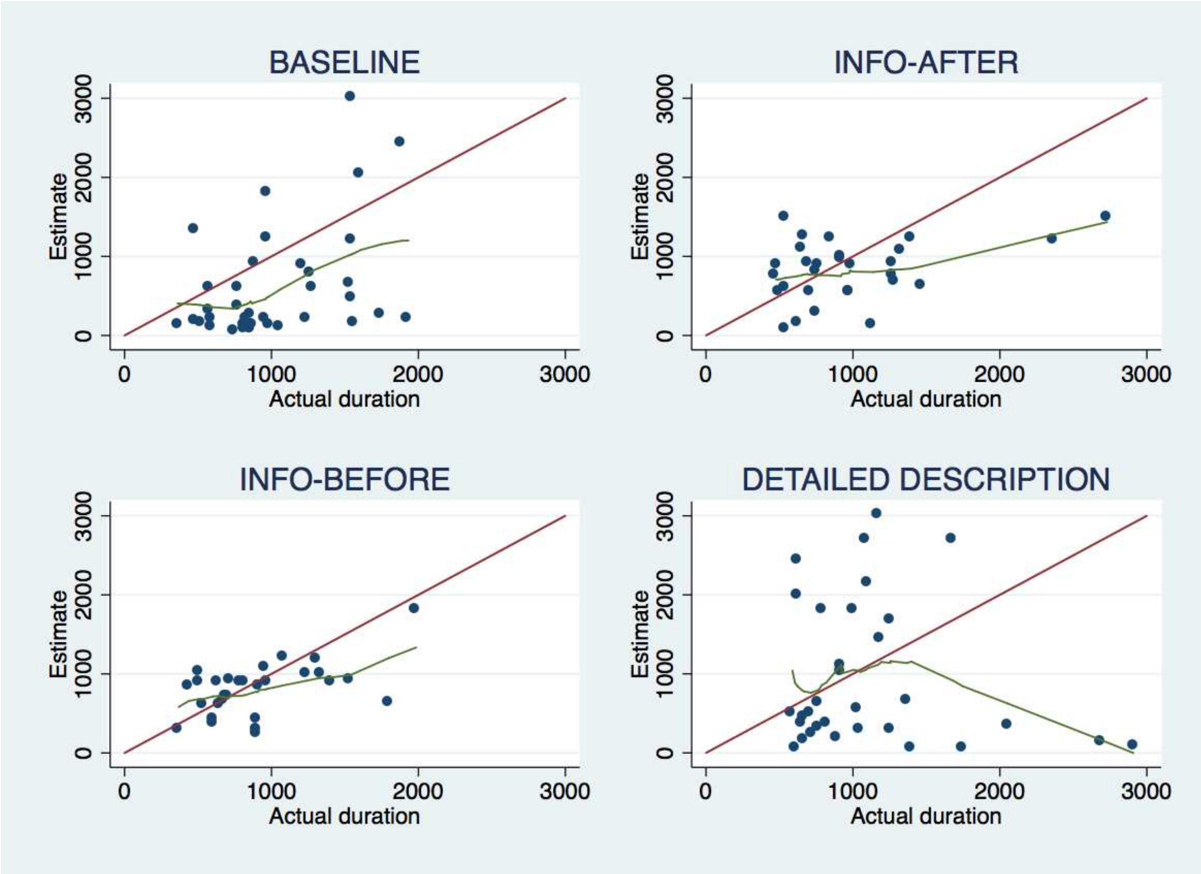

Figure 2: Individual-level estimates and the actual task duration

Notes: Figure 2 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. Any dot above the red line indicates overestimation, while a dot below the red line indicates underestimation. The green line represents the Lowess smoothing of estimates on actual duration with the weight of running-line least squares. For presentational clarity purposes, 9 outliers were removed (4 subjects from Baseline, 1 from Info-After, 1 from Info-Before, and 1 from Detailed Description who spent more than 3000 seconds to complete the task, and 2 subjects from Detailed Description whose estimates were higher than 3000 seconds).

Detailed description

Next, we compare the behavior in the Baseline treatment to the behavior in the Detailed Description treatment. In line with our Hypothesis 3, 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 (M-W, $p = 0.049$). The actual task duration does not differ between the two treatments (M-

W, p = 0.66), and the same holds for the number of incorrect answers and hence also for the task performance earnings.

The subjects in the Detailed Description treatment exhibit a mean estimation bias of only 5 seconds, which is significantly lower than in the Baseline treatment (M-W, p = 0.04). However, the small bias itself does not necessarily imply high estimation accuracy, which depends on the severity of over- and underestimates. In the Detailed Description treatment, 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 2). Although the subjects provide unbiased estimates on average, their estimation accuracy appears slightly lower than in the Baseline treatment, but the difference is not statistically significant (M-W, p = 0.33). Our Hypothesis 4, stating that providing a more detailed description leads to less biased and more accurate estimates is supported only partially. Interestingly, we also find that it takes on average 48 seconds for the subjects in the Detailed Description treatment to provide their estimates, which is 10-17 seconds longer than in any other treatment. The difference is weakly statistically significant compared to the Baseline treatment (M-W, p = 0.06). The result indicates that subjects might have hard time to grasp the complexity of the task based on the detailed description, resulting in large estimation errors.

Result 3: Providing a more detailed task description mitigates the underestimation bias but does not improve the estimation accuracy.

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

The common Baseline treatment allows us to directly compare the effect of the two implemented interventions. We find that both providing historical information and providing a more detailed task description mitigates the underestimation of the time necessary to complete the task. Compared to the Baseline, the estimation bias is significantly reduced in the Info-Before treatment (M-W, p <0.01), the Info-After treatment (M-W, p = 0.02) as well as in the Detailed Description treatment (M-W, p = 0.04). In contrast, we find a similar estimation bias in all comparisons across the three treatments with an intervention (M-W, p = 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).

Regarding the estimation accuracy, we find that the historical information intervention is effective, while the detailed description intervention is not. The absolute estimation error is reduced (against the Baseline) in the Info-Before treatment (M-W, p <0.01) and the Info-After treatment (M-W, p <0.01), but not in the Detailed Description treatment (M-W, p = 0.33). Furthermore, we find no statistically significant differences in the estimation accuracy between the Info-Before and the Info-After treatments (M-W, p = 0.09). We do, however, find that subjects in the Detailed Description treatment are less accurate than subjects in both the Info-Before treatment (M-W, p <0.01) and the Info-After treatment (M-W, p <0.01). Thus, in terms of the estimation accuracy, the effect of historical information significantly outperforms the effect of more detailed task description.

Result 4: Providing historical information as well as providing detailed task description significantly reduces the underestimation bias. However, only the provision of historical information also significantly improves the estimation accuracy.

Robustness

To verify the robustness of the effect of our interventions, we conduct a regression analysis, controlling for risk attitudes, time spent on estimation, time spent on indicating confidence, subjective confidence in own estimate, and demographics (age, gender, education, employment status and self-reported math skill). The regression results (presented in the appendix, Table 4) 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 of historical information significantly improves the estimation accuracy. Again, we find no effect of any intervention on the actual task duration. Furthermore, we find that higher confidence is associated with lower estimates but has no effect on the actual task duration and estimation accuracy. Finally, we find a significant negative correlation between the self-reported math skill and the actual task duration. This observation is in contrast with Mazar et al. (2008, p. 636) claim that subjects "did not view this task as one that reflected their math ability or intelligence". 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, possibly reversing the causality.

7. Subjective confidence in estimates and willingness to pay

Are people able to differentiate between various degrees of ambiguity embedded in different specifications of the same project? And how valuable do they perceive the historical information to be? To provide additional behavioral insights, we analyze the elicited subjective confidence in estimates and willingness to pay for historical information across all four treatments.

Subjective confidence in estimates

Intuitively, estimates based on a more detailed task description, or supported by historical information, could be produced with higher confidence. Thus, one might expect subjects in the Baseline treatment to be less confident in their estimates than in any other treatment. However, our across-subject design makes it difficult for subjects to realize that some essential information might be missing. Hence, the "what you see is all there is" rule predicts subjects to focus only on the tangible information and remain unaware of what they do not know, resulting in similarly high confidence in estimates in all treatments. To test the rule of "what you see is all there is", we asked subjects to indicate their subjective confidence in the accuracy of their 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 (with the assigned value of 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)	Detailed Description (N = 34)
		Initial est.	Revised est.		
Mean confidence (SD)	3.7 (0.8)	3.5 (0.7)	3.8 (0.7)	3.7 (0.8)	3.5 (0.9)
Median confidence	4	4	4	4	4

Notes: SD refers to standard deviation.

Summary statistics are presented in Table 3. The subjective confidence in estimates is similar across all treatments (Kruskal-Wallis test, $p = 0.52$). In general, subjects report relatively high confidence in their estimates, as the median confidence in all treatments is 4 out of the maximum of 5, which supports the "what you see is all there is" rule. Subjects display similar confidence in estimates irrespectively of whether they received historical information prior to the estimation, and also irrespectively of how detailed the task description was. Importantly, the confidence in estimates does not significantly correlate with estimation inaccuracy (errors) in any treatment. For pooled data, the Pearson correlation coefficient yields $r = -0.1$, and $p = 0.27$.

Result 5: Subjective confidence in estimates is not affected by the amount or detail of available information. Subjects display similar level of confidence regardless of what they know about the task.

Willingness to pay for historical information

Finally, to investigate whether 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, we asked subjects to consider that historical information was not given for free and requested to state the maximum amount they would be willing to pay in order to obtain such information. In the Baseline treatment and the Detailed Description treatment, we asked subjects to consider that there was historical information available before the estimation and state the maximum amount they would be willing to pay for it. From the analysis we eliminated 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. The median willingness-to-pay is AUD 5.00 (the average is AUD 5.45) in treatments with historical information (pooled Info-Before and Info-After) and AUD 3.00 (AUD 3.53) in treatments without historical information (pooled Baseline and Detailed Description). The difference in willingness-to-pay is weakly statistically significant (M-W, $p = 0.06$). With this caveat, we speculate that the subjects in treatments with historical information are willing to pay more because they have experienced the benefits of the information.[7]

We also report an interesting observation. Subjects in treatments with historical information earn on average AUD 4.81 more for their estimation accuracy than subjects in treatments without historical information. These "additional" accuracy earnings are similar to the median amount of AUD 5.00 that the subjects in treatments with historical information are willing to pay for the information. While this result might be due to chance, it points out that although the historical information is costly, it often has positive return on investment and that people who have actually used the historical information are more aware of what it is worth.

8. Discussion

An adequate business project schedule is essential for project success and plays a key role in effective allocation and utilization of resources in an organization. In this paper, we investigate the effectiveness of two interventions designed to induce more accurate duration estimates within the

[7] We note that the difference is not significant if we include the eliminated subjects.

project planning process: providing historical information and providing a more detailed project specification. In the task description, 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 large gap between the intuitive estimate and the time necessary to complete the task. We show that the utilization of historical information in the planning process can significantly mitigate the underestimation bias and improve the estimation accuracy. We further find that the timing when the information is provided does not play a role in our experimental setting. 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 initial estimates towards historical averages may be seen as poor competence of the planner.

One could argue that not disclosing crucial information regarding the nature of the task makes its description perhaps too uninformative. While it is possible, we note that virtually any project specification is a simplification of the actual deliverables as organizations 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 a treatment in which a sample matrix is added to the task description. We find that a more detailed specification eliminates the estimation bias (in particular underestimation), which becomes almost zero when averaged across all subjects, 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 compared to the situation when only crude specification is provided, 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, encouraging planners to utilize reliable historical information and nudging them towards the 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 not a reliable predictor of estimation accuracy. Our subjects provided essentially identical 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 (especially the "unknown unknowns") 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.

In our experiment, we have taken a conservative approach of creating weak interventions by providing only a single piece of additional information in each treatment, designed to pick up the lower bound of their effect on estimation bias and (in)accuracy. Remarkably, we observe that both minimal interventions successfully eliminate the underestimation bias. While our results are promising, we note that this area of research deserves more attention. It would be worthwhile to investigate additional enhancements of our interventions and test whether they further improve the estimation accuracy. For example, it would be interesting to explore whether the effectiveness of utilizing historical information can be strengthened by disclosing also the variance, quartiles or even the whole distribution of outcomes in addition to the class average. Future research could also shed light on the interaction between the estimation accuracy and experience, e.g., by giving subjects an opportunity to practice the task before the estimation. Furthermore, since a project is usually executed by more than one person and past research (Staats, Milkman, & Fox, 2012) shows that the underestimation of effort needed to deliver a project tends to increase with larger project team size, testing the effectiveness of historical information when team size is varied could be another natural extension of the current study.

Another limitation 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 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 identical. To consult historical averages in such environment, planners must first carefully select a meaningful reference class of past projects.

In case the project is so unique that there is no prior undertaking to compare it to, or if the historical information from similar projects is unavailable, it might be helpful to assess the effect of providing information about the differences between the estimates and the actual duration from other projects. Acquiring historical information about related projects may be costly (e.g., because of search costs) and if planners do not consider the information valuable, they may be reluctant to seek it. In the current study, we elicit the willingness to pay for historical information ex-post and observe that subjects who have experienced the benefits of using such information value the information more. A deeper scientific inquiry into the process of reference class selection and a salient elicitation of willingness to pay for historical information are other 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, R., 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

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

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*.

Eubanks, D. L., Read, D., & Grushka-Cockayne, Y. (2015). Biases as constraints on planning performance. In M. D. Mumford & M. Frese (Eds.), *The Psychology of Planning in Organizations: Research and Applications* (pp. 229–242). Routledge.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2008.11.010

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

Forsyth, D. K., & Burt, C. D. B. (2008). Allocating time to future tasks: the effect of task segmentation on planning fallacy bias. *Memory & Cognition*, *36*(4), 791–798. https://doi.org/10.3758/MC.36.4.791

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

Goodwin, P., Moritz, B., & Siemsen, E. (2019). Forecast decisions. In *The Handbook of Behavioral Operations*. https://doi.org/10.1002/9781119138341.ch12

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

Grushka-Cockayne, Y., Erat, S., Wooten, J., Donohue, K., Katok, E., & Leider, S. (2018). New product development and project management decisions. In *In The Handbook of Behavioral Operations* (pp. 367–392). NJ: Wiley.

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

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. (1977). Intuitive prediction: biases and corrective procedures. *Technical Report. Advanced Decision Technology*, *12*. https://doi.org/citeulike-article-id:3614496

Kelley, J. E. (1961). Critical-Path Planning and Scheduling: Mathematical Basis. *Operations Research*. https://doi.org/10.1287/opre.9.3.296

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

Kremer, M., Moritz, B., & Siemsen, E. (2011). Demand forecasting behavior: System neglect and change detection. *Management Science*. https://doi.org/10.1287/mnsc.1110.1382

Kruger, J., & Evans, M. (2004). If you don't want to be late, enumerate: Unpacking reduces the planning fallacy. *Journal of Experimental Social Psychology*. https://doi.org/10.1016/j.jesp.2003.11.001

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

Lorko, M., Servátka, M., & Zhang, L. (2020). *Hidden inefficiency: Strategic inflation of project schedules* (Working paper).

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

Massey, C., & Wu, G. (2005). Detecting regime shifts: The causes of under- and overreaction. *Management Science*. https://doi.org/10.1287/mnsc.1050.0386

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

Newby-Clark, I. R., Ross, M., Koehler, D. J., Buehler, R., & Griffin, D. (2000). People focus on optimistic scenarios and disregard pessimistic scenarios while predicting task completion times. *Journal of Experimental Psychology: Applied*. https://doi.org/10.1037/1076-898X.6.3.171

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., 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
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
Staats, B. R., Milkman, K. L., & Fox, C. R. (2012). The team scaling fallacy: Underestimating the declining efficiency of larger teams. *Organizational Behavior and Human Decision Processes*. https://doi.org/10.1016/j.obhdp.2012.03.002
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

Appendix

Table 4: Linear regression analysis (Ordinary least squares regeression)

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

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

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.

<u>**Task**</u>

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.

<u>**Earnings**</u>

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:

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

$^{\times}$ 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.

Instructions

(note: used for the 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.

48.47	54.94	74.77	34.22
56.26	87.77	69.78	75.36
72.86	30.22	60.15	79.39
23.01	72.09	26.34	84.94

Correct answer for this sample matrix

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:

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

$^{\times}$ 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.