Replication of "Procrastination, Deadlines, And Performance: Self-Control by Precommitment"

Kyle Hyndman[1] and Alberto Bisin[2]
[1]Naveen Jindal School of Management
The University of Texas at Dallas
[2]Department of Economics
New York University

Abstract

We present the results of a replication of Study 2 from Ariely and Wertenbroch (2002) on the role of deadlines in mitigating procrastination. We show that our results do not replicate those of the original paper. In our replication, changes in the deadlines have a negligible effect on the three performance metrics and several survey metrics used in the original study. Evenly spaced deadlines, externally imposed on participants by the experimenters, do not stand out for their effectiveness in reducing procrastination in participants. Our data further show consistency with several patterns of participant behavior that should be expected irrespective of our main replication results.

Keywords: Replication, Deadlines, Procrastination

Research Transparency Statement

General disclosures

Conflicts of interest: All authors declare no conflicts of interest. **Funding:** This research was supported by the University of Texas at Dallas and New York University. **Artificial intelligence:** Artificial intelligence was used for grammatical checking and general proofreading. **Ethics:** This research received approval from one author's Institutional Review Board (ID: IRB-24-151).

Study disclosures

Preregistration: The hypotheses and methods were preregistered on January 29, 2025 (the main study) and on January 30, 2024 (the supplemental study). In both cases, pre-registration was done prior to any data collection. **Materials:** All study materials are publicly available. **Data and Analysis Scripts:** All primary data and analysis scripts are publicly available. All data, code, study materials, and preregistration can be found at https://researchbox.org/3063&PEER_REVIEW_passcode=BDSOCK.

Introduction

Procrastination is ubiquitous in modern society, with self-help books (Steel, 2011) and dozens of academic papers in psychology, economics, and numerous other disciplines.[1] Given its prevalence, researchers have devoted considerable attention to understanding how people can overcome procrastination. Psychologists and management scholars have pointed to goal setting as one way to overcome procrastination (e.g., Locke & Latham, 2006). In this literature, psychologists and economists have also studied deadlines and whether they may be effective tools to overcome procrastination. According to Temporal Motivation Theory (Steel, 2007) – a leading theory for procrastination in psychology – deadlines may help overcome procrastination because one's motivation increases as the deadline approaches. Economists have also modeled how deadlines may mitigate procrastination by decision makers with self-control problems, typically modeled by the $\beta - \delta$ model of hyperbolic discounting (O'Donoghue & Rabin, 1999, 2001).

While it is theoretically possible that deadlines may increase task completion, they are not guaranteed to do so and could, in fact, hurt. Most concretely, a deadline reduces the option value of completing a task at a later date, where it could be done under potentially more favorable terms. Whether a deadline helps or hurts depends intricately on this option value, which itself depends on various psychological and behavioral factors such as the extent of hyperbolic discounting and a decision maker's knowledge of their own self-control problems (Hyndman & Bisin, 2022; O'Donoghue & Rabin, 2001). Ordóñez et al. (2009) provided a more psychological perspective on some of the potential pitfalls of an over-reliance on goals, and Self-Determination Theory (Deci & Ryan, 2000; Ryan & Deci, 2000) spoke to some potentially negative effects of deadlines – particularly externally imposed deadlines – whereby a deadline reduces one's sense of autonomy, which could reduce intrinsic motivation and lead to worse outcomes.

Not surprisingly, within this large literature, experimenters have examined the role of deadlines in overcoming procrastination. Arguably, the most well-known and impactful paper addressing this trade-off experimentally is Ariely and Wertenbroch (2002).[2] The authors opened their paper with three general observations: (1) most of us procrastinate; (2) deadlines are a common feature of life; and (3) many people seek commitment devices to help them achieve their goals. Ariely and Wertenbroch then asked three questions. First,

---

[1] A Google Scholar search conducted on March 28, 2025, returned over 150 papers with 'procrastination' in the title.

[2] The paper has been cited over 2000 times as of April 2025. The Google Scholar page for Ariely and Wertenbroch (2002) is https://scholar.google.com/citations?view_op=view_citation&hl=en&citation_for_view=Z1G9Lk4AAAAJ:qjMakFHDy7sC.

will participants self-impose deadlines on themselves? Second, are self-imposed deadlines an effective tool to induce participants to complete tasks on time? Finally, do participants optimally self-impose deadlines? The original paper presented a series of studies to answer these questions, but the one we focus on is Study 2 – the Proofreading Study.

In the Proofreading Study, participants had three weeks to complete three tasks. Each task consisted of finding a number of spelling and grammatical errors that were inserted into texts that were generated by a post-modern text generator. Participants were rewarded according to a piece-rate (10 cents per correctly identified error), but also faced penalties of $1 per day that a task was submitted after its deadline. Participants were assigned to one of three treatments: evenly spaced deadlines (imposed by the experimenters), self-imposed deadlines, or no intermediate deadlines. Ariely and Wertenbroch (2002) presented data showing that a substantial number of people were willing to self-impose deadlines on themselves and that the deadlines were partially effective. However, their main result was that performance was highest in the evenly spaced deadlines treatment, where the deadlines were externally imposed.

Two subsequent papers, written by economists, have provided experimental evidence that deadlines may not be as effective a tool to overcome procrastination as suggested by Ariely and Wertenbroch (2002). Implementing the same three treatment conditions, but with a different task and a different incentive structure, Bisin and Hyndman (2020) showed that performance is actually worse in the presence of deadlines. The authors argued that this was because, in the presence of deadlines, participants delayed work until near the deadline and were unable to complete the task on time due to poorly formed expectations about task difficulty. Similar to Bisin and Hyndman (2020), Burger et al. (2011) showed that binding intermediate deadlines actually harmed participants' ability to record 75 hours of studying time inside a monitored location over a five-week period.

Beyond these two closely related experiments, Epton et al. (2017) conducted a meta-analysis of the effectiveness of goal setting in general. While they found a positive overall effect, their analysis contained some important moderating factors on their potential effectiveness that are directly relevant to the experimental setting of Ariely and Wertenbroch (2002). In particular, they documented that goals are most likely to be effective when the task is difficult, the goals are set publicly, and the goal is a team goal. Additionally, goals are less effective in a university setting. Interestingly, despite the slight differences in tasks and incentives between Ariely and Wertenbroch (2002) and Bisin and Hyndman (2020), all of the moderators identified by Epton et al. (2017) are broadly similar, making the starkly different results on the effectiveness of deadlines worthy of further investigation.

All of this discussion, in our opinion, makes an attempted close replication more important. The differences in designs can merely suggest a potential lack of robustness to changes in the structure of tasks, incentives, and deadlines, while a close replication can speak more forcefully to the robustness of the result itself.

We approach our investigation within the context of the broader questions about both the replicability of many results in psychology, economics, and other disciplines (e.g., Camerer et al., 2016, 2018; Open Science Collaboration, 2015, among others) and the potential use of questionable research practices (Lewis-Kraus, 2023; Stern, 2023) in some instances. Like many social scientists, we believe that it is important to test the replicability/reproducibility of important papers that are influential both within academic circles and public opinion more broadly. With its high citation count and its numerous representations reported in the press (examples include Jaffe, 2014; Surowiecki, 2010; Thompson, 2014), we believe that Ariely and Wertenbroch (2002) clearly merits the scrutiny of a replication.

In the rest of the paper we show that our results do not replicate those of the original paper. In our replication, changes in the deadlines have a negligible effect on the three performance metrics and several survey metrics used in the original study. Evenly spaced deadlines, externally imposed on participants by the experimenters, do not stand out for their effectiveness in reducing procrastination in participants.

Method

We decided to replicate Study 2 from the original paper as closely as possible. Study 1 and the pilot study were both classroom studies, making any attempted replication virtually impossible. In Study 2, participants were asked to read and provide spelling and grammatical corrections for three different texts that were generated with a post-modern text generator[3] and had 100 errors inserted into them. In January 2014, we contacted the original authors requesting the study's instructional materials. Wertenbroch replied that he didn't "know the experimental details as Dan ran the studies at MIT," and Ariely replied, "Sorry but I don't have the instructions."

Faced with a lack of original materials, we necessarily faced choices and cannot guarantee that every procedure was identical to the original study. We outline our methods now. First, following the original authors, we used the same post-modern text generator to generate texts, and we inserted 100 spelling and grammatical errors in each text. Our texts were selected to be between 2000 and 2400 words, divided over 8 or 9 pages. Second,

[3] The post-modern text generator is available at https://www.elsewhere.org/pomo/. The information on the website suggests that the text-generator has not been modified subsequent to launch in February 2000, meaning that the texts we used are substantially similar to those in the original study.

participants were recruited to the study via an email from the participant pool at one author's home institution. They were informed that there would be a short in-person meeting, followed by work that would be done outside of the lab on their own time. As in the original paper, participants were informed that they must be native/fluent English language speakers, and they were required to complete a short survey before registering to participate, where they verified their language proficiency.

Upon arrival to the initial lab session, registered participants were given a login code to the online platform (SoPHIE), where the experiment would take place both within and outside the lab. Once on the platform, participants received instructions about the tasks, their compensation, and deadlines. Participants were randomly assigned to one of three deadline treatments: 1. no intermediate deadlines (End deadline), 2. exogenous and evenly spaced deadlines (Evenly spaced deadlines), and 3. self-imposed deadlines.[4] This randomization was done at the participant level, via the login codes that they received.

After reading the instructions, participants in the evenly spaced deadlines and end deadline treatments were given the deadline (date and time) for each of the three tasks. For example, participants in the evenly spaced deadlines treatments saw:[5]

- Task 1 Deadline: February 11, 2025 at 07:00pm
- Task 2 Deadline: February 18, 2025 at 07:00pm
- Task 3 Deadline: February 25, 2025 at 07:00pm.

Participants in the end deadline treatment would see the same date and time for all three tasks. Participants in the self-imposed deadlines treatment were given the option but not the obligation to impose binding deadlines on any or all of the three tasks. Except for the earliest and the latest allowable deadlines, deadlines were otherwise unrestricted. Participants set deadlines separately for each task (any date and time within the allowed bounds), and deadlines for different tasks could be the same or different from each other according to their preference. See Figure A1 for the deadline choice screen used in the study.

After completing all of the in-person requirements, participants received an email that contained their login code to access the experimental platform later, as well as a

[4] For ease of comparison, we use the same terminology to describe the treatment conditions as the original authors.

[5] Sessions were conducted over two separate days. The dates given correspond to the deadlines for one of the session days. Participants who came on the other date would have seen appropriately adjusted dates.

reminder of the deadlines that they faced.[6] Absent any intermediate deadlines determined by the treatment (i.e., evenly spaced deadlines treatment) or by a particular participant (i.e., self-imposed deadlines treatment), there was an overall deadline of three weeks from the start of the experiment, after which point penalties would start to accrue. The experimental platform remained open for 10 days beyond this three-week deadline. For example, in one of the sessions, the instructions noted:

> This study runs from **February 04, 2025 at 07:00pm** to **February 25, 2025 at 07:00pm**, although the platform will remain open until **March 08, 2025 at 07:00pm** for any late submissions and a brief final survey.

We chose to keep it open for 10 days after the three-week run time because this is the point at which it would be impossible to earn money from completing a task.[7] This information was displayed on the deadlines screen (see again Figure A1).

For those participants who successfully submitted all three tasks, they were asked a short series of questions. Following Ariely and Wertenbroch (2002), we asked them to provide a subjective evaluation of the tasks on five dimensions (e.g., how much they liked it, the quality of the texts, etc.). Participants were asked to enter an integer between 0 and 100, which we believe is identical to the original authors (Ariely & Wertenbroch, 2002, p. 223). We also asked participants to provide an estimate of the time they spent on each task (in minutes), and we restricted responses to being between 0 and 180 minutes.[8] Lastly, we asked participants to respond in a text box to the following prompt about how they perceived deadlines to affect their performance:

> Please comment on how the deadlines for the tasks that you faced either helped or hindered your ability to successfully complete the tasks. If you were given a chance to change the deadlines, which deadlines would you choose?

In Appendix A, we provide a description of all of the variables that we directly collected or derived from participants' behavior as part of our studies, including the relevant scales. For

[6] All participants individually provided verbal confirmation that they received the email before they left the lab.

[7] The maximum earnings from a task were $20 and if a task was submitted exactly 10 days after the end date, penalties would have been $20 if the deadline was the latest possible date or higher if the deadline were earlier.

[8] We do not know whether the original authors imposed an upper bound on the allowable responses for participants. Nevertheless, to avoid the potential influence of outlier responses, we decided to impose bounds. Ariely and Wertenbroch (2002) report that the highest average total time (i.e., time on all three tasks) was 84 minutes (p. 223). Therefore, we did not expect that our per-task limit of 180 minutes would be binding.

survey questions, we also include how participants provided their responses. After the close of the experimental platform, participants' work was evaluated, and payments were transferred to a debit card that they were given during their initial visit to the lab.

Potential differences with the original study

As noted, we did not have access to the original study materials, which means differences between our study and the original are possible. Additionally, technological changes, in our opinion, necessitated further changes in study design.

1. We chose to fully computerize our study. We believed that – with modern advances in scanning, character recognition, and spelling/grammar checking – a paper-based study would lead participants to exploit modern technologies to an extent that was not possible when the original study was done. We attempted to make it difficult for participants to enter the text into a computerized grammar/spell checker.[9]

2. Participants were asked to record their corrections in a text box on the computer interface using the format "Line XX: beleive → believe;". It is possible that this way of reporting errors differs from how participants reported errors in the original study.

3. We created six substantially identical tasks, and participants were given three randomly selected tasks.

4. We doubled the monetary incentives: participants earned \$0.20 per error found and faced \$2 penalties per day late, compared to \$0.10 per error and \$1 per day late in the original study. The late penalties were prorated according to the fraction of the day that they were late, and penalties were applied separately for each task, as appropriate.

5. In order to avoid the possibility of losses, it was agreed with our Institutional Review Board that participants would be offered a \$10 participation fee, \$5 of which would be given up front, and the remaining \$5 would be held in reserve against any late

[9] Specifically, we used JavaScript to disable right-clicking, copy and paste commands, and printing. We also checked that JavaScript was enabled in the browser. If JavaScript was disabled, the tasks were not displayed, and a message asking participants to enable JavaScript was displayed. The texts were presented as images with line numbers in the margins and were rotated by between -7 and $+7$ degrees to make cropping the image and inputting into an optical character recognition package less likely to succeed. It was possible for a participant to take a picture and upload it to an artificial intelligence tool such as ChatGPT. Our tests suggested that such an approach would do a poor to adequate job (depending on the tool used), but that it would still make mistakes and require time and effort to review the output, decide what was useful, and input the suggestions in the manner required. While a sophisticated and dedicated participant could circumvent our precautions, we felt that our measures were sufficiently challenging that it would be just as easy to actually do the task as intended as it would be to try to circumvent it.

> penalties that participants incurred. If penalties exceeded the sum of their task earnings and the $5 held in reserve, then participants would not receive any additional payment, nor would they be liable for the penalties.[10]

While the above items represent potential or actual differences from the original procedures, since the differences apply across all three treatment conditions, we have no reason to expect that they would affect the efficacy (or lack thereof) of deadlines on performance.

Note also that in the Supplementary Materials we report the results of a nearly identical replication. The only difference was that, in that replication, participants were randomly assigned to treatment at the session level. For this replication, 120 participants participated. The same experimenter and lab assistant conducted the experiments and scored the participants' performance on tasks in both the main and supplemental studies. The results of this supplemental study are virtually identical to those of the main study.

Power analysis

Our goal was to achieve 90% power at the $\alpha = 0.05$ level of significance. There are three treatments and three performance measures of interest, leading to nine potential hypothesis tests. For each performance measure, we conducted a one-way ANOVA of performance on treatment, giving us the variance of the error for the original data as well as the average performance for each treatment. We then used the `power oneway` command in Stata to generate the required sample size. The highest required sample size was $N = 33$ (or 11 participants per treatment). Although the original paper only conducts joint tests of the null hypothesis that performance is the same in all three treatment conditions, we were also interested in having sufficient power to potentially conduct pairwise treatment tests. To calculate the required sample size, we used the `contrast` option to account for pairwise treatment tests. In this case, the highest required sample size was $N = 105$ (or 35 participants per treatment). We ultimately decided that we would recruit 40 participants per treatment to the initial in-person session.

Another factor in our recruitment decision was the expected attrition of participants. In particular, Bisin and Hyndman (2020) showed that many participants who

[10] Applying the reported rule for payments to participants (10 cents per error correctly identified minus $1 per day late) to the first two components of Figure 2a, it is apparent that the average payments do not match in the original paper. In fact, many participants should have lost money. In response to an inquiry in 2014 by the authors to clarify this issue, Dan Ariely stated that to his recollection the payment for the experiment was all done at the end, there was no show-up fee, the authors did not say anything about the fact that the payment rule could mean that participants will lose money, and they never had to deal with this problem. Nevertheless, because we were concerned with the possibility of losses, we entered into discussions with our IRB, where the above solution was agreed to and, therefore, implemented.

initially started the experiment did not finish. Other authors have explicitly planned for attrition in their study design (Wilcox et al., 2016).[11] Although Ariely and Wertenbroch (2002) did not report that any participants failed to finish, we wanted to have data to analyze performance conditional on finishing as a supplemental analysis. We conducted 9 initial lab sessions in total, with four on January 30, 2025, and another five on February 4, 2025, recruiting 124 participants in total, with at least 40 participants in each treatment.

Results

Main performance metrics

We begin our discussion of results for the main study by examining participants' choice to set binding deadlines in the self-imposed deadlines treatment (i.e., a deadline strictly earlier than the latest possible deadline). 93.33% of the participants self-imposed a binding deadline on themselves for both tasks 1 and 2, while 89.33% of participants self-imposed a binding deadline on themselves for task 3. On average, conditional on setting a binding deadline, the self-imposed deadlines are, respectively, 11.6, 7.8, and 4.2 days before the overall deadline for the experiment. This appears to be consistent with the original paper, which reports that many participants self-imposed deadlines and that the deadlines were spaced out over the three-week duration of the experiment.

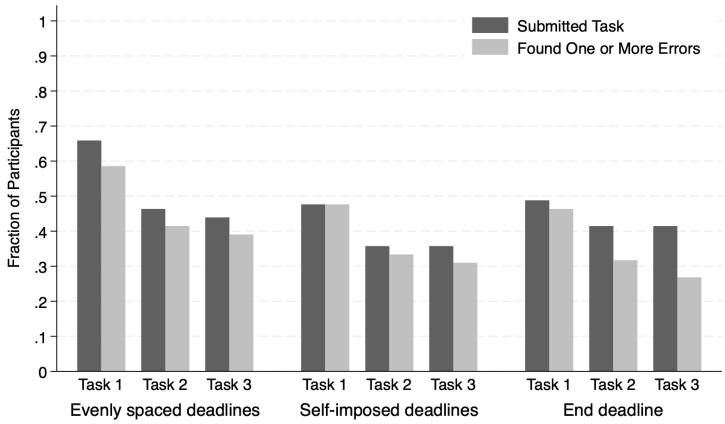
Turning our attention to actual performance in the experiment, we first address the question of whether participants complete the task at all. We do so because of an apparent stark contrast in results between Ariely and Wertenbroch (2002) and both Bisin and Hyndman (2020) and Burger et al. (2011). The former did not report any attrition, while the latter studies found it to be common. In Figure 1, we report – for each task in our replication data – the fraction of participants who submitted the task and the fraction who submitted the task *and* correctly found at least one error. As is visually quite apparent, in our replication, many participants failed to submit some or all of the tasks. This contrasts starkly with the original result that all participants completed all tasks. That said, there are no apparent treatment differences in task submissions. One-way ANOVAs failed to find

[11] Indeed, attrition is common and generally expected in experiments taking place whenever participants participate on their own time and outside of a more controlled lab environment. Augenblick et al. (2015) and Augenblick and Rabin (2019) reported attrition rates of 10-20% in their studies that involve participants performing work over a period of several weeks (see also Acland & Levy, 2015; Charness & Gneezy, 2009; Halevy, 2015, among others). In Psychology, Flick (1988) and Zhou and Fishbach (2016) have discussed issues concerning attrition. We would even go so far as to argue that attrition should be expected on theoretical grounds. Hyndman and Bisin (2022) considered a model in which decision makers must complete a single task before some final deadline, and the cost to complete the task varies stochastically over time. They showed that decision makers adopt a cost threshold rule (which possibly depends on the time to the deadline) so that they complete the task in time t if their cost is less than the threshold cost at that time. It is possible – say, because the decision maker received adverse cost shocks – for a decision maker to *never finish* the task.

significant evidence of an effect of deadline condition on either the fraction of tasks submitted ($F(2, 121) = 0.74$; $p = .478$; $\eta^2 = .012$) or the fraction of tasks submitted that also found one or more errors ($F(2, 121) = 0.77$; $p = .466$; $\eta^2 = .013$).

Figure 1

Fraction of participants who submitted each task by treatment condition



Note. For each participant we created indicator variables that take value 1 if the participant submitted the given task. The bar heights represent the average of each such variable.

The original finding that all participants completed all tasks appears to lack robustness. Participants who fail to submit a task could be those with very high present-bias and/or participants with specific psychological characteristics, e.g., low conscientiousness. Such participants are arguably generally present in the wider population and in the selected participant pools of university students (students routinely fail to submit homework assignments). Furthermore, the post-modern texts are, in our opinion, quite frustrating to read, making it plausible to expect that some students would judge the task not worth completing.

We turn now to the main performance results. In Figure 2, we illustrate the performance of the participants on three metrics: the number of errors found, the number of days late, and the payment. Panel (a) is taken directly from Ariely and Wertenbroch (2002). Panels (b) and (c) contain our replication data regarding the effects of different deadline structures on the original three dimensions of participants' performance. Panel (b) contains all participants, whether they submitted a task or not, while panel (c) contains only those participants who submitted all three tasks.

Comparing panel (a) with panels (b) and (c) shows clear visual evidence that our data do not replicate the results reported in Ariely and Wertenbroch (2002). The directional effects that we find do not align with the original paper, and our error bars are substantially larger. To provide statistical evidence for what is visually apparent, we run a

Figure 2

Comparison of performance metrics between Ariely and Wertenbroch (2002) and our main study

(a) *Figure 2 from Ariely and Wertenbroch (2002)*

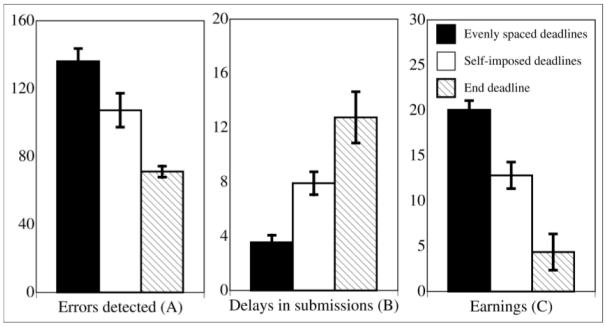
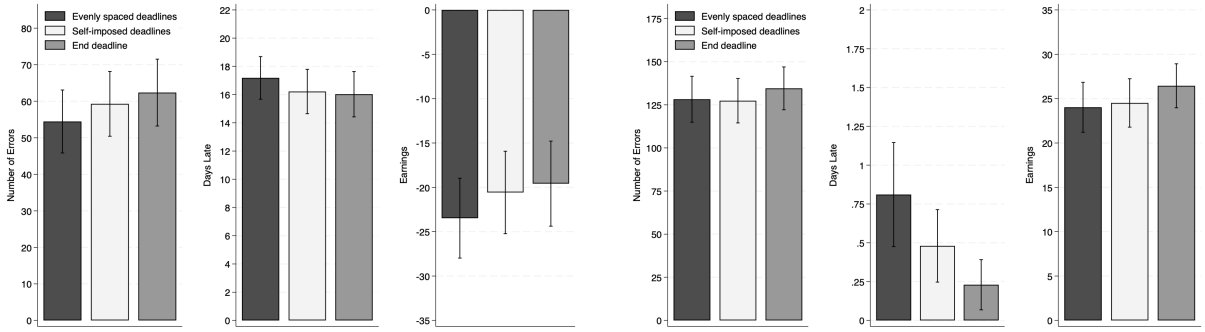


Fig. 2. Mean errors detected (a), delays in submissions (b), and earnings (c) in Study 2, compared across the three conditions (error bars are based on standard errors). Delays are measured in days, earnings in dollars.

(b) *Main study: All participants*

(c) *Main study: Submitted all three tasks*



Note. Panel (a) is taken directly from Ariely and Wertenbroch (2002), while panels (b)–(c) correspond to our replication data. Following Ariely and Wertenbroch (2002), error bars in panels (b)–(c) are based on standard errors.

separate one-way ANOVA for each performance metric to test the effect of treatment. The $F-$statistic, $p-$value, and effect size (η^2) from these analyses are presented in Table 1. Ariely and Wertenbroch (2002) reported that "[a]ll differences were statistically significant (all ps< 0.01)" (p. 222). However, as can be seen from Table 1, our replication data – regardless of which cut of the data we examine – show no statistical evidence that performance is influenced by the deadline condition.

Analysis of subjective metrics

In addition to their analysis of the main performance metrics, the original authors provided analysis of various post-completion survey questions (refer to the Method section and Table A1). They reported that participants enjoyed the task least in the evenly spaced deadlines treatment and most in the end deadline treatment, while the time spent on the tasks followed a mirror-image pattern.

Table 1

Summary results from multiple one-way ANOVA for the main study

Performance Metric	Reported in AW (2002)	All Participants			Sub. All Three Tasks		
		$F(2, 121)$	p	η^2	$F(2, 49)$	p	η^2
Errors (Total)	Unknown:	0.18	.835	.003	0.70	.503	.029
Days Late (Total)	Reported "All	0.35	.703	.006	1.11	.339	.045
Payment	$ps < 0.01$"	0.22	.806	.004	1.10	.340	.045

Note. "Reported in AW (2002)" means that we have extracted the information from the original paper. "All Participants" means that we conducted the analysis using all participants who came to the initial lab session, while "Sub. All Three Tasks" means that we conducted the analysis only on the subset of participants who completed all three tasks in our replication. Following Ariely and Wertenbroch (2002), we conducted separate one-way ANOVAs for each performance metric, with treatment being the explanatory variable. Unlike the original study, we present results both for all participants and for only those who completed all three tasks.

For participants who submitted all three tasks, we collect similar subjective measurements. Table 2 contains averages and standard deviations by treatment for each metric. We also report the $F-$statistic and effect size (η^2) from a series of one-way ANOVA of each subjective metric on the treatment condition. As with performance data, we find a general lack of significance across all six measures.[12] That is, unlike Ariely and Wertenbroch (2002), we find no relationship between deadlines and either subjective evaluation of the tasks or time spent on the tasks.

Indeed, Ariely and Wertenbroch (2002) went further, interpreting their post-completion survey data causally: On page 223, Ariely and Wertenbroch (2002) wrote, "Taken together, the results show that when deadline constraints increased, performance improved, time spent on the task increased, and enjoyment of the task decreased (because of enhanced recognition of the true low quality of the texts)." The suggested causal mechanism was that deadlines cause changes in time allocation, which in turn cause changes in performance and subjective evaluation. They demonstrated each link in this chain by separately conducting an ANOVA of the given metric on the treatment. Our results in Table 2 already suggest that the causal chain espoused by the original authors is not supported in our data. If we take the quotation at face value, it suggests that we should see a negative relationship between time spent on the task and subjective enjoyment of the task. However, in all cases, we find a positive (and often significant) relationship. These results are given in Table 3. Upon reflection, we argue that this makes sense. The promise of monetary compensation could induce more people to start the task, but to actually finish the task and be willing to engage with it for a substantial amount of time,

[12] Ariely and Wertenbroch (2002) actually reported the results of a one-way ANOVA based on the average of the five subjective measures, finding significant differences across treatments. A similar analysis based on our replication data yields no significant differences ($F(2, 49) = 0.33$, $p = .723$ and $\eta^2 = .014$).

Table 2

Summary statistics on subjective metrics for participants who submitted all three tasks in the main study

Measure	Evenly spaced deadlines		Self-imposed deadlines		End deadline		$F(2, 49)$	η^2
	M	SD	M	SD	M	SD		
Time (All)	226.44	143.39	305.00	141.68	187.71	162.36	2.40	.095
Like	58.50	36.08	64.79	29.73	52.65	35.09	0.49	.021
Interesting	57.61	38.66	55.71	33.16	51.12	38.10	0.14	.006
Quality	63.94	31.40	71.57	23.30	62.71	27.23	0.45	.019
Grammar	61.89	30.84	69.86	21.65	58.24	25.98	0.74	.031
Ideas	65.44	31.35	58.43	27.32	56.47	33.85	0.40	.017

Note. The variable Time (All) represents the sum of the self-reported time spent on each task. See Table A1 for details on how each of the subjective measures was elicited. For the ANOVA results, the dependent variable is the given "Measure", and the results capture the effect of the treatment conditions on the dependent variable.

one should enjoy it.

Within-participants behavior

While our replication fails to reproduce the original findings, we can examine whether our data exhibit some fundamental behavioral relationships that should be present regardless of treatment effects. Specifically, we focus on associations that we believe *should be* present and whose absence would raise questions. These associations were not reported in Ariely and Wertenbroch (2002), which prevents us from drawing further inferences regarding our failure to replicate their reported results.[13]

We begin with the question of the relationship between effort and errors found, discussed above. As noted, Ariely and Wertenbroch (2002) appeared to argue that deadlines cause participants to spend more time on the tasks, which leads to a greater number of errors found. We already showed that, in our data, deadlines neither cause participants to devote more time to the tasks, nor does performance increase in the presence of deadlines. Nevertheless, we argue that there is a more fundamental relationship that should exist regardless of treatment. In particular, time spent on the tasks can be viewed as effort, and the number of errors found should be positively associated with effort,

[13] In 2006, while working on another project, we emailed the authors asking for their data. Shortly after, we received an email from Dan Ariely's MIT email address with three Microsoft Excel files attached (purportedly one for each study reported in the original paper). In October 2024, at the request of the editors, we shared with Dan Ariely an analysis of the contents from the file purportedly for their Study 2 and asked for permission to include a summary of it in the paper. Dan Ariely denied our request, arguing, among other things, that the files we received may not be the actual data. He did not subsequently provide us with any additional data from the original paper. Consequently, we are unable to supplement our replication exercise with any additional analysis of the files we received in 2006 or any other data.

Table 3

OLS regression results capturing the relationship between subjective measures and total time spent on tasks in the main study

Variables	Estimate	*SE*	*p*
DV: Like			
Intercept	29.086	11.859	$< .001$
Time (All Tasks)	0.117	0.029	.018
DV: Interesting			
Intercept	19.000	13.122	.155
Time (All Tasks)	0.12	0.032	.001
DV: Quality			
Intercept	56.824	10.929	$< .001$
Time (All Tasks)	0.048	0.027	.001
DV: Grammar			
Intercept	44.403	9.637	$< .001$
Time (All Tasks)	0.083	0.024	.001
DV: Ideas			
Intercept	36.17	11.87	.004
Time (All Tasks)	0.073	0.029	.016

Note. DV denotes "Dependent Variable". The variable Time (All Tasks) represents the sum of the self-reported time spent on each task. See Table A1 for details on how each of the subjective measures was elicited. Treatment controls were included in the regressions, but not reported to conserve space. In none of the regressions does any treatment indicator reach statistical significance; in all cases, $p \geq .243$. In all regressions, the number of observations is 49.

regardless of one's treatment. We test for this with a linear random-effects regression. The dependent variable is the number of errors found on each task, while the main explanatory variable is the self-reported time spent on each task. We estimate the model both with and without treatment controls as additional explanatory variables. The results are in Table 4. As can be seen, both with and without treatment fixed effects, the number of errors found is significantly positively associated with the self-reported time spent on each task. Thus, in our data, there is a direct relationship between time spent on tasks and the number of errors found.

Knowing whether this direct relationship was present in Ariely and Wertenbroch's (2002) data would provide important insights. They established via ANOVA that deadlines affected both errors found (their Figure 2) and time spent on task (p. 223, "$F(2, 57) = 45.76$, $p < .001$"), but did not test the direct relationship between time spent and errors found. This leaves open the possibility of Simpson's Paradox (Pearl, 2014; Simpson, 1951): their ANOVA results might reflect only between-treatment differences, with no within-treatment correlation between time and performance. Analyzing their

original data to examine this relationship – both within and across treatments – would be valuable for understanding the underlying causal mechanism.

Table 4

Linear random-effects regression of the effect of time spent on task on errors found in the main study

Effect	Estimate	*SE*	*p*
Spec. 1			
Fixed effects			
Intercept	18.734	5.841	.001
Time on Task [a]	0.271	0.039	< .001
Treatment controls		Included	
Random effects			
Within-study variance	114.831	16.409	< .001
Between-study variance	423.563	93.524	< .001
Spec. 2			
Fixed effects			
Intercept	24.506	4.384	< .001
Time on Task [a]	0.261	0.038	< .001
Treatment controls		Not included	
Random effects			
Within-study variance	114.779	16.397	< .001
Between-study variance	459.438	100.699	< .001

Note. For each participant, we have three observations – one for each task; hence, we adopt a panel-data regression framework. The dependent variable is the number of errors found on each task. There are 49 participants, making for a total sample size of 147.

[a] Time on Task was a self-reported measure between 0 and 180 minutes. See Table A1 for more details.

In addition to studying the relationship between errors found and time, we also analyze the within-participants correlation between the errors found on any two tasks i and j. For example, a participant who finds a high number of errors on task 1 might be likely to find a high number of errors on task 2. This would be the case if high performance on one task is representative of either high effort or high ability, which should hold across tasks. As in the case above, we expect correlations in errors found to hold regardless of treatment assignment. In Table 5, for our replication sample, we report correlation matrices (conditional on treatment) for the number of errors found on each task. The smallest pairwise correlation is .793, and in all cases, the correlations reached significance at the $p < .001$ level.[14]

As with the time-performance relationship discussed above, whether such

[14] If we restrict attention to only those participants who submitted all three tasks, the same result holds in terms of significance, but the correlations are somewhat smaller (though always above 0.66).

within-participant correlations held in the original data cannot be determined without access to it. If high task-to-task correlations were absent in Ariely and Wertenbroch's (2002) data, this would suggest their results were driven primarily by between-treatment differences rather than stable individual differences in effort or ability. Conversely, if similar correlations existed in their data, this would help isolate other factors that might explain the divergence in our treatment effect findings.

Table 5

Correlation matrices for errors found per task in the main study

(a) *End deadline*

Variables	n	M	SD	1	2	3
1. Errors (Task 1)	41	25.390	31.698	–		
2. Errors (Task 2)	41	17.707	30.347	.865***	–	
3. Errors (Task 3)	41	18.049	30.685	.793***	.901***	–

(b) *Self-imposed deadlines*

Variables	n	M	SD	1	2	3
1. Errors (Task 1)	42	18.548	26.006	–		
2. Errors (Task 2)	42	17.810	28.600	.935***	–	
3. Errors (Task 3)	42	16.786	29.020	.835***	.918***	–

(c) *Evenly spaced deadlines*

Variables	n	M	SD	1	2	3
1. Errors (Task 1)	41	19.317	24.720	–		
2. Errors (Task 2)	41	18.195	27.598	.872***	–	
3. Errors (Task 3)	41	16.951	27.069	.911***	.975***	–

Note. Errors (Task i) denotes the number of errors found in Task i, for $i = 1, 2, 3$; it is an integer between 0 and 100. See Table A1 for more details.
*** $p < 0.001$.

General Discussion

Our results do not replicate the main finding of Ariely and Wertenbroch (2002). We find no evidence that evenly spaced deadlines improve performance on any of the three key metrics: errors found, days late, or overall payment. We also find evidence of substantial attrition, unlike the original authors. These findings align with other studies questioning the robustness of deadline effects, including Bisin and Hyndman (2020) and Burger et al. (2011), and the more general cautionary note about goal-setting discussed in Ordóñez et al. (2009).

There are several reasons why any particular experimental study may fail to replicate. First, statistical variation (e.g., sampling error, measurement noise, and chance) combined with publication bias for positive results can lead initial findings to be statistical outliers that overstate the true treatment effect. Second, subtle contextual differences in participant populations across time periods or locations may impact results.

Given the simplicity of the original data analysis, overt forms of $p-$hacking are unlikely to be the source of our failure to replicate the original result. However, other forms of questionable research practices – such as selective stopping or selective reporting of dependent variables (e.g., the failure to report additional analysis of within-participants behavior) – would not be apparent from the published analysis and cannot be ruled out as potential explanations for our failure to replicate the original result.

Beyond these general factors, our replication faces specific limitations. Over 20 years elapsed between studies. As a consequence, participants may have responded differently to the same experimental tasks due to cultural shifts, changes in social norms, increased familiarity with experimental procedures, or shifts in student attitudes toward academic experiments. Furthermore, the global pandemic that may have altered how deadlines influence behavior.[15] Additionally, lack of access to original materials and technological changes necessitated protocol differences, and there may be unknown differences in experimental procedures. Any of these factors could contribute to the divergent results.

In our opinion, the design changes we implemented are relatively minor. We claim that these changes would have, if anything, encouraged more tasks to be completed: We allowed participants to complete the work online, without requiring participants to return any tasks in person, nor did we require them to collect any payments due in person. Furthermore, given the way we implemented the participation fee, participants actually had stronger incentives to submit the tasks – even poor quality work – in order to receive the additional payment. At minimum, therefore, our lack of replication suggests that the result may not be robust to slight modifications in experimental protocols.

To take a broader perspective, there are additional reasons to believe that the original result may not be robust. One can interpret deadlines as goals, which means that we can view Ariely and Wertenbroch's (2002) results from the standpoint of Goal Theory in psychology (Locke & Latham, 1990, 2002, 2006). A recent meta-analysis conducted by Epton et al. (2017) presented empirical evidence that provides some support for goal theory, though with important moderators that speak directly to our setting and Ariely

[15] That said, as noted, Burger et al. (2011) and Bisin and Hyndman (2020) reported similar experiments – conducted both before the pandemic and closer in time to the original study – that also questioned the effectiveness of deadlines, suggesting temporal factors alone cannot explain the divergence.

and Wertenbroch's setting.[16] Epton et al. (2017) found that goals are most effective when the goal is difficult, it is set publicly, and it is a group goal. Arguably, viewing the deadline(s) as the goal(s), then none of these three criteria for the beneficial effect of goals is met in either our or the original authors' setting. Additionally, they showed that goals are significantly less effective in a university environment, as is the case for both our study and the original study. Therefore, one could argue that the original result is the outlier, while our replication results provide evidence consistent with the meta-analytic review.

In the end, and in the context of the moderators identified by Epton et al. (2017), we would stress that our lack of replication of the original result of Ariely and Wertenbroch (2002) implies that the commonly received wisdom that (evenly spaced) deadlines are an effective mechanism to limit procrastination should be met with some skepticism. While Epton et al. (2017) have already begun important work mapping out boundary conditions for when goal setting and deadlines are more or less effective, our understanding remains incomplete. Indeed, both psychologists and behavioral economists have identified theoretical reasons and provided empirical evidence that deadlines (and goals more broadly) may be detrimental. The mechanisms underlying these negative effects appear to focus on two key factors: reduced flexibility due to pre-commitment and the possibility that deadlines make tasks more aversive. Drawing on Self-Determination Theory (Deci & Ryan, 2000; Ryan & Deci, 2000), the imposition of deadlines can undermine positive behavior by reducing individuals' sense of autonomy, which may decrease intrinsic motivation and reduce engagement. This leads us to believe that it is important to go back to the theory (goal setting, self-determination, behavioral economics, etc.) and empirical results to identify new testable hypotheses for whether and in what context deadlines might work. Deadlines play an important role in both educational institutions and many business practices. A better understanding of their effectiveness, including the conditions under which they can backfire, could lead to significant improvements in these areas.

Acknowledgments

We thank an editor and two anonymous reviewers for their many helpful suggestions and feedback. In addition, we are grateful to Gary Bolton, Leif Nelson, and Uri Simonsohn for helpful discussions, as well as internal seminar participants at the Center & Laboratory for Behavioral Operations and Economics at the University of Texas at Dallas. Finally, we acknowledge Ella Lindsay and Tiffany Matthews for their help running the experiments, and the University of Texas at Dallas and New York University for financial support.

---

[16] We would like to thank an anonymous reviewer for pointing us to this important reference.

References

Acland, D., & Levy, M. R. (2015). Naiveté, projection bias, and habit formation in gym attendance. *Management Science*, *61*(1), 146–160.

Ariely, D., & Wertenbroch, K. (2002). Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological Science*, *13*(3), 219–224.

Augenblick, N., Niederle, M., & Sprenger, C. (2015). Working over time: Dynamic inconsistency in real effort tasks. *Quarterly Journal of Economics*, *130*(3), 1067–1115.

Augenblick, N., & Rabin, M. (2019). An experiment on time preference and misprediction in unpleasant tasks. *Review of Economic Studies*, *86*, 941–975.

Bisin, A., & Hyndman, K. (2020). Present-bias, procrastination and deadlines in a field experiment. *Games and Economic Behavior*, *119*, 339–357.

Burger, N., Charness, G., & Lynham, J. (2011). Field and online experiments on self-control. *Journal of Economic Behavior & Organization*, *77*, 393–404.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., ... Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behavior*, *2*, 637–644.

Charness, G., & Gneezy, U. (2009). Incentives to exercise. *Econometrica*, *77*(3), 909–931.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Epton, T., Currie, S., & Armitage, C. J. (2017). Unique effects of setting goals on behavior change: Systematic review and meta-analysis. *Journal of consulting and clinical psychology*, *85*(12), 1182.

Flick, S. N. (1988). Managing attrition in clinical research. *Clinical Psychology Review*, *8*, 499–515.

Halevy, Y. (2015). Time consistency: Stationarity and time invariance. *Econometrica*, *83*(1), 335–352.

Hyndman, K., & Bisin, A. (2022). Procrastination, self-imposed deadlines and other commitment devices. *Economic Theory*, *74*(3), 871–897.

Jaffe, E. (2014). Self-imposed deadlines don't stop procrastination. Here's what might. *Fast Company, March 26*.

Lewis-Kraus, G. (2023). They studied dishonesty. Was their work a lie? *The New Yorker, September 30*.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance.* Prentice-Hall, Inc.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American psychologist*, *57*(9), 705.

Locke, E. A., & Latham, G. P. (2006). New directions in goal-setting theory. *Current Directions in Psychological Science*, *15*(5), 265–268.

O'Donoghue, T., & Rabin, M. (1999). Incentives for procrastinators. *Quarterly Journal of Economics*, *114*, 769–816.

O'Donoghue, T., & Rabin, M. (2001). Choice and procrastination. *Quarterly Journal of Economics*, *116*(1), 121–160.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Ordóñez, L. D., Schweitzer, M. E., Galinsky, A. D., & Bazerman, M. H. (2009). Goals gone wild: The systematic side effects of over-prescribing goal setting. *Academy of Management Perspectives*, *23*(1), 6–16.

Pearl, J. (2014). Understanding simpson's paradox. *The American Statistician*, *68*(1), 8–13.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(2), 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x

Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, *133*(1), 65–94.

Steel, P. (2011). *The procrastination equation.*

Stern, J. (2023). An unsettling hint at how much fraud could exist in science. *The Atlantic, August 2*.

Surowiecki, J. (2010). Later. *The New Yorker*, *86*(31), 210–210.

Thompson, D. (2014). The procrastination doom loop – and how to break it. *The Atlantic, August 26*.

Wilcox, K., Laran, J., Stephen, A. T., & Zubcsek, P. P. (2016). How being busy can increase motivation and reduce task completion time. *Journal of Personality and Social Psychology*, *110*(3), 371–384.

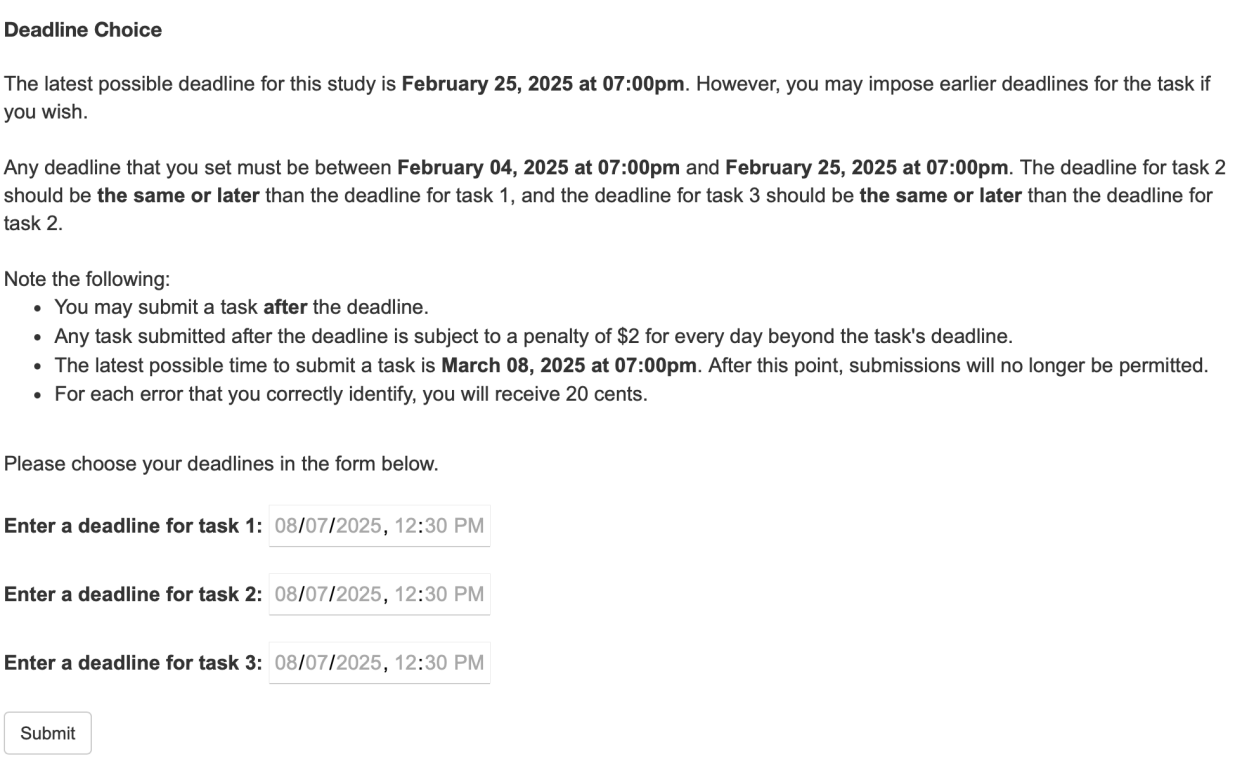
Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504.

Appendix

Description of Variables and Experimental Interface Screenshot

Figure A1

Deadline choice screen for the self-imposed deadlines treatment



Deadline Choice

The latest possible deadline for this study is **February 25, 2025 at 07:00pm**. However, you may impose earlier deadlines for the task if you wish.

Any deadline that you set must be between **February 04, 2025 at 07:00pm** and **February 25, 2025 at 07:00pm**. The deadline for task 2 should be **the same or later** than the deadline for task 1, and the deadline for task 3 should be **the same or later** than the deadline for task 2.

Note the following:

- You may submit a task **after** the deadline.
- Any task submitted after the deadline is subject to a penalty of $2 for every day beyond the task's deadline.
- The latest possible time to submit a task is **March 08, 2025 at 07:00pm**. After this point, submissions will no longer be permitted.
- For each error that you correctly identify, you will receive 20 cents.

Please choose your deadlines in the form below.

Enter a deadline for task 1: 08/07/2025, 12:30 PM

Enter a deadline for task 2: 08/07/2025, 12:30 PM

Enter a deadline for task 3: 08/07/2025, 12:30 PM

Submit

Table A1
Description of variables used in our analysis

Variable	Description	Scale
Deadline Set on Task i	Self-imposed deadlines condition only. Indicator for whether participant set a deadline strictly before the final deadline	0 if deadline is the final deadline; 1 if deadline is strictly before the final deadline
Deadline on Task i	Self-imposed deadlines condition only. The deadline for Task i, after which penalties will commence.	A date and time between the start and end date/time of the experiment.
Task i Submitted	Indicator for whether task i ($i = 1, 2, 3$) was submitted.	0 if not submitted; 1 if submitted
Number of Errors on Task i	Number of errors correctly identified on task i	Integer between 0 and 100
Number of Errors (Total)	Sum of number of errors correctly identified on all tasks	Integer between 0 and 300
Days Late on Task i	The difference (in days) between the time task i was submitted and the deadline for task i.	Non-negative real number. E.g., 3 hours late implies: $3/24 = 0.125$. Submit times were precise to the nearest second.
Days Late (Total)	Sum of Days Late on Task i for all tasks	
Earnings	The payment earned by the participant according to the payment function: Earnings = $0.2\times$(Number of Errors (Total)) $- 2\times$(Days Late (Total))	Can be negative if penalties exceeded payment for errors found; actual earnings to participants coluld not be negative.
Like	Survey: How much did you like the task?	Integer between 0 and 100 (Input #)
Interesting	Survey: How interesting did you find the task?	Integer between 0 and 100 (Input #)
Quality	Survey: How good would you rate the quality of writing?	Integer between 0 and 100 (Input #)
Grammar	Survey: How good would you rate the grammatical quality?	Integer between 0 and 100 (Input #)
Ideas	Survey: How effectively did the text communicate the ideas contained in it?	Integer between 0 and 100 (Input #)
Time Spent on Task i	Survey: How much time (in minutes) did you spend on Task i?	Integer between 0 and 180 (Input #)
Time (All Tasks)	Sum of Time Spent on Task i for all tasks	Integer between 0 and 540

Replication of

"Procrastination, Deadlines, And Performance: Self-Control by Precommitment"

Supplementary Materials

Kyle Hyndman[1] and Alberto Bisin[2]
[1]Naveen Jindal School of Management
The University of Texas at Dallas
[2]Department of Economics
New York University

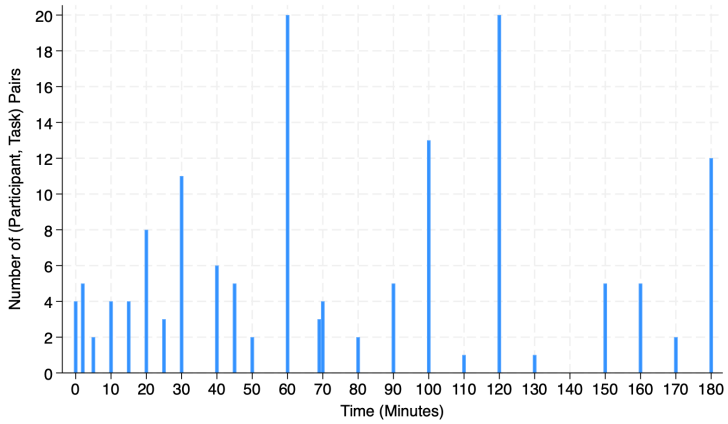
Supplementary Material A: Additional Data Analysis of the Main Study

Focal Points in Reporting Time

In Figure S1 we report a histogram of time estimates that participants from the main study provided for each task (assuming that they finished all tasks and, therefore, the survey). In our replication, 139 out of 147 (Participant, Task) pairs reported time estimates that end in a 0 or a 5 and the time estimates ranged from 0 to 180 minutes (which are hard-coded bounds on allowable responses).[1]

Figure S1

Histogram of Participants' self-reported time to complete each task.



Note. Data are from the main study. The time that participants could report as having spent on each task was capped at 180 minutes. The most money that a participant could have earned from perfectly executing a single task is $20. Therefore, spending more than 3 hours on a task would have yielded an hourly pay of less than $7 per hour.

Correlations on Subjective Measures

Table S1 looks at the correlations between subjective measures we elicited from participants in the main study about the task, such as grammar, the quality of writing, how interesting the tasks were, etc. We have no clear prior on whether the correlations should be positive or negative for any pairwise correlation, and report them here simply because it would be instructive to compare the correlations in our data with the original data. In such a case, we should expect a similar pattern of correlations. In our main study data, 25 of 30 correlations are significant at $p < .05$. In terms of the signs of the correlations, all 30 correlations are positive in the replication data.

---

[1] 8.2% of response times are reported to be 180 minutes – the maximum allowable time. In the supplemental study, reported in the Supplementary Material B, this is 12.3%. In order to make our results comparable across the main and the supplemental studies, we did not increase the maximum allowable time. There are no discernible or statistically significant patterns between treatments regarding the likelihood of reporting the maximum allowable time in either replication.

Table S1

Correlation matrices on subjective measures of tasks

(a) *End deadline*

Variables	n	M	SD	1	2	3	4	5
1. Ideas	17	56.471	33.855	–				
2. Grammar	17	58.235	25.977	0.373	–			
3. Interesting	17	51.118	38.095	.607**	0.471	–		
4. Like	17	52.647	35.094	.686**	0.448	.870***	–	
5. Quality	17	62.706	27.232	.707**	.526*	0.472	0.448	–

(b) *Self-imposed deadlines*

Variables	n	M	SD	1	2	3	4	5
1. Ideas	14	58.429	27.315	–				
2. Grammar	14	69.857	21.651	.695**	–			
3. Interesting	14	55.714	33.158	.634*	.731**	–		
4. Like	14	64.786	29.727	.771**	.818***	.905***	–	
5. Quality	14	71.571	23.303	.854***	.905***	.670**	.781***	–

(c) *Evenly spaced deadlines*

Variables	n	M	SD	1	2	3	4	5
1. Ideas	18	65.444	31.352	–				
2. Grammar	18	61.889	30.841	.930***	–			
3. Interesting	18	57.611	38.662	.748***	.781***	–		
4. Like	18	58.500	36.077	.790***	.838***	.977***	–	
5. Quality	18	63.944	31.400	.904***	.931***	.736***	.768***	–

Note. Data are from the main study. Refer to Table A1 for more details on these variables.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

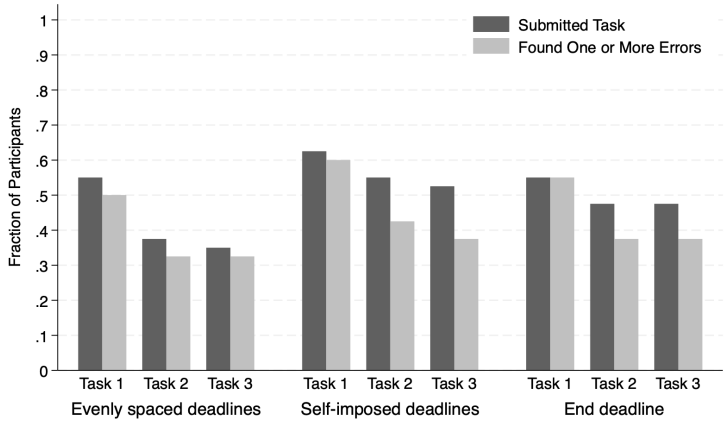
Supplementary Material B: Analysis of the Supplemental Study with Randomization at the Session Level

As noted, we originally conducted this experiment where participants were randomly assigned to treatment at the session level. Since this goes against accepted norms in the psychology literature, we decided to redo the experiment. Nevertheless, for completeness, for transparency and methodological purposes (i.e., to see if the method of random assignment to treatment matters), we report the same tables and figures as in the main text.

Rather than discussing each figure and table in detail, we note that this supplemental study yields results consistent with our main study: substantial attrition, no beneficial effect of deadlines, and the same within-participants correlations reported earlier. These findings further strengthen our conclusion that the original result lacks robustness.[2]

Figure S2

Fraction of participants who submitted each task (by treatment condition) in the Supplemental Study



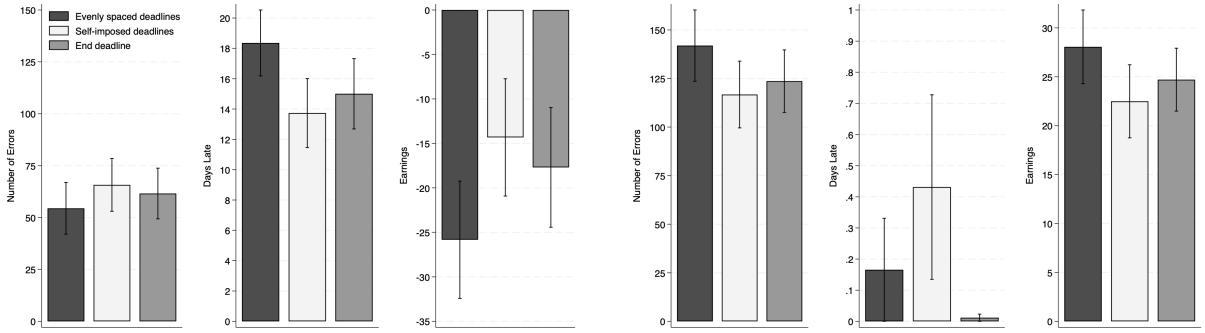
Note. For each participant we created indicator variables that take value 1 if the participant submitted the given task. The bar heights represent the average of each such variable.

[2] When speaking of treatment differences for subjective measures, Ariely and Wertenbroch (2002) actually reported the results of a one-way ANOVA based on the average of the five subjective measures, finding significant differences across treatments. A similar analysis based on our supplemental study (reported in Table S5) yields no significant differences ($F(2, 51) = 1.01$, $p = .371$ and $\eta^2 = .038$).

Figure S3

Comparison of performance metrics for the Supplemental Study.

(a) *Supplemental study: All participants* (b) *Supplemental study: Submitted all three tasks*



Note. Following Ariely and Wertenbroch (2002), error bars are based on standard errors.

Table S2

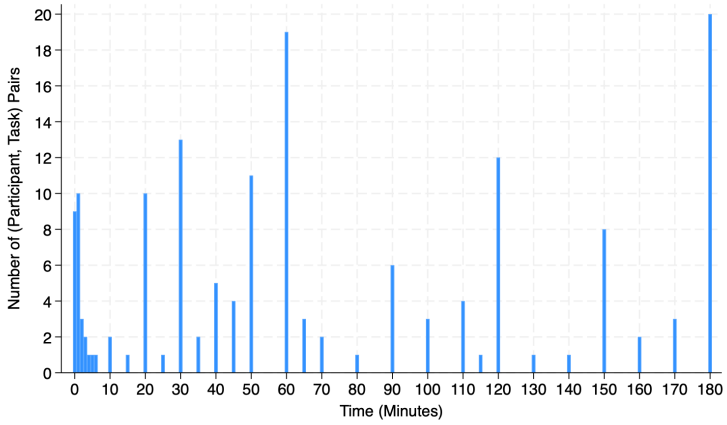
F−Statistics from multiple one-way ANOVA for the supplemental study

Performance	Reported in	All Participants			Sub. All Three Tasks		
Metric	AW (2002)	$F(2, 121)$	p	η^2	$F(2, 49)$	p	η^2
Errors (Total)	Unknown:	0.21	.811	.004	0.51	.605	.019
Days Late (Total)	Reported "All	1.12	.330	.019	1.10	.342	.041
Payment	$ps< 0.01$"	0.79	.454	.013	0.55	.578	.021

Note: "Reported in AW (2002)" means that we have extracted the information from the original paper. "All Participants" means that we conducted the analysis using all participants who came to the initial lab session, while "Sub. All Three Tasks" means that we conducted the analysis only on the subset of participants who completed all three tasks in our replication. Following Ariely and Wertenbroch (2002), we conducted separate one-way ANOVAs for each performance metric, with treatment being the explanatory variable. Unlike the original study, we present results both for all participants and for only those who completed all three tasks.

Figure S4

Histogram of Participants' self-reported time to complete each task for the supplemental study



Note. The time that participants could report as having spent on each task was capped at 180 minutes. The most money that a participant could have earned from perfectly executing a single task is $20. Therefore, spending more than 3 hours on a task would have yielded an hourly pay of less than $7 per hour.

Table S3

Linear random-effects regression of the effect of time spent on task on errors found

Effect	Estimate	SE	p
Spec. 1			
Fixed effects			
Intercept	37.232	6.127	$< .001$
Time on Task	0.140	0.035	$< .001$
Treatment controls	Included		
Random effects			
Within-study variance	104.628	14.372	$< .001$
Between-study variance	399.769	85.470	$< .001$
Spec. 2			
Fixed effects			
Intercept	31.656	3.899	$< .001$
Time on Task	0.139	0.036	$< .001$
Treatment controls	Not included		
Random effects			
Within-study variance	104.557	14.359	$< .001$
Between-study variance	412.090	87.837	$< .001$

Note. For each participant, we have three observations – one for each task; hence, we adopt a panel-data regression framework. The dependent variable is the number of errors found on each task. There are 54 participants, making for a total sample size of 162.

[a] Time on Task was a self-reported measure between 0 and 180 minutes. See Table A1 for more details.

Table S4

Correlation matrices on errors found per task for the Supplemental Study

(a) *End deadline*

Variables	n	M	SD	1	2	3
1. Errors (Task 1)	40	23.375	26.203	–		
2. Errors (Task 2)	40	18.525	26.519	.821***	–	
3. Errors (Task 3)	40	19.675	27.979	.842***	.953***	–

(b) *Self-imposed deadlines*

Variables	n	M	SD	1	2	3
1. Errors (Task 1)	40	24.125	26.931	–		
2. Errors (Task 2)	40	21.775	28.123	.922***	–	
3. Errors (Task 3)	40	19.850	28.355	.871***	.894***	–

(c) *Evenly spaced deadlines*

Variables	n	M	SD	1	2	3
1. Errors (Task 1)	40	19.925	27.220	–		
2. Errors (Task 2)	40	17.125	27.101	.907***	–	
3. Errors (Task 3)	40	17.400	27.983	.868***	.850***	–

Note. Errors (Task i) denotes the number of errors found in Task i, for $i = 1, 2, 3$; it is an integer between 0 and 100.
*** $p < 0.001$.

Table S5

Summary statistics on subjective metrics for the Supplemental Study (Submitted all three tasks)

Measure	Evenly spaced deadlines		Self-imposed deadlines		End deadline		$F(2, 51)$	η^2
	M	SD	M	SD	M	SD		
Time (All)	216.07	187.00	215.95	179.98	232.47	174.15	0.05	.002
Like	75.29	26.82	57.86	38.65	55.05	26.71	1.83	.067
Interesting	70.14	29.24	53.33	39.35	47.84	30.09	1.84	.067
Quality	64.64	26.08	62.14	32.86	53.21	29.14	0.73	.028
Grammar	63.71	22.94	60.43	32.86	55.58	29.14	0.33	.013
Ideas	64.93	29.08	57.95	35.29	56.68	29.32	0.30	.012

Note. The variable Time (All) represents the sum of the self-reported time spent on each task. See Table A1 for details on how each of the subjective measures was elicited. For the ANOVA results, the dependent variable is the given "Measure", and the results capture the effect of the treatment conditions on the dependent variable.

Table S6

OLS Regression results capturing the relationship between subjective measures and time spent on tasks for the Supplemental Study

Variable	Estimate	SE	p
DV: Like			
Intercept	45.705	8.433	$< .001$
Time (All Tasks)	0.056	0.024	.022
DV: Interesting			
Intercept	45.573	9.241	$< .001$
Time (All Tasks)	0.036	0.026	.175
DV: Quality			
Intercept	54.244	7.982	$< .001$
Time (All Tasks)	0.0366	0.023	.111
DV: Grammar			
Intercept	49.530	7.746	$< .001$
Time (All Tasks)	0.050	0.022	.025
DV: Ideas			
Intercept	48.507	8.556	$< .001$
Time (All Tasks)	0.044	0.024	.077

Note. DV denotes "Dependent Variable". The variable Time (All Tasks) represents the sum of the self-reported time spent on each task. See Table A1 for details on how each of the subjective measures was elicited. Treatment controls were included in the regressions, but not reported to conserve space. In none of the regressions does any treatment indicator reach statistical significance; in all cases, $p \geq .105$. In all regressions, the number of observations is 54.

Table S7

Correlation matrices on subjective measures of tasks for the supplemental study

(a) *End deadline*

Variables	n	M	SD	1	2	3	4	5
1. Ideas	19	56.684	29.318	–				
2. Grammar	19	55.579	29.140	.836***	–			
3. Interesting	19	47.842	30.089	.598**	.483*	–		
4. Like	19	55.053	26.709	.533*	0.335	.814***	–	
5. Quality	19	53.211	28.436	.746***	.777***	.669**	.616**	–

(b) *Self-imposed deadlines*

Variables	n	M	SD	1	2	3	4	5
1. Ideas	21	57.952	35.292	–				
2. Grammar	21	60.429	32.861	.848***	–			
3. Interesting	21	53.333	39.349	.876***	.725***	–		
4. Like	21	57.857	38.651	.838***	.760***	.967***	–	
5. Quality	21	62.143	32.232	.901***	.959***	.757***	.758***	–

(c) *Evenly spaced deadlines*

Variables	n	M	SD	1	2	3	4	5
1. Ideas	14	64.929	29.079	–				
2. Grammar	14	63.714	22.939	.768**	–			
3. Interesting	14	70.143	29.239	.840***	.742**	–		
4. Like	14	75.286	26.811	.793***	.837***	.897***	–	
5. Quality	14	64.643	26.079	.785***	.817***	.854***	.898***	–

Note. Refer to Table A1 for more details on these variables.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.