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

Kyle Hyndman[1] Alberto Bisin[2]

[1] University of Texas at Dallas
kyleb.hyndman@utdallas.edu
[2] New York University
alberto.bisin@nyu.edu

May 27, 2024

Abstract

We present the results of a replication of Study 2 from Ariely and Wertenbroch (2002), as well as a comparison of the replication data and the original data, which was kindly given to us in 2006 by one of the original authors. We show that the results of the paper do not replicate. In particular, in the replication, changes in the deadlines have a negligible effect on the three performance metrics and several survey metrics that were used in the original study. In particular, evenly spaced deadlines exogenously imposed on subjects do not stand apart for their effectiveness in reducing procrastination in subjects. Beyond failing to replicate the original finding, we also document that several patterns that arguably should be present in both datasets are not found in the original data, suggesting a further lack of robustness.

Keywords: Replication, Deadlines, Procrastination

1 Introduction

Deadlines may help overcome procrastination. Indeed, a deadline on a task arguably creates a sense of immediacy in finishing the task. However, the cost of deadlines is a loss of flexibility. With more time, a person has more options to complete a task when conditions are favorable. Arguably the most well-known and most impactful paper addressing this trade-off experimentally is Ariely and Wertenbroch (2002).[1] The authors open their paper with three general observations: (1) most of

*We would like to thank 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. We are also grateful to Ella Lindsay and Tiffany Matthews for their help running the experiments, and to UT Dallas and New York University for financial support.

[1]The paper has been cited over 1900 times as of May 2024. 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.

us procrastinate; (2) deadlines are a common feature of life; and (3) many people seek commitment devices to help them accomplish their goals. Ariely and Wertenbroch then ask three questions. First, will people self-impose deadlines on themselves? Second, are self-imposed deadlines effective? Finally, do people optimally self-impose deadlines? The original paper presented a series of studies which showed that a substantial number of people are willing to self-impose deadlines on themselves and that the deadlines were partially effective. Their main, result, however, is that performance is highest when evenly spaced deadlines are *exogenously* imposed on subjects.

On the contrary, two studies conducted subsequently to Ariely and Wertenbroch (2002) provide experimental evidence that deadlines are not effective. In Bisin and Hyndman (2020), we study the role of deadlines both in a single task and, like Ariely and Wertenbroch (2002), in a multiple repeated task environment; however, there are two important differences. First, that paper considered a "0/1" task (i.e., the task is either complete or it is incomplete and partial work cannot be saved or submitted). Second, Bisin and Hyndman (2020) considered the case of "hard" deadlines (i.e., if the task is not completed by the deadline, then late submission is not possible). Our results in that paper showed that performance is actually worse in the presence of deadlines. We argued that this is because, in the presence of deadlines, subjects delay work until near the deadline and are unable to complete the task in time due to poorly formed expectations about task difficulty. Burger et al. (2011) also show that subjects who face intermediate performance targets before reaching an ultimate goal perform significantly worse than subjects without intermediate performance targets.

It is possible that the different nature of tasks could be an important factor in explaining the differences. For instance, partial work can be submitted to receive some credit in Ariely and Wertenbroch (2002), while it cannot in Bisin and Hyndman (2020). Indeed, the ability to trade-off effort/time for quality could materially affect the role that deadlines have in performance. This makes an attempted close replication all the more important because the differences in designs can merely suggest a potential lack of robustness to changes in the structure of tasks and deadlines, while a close replication can speak more forcefully to the robustness of the result.

More generally, there are broad questions about the replicability of many results in economics, psychology and other disciplines (e.g., Open Science Collaboration 2015, Camerer et al. 2016, 2018, among others). Like many social scientists, we believe that it is important to test the replicability/reproducibility of important papers that are influential within academic circles, especially when they also have a large impact on public opinion. With its high citation count and its numerous representations reported in the press (examples include Surowiecki 2010, Jaffe 2014, Thompson 2014), we believe that Ariely and Wertenbroch (2002) merits undergoing the scrutiny of a replication.

2 Methods

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, subjects were asked to read and provide spelling and grammatical corrections for three

different texts that were generated with a post-modern text generator[2] and had errors inserted into them. To this end, we used the same post-modern text generator to generate texts and then 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. Participants were recruited to the study via an email from the subject 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, subjects 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.

Registered subjects came to the lab where they received instructions about the tasks (providing grammatical/spelling corrections to three texts over a three-week period), their compensation and deadlines. Subjects were randomly assigned to one of three deadline treatments: 1. no intermediate deadlines (No Deadlines), 2. exogenous and evenly spaced deadlines (Exogenous Deadlines), and 3. self-imposed deadlines (Endogenous Deadlines). After reading the instructions, subjects in the Endogenous Deadlines treatment were given the option but not the obligation to impose binding deadlines on any or all of the three tasks. After completing all of the in-person requirements, subjects received an email which contained their login code to access the experimental platform later, as well as a reminder of the deadlines that they faced.[3] Absent any intermediate deadlines determined by the treatment (i.e., Exogenous Deadlines treatment) or by a particular subject (i.e., Endogenous 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 approximately 10 days beyond this three week deadline.[4] After the close of the experimental platform, subjects' work was evaluated and payments were transferred to a debit card that they were given during their initial visit to the lab.

There are several unavoidable differences between our replication and the study in the original paper, partially due to the large time passed since the original study:

1. Our study is computerized and precautions were taken to make it difficult for participants to enter the text into a computerized grammar/spell checker.[5] Subjects were asked to record their corrections in a text box on the computer interface using the format "Line XX: beleive $\rightarrow$ believe;". It is possible that this way of reporting errors differs from how subjects reported errors in the original study.

2. We created six substantially identical tasks and subjects were given three randomly selected

[2]The post-modern text generator is available at https://www.elsewhere.org/pomo/.

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

[4]After this point, it would be difficult to impossible to earn money net of the penalties for being late.

[5]Specifically, we used Javascript to disable right-clicking, copy and paste commands, and printing. We also checked that Javascript was enabled on the browser and if not, did not display the tasks. 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. While a sophisticated and dedicated person 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.

tasks.

3. Incentives were doubled. Specifically, subjects earned \$0.20 per error found and were penalized \$2 for each day late, while the incentives were \$0.10 per error and \$1 per day late in the original.

4. In order to avoid the possibility of losses, it was agreed with our Institutional Review Board that subjects 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 penalties that subjects incurred. If penalties exceeded the \$5 held in reserve, then subjects would not receive any additional payment, nor would they be liable for the penalties.[6]

2.1 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 sized. The highest required sample size was $N = 33$ (or 11 subjects per treatment). If we use the `contrast` option to account for pairwise treatment tests, then the highest required sample size was $N = 105$ (or 35 subjects per treatment).[7] Although we are not specifically interested in repeating all pairwise comparisons (indeed, such pairwise comparisons were not reported in the original paper), we decided that we would recruit 40 subjects per treatment to the initial in-person session to try to meet this higher standard. Another factor in our recruitment decision is because Bisin and Hyndman (2020) show that a substantial number of subjects who initially start the experiment do not finish. Although Ariely and Wertenbroch (2002) do not report that any subjects failed to finish and subjects who fail to finish should remain in the dataset, we wanted to have data to analyze performance conditional on finishing as a supplemental analysis. We conducted six initial lab sessions (two for each treatment) on February 2, 2024 and February 6, 2024, recruiting 120 subjects in total and 40 subjects for each treatment.[8] We pre-registered our replication.

[6] In our analysis of the original data, which was provided to us in 2006 by Dan Ariely (see Figure 2b and the discussion therein), it appeared that many participants *should have* lost money. On January 7, 2014, we wrote to both of the original authors and asked if a participation fee was given to guard against potential losses. On January 7, 2014, Klaus Wertembroch replied that, "I am sorry to say that I don't know the procedural details as Dan ran the studies at MIT." On January 8, 2014, Dan Ariely responded, "As far as I remember the payment for the experiment was all done at the end. There was no show up fee and we did not say anything about the fact that the payment rule could mean that subjects will lose money (and as far as I remember we 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.

[7] This would be to test the difference between days late in the Exogenous and Endogenous Deadlines treatments.

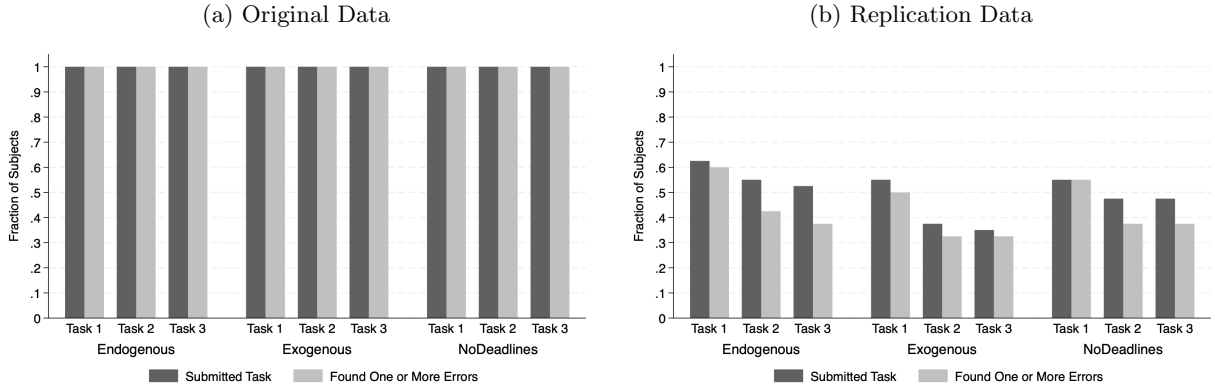
[8] For the main performance metrics (errors found, payment and days late), we did not find any differences between the sessions. There is, however, one notable difference between the two sessions for the Endogenous Deadlines treatment. Overall, deadlines were slightly more common for the first two tasks in Session 1 (21 of 21) versus Session 2 (16 of 19). For the third task it was 19 of 21 and 16 of 19, respectively. More significantly, the deadlines for the

3 Results

It is natural to believe that deadlines may affect both the extensive (i.e., whether the task gets done at all) margin and the intensive (i.e., how well the task gets done) margin. Figure 1 contains summary information on task completions for both the original data in Ariely and Wertenbroch (2002) (panel (a)) and our replication data (panel (b)). For each task, we report the fraction of subjects who submitted the task, and the fraction who submitted the task *and* correctly found at least one error. To give a first indication of some key differences in results with the original study, first notice that in the original study *all* subjects submitted *all* tasks and, indeed, found at least one error. That is, in the original study, there is no extensive margin effect because all subjects complete all tasks.

In contrast, in our replication, many subjects fail to submit some or all of the tasks. Therefore, there is potentially scope for effects of deadlines at both margins. However, upon examining Figure 1(b), whether we look at subjects who submitted a task irrespective of finding any errors or those who submitted the task and found at least one error, there are only very small differences between treatments and, moreover, for all tasks, subjects completed the fewest tasks in the Exogenous Deadlines treatment.

Figure 1: Summary of Task Completions



Note: Recall that "Endogenous" corresponds to the treatment where subjects self-imposed their own deadlines, "Exogenous" corresponds to the treatment where subjects faced evenly-spaced intermediate deadlines and "NoDeadlines" corresponds to the treatment where there were no intermediate deadlines.

This property of the original data – that all subjects complete all tasks – appears non-robust. Subjects who fail to submit a task could be subjects with very high present-bias and/or subjects with specific psychological characteristics, e.g., low conscientiousness. Such subjects are, arguably, generally present in the population and in the selected subject pools of university students (students routinely fail to submit homework assignments). Furthermore, the post-modern texts are, in our

first two tasks were significantly more strict in Session 1 (12.6 and 7.6 days before the final deadline) versus Session 2 (4.7 and 3.0 days before the final deadline). The same two people administered both the first and second sessions and the same procedures were used. The only apparent difference between the sessions is that the first session occurred on a Friday and the second session occurred on a Tuesday, though it is unclear how this could affect the severity of deadlines by more than a day (e.g., avoid a Friday deadline).

opinion, quite frustrating to read, making it plausible to expect some students would judge the task not worth completing.

Remark 1 *The finding that not all subjects complete all tasks is, however, similar to our earlier work (Bisin and Hyndman 2020), despite the different nature of the tasks and the deadlines. Further similarities to Bisin and Hyndman (2020) become apparent once we dig into the 51 subjects who did not submit Task 1 (and so could not have submitted subsequent tasks). Like in that paper, some subjects (19) never logged in after the initial lab session, while an even larger number (32) logged in at least once to see the first page of the first task. Indeed, 19 of these 32 subjects completed at least one page of the first task and 15 of them found at least one error (average 9.8, min 0, max 28).*[9]

Turn now to the main performance results. In Figure 2 we illustrate subjects' performance 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) and panel (b) is from our reanalysis of the original data (discussed below). Panels (c) and (d) contain our replication data regarding the effects of different deadline structure on the original three dimensions of subject's performance in Ariely and Wertenbroch (2002). Panel (c) contains all subjects, whether they submitted a task or not, while panel (d) contains only those subjects who submitted all three tasks.

Comparing panel (a) with panels (c) and (d), there is strong visual evidence that the results report in Ariely and Wertenbroch (2002) do not replicate. The directional effects that we find do not align with the original paper and our error bars are substantially larger. To provide some statistical evidence to support what is visually very clear, in Table 1, we provide the $F-$statistics from an analysis of variance of performance metric by treatment. Ariely and Wertenbroch (2002) report that "[a]ll differences were statistically significant (all ps<0.01)", and our analysis of their data confirm this. However, as can be seen from Table 1, our replication data – regardless of which cut of the data we examine – shows no statistical evidence that performance is influenced by the deadline condition.[10] In Appendix A, we also provide results on the survey responses from the original study and our replication.[11] As with the performance results, there is almost no correspondence between Ariely and Wertenbroch (2002) and our replication.

[9] It is perhaps surprising that the students who actually logged in and attempted the tasks did not submit them. In designing the experiment, we tried to keep "transactions costs" of participating in our study low. To this end, we reminded subjects on each task page that "you can submit with no input if you no longer wish to work on the task." Therefore, with approximately 30 clicks of a mouse, they could have submitted all the tasks and they would subsequently earn the additional $5 participation fee that was held in reserve against potential penalties. Additionally, there was no incremental cost to receive the money as participants were given debit card in the lab session and were told that additional earnings would be loaded onto the cards at the conclusion of the study. Perhaps these subjects thought that they would return to the task later, but did not, or that there is some kind of norm against receiving payment for work that was not done.

[10] The critical value for significance at the 5% level is at least 3.07 (depending on whether all subjects or only a subset are included). In all cases, the $F-$statistic reported is substantially below the threshold, indicating a lack of any meaningful treatment effect.

[11] Only subjects who completed all three tasks completed this survey.

Figure 2: Comparison of Treatments

(a) Figure 2 From Ariely and Wertenbroch (2002)

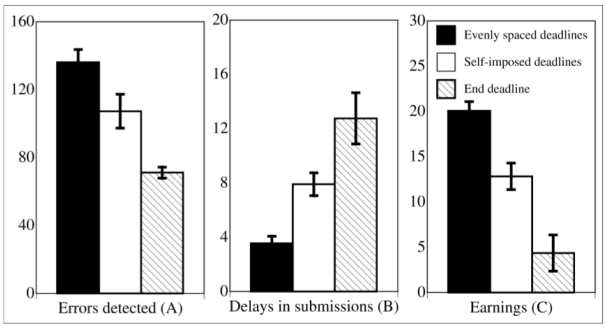
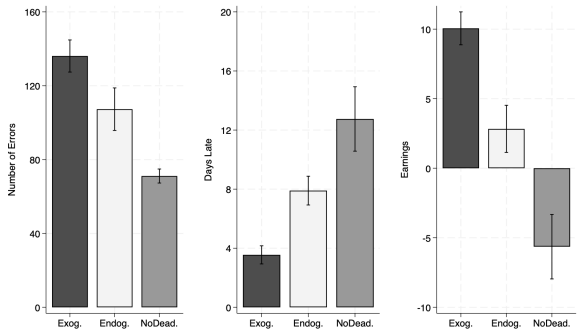
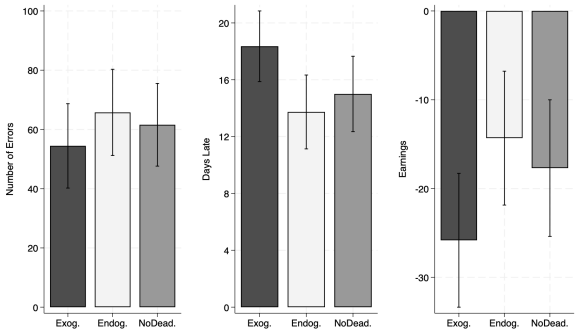


Fig. 2. Mean errors detected (a), delays in submissions (b), and earnings (c) in Study 2, compared across the three conditions (error bars are based on standard errors). Delays are measured in days, earnings in dollars.

(b) Figure 2 From Our Analysis of Original Data



(c) Replication Data: All Subjects



(d) Replication Data: Submitted All Three Tasks

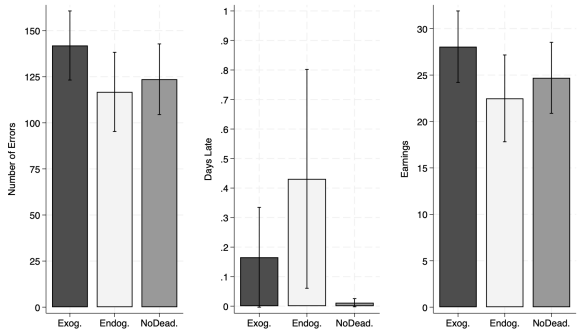


Table 1: $F-$Statistics From ANOVA (Performance Metric on Treatment)

	Reported in	Our Analysis of	Replication Data	
Performance Metric	AW (2002)	Original Data	All Subjects	Sub. All Three Tasks
Errors	Unknown:	19.081	0.210	0.507
Days Late	Reported "All	13.953	1.120	1.096
Payment	ps< 0.01"	25.623	0.794	0.553
Degrees of Freedom		$(2, 57)$	$(2, 117)$	$(2, 51)$

Recall that the original authors reported that all subjects earned positive amounts of money (Figure 2a). As we see in Figure 2b, which contains our analysis of the original data, it appears that earnings for many were negative (see also Footnote 6). In part, because of this apparent discrepancy, we will delve deeper into a reanalysis of the original data below.

3.1 Within-subjects Behavior

In this section, we perform a further comparative analysis of the original Ariely and Wertenbroch (2002) data and our replication data with respect to several important associations that one could argue should be present in subjects' behavior given the results reported in the original paper.

Ariely and Wertenbroch (2002) show three things: (i) performance is highest under exogenous deadlines, (ii) subjects spend more time on tasks under exogenous deadlines and (iii) subjects like the tasks least under exogenous deadlines. The suggested causal mechanism is that deadlines cause changes in time allocation, which in turn cause changes in performance and subjective evaluation. They demonstrate each of their results separately by conducting ANOVA of the given metric on the treatment. However, if time spent on the task affects performance, then we would expect this to be true even after controlling for assignment to treatment.

We can test for this with a random effects regression of errors found on each task on self-reported time spent on each task. We report the results of this exercise for both our data and the original data in Table 2. As can be seen, in the replication data, controlling for treatment, we find that the number of errors found is significantly positively associated with the self-reported time spent on each task. In contrast, and surprisingly, after controlling for treatment, there is no such association in the original data.[12]

Table 2: Relationship Between Errors Found and Time on Task

	Replication Data		Original Data	
Time Spent on Task	0.138***	(0.036)	0.128	(0.203)
Constant	37.363***	(6.300)	41.792***	(6.191)
Treatment Controls	Yes		Yes	
Num Subjects	54		60	
R^2 (Overall)	0.233		0.212	

Notes: e_i denotes the number of errors found in Task i, for $i = 1, 2, 3$. ** $p < 0.05$, *** $p < 0.01$.

In addition to the relationship between errors found and time, we would also expect to observe a within-subjects correlation between the errors found on any two tasks i and j. For example, we would expect a subject who finds a high number of errors on task 1 to be more likely to find a high number of errors on task 2 – either because of high effort or high ability. Moreover, we would expect this correlation to be present regardless of their treatment assignment.

[12] If we *do not* control for treatment assignment, then there is a significant positive association in the original data. However, as argued in the text, the association *should* be present regardless of one's assignment to treatment.

In Table 3 we report the correlations matrices (conditional on the treatment) for both the original data and also our replication sample for the number of errors found on each task. As can be seen, in the original data, the correlations are relatively low, sometimes negative and in only one case is the correlation (between Task 1 and Task 2) significant at the 5% level. On the other hand, in our replication data, the smallest correlation is 0.821 and in all cases the correlations are significant at the 1% level or better.[13]

Table 3: Correlation Matrices on Errors Found Per Task

(a) Replication, No Deadlines

	e_1	e_2	e_3
e_1	1		
e_2	0.821***	1	
e_3	0.842***	0.953***	1

(b) Replication, Endogenous

	e_1	e_2	e_3
e_1	1		
e_2	0.922***	1	
e_3	0.871***	0.894***	1

(c) Replication, Exogenous

	e_1	e_2	e_3
e_1	1		
e_2	0.907***	1	
e_3	0.868***	0.850***	1

(d) Original, No Deadlines

	e_1	e_2	e_3
e_1	1		
e_2	0.199	1	
e_3	-0.167	0.197	1

(e) Original, Endogenous

	e_1	e_2	e_3
e_1	1		
e_2	0.483*	1	
e_3	0.427	0.296	1

(f) Original, Exogenous

	e_1	e_2	e_3
e_1	1		
e_2	-0.021	1	
e_3	-0.209	0.244	1

Notes: e_i denotes the number of errors found in Task i, for $i = 1, 2, 3$. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The above results show that our replication does produce important and sensible associations in subjects' behavior, which are not present in the original Ariely and Wertenbroch (2002) data. This provides us with greater confidence about the robustness of our lack of replication, and suggests that the lack of replication is not because our data is too noisy or otherwise flawed.

3.2 Further Reanalysis of the Original Data

Give the differences identified thus far, we conducted a deeper reanalysis of the original data and compared it to our replication data. The details are provided in Appendix B. We simply note here that there are very stark differences in how subjects report the time spent working on each task. We find that subjects overwhelmingly report round numbers, while the original data contains very precise numbers. We also find stark differences in correlations between the subjective measures elicited after subjects completed the tasks. Specifically, we find consistently positive and highly significant correlations, while the original data suggests few significant correlations, and correlations equally likely to be positive or negative.

[13] Although not shown, if we restrict attention to only those subjects who submitted all three tasks, the same result holds in terms of significance, but the correlations are somewhat smaller (though always above 0.68).

4 Conclusions

Our results show that the original result of Ariely and Wertenbroch (2002) does not replicate. To be precise, in our replication, exogenous and evenly spaced deadlines do not improve task performance. This result complements other studies, notably Bisin and Hyndman (2020) and Burger et al. (2011), which already hinted at a lack of robustness of the effect of deadlines in Ariely and Wertenbroch (2002). Furthermore, we show that, in comparison with our replication, the data in Ariely and Wertenbroch (2002) display some relevant lack of robustness while missing several sensible important associations in subjects' behavior.

In fairness, over 20 years have passed between the original study and our replication. As we noted, because of the lack of original material and further necessitated by technological changes in the intervening years, there are differences between our replication and the original study. It is possible that one or more of these differences is responsible for the different results. However, the changes in design we instituted are, in our opinion, relatively minor. We claim that the changes in design we instituted would – if anything – encourage more tasks to be completed. Indeed, we allowed subjects to complete the work online, without requiring subjects 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, subjects actually had stronger incentives to submit the tasks – even poor quality work – in order to receive the additional payment. Most importantly, even considering these potential mitigating factors, our lack of replication still means that the result is not robust to slight modifications in experimental protocols.

In the end, we would stress that the lack of replication of original result of Ariely and Wertenbroch (2002) implies that the commonly received wisdom that exogenous spaced deadlines are an effective mechanisms to limit procrastination is possibly false. This leads us to believe that it is important to go back to the drawing board to develop new testable hypotheses for why deadlines might work and, if so, in what contexts. Deadlines play an important role in both educational institutions and many business practices. A better understanding of their effectiveness could lead to meaningful improvements in these areas.

References

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

Bisin, Alberto, Kyle Hyndman. 2020. Present-bias, procrastination and deadlines in a field experiment. *Games and Economic Behavior* **119** 339–357.

Burger, Nicholas, Gary Charness, John Lynham. 2011. Field and online experiments on self-control. *Journal of Economic Behavior & Organization* **77** 393–404.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke

Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, Hang Wu. 2018. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behavior* **2** 637–644.

Camerer, Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, Hang Wu. 2016. Evaluating replicability of laboratory experiments in economics. *Science* **351**(6280) 1433–1436.

Jaffe, Eric. 2014. Self-imposed deadlines don't stop procrastination. here's what might. *Fast Company* **March 26**.

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

Surowiecki, James. 2010. Later. *The New Yorker* **86** 210–210.

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

A Comparison of Subjective Metrics between Original and Replication Data

In the original paper, the authors had subjects complete a survey about how much they enjoyed the task, how much time they spent on the task, and other subjective measures. For subjects who submitted all three tasks, we collected similar subjective measurements.[14] Table 4a contains averages by treatment for each metric, while Table 4b conducts the same analysis for the replication data. In the last row of each sub-table, we report the $F-$statistic from an ANOVA analysis of the subjective metric by treatment. As was the case with performance data, there is virtually no correspondence between the original and replication data sets and, again, we find a general lack of any significance across all six measures, where the original paper reports significant treatment effects for four of the six measures.

Table 4: Comparison of Treatments: Subjective Metrics

(a) Original Data

Treatment	Time	Like	Interesting	Quality	Grammar	Ideas
Exogenous	83.95	29.50	19.25	17.25	19.25	25.00
Endogenous	69.90	23.55	23.70	26.30	34.45	32.60
NoDeadlines	50.80	32.50	38.00	41.00	45.00	33.00
ANOVA $F(2, 57)$	37.817	1.343	5.670	9.677	9.600	0.898

(b) Replication Data (Submitted All Three Tasks)

Treatment	Time	Like	Interesting	Quality	Grammar	Ideas
Exogenous	216.07	75.29	70.14	64.64	63.71	64.93
Endogenous	215.95	57.86	53.33	62.14	60.43	57.95
NoDeadlines	232.47	55.05	47.84	53.21	55.58	56.68
ANOVA $F(2, 51)$	0.052	1.826	1.839	0.732	0.326	0.304

Finally, note that the original paper discusses conducting their analysis on the average of all subjective measures:

> An analysis of the average subjective evaluation across the five questions revealed a pattern that was the opposite of the performance results, $F(2, 57) = 17.06$, $p < .001$. [p. 223]

The original dataset contains all five subjective variables as well as a variable called "All Subjec", which we would expect to be the average over all subjective measures. However, when we compute the average directly over the subjective measures, for the 51 of 60 of subjects the absolute difference

[14] Without access to the source material, it is possible that we did not ask precisely the same questions, but every effort was made to match the language that was reported in the original paper.

with the variable "All Subjec" is strictly larger than 0.2. Moreover, whether we use the variable "All Subjec" or the average that we compute from the data, we cannot replicate the $F-$statistic reported above, getting instead 13.91 and 19.04, respectively.

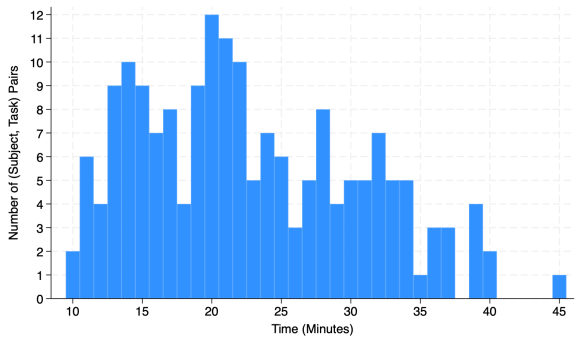
B Deeper Reanalysis of the Original Data

B.1 Focal Points in Reporting Time

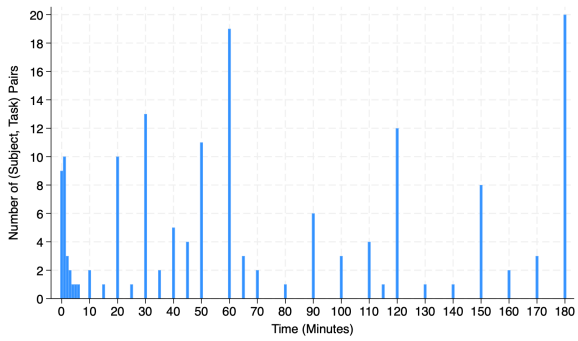
In Figure 3 we report histograms of time estimates that subjects provided for each task (assuming that they finished all tasks and, therefore, the survey. As can be seen in panel (a), subjects in the original data reported very precise time estimates and only 38 out of 180 (Subject, Task) pairs ended in a 0 or a 5. In contrast, 145 out of 162 (Subject, Task) pairs reported time estimates that ended in a 0 or a 5. Another notable difference is that the time estimates are very different between the original and replication data. In the former, time estimates always ranged from 10 to 45 minutes, while in our study, the time estimates range from 0 to 180 minutes.

Figure 3: Subjects' Self-Reported Time to Complete Each Task

(a) Original Data



(b) Replication Data



B.2 Correlations on Subjective Measures

Table 5 looks at the correlations between subjective measures about the task such as grammar, the quality of the writing, how interesting the tasks were, etc. Although we have no clear prior on whether the correlations should be positive or negative for any pairwise correlation, a natural hypothesis is to expect a similar pattern of correlations between the replication data and the original data. As can be seen from the table, in terms of significance, the patterns are very different. In the original data, there are only 4 of 30 possible pairwise correlations that are significant at the 5% level.[15] In contrast, for the replication data, 29 of 30 correlations are significant. In terms of the signs of the correlations, in the original data, 15 out of 30 correlations are negative, while all 30 correlations are positive in the replication data.

[15] Just by chance, we would expect about 2 significant correlations out of 30.

Table 5: Correlation Matrices on Subjective Measures of Tasks

(a) Original, No Deadlines

	ideas	gram.	int.	like	qual.
ideas	1				
gram.	-0.210	1			
int.	0.226	-0.008	1		
like	-0.0161	-0.548*	-0.134	1	
qual.	0.101	-0.023	-0.321	0.104	1

(b) Replication, No Deadlines

	ideas	gram.	int.	like	qual.
ideas	1				
gram.	0.836***	1			
int.	0.598**	0.483*	1		
like	0.533*	0.335	0.814***	1	
qual.	0.746***	0.777***	0.669**	0.616**	1

(c) Original, Endogenous

	ideas	gram.	int.	like	qual.
ideas	1				
gram.	0.238	1			
int.	0.093	0.771***	1		
like	0.571**	0.214	-0.030	1	
qual.	-0.278	0.0318	0.139	-0.250	1

(d) Replication, Endogenous

	ideas	gram.	int.	like	qual.
ideas	1				
gram.	0.848***	1			
int.	0.876***	0.725***	1		
like	0.838***	0.760***	0.967***	1	
qual.	0.901***	0.959***	0.757***	0.758***	1

(e) Original, Exogenous

	ideas	gram.	int.	like	qual.
ideas	1				
gram.	0.304	1			
int.	0.097	-0.241	1		
like	0.295	-0.088	-0.100	1	
qual.	0.0637	-0.349	0.479*	-0.201	1

(f) Replication, Exogenous

	ideas	gram.	int.	like	qual.
ideas	1				
gram.	0.768**	1			
int.	0.840***	0.742**	1		
like	0.793***	0.837***	0.897***	1	
qual.	0.785***	0.817***	0.854***	0.898***	1

Notes: "gram." indicates "grammar", "int." indicates "interesting" and "qual." indicates "quality". Refer to Appendix A for more details on these variables. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$