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

Kyle Hyndman[1] and Alberto Bisin[2]

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

April 9, 2025

Abstract

We present the results of a replication of Study 2 from Ariely and Wertenbroch (2002). 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. Evenly spaced deadlines exogenously imposed on participants do not stand apart for their effectiveness in reducing procrastination in participants. While our replication of the main result fails, our data do show several patterns of participant behavior which should be expected. This suggests to us that our replication attempt is of sufficient quality, but the result itself is not robust.

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 authors' respective universities. Artificial intelligence: No artificial intelligence assisted technologies were used in this research or the creation of this article. Ethics: This research received approval from an author's Institutional Review Board (ID: IRB-24-151).

Preregistration: The hypotheses and methods were preregistered on January 30, 2024 for the original experiment (now in supplemental materials) and January 29, 2025 for the new experiment. 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.

1 Introduction

Procrastination is ubiquitous in modern society, with self-help books (Steel 2011) and dozens academic papers across psychology, economics and numerous other disciplines.[1] Given its prevalence, it is no surprise that considerable attention has been devoted to how people can overcome procrastination. Psychologists and management scholars have pointed to goal-setting as one way to overcome procrastination (e.g., Locke and Latham 2006). Within this literature, psychologists and economists have also studied deadlines and whether they may be effective tools to overcome procrastination. A leading theory for procrastination is temporal motivation theory (Steel 2007). According to this, 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 and 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 presumably be done under 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 (O'Donoghue and Rabin 2001, Hyndman and Bisin 2022). Ordóñez et al. (2009) provides a more psychological perspective on some of the potential pitfalls of an over-reliance on goals.

Not surprising, within this large literature, experimenters have examined the role of deadlines in overcoming procrastination. Arguably the most well-known and most impactful paper addressing this trade-off experimentally is Ariely and Wertenbroch (2002).[2] The authors open 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 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 to answer these questions, but the one we will focus on is Study 2 – The Proofreading Study.

In this 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: exogneous and evenly spaced deadlines, self-imposed (i.e., endogenous) deadlines or no intermediate deadlines. Ariely and

[1] A Google Scholar search (on Mach 28, 2025) for papers with the word procrastiation in the title yielded over 150 results.

[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.

Wertenbroch (2002) showed that a substantial number of people are willing to self-impose deadlines on themselves and that the deadlines were partially effective. However, their main result is that performance is highest in the exogenous deadlines treatments.

Subsequently, two papers have provided experimental evidence that deadlines may not be effective. While neither paper represents an attempt to replicate the original paper, the closest paper is Bisin and Hyndman (2020). Like Ariely and Wertenbroch (2002), they study the role of deadlines in a multiple-repeated task environment. There are two important differences, which were made to simplify the environment to facilitate a comparison to theoretically derived predictions from the canonical $\beta - \delta$ model of hyperbolic discounting used in economics. First, Bisin and Hyndman (2020) consider a task that is either finished or not finished, and participants received a monetary reward for each task that was finished. This differs from Ariely and Wertenbroch (2002) because in that paper participants could submit partially completed work for a lower payment. Second, unlike the "soft" deadlines of Ariely and Wertenbroch (2002), Bisin and Hyndman (2020) considered the case of "hard" deadlines. That is, the task could not be submitted after the deadline. The results in Bisin and Hyndman (2020) paper showed that performance is actually worse in the presence of deadlines. The authors argued that this is because, in the presence of deadlines, participants delay work until near the deadline and are unable to complete the task on time due to poorly formed expectations about task difficulty.

The second closely related paper is Burger et al. (2011). They consider a setting in which participants were required to log 75 hours of studying inside a specific location over a five-week period. Successful completion generated a payment of $95, while unsuccessful completion led to no payment. The authors considered two treatments: no intermediate deadlines and evenly-spaced intermediate deadlines, which required participants to log approximately 12 hours for each of the first four weeks. The completion rates for participants in the intermediate deadlines treatment was approximately 40%, while over 60% of participants in the no deadlines treatment completed the task. That is, performance was *lower* in the presence of binding intermediate deadlines.

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 itself.

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 both within academic circles and public opinion more broadly. 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. This scrutiny seems particularly warranted given that one of the authors – Dan Ariely – has been repeatedly the subject of discussions regarding research practices in recent years has had at least one influential paper retracted (Stern 2023, Lewis-Kraus 2023).

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, 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 errors inserted into them. In January 2014, while working on another project, we emailed the original authors to ask for instructional materials related to the study. Wertembroch 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, 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. Second, 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.

Registered participants came to the lab where they were given a login code to the online platform where the experiment would take place both within and outside of the experimental 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 (No Deadlines), 2. exogenous and evenly spaced deadlines (Exogenous Deadlines), and 3. self-imposed deadlines (Endogenous Deadlines). This randomization was done at the participant level, via the login codes that they received.

After reading the instructions, participants in the Exogenous and No Deadline treatments were given the deadline (date and time) for each of the three tasks. For participants in the Endogenous Deadlines treatment, they were given the option but not the obligation to impose binding deadlines on any or all of the three tasks. Except for an earliest and a latest allowable deadline, deadlines were otherwise unrestricted. Participants set deadlines separately for each task (any date and time

[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.

within the allowed bounds) and deadlines for different tasks could be the same or different from each other according to their preference.

After completing all of the in-person requirements, participants 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.[4] Absent any intermediate deadlines determined by the treatment (i.e., Exogenous Deadlines treatment) or by a particular participant (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.[5]

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.). We used the same 100 point scale as the original authors. 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.[6] Lastly, we asked participants to comment on how the deadlines they faced either helped or hindered their performance. 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.

There are several unavoidable differences between our replication and the study in the original paper. This is because we did not have access to the original study materials and also partially due to the large time passed since the original study:

1. We made the conscious decision to fully computerize our study. We believed that with modern advances in scanning, character recognition and spelling/grammar checking that 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.[7] 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

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

[5] After this point, it would be exceedingly difficult, if not impossible, to earn money net of the penalties for being late.

[6] We do not now whether the original authors imposed an upper bound on the allowable responses for participants. Nevertheless, to avoid mistakes that might cloud the analysis, 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.

[7] 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. It was possible for a participant to take a picture and upload 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 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.

how participants reported errors in the original study.

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

3. Incentives were doubled. Specifically, participants 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 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 penalties that participants incurred. If penalties exceeded the \$5 held in reserve, then participants would not receive any additional payment, nor would they be liable for the penalties.[8]

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 2024 we conducted a nearly identical experiment with the only difference being that participants were randomized to treatment at the session level. For this experiment, 120 participants participated and the same experimenter and lab assistant conducted the experiments and scored the participants' performance on tasks, making a valid comparison between the 2024 and 2025 samples possible. We report the results in the appendix and note here that the results are virtually identical.

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 participants 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 participants per treatment).[9] Although we are not specifically interested in repeating all pairwise comparisons (indeed, such pairwise comparisons were not reported in the original paper),

[8] 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 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.

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

we decided that we would recruit 40 participants 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 participants who initially start the experiment do not finish. Although Ariely and Wertenbroch (2002) do not report that any participants failed to finish and participants 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 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. Randomization to treatment was done at the participant level. We pre-registered our replication.

3 Results

We begin our discussion of results by examining participants' choice to self-impose deadlines in the Endogenous Deadlines treatment. 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. Conditional on setting a binding deadline, the self-imposed deadlines were, 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.

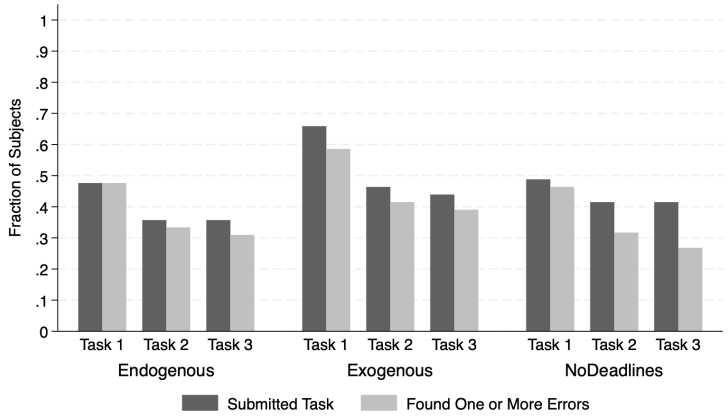
Turning our attention to actual performance in the experiment, we first address the question of whether participants complete the task at all. We do so because Ariely and Wertenbroch (2002) implicitly indicates that all participants complete all tasks. In contrast, related experiments on the role of deadlines by Bisin and Hyndman (2020) and Burger et al. (2011) suggest that a substantial number of participants fail to complete the tasks and effectively drop out. Wilcox et al. (2016) specifically planned for attrition in their study design.[10] 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 fail to submit some or all of the tasks. This contrasts starkly with the original result that all participants completed all tasks.

Upon closer examination of Figure 1, it appears that participants complete the most tasks in the Exogenous Deadlines treatment (overall rate of 52% for submissions and 46% for finding one

[10] 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) report attrition rates of 10-20% in their studies that involve participants performing work over a period of several weeks (see also, Charness and Gneezy 2009, Halevy 2015, Acland and Levy 2015, among others). In Psychology, Flick (1988) and more recently 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) consider a model in which decision makers must complete a single task before some final deadline, and their cost to complete the task varies stochastically over time. They show that decision makers, whether they have exponential or hyperbolic time discounting preferences, generally 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.

or more error). However, an ANOVA fails to reject a treatment effect for either of the metrics ($F(2, 121) = 0.74$ for submissions and $F(2, 121) = 0.77$ for finding one or more errors; in both cases $p > 0.1$).

Figure 1: Summary of Task Completions



Note: Recall that "Endogenous" corresponds to the treatment where participants self-imposed their own deadlines, "Exogenous" corresponds to the treatment where participants 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 participants complete all tasks – appears non-robust. 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.

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 structure on the original three dimensions of participants' performance in Ariely and Wertenbroch (2002). Panel (b) contains all participants, whether they submitted a task or not, while panel (c) contains only those participants who submitted all three tasks.

When panel (a) is compared to panels (b) and (c), there is strong visual evidence that the results report in Ariely and Wertenbroch (2002) are not replicated. 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)" (p. 222). However, as can be seen from Table 1, our replication data – regardless of which cut of the data we examine – show

Figure 2: Comparison of Treatments

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

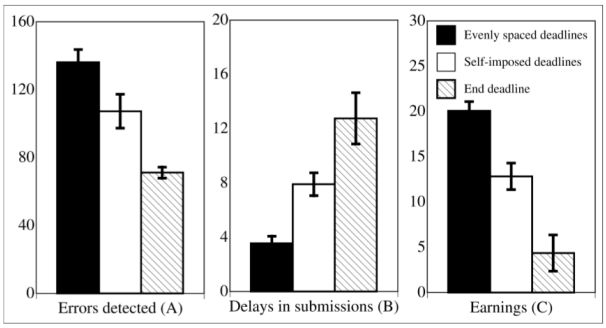
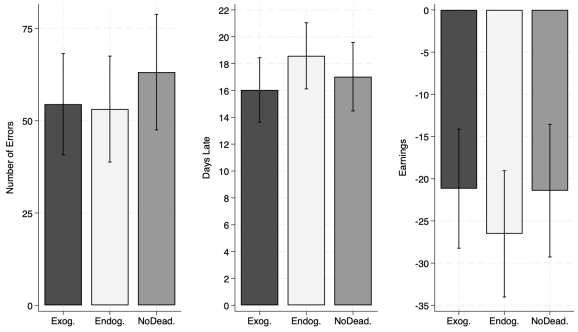
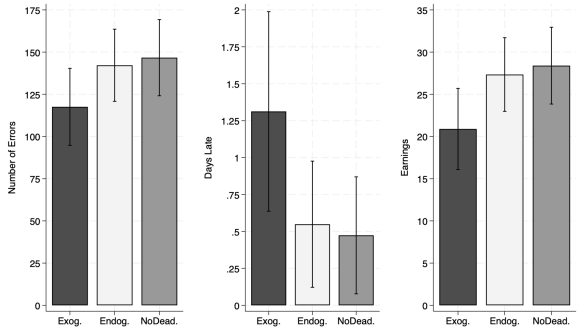


Fig. 2. Mean errors detected (a), delays in submissions (b), and earnings (c) in Study 2, compared across the three conditions (error bars are based on standard errors). Delays are measured in days, earnings in dollars.

(b) Replication Data: All Participants



(c) Replication Data: Submitted All Three Tasks



Note: Panel (a) is taken directly from the original publication. For panels (b)–(d), "Exog" corresponds to "Evenly spaced deadlines" in panel (a), while "Endog" and "NoDead" correspond to "self-imposed deadlines" and "End deadline", respectively. Error bars are based on standard errors

no statistical evidence that performance is influenced by the deadline condition.[11] In Appendix A, we also provide results on the survey responses our replication.[12] In the original paper, the authors reported significant differences across treatments for self-reported time spent on the task and the average of the five subjective measures. However, as we show, we find no differences for any of the measures and the across-treatment comparative statics do not match those in the original paper.

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

	Reported in	Replication Data	
Performance Metric	AW (2002)	All Participants	Sub. All Three Tasks
Errors	Unknown:	0.181	0.697
Days Late	Reported "All	0.353	1.107
Payment	ps<0.01"	0.215	1.103
Degrees of Freedom		$(2, 121)$	$(2, 49)$

Note: "Reported in AW (2002)" means that we have extracted the information from the original paper. "All Participant" 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.

3.1 Within-subjects Behavior

As with any replication effort, it is important for us to convince the reader that a high-quality replication was conducted. This is all the more important given the strong lack of replication of the original result. To provide support for the quality of our replication we will delve into other important associations which one could argue should be present in the participants' behavior and show that they are present in our replication data.

Ariely and Wertenbroch (2002) show three things: (i) performance is highest under exogenous deadlines, (ii) participants spend more time on tasks under exogenous deadlines and (iii) participants 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 in Table 2. As can be seen, controlling for treatment, we find that the number of errors found is significantly positively associated with the self-reported time spent on each task.[13] What is more, the relationship demonstrated in Table

[11] The critical value for significance at the 5% level is at least 3.07 (depending on whether all participants 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.

[12] Only participants who completed all three tasks completed this survey.

[13] The correlation between errors found and time on task remains significantly positive even if we do not include treatment controls.

2 holds even though *none* of the ANOVA comparisons reported by the original authors hold true in our data set: (i) there is no treatment effect on performance (Table 1), (ii) there is no treatment effect on time spent on task (Table 4), and (iii) there is no treatment effect on how much participants like the task (also Table 4).

Table 2: Relationship Between Errors Found and Time on Task

	Replication Data	
Time Spent on Task	0.270^{***}	(0.039)
Constant	18.769^{***}	(6.037)
Treatment Controls	Yes	
Num Participants	49	
R^2 (Overall)	0.306	

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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 participant 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. In Table 3 we report the correlations matrices (conditional on treatment) for our replication sample for the number of errors found on each task. The smallest correlation is 0.793 and in all cases the correlations are significant at the 1% level or better.[14]

Table 3: Correlation Matrices on Errors Found Per Task

(a) Replication, No Deadlines

	e1	e2	e3
e_1	1		
e_2	0.865^{***}	1	
e_3	0.793^{***}	0.901^{***}	1

(b) Replication, Endogenous

	e_1	e_2	e_3
e_1	1		
e_2	0.935^{***}	1	
e_3	0.835^{***}	0.918^{***}	1

(c) Replication, Exogenous

	e_1	e_2	e_3
e_1	1		
e_2	0.872^{***}	1	
e_3	0.911^{***}	0.975^{***}	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 participants' behavior. 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.

[14] Although not shown, 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).

4 General Discussion

Our results fail to replicate the main finding of Ariely and Wertenbroch (2002). Specifically, we find no evidence that exogenous, evenly spaced deadlines improve performance on three metrics of interest: number of errors found, number of days late and overall payment. This aligns 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).[15]

There are several reasons why any particular experimental study may fail to replicate: statistical variation, combined with a publication bias for positive results, can lead initial findings to represent outliers rather than robust effects; subtle contextual differences in participant populations across time periods or locations may impact results; and the passage of time can alter how participants engage with experimental tasks. Additionally, research practices such as p-hacking can lead to published findings that do not hold up in subsequent replication attempts.[16]

In fairness, over 20 years have passed between the original study and our replication, and this period included a global pandemic which may have altered whether and how deadlines work.[17] As we noted, because of the lack of original material and further necessitated by technological changes in the intervening years, there are known differences between our replication and the original study. Additionally, there may be unknown differences between the experimental protocols that we followed and those that Dan Ariely followed. It is possible that one or more of these factors is responsible for the different results we observe.

In our opinion, the changes in design we have instituted are relatively minor. We claim that the changes in design that we instituted would, if anything, encourage more tasks to be completed. Indeed, 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. 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. Intermediate deadlines do cause

[15] Given this lack of replication, additional analysis of the original data could help illuminate the differences between our findings and the original study. To this end, we note that 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 the attached data according to our request. In writing an earlier draft of this paper, at the request of the editors - in October 2024 - we shared with Dan Ariely our analysis of the original data and asked permission to include a summary of this analysis in the paper. Dan Ariely denied us this request and, consequently, we do not report in the paper our analysis of the original data.

[16] Given the simplicity of the original data analysis, we do not suspect p-hacking as a likely reason for why we fail to replicate the original study.

[17] That said, as noted, Burger et al. (2011) and Bisin and Hyndman (2020) report similar experiments – conducted both before the pandemic and closer in time to the original study – that questioned he effectiveness of deadlines.

people to complete tasks earlier than the ultimate deadline, but do not lead to better performance overall. 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.

Acknowledgments:

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.

References

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

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

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

Augenblick, Ned, Matthew Rabin. 2019. An experiment on time preference and misprediction in unpleasant tasks. *Review of Economic Studies* **86** 941–975.

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.

Charness, Gary, Uri Gneezy. 2009. Incentives to exercise. *Econometrica* **77**(3) 909–931.

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

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

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

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

Lewis-Kraus, Gideon. 2023. They studied dishonesty. was their work a lie. *The New Yorker* **September 30**.

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

O'Donoghue, Ted, Matthew Rabin. 1999. Incentives for procrastinators. *Quarterly Journal of Economics* **114** 769–816.

O'Donoghue, Ted, Matthew Rabin. 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, Lisa D., Maurice E. Schweitzer, Adam D. Galinsky, Max H. Bazerman. 2009. Goals gone wild: The systematic side effects of over-prescribing goal setting. *Academy of Management Perspectives* **23**(1) 6–16.

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

Steel, Piers. 2011. *The Procrastination Equation*.

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

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

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

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

Zhou, Haotian, Ayelet Fishbach. 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.

A Analysis of Subjective Metrics

In the original paper, the authors had participants complete a survey about how much they enjoyed the task, how much time they spent on the task, and other subjective measures. For participants who submitted all three tasks, we collected similar subjective measurements.[18] Table 4 contains averages by treatment for each metric. In the last rows , we report both the $F-$statistic and $p-$value from an ANOVA analysis of the subjective metric by treatment. As was the case with performance data, we find a general lack of any significance across all six measures.

Table 4: Subjective Metrics (Submitted All Three Tasks)

Treatment	Time	Like	Interesting	Quality	Grammar	Ideas
Exogenous	226.44	58.50	57.61	63.94	61.89	65.44
Endogenous	305.00	64.79	55.71	71.57	69.86	58.43
No Deadlines	187.71	52.65	51.12	62.71	58.24	56.47
ANOVA $F(2, 49)$	2.405	0.489	0.141	0.445	0.742	0.398
$p-$value	0.102	0.617	0.869	0.644	0.482	0.674

Note: We used the same scale for the Time spent on tasks (minutes) and the subjective measures (100-point scale) as in the original paper.

On page 223, Ariely and Wertenbroch (2002) write, "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)." We already know that deadlines did not improve performance, nor did it appreciably influence the time spent on tasks. Therefore, the causal chain is already suspect. Nevertheless, 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 5. Upon reflection, we would 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, one should enjoy it.

B Deeper Analysis Data

B.1 Focal Points in Reporting Time

In Figure 6 we report a histogram of time estimates that participants provided for each task (assuming that they finished all tasks and, therefore, the survey. In our replication, 142 out of 147 (Participant, Task) pairs reported time estimates that ended in a 0 or a 5 and the time estimates range from 0 to 180 minutes (which were hard-coded bounds on allowable responses).

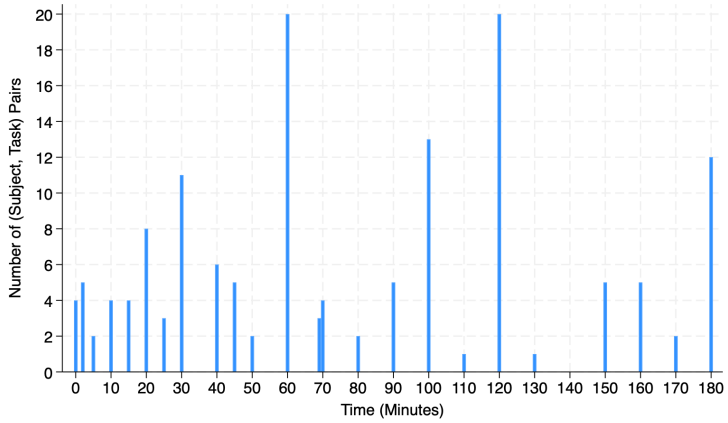
[18] Without access to the source material, it is possible that we did not ask precisely the same questions, but every effort was made to match the language that was reported in the original paper.

Table 5: Relationship Between Time on Task and Subjective Measures

	Like	Interesting	Quality	Grammar	Ideas
Time	0.117***	0.120***	0.048*	0.083***	0.073**
	(0.029)	(0.032)	(0.027)	(0.024)	(0.029)
Constant	29.086**	19.000	56.824***	44.403***	36.170***
	(11.859)	(13.122)	(10.929)	(9.637)	(11.870)
N	49	49	49	49	49
R^2	0.281	0.242	0.085	0.242	0.138

Notes: Standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Treatment controls were included in the regressions, but not reported.

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



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 earn from perfectly executing a single task was $20. Therefore, spending more than 3 hours on a task would yield an hourly pay of less than $7 per hour.

B.2 Correlations on Subjective Measures

Table 6 looks at the correlations between subjective measures 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. For the replication data, 25 of 30 correlations are significant at the 5% level or better. In terms of the signs of the correlations, all 30 correlations are positive in the replication data.

Table 6: Correlation Matrices on Subjective Measures of Tasks

(a) Replication, No Deadlines

	ideas	gram.	int.	like	qual.
ideas	1				
gram.	0.373	1			
int.	0.607^{**}	0.471	1		
like	0.686^{**}	0.448	0.870^{***}	1	
qual.	0.707^{**}	0.526^{*}	0.472	0.448	1

(b) Replication, Endogenous

	ideas	gram.	int.	like	qual.
ideas	1				
gram.	0.695^{**}	1			
int.	0.634^{*}	0.731^{**}	1		
like	0.771^{**}	0.818^{***}	0.905^{***}	1	
qual.	0.854^{***}	0.905^{***}	0.670^{**}	0.781^{***}	1

(c) Replication, Exogenous

	ideas	gram.	int.	like	qual.
ideas	1				
gram.	0.930^{***}	1			
int.	0.748^{***}	0.781^{***}	1		
like	0.790^{***}	0.838^{***}	0.977^{***}	1	
qual.	0.904^{***}	0.931^{***}	0.736^{***}	0.768^{***}	1

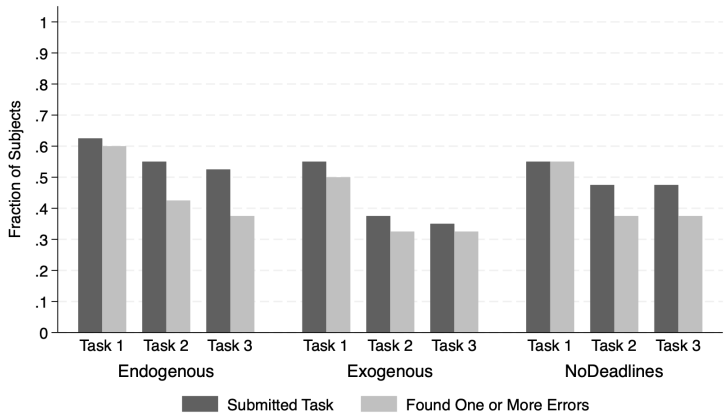
Notes: "gram." indicates "grammar", "int." indicates "interesting" and "qual." indicates "quality". Refer to Appendix A for more details on these variables. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

C Analysis of 2024 Experiment 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, transparency and methodological purposes, we retort the same tables and figures as in the main text and Appendix A.

We refrain from providing direct discussion of each figure and table. We simply note here that the results from this experiment are qualitatively identical to our participant-level randomization experiment reported in the main text and further strengthen our conclusion about the lack of robustness of the original paper: there is substantial attrition, deadlines do not help, and the same (sensible) within-subjects correlations reported in the main text are also present in this replication.

Figure 4: Summary of Task Completions

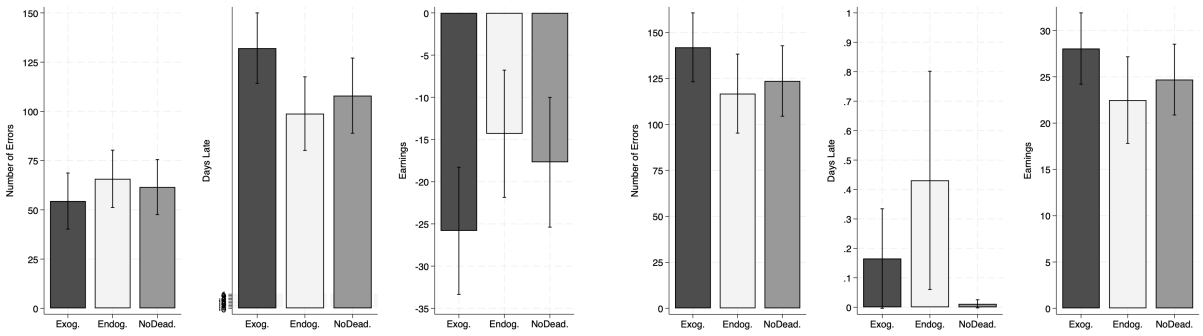


Note: Recall that "Endogenous" corresponds to the treatment where participants self-imposed their own deadlines, "Exogenous" corresponds to the treatment where participants faced evenly-spaced intermediate deadlines and "NoDeadlines" corresponds to the treatment where there were no intermediate deadlines.

Figure 5: Comparison of Treatments

(a) Replication Data: All Participants

(b) Replication Data: Submitted All Three Tasks



Note: Panel (a) is taken directly from the original publication. For panels (b)–(d), "Exog" corresponds to "Evenly spaced deadlines" in panel (a), while "Endog" and "NoDead" correspond to "self-imposed deadlines" and "End deadline", respectively.

Table 7: $F-$Statistics From ANOVA (Performance Metric on Treatment)

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

Table 8: Relationship Between Errors Found and Time on Task

	Replication Data	
Time Spent on Task	0.138^{***}	(0.036)
Constant	37.363^{***}	(6.300)
Treatment Controls	Yes	
Num Participants	54	
R^2 (Overall)	0.233	

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

Table 9: 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

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

Table 10: Subjective Metrics (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

Table 11: Relationship Between Time on Task and Subjective Measures

	Like	Interesting	Quality	Grammar	Ideas
Time	0.056**	0.036	0.037	0.050**	0.044*
	(0.024)	(0.026)	(0.023)	(0.022)	(0.024)
Constant	45.705***	45.573***	54.244***	49.530***	48.507***
	(8.433)	(9.241)	(7.982)	(7.746)	(8.556)
N	54	54	54	54	54
R^2	0.160	0.101	0.076	0.107	0.072

Notes: Standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Treatment controls were included in the regressions, but not reported.

Figure 6: Participants' Self-Reported Time to Complete Each Task

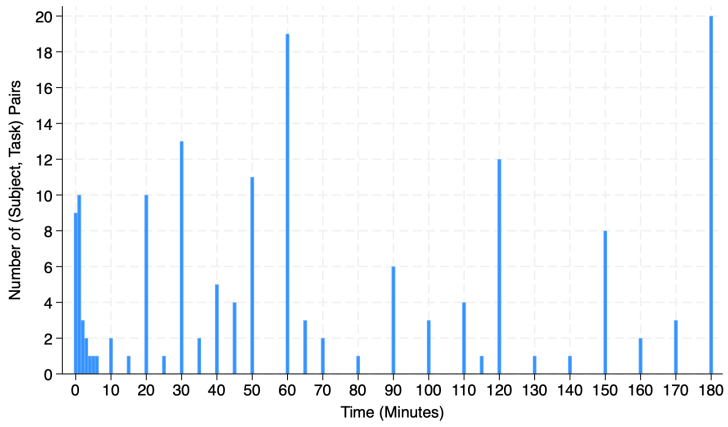


Table 12: Correlation Matrices on Subjective Measures of Tasks

(a) 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

(b) 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

(c) 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. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$