Three Field Experiments on Procrastination and Willpower

Nicholas Burger, Gary Charness, and John Lynham*

September 17, 2008

Abstract: We conducted three field experiments to investigate how people schedule and complete tasks, providing some of the first data concerning procrastination and willpower under financial incentives. In our first study, we paid students $95 if they completed 75 hours of monitored studying over a five-week period. We also required people to meet interim weekly targets in one treatment, but not in the other. In a second study, the task consisted of answering multiple-choice questions on seven consecutive days, with staggered start dates and an endogenous task ordering (tasks varied by number of questions). In our third study, participants answered 20 multiple-choice questions over two consecutive days, varying whether this was during the week or on the weekend. Participants were assigned to either an easy or difficult Stroop test (used by psychologists to deplete willpower) on the first day, before any questions could be answered. We find evidence of procrastination and willpower depletion/replenishment, as well as evidence suggesting a self-reputation interpretation. And yet the behavioral interventions we used led to outcomes that surprised us in all three studies, although these outcomes are largely consistent with the standard neo-classical model.

Keywords: Field experiment, Incentives, Procrastination, Studying, Willpower

JEL Codes: A13, A22, B49, C93, D0

* Contact: Nicholas Burger, Rand Corporation, nburger@rand.org, Gary Charness, Dept. of Economics, University of California at Santa Barbara, charness@econ.ucsb.edu, John Lynham, Dept. of Economics, University of Hawai'i at Manoa, lynham@hawaii.edu

1. INTRODUCTION

People experience self-control problems when their preferences are not consistent across time. One form of self-control problem concerns persistent bad habits or addictions, such as overeating or cigarette smoking. An individual knows that he or she will later regret a current choice of self-indulgence, but nevertheless engages in the activity. The other side of the coin is a situation where an individual is faced with an activity that will lead to future benefits, but is unappealing at the moment. This often leads to procrastination, common in everyday life. People vow to stop smoking, stop eating ice cream, or start exercising tomorrow. Procrastination has been found to be quite pervasive among students: Ellis and Knaus (1977) find that 95% of college students procrastinate, while Solomon and Rothblum (1984) find that 46% nearly always or always procrastinate in writing a term paper.

There have been at least a handful of studies that consider how one might overcome self-control problems. Aside from exerting willpower in the face of a disagreeable task, one approach is to bind one's own behavior with costly restrictions. Wertenbroch (1998) presents anecdotal examples of binding behavior, including tactics such as putting savings into a Christmas-club account that does not pay interest or buying small packages of goods such as cigarettes or ice cream.[1] Schelling (1992) mentions reforming drug addicts who send out self-incriminating letters, to be divulged in the case of a relapse into drug use.

Empirical studies of habits and procrastination are new in economics. Recently, there have been some field interventions, which attempt to study these issues in a controlled environment. Angrist and Lavy (2002) offer substantial cash incentives in Israel for matriculation; while this is ineffective when individual students are selected for the treatment, matriculation rates do increase when this program is school-wide. Charness and Gneezy (2006)

[1] See Ariely and Wertenbroch (2002) for a more complete literature review.

pay students at two American universities to attend a gym during a period of time, finding that attendance rates increase substantially not only during this period, but also after the intervention ends. Angrist, Lang, and Oreopoulos (2007) offer merit scholarships to undergraduates at a Canadian university, with some success in improving performance, but mixed results overall.

Another device is to set deadlines for one's self; for example, many a researcher has agreed to present a yet-unwritten paper in the future, in the hopes that the embarrassment of being forced to cancel or make a change will be a strong motivation for writing the paper prior to the presentation. In fact, many activities seem deadline-driven, particularly in our contemporary society in which people seem to be short on time. Ariely and Wertenbroch (2002) assign three tasks to be completed over a three-week period and find that externally-imposed costly deadlines during this period are more effective than self-imposed (and binding) costly deadlines, which in turn are more effective than having no additional deadlines. Burger and Lynham (2007) examine weight-loss bets in England, where one could bet on achieving a weight goal by a deadline; however, the vast majority of bettors lost their bets with the agency.

In this paper, we report the results of three field experiments designed to provide data on procrastination and willpower. The primary goal of our research is to identify patterns in behavior that will both aid other researchers, and perhaps policy-makers, in designing mechanisms that are effective in overcoming obstacles to performance and inform theorists so that more descriptive models can be developed.[2] We also discuss our results with respect to several models of self-control and willpower.

In our first study, we paid students $95 to complete 75 hours of studying at a monitored location in the campus library over a five-week period. In one treatment, participants were

[2] In this respect, this paper falls into all categories ("Searching for Facts", "Whispering in the Ears of Princes", and "Speaking to Theorists") of the Roth (1995) taxonomy.

required to complete at least 12 hours during the first week, at least 24 hours by the end of the second week, etc., while there were no interim requirements in the second treatment. We expected people to procrastinate with their timing, leaving the bulk of the required studying until the end; in line with the results in Ariely and Wertenbroch (2002), we expected that externally-imposed costly deadlines would be effective, so that the group with the weekly studying requirement would be more likely to complete the task.[3]

However, completion rates were actually 50% *higher* with no interim requirements, as would be predicted by a standard neoclassical model. The patterns of study time show a pronounced weekly cycle, even in the no-weekly-requirement treatment, with little difference in the aggregate from week to week; however, individual analysis reveals substantial heterogeneity, with some people logging the bulk of the hours in the early weeks and some other people doing so in the late weeks. We find evidence that, over time, students who achieve the studying goal improve their performance in the course relative to those students who did not. Finally, women complete the studying task more often than men do.

Having observed different behavior on weekends and weekdays, we designed a second study in which the task consisted of answering different numbers of multiple-choice questions (the order was endogenous) on each of seven consecutive days; seven groups started on different days of the week. We asked people to designate in advance their plans for task completion and then observed their actual behavior. We offered people lottery draws for iPods for submitting their plan, for completing the study, and for answering enough questions correctly.

We find that completion rates vary substantially across the starting-day groups, even though everyone must perform a (chosen) task on each day of the week. Further, people are the

[3] Fischer (2001) also presents arguments for breaking a task into smaller components. On p. 261, she states: "Therefore, the best way for a supervisor to …reduce the risk of missing the ultimate deadline may be to break it into smaller tasks with more deadlines to better compete with the other demands on the student's time."

least likely to drop out on a Monday or Tuesday, regardless of the start date. We do see evidence of procrastination among even the most disciplined people in the population, as the average number of questions answered for people who succeed at the task increases steadily over the course of the seven days. People who stick with their plan are no more likely to complete the task than people who do not. Finally, women are twice as likely to succeed at the task as men.

For our third study, we reduced the duration of the task (answering 20 multiple-choice questions) to two consecutive days and allowed the participants to complete the entire task in one sitting, if desired. We promised definite amounts of money for completing the task and for each correct answer, rather than lottery chances. We also required people to complete either an easy or difficult Stroop test on the first day. The difficult version features cognitively-discordant tasks, and is considered by psychologists to be willpower-depleting.[4]

Most people (63%) answered all 20 questions during the two days, and about 40% of these people answered all the questions on the first day. Indeed, among those who completed the task, those people assigned the difficult Stroop answered significantly fewer questions on the first day. However, a real surprise is that people who were assigned the difficult Stroop were somewhat more likely to finish during the allotted two days than those people assigned the easy Stroop. There is also evidence that people didn't try as hard on the second day, as the percentage of correct answers was significantly lower then. Finally, in contrast to our first two studies, males are significantly more likely to finish this task of more limited duration.

The remainder of this paper is organized as follows: We provide details of our experimental design in section 2, and we present some theoretical models and their predictions in section 3. We describe our experimental results in section 4, offer some discussion in section 5 and conclude in section 6.

[4] We thank Emre Ozdenoren, Stephen Salant and Dan Silverman for suggesting many of these design changes.

2. THE FIELD EXPERIMENTS

Study 1

Our experiment was conducted at the University of California at Santa Barbara. We obtained permission to have anonymous access to the records of students in a large introductory class and then recruited as many as possible from this class. We then advertised the session to first-year students in the general experimental subject pool.[5] All students were told that they could attend an introductory meeting about an experiment that would involve a non-trivial amount of money to be earned over time. Interested students were randomly assigned to an introductory meeting.[6] Participation was voluntary and everyone who showed up was guaranteed $5 if they were not interested in participating. At the meetings, we explained the nature and rules of the experiment. This process lead to a total of 73 eventual participants (out of 87 students who showed up to the meetings); 42 were from the class and 31 were from the campus-wide experimental subject pool. As we shall see later, there was no appreciable difference in behavior across these two sets of participants.

We chose the task of studying because it is a common activity for students, but one that is susceptible to procrastination. Studying has obvious long-term benefits, but is costly in the short-run insofar as other activities have more immediate appeal.[7] Nevertheless, there are already incentives in place for the studier; thus, we did not pay the usual average per-hour rates for experiments, but chose to pay $95 (not as salient as $100) for 75 hours of monitored studying.

[5] We thank ORSEE for the free recruiting software, which allowed selective invitations.
[6] All the students in a particular informational meeting were assigned to the same treatment group. This was done to reduce social interaction threats (Cook and Campbell, 1979).
[7] Students may therefore wish to do more studying than they actually manage; this is similar to self-control problems such as dieting or smoking.

We showed participants the studying location, a room in the library that was frequently (but intermittently) monitored.[8] This study area was available for between 12 and 16 hours each day. Subject to the availability constraint, students were free to log in and out by handing over an ID card to the monitor who would then log the student in or out on a computer. In addition, students were each given a large identifying number, unique to each individual. This was visible to the monitor at all times. The studying area was monitored hourly at a varying time each hour to ensure students were present at the studying location when signed in.

Each student was assigned a web page where he or she could check on the number of hours logged, and could then contact us in the case of any discrepancy. In addition, students who satisfied weekly studying requirements 'banked' their contingent earnings; their web pages had a check-like graphic showing the credit already amassed (of course, this credit was only to be paid if the student completed the overall 75-hour requirement); students who failed to meet a weekly requirement were notified at the end of the applicable week that they were no longer eligible to earn the $95. At the end of the five-week period, those students who had completed the requirement(s) received their earnings and filled out a short questionnaire.

We would like to immediately address two possible concerns. First, students had access to both computers and wireless Internet, so we cannot be certain how much of their time in the library was devoted to studying. However, the anecdotal evidence from the monitors is that, although students were occasionally just sitting and checking e-mail or Facebook, etc., studying was by far the most common activity.[9] Second, one might also be concerned about contamination, since people from both treatments studied in the same area. Again, we have only anecdotal evidence against this: 1) Monitors did not observe students in conversation with one

[8] We thank the UCSB Library staff (and in particular, Eric Forte) for helping to arrange this for us.
[9] And, as we shall see, our results indicate some improvement in grades among those students who completed the studying task, so there is also an inference that significant studying occurred.

another, and 2) Exit interviews of the people who completed the study task indicate that students who spent over 75 hours in each other's company weren't aware of the other treatment group.

Study 2

For our second study, we recruited 181 participants from a micro principles class and two intermediate micro classes at UCSB and described the task in general terms, as well as the prize. The task involved answering multiple-choice questions (drawn from test banks and previous exams in these classes) that were closely related to the material in their class. Each participant was asked to complete seven sub-tasks on seven consecutive days. One sub-task consisted of answering two multiple-choice questions online; the second consisted of answering four multiple-choice questions online, etc.[10]

Each participant received one entry for submitting a non-binding plan for the order in which they would complete the sub-tasks starting a week or so later, two entries for completing the task, and four entries for answering at least 75% of the questions correctly. In order to investigate the day-of-the-week effects we found in Study 1, we assigned each participant randomly to one of seven groups; one of these groups started on a Monday, one group started on a Tuesday, etc.

As in Study 1, we felt that incentives (in this case, to do well in the specific class) were already in place for the participant and so did not pay the usual average per-hour rates for experiments. The prize consisted of one state-of-the-art iPod (which cost $400) for every 50 participants; we awarded four iPods, with the winners determined by drawing entries in a lottery;

[10] Clearly there are advantages and disadvantages to online experiments. For a detailed discussion, see section 2.2 of Charness, Haruvy, and Sonsino (2007).

we gave more weight to correct answers, as otherwise people would be tempted to choose answers randomly.

Study 3

Our third study was also conducted online with multiple-choice economics questions. We recruited participants from a micro principles class and two intermediate micro classes at UCSB; we supplemented the 134 people from these classes who signed up with 23 economics or business economics majors who were not in these classes, but who were in the campus-wide subject pool. In all cases, we described the task in general terms and the payment scheme. In Study 3, the task consisted of answering 20 multiple-choice questions similar to the questions in Study 2 (no one participated in more than one of our three studies) over a two-day period. The participant could answer all 20 of these on the first day or spread them over the two days. We chose to avoid lotteries in our payoff scheme in Study 3. Instead, we paid each participant $7.50 for completing the Stroop exercise and 20 multiple-choice questions, with each correct answer earning the participant an additional $0.75.

Each participant was randomly assigned to either a Tuesday-Wednesday group or a Friday-Saturday group; based on the results of Study 1, these different two-day periods presumably reflect different stocks and/or flows of willpower. We also required people to do 250 rounds of the Stroop exercises on the first day before proceeding to answer the multiple-choice questions. Difficult (or "Discordant") Stroop exercises are used in psychology to deplete willpower; this consists of showing words that are the names of colors, although the actual words are printed in a color of ink different from the color name they represent. One is asked to respond by typing the color seen and ignoring the word itself. It turns out that this is much harder than it

sounds. We randomly assigned each participant to process either difficult Stroop exercises or easy ones (the word color always matches the ink color).

3. THEORETICAL BACKGROUND

A recent literature in economics has explored time-inconsistent behavior. Overall, one of the causes for apparent reversals in preferences over time seems to be the change in the saliency of the costs and benefits of the activity in question (Akerlof 1991). This type of systematic preference reversal is often described by quasi-hyperbolic time discounting, under which immediately available rewards have a disproportionate effect on preferences relative to more delayed rewards, causing a time-inconsistent taste for immediate gratification. Strotz (1956), Ainslie (1992), Laibson (1997) and O'Donoghue and Rabin (1999) discuss present-biased (quasi-hyperbolic) preferences as an explanation for persistent bad habits and addictions.[11] The idea here is that the present is qualitatively different than any future date, so that the present "self" is drawn to immediate gratification. The formulation from O'Donoghue and Rabin (1999) for preferences across a stream of utilities where u_t is a person's instantaneous utility in period t is:

$$\forall t, U^t(u_t, u_{t+1}, \ldots, u_T) \equiv u_t + \beta \sum_{\tau=t+1}^{T} \delta^{\tau} u_{\tau}, \text{ where } 0 < \beta \text{ and } \beta \leq 1.$$

Here δ is the standard, time-consistent, exponential discount rate, whereas $\beta < 1$ indicates a "bias for the present". The quasi-hyperbolic discounting model is consistent with considerable

[11] Frederick, Loewenstein, and O'Donoghue (2002) provide a comprehensive review of empirical research on intertemporal choice, as well as an overview of related theoretical models. We would also like to mention two very new papers. Bisin and Hyndman (2008) investigate stopping-time problems and characterize behavior for exponential, naïve-hyperbolic and sophisticated-hyperbolic discounters. They show that an agent with standard time preferences who suffers from "temptation and self-control" would never be willing to self-impose a deadline. Suvorov van de Ven (2008) develop a theory of self-regulation based on goal setting. They derive a condition under which proximal short-term goals are better than distal long-term goals.

experimental evidence.[12]

Bénabou and Tirole (2004) develop a theory of internal commitments, wherein one's self-reputation leads to self-regulation; the key assumption (p. 859) is that people have only imperfect knowledge of their *strength of will*, learning it only through experience and then having difficulty later remembering it. Their model is embedded in a quasi-hyperbolic framework. There are two periods in the model, each with two sub-periods; a standard discount rate applies between the two periods. In sub-period 1, one decides whether or not to attempt a willpower activity. Indulging gives an immediate benefit, giving up after starting yields a delayed benefit, whereas persevering to complete the task provides a larger delayed benefit. If one decides to attempt the willpower activity, he or she is faced with a (stronger) temptation in the second sub-period and experiences a craving cost.[13]

Decisions in the first period impact those in the second period, although through imperfect recall. Since people have only imperfect knowledge of their own willpower, their choices serve as 'signals' about their types; a negative signal (precedent) undermines self-confidence and willpower.[14] Their basic result (see Figure 2) is that some people will not even try the willpower task, some people will take it on and give up, while other people persevere (this is presumed to be the *ex post* optimal choice); what a particular individual does will depend on his or her parameter values. In general, if one has undertaken the willpower task, he or she

[12] However, see Rubinstein (2003) for experimental evidence not supportive of the quasi-hyperbolic-discounting models.

[13] When making the initial choice in subperiod 1, the delayed benefit from attempting the willpower activity is discounted by $\beta \leq 1$. All future payoffs (i.e., the delayed benefits) are discounted by $\delta < 1$. One of the key differences between β and δ is that β is always known but β is revealed only through the experience of putting one's will to the test.

[14] They state (p. 850): "The fear of creating precedents and losing faith in oneself then creates an incentive that helps counter the bias toward instant gratification."

will exert self-control and persevere if the disutility of resisting temptation is less than the value of self-reputation foregone if a lapse is recalled next period.[15]

Fudenberg and Levine (2006) present a model in which a person is considered to have both a series of myopic short-run selves and a long-run self whose utility is the sum of the utilities of all short-run selves; thus, 'everyone' has identical stage-game preferences but they differ in how they view the future. In the first phase, the long-run self (who is a sophisticated agent aware of his or her control issues) chooses a self-control action that affects any short-run self's utility (e.g., how much cash to bring to the nightclub); in the second phase, the short-run self then chooses the action (e.g., how much to spend). They provide tightly-parameterized specific examples of how this approach can explain behavior in economic contexts such as deciding how much to save and when to take an action that is always currently unattractive, but will provide a flow of utility once taken. In many instances, this model delivers predictions similar to the quasi-hyperbolic model although it also predicts that increased cognitive load makes temptations harder to resist.

Ozdenoren, Salant, and Silverman (2007) provide an explanatory model based on willpower as a depletable (but renewable) resource. Willpower depletion provides an alternative explanation for a taste for commitment, intertemporal preference reversals, and procrastination. This model predicts that an agent with considerable willpower will smooth consumption of leisure over time; however, if the agent's willpower is constrained, he or she may choose to procrastinate (this decision depends on whether willpower is being naturally replenished or depleted).[16] In addition, a willpower-constrained individual will regard seemingly unrelated

[15] For more detail, see their equation (3) on p. 863 and the setup leading to it.
[16] For more detail (when willpower does not have alternative uses), see their Proposition 1 and the setup to the model on the preceding pages.

activities as linked because he or she uses the same cognitive resources to exercise self-control in different activities.

What do the models predict in our experiments? The standard neoclassical model predicts that, in the absence of uncertainty, an agent should be indifferent to weekly requirements as long as the optimal amount of weekly studying is not less than the requirement. If the optimal allocation of studying is less than the requirements, then these requirements are effectively additional hurdles that must be overcome. If the cost of studying is subject to stochastic shocks, then the weekly requirements reduce the flexibility that students have to recover from an early shock (in addition being extra hurdles). In short, we should expect a higher success rate from participants in the treatment without weekly requirements. In terms of allocation of studying hours across weeks, in a deterministic setting, we expect students to study less early on and more later (assuming they have a daily or weekly discount rate). However, with uncertainty about the future (e.g, one's own health later in the study), a prudent individual would log more study hours early on, thereby preserving flexibility. The intra-week pattern could either be increasing or declining depending on uncertainty; however, important differences in the opportunity cost of studying on different days could also be reflected in intra-week studying patterns.

Without additional assumptions, this model makes no clear predictions in Study 2 about the order of the sub-tasks.[17] Again, opportunity-cost considerations could affect the endogenous sub-task ordering. In Study 3, agents might avoid the additional transaction cost of logging in on the second day, thus answering all 20 questions on the first day; however, we could observe smoothing over the two days with sufficiently convex costs. In terms of performance, there

[17] Recall that participants were asked to complete sets of 2, 4, 6, 8, 10, 12, and 14 questions over seven consecutive days, with the order chosen by the participant.

should be no difference across days in the percentage of questions answered correctly. Finally, the neoclassical model predicts that the Stroop test should have no effects.

The quasi-hyperbolic discounting model (with present-biased preferences) and the dual-self model make similar predictions in our environments: all else equal, agents should delay the work until the deadline approaches. In Study 1, with weekly requirements, we would expect study hours to be logged near the end of each week; without weekly requirements, we would expect a surge in hours logged late in the five-week period. However, with uncertainty about the future a prudent person will tend to log study hours earlier and this could even overwhelm the tendency to procrastinate. Thus, the predictions for the studying profile in Study 1depend on assumptions about parameter values. In the absence of uncertainty, both models predict that students with weekly constraints should be more likely to complete the task.

In Study 2, in principle we should observe an increasing profile over time for the number of questions in the sub-task; while this could be mitigated by uncertainty about the future, a really bad day is fatal regardless of the chosen sub-task order. Predictions in Study 3 depend on parameter values, so once again there are no clear predictions. If β in the quasi-hyperbolic-discounting model is linked to willpower or cognitive resources then reducing willpower should induce procrastination. To date, elucidating the factors that determine an individual's β has not been a focus of the quasi-hyperbolic-discounting literature. On the other hand, Fudenberg and Levine (2006, p. 1449) explicitly state: "increased cognitive load makes temptations harder to resist". Thus, their model predicts that students who are assigned the difficult Stroop should procrastinate.

In general, the predictions of the Bénabou and Tirole (2004) model of self-reputation depend on parameter values. However, one important aspect of the model is that (in their two-

period setup) if one has been subject to external controls, these controls (weakly) reduce the likelihood that the agent puts his will to the test in period 2, as he or she doesn't gain the needed self-confidence. In their own words: "The degree of self-control an individual can achieve is shown to … decrease with prior external constraints" (Bénabou and Tirole, 2004, p. 848). A natural interpretation of this result in our context is that a participant with weekly requirements in Study 1 will be less likely to complete the studying task. In Study 2, an increasing profile might make sense, since completing sub-tasks successfully leads to positive signals about one's self. In Study 3, the Bénabou and Tirole model makes no clear prediction, since strength of will is fixed over time. However, if the difficult Stroop test enhances a student's recall of their own strength of will then the model predicts students with the difficult Stroop will be less likely to procrastinate or at least be more successful on the second day (see p. 865).

The predictions of the Ozdenoren, Salant, and Silverman (2007) model depend on the stock of willpower and the relative depletion/replenishment rate. Agents with a sufficient stock of willpower will smooth their hours over time, while other agents would front-load effort provision if willpower is being depleted, or back-load effort if willpower is being replenished. Thus, we could observe differing profiles of logged studying hours over time, depending on the parameters. We could observe a declining studying-hour profile over time, with this tendency augmented by the presence of uncertainty about the future. It is interesting that if willpower is depleted during one part of the week and replenished during another, this model predicts weekly cycles in Study 1, even without opportunity costs that vary over the days of the week. In terms of which treatment group will be more successful in Study 1, the Ozdenoren, Salant, and Silverman (2007) model predicts that students with weekly requirements should do better since the requirements substitute for using an student's own willpower to stay on task (p. 19).

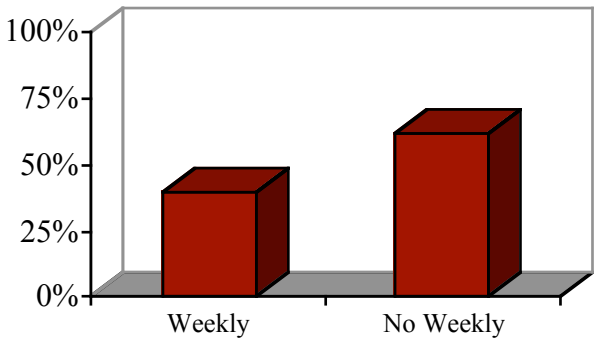
In Study 2, behavior will reflect depletion/replenishment patterns, with longer sub-tasks chosen when the stock of willpower is relatively high. If willpower is depleted during the week and replenished on the weekend then this should influence when students complete tasks. In Study 3, students who have their willpower depleted by the hard Stroop should choose to answer more questions on the second day. The larger the transaction cost of logging in on the second day, the more likely that one answers all 20 questions on the first day. We would expect more success for the weekday group, as their willpower is presumably higher than the weekend group. Finally, if willpower is lower on the second day, the percentage of questions answered correctly might also diminish then.

4. EXPERIMENTAL RESULTS

Study 1

Figure 1 shows the proportion of students who successfully completed the requirements:

Figure 1 - Success rates, by treatment



In the weekly-requirements treatment, 15 of 37 students (40.5%) were successful. When there were no weekly requirements, 22 of 36 students (61.1%) completed the mandated 75 hours. Clearly, there is no support for the view that externally-imposed restrictions helped students to

achieve the goal. In fact, the test of the difference of proportions (see Glasnapp and Poggio 1985) gives $Z = 1.76$, so that there is a marginally-significant difference in success rates across treatments ($p = 0.078$, two-tailed test).[18]

Recall that some participants were from an introductory class and others were first-year students from the general subject pool. Thirteen of 20 students from the class succeeded in the no-weekly treatment, compared to 9 of 16 students from the subject pool ($Z = 0.54$), while eight of 22 students from the class succeeded in the weekly-requirements treatment, compared to 7 of 15 students from the subject pool ($Z = 0.63$). Neither difference (nor the overall comparison of 21 of 42 versus 16 of 31) is close to statistical significance.

We also find that the completion rate for the 45 female participants was 50% higher than for the 28 male participants.[19] However, perhaps since the number of observations is relatively small, this difference is not quite significant; the test of the difference of proportions gives $Z = 1.54$, for $p = 0.124$ with a two-tailed test. While we had no *ex ante* hypothesis regarding gender, this result is similar in flavor to the finding in Angrist, Lang, and Oreopoulos (2007) that providing merit scholarships is much more effective for female students.[20]

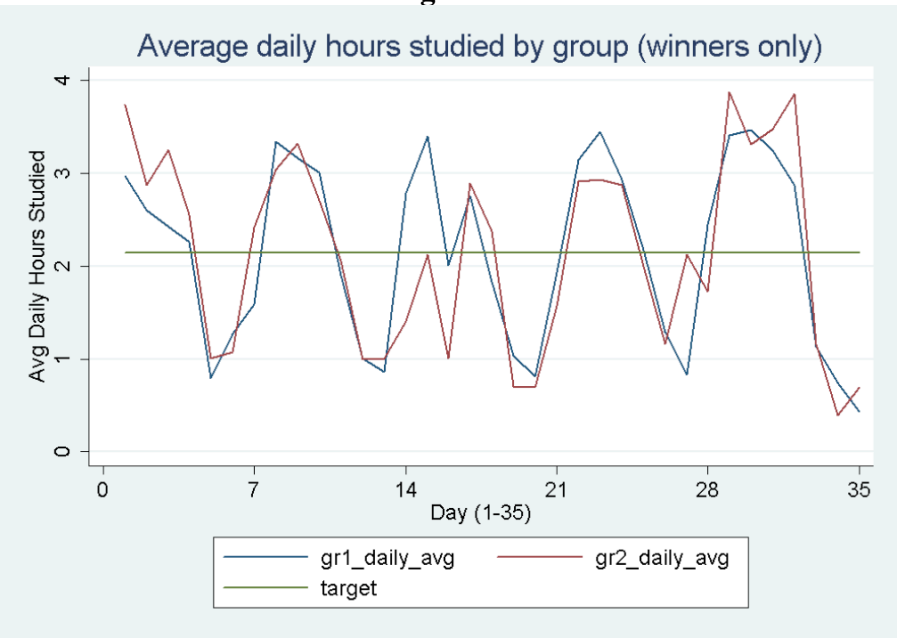
It may be interesting to speculate on what the completion rate for the students with no weekly requirements would have been had they faced the requirements. While it is impossible to construct the appropriate counterfactual, by simply imposing *ex post* the weekly requirements on the no-requirements treatment, we find that the completion rates are quite similar. The completion rate for the no-requirements treatment would now be 14 of 36 (38.9%), compared to

[18] After the experiment had started, we received an email from one student indicating that he did not plan on participating. We have omitted this student from our analysis, but if we include him (on the premise that he is no different than a student who didn't study), this result is slightly stronger ($Z = 1.86$ and $p = 0.062$, two-tailed test).
[19] Twenty-six of the 45 female participants (57.8%) completed the requirement(s), compared to 11 of the 28 male participants (39.3%).
[20] And, as we have since learned from discussions with students, there is a near-consensus view that female students are better at managing their study time.

15 of 37 (40.5%) for the weekly-requirements treatment. This gives some support for the hypothesis that flexibility is preferred to constraints because it allows students to recover from expected or unexpected shocks to the cost of studying.

Perhaps the most interesting aspect of our data is the cyclical (weekly) patterns in the number of study hours logged.[21] Figure 2 shows these patterns for those students who successfully completed the studying project:

Figure 2



"gr1" ("gr2") refers to the group without (with) weekly requirements

While it may not be surprising to see a weekly, cyclical pattern when students face weekly study requirements, we had not anticipated this in the unstructured regime; if anything, the pattern is stronger for the group without weekly requirements. Cumulatively, the study hours logged for winners were very close to a target line of 15 hours per week. The average number of

[21] Individual weekly study rates are shown in Appendix A.

study hours (75.35) for the winners was close to the minimum of 75, ranging from 75.02 to 76.74. This suggests that students did not find this studying task to be innately pleasurable.

We turn to whether there is a difference in study hours across weeks for those who completed the 75 hours of study. Table 1 presents the average number of hours for the winners by week and by group:

Table 1: Average weekly study hours (winners), by group

Week	Weekly requirements	No weekly requirements
1	16.92	13.91
2	14.53	16.08
3	11.38	13.78
4	15.75	16.27
5	16.78	15.31

There is no clear trend over time for either treatment. Regressions of hours against weeks yield insignificant coefficients for the time trends (0.09 and 0.30, respectively, with corresponding *t*-statistics of 0.11 and 0.76). This does not appear to be evidence of procrastination. However, if we look at the study patterns for each individual, there is evidence that some people front-load studying hours while others delay the bulk of their studying hours.[22] Appendix B1 shows the study hours for each person by week, while Appendix B2 breaks this out for people who completed the 75 hours. We classify people who finished as front-loaders (back-loaders) if they logged at least half of their hours in the first (last) two weeks of the five-week period. Consistent with precautionary 'savings', four of the 37 people who finished were front-loaders, while consistent with procrastination, 11 people were back-loaders.

We present OLS regressions showing study patterns for both treatments in Table 2:

[22] In addition, one might classify those people who didn't finish (or didn't even start) as more serious procrastinators (we thank Jeroen van de Ven for this insight). Indeed, more extreme procrastinators meant to sign up, but didn't get around to it (or didn't even get around to submitting their college applications on time).

Table 2 – Regressions for weekly studying hours (winners only)

Variables	Group (1) Weekly requirements	(2) No weekly requirements	(3) Pooled data
Week 2	-2.39 [1.57]	2.17 [1.41]	0.32 [1.09]
Week 3	-5.54*** [1.62]	-0.13 [1.36]	-2.32* [1.11]
Week 4	-1.18 [2.24]	2.36 [2.82]	0.92 [1.89]
Week 5	-0.14 [2.82]	1.40 [2.93]	0.78 [2.04]
Constant	16.92*** [1.15]	13.91*** [1.36]	15.13*** [0.95]
# Observations	75	110	185
R^2	0.12	0.02	0.03

Week 1 is the omitted variable in these regressions. Robust standard errors clustered by subject are in brackets. * and *** indicate significance at the 10% and 1% level, respectively (two-tailed tests). The pooled regression includes a control for treatment group (not reported).

The regressions confirm that there is no clear increasing or decreasing profile over the course of the experiment. Only the coefficient for the dummy for Week 3 has any statistical significance, and this may reflect the effect of a closure of the library during evening peak study time due to a power outage (see the dip around day 16 in Figure 2).

Table 3 shows the average hours of studying logged by winners on each day of the week:

Table 3: Average study hours (winners), by day and group

Day	Weekly requirements	No weekly requirements	Pooled data
Monday	3.14	3.25	3.20
Tuesday	2.69	2.94	2.84
Wednesday	3.04	2.87	2.94
Thursday	2.57	2.21	2.36
Friday	1.01	1.05	1.03
Saturday	1.06	0.90	0.96
Sunday	1.56	1.84	1.73

The number of study hours shows a tendency to decrease over the course of the week (from Monday to Thursday), with a dramatic drop on Friday and Saturday, and some recovery on Sunday.[23] The patterns are essentially similar across the two treatments. Students appear to start the week fairly fresh and run out of steam as it progresses. The weekend appears to be the time when students 're-charge'.[24] Table 4 offers a regression analysis:

Table 4 – Regressions for intra-week studying hours (winners only)

Variables	(1) Weekly requirements	(2) No weekly requirements	(3) Pooled data
Tuesday	-0.45 [0.30]	-0.31 [0.34]	-0.37 [0.23]
Wednesday	-0.09 [0.37]	-0.38 [0.30]	-0.26 [0.23]
Thursday	-0.57 [0.53]	-1.05* [0.41]	-0.85*** [0.32]
Friday	-2.13*** [0.22]	-2.20*** [0.28]	-2.17*** [0.19]
Saturday	-2.08*** [0.37]	-2.35*** [0.29]	-2.24*** [0.23]
Sunday	-1.57*** [0.37]	-1.42*** [0.34]	-1.48*** [0.25]
# Observations	525	770	1295
R^2	0.152	0.152	0.149

Monday is the omitted variable in these regressions. Robust standard errors clustered by subject are in brackets. * and *** indicate significance at the 10% and 1% level, respectively (two-tailed tests). The pooled regression includes a control for the treatment group (not reported).

The regressions confirm the decrease in study hours logged over the course of the week, dropping dramatically on Friday. There is virtually no difference between the estimated

[23] Most of the studying on Sunday took place in the evening.

[24] In fact, our data also shed light on a central mystery of UCSB – how is it possible to simultaneously be both a party school and one where the students are fairly capable and responsible? The answer may not be between-student heterogeneity as much as it is 'within-subject heterogeneity', as it appears that the UCSB students in our field experiment tended to compartmentalize their time into party time on the (extended) weekends and school time otherwise.

coefficients for the two groups (when a dummy variable for group is added, its coefficient is 0.00; when we include week*group interaction dummies, none of these has a coefficient that is close to statistical significance – the lowest *p*-value is 0.47). Thus, we see a strong and significant cyclical pattern, common to both treatments. To the extent that we observe front-loading in each week, this appears more consistent with a willpower explanation than one of quasi-hyperbolic discounting.

However, the winners only represent 51% of the participants (37 of 73); perhaps there is more procrastination among those students who did not manage to complete the task. Naturally, the data for this group is far less complete, as most people who did not complete the studying task stopped logging hours early on.[25] Contrary to what we might expect from quasi-hyperbolic preferences, we do not see many people desperately trying to catch up in the last days or week of the unstructured treatment (although three people in this group did struggle on to the final week without success).[26]

A final question of importance is whether completing (or even attempting) the study task was helpful in terms of performance. As mentioned earlier, 42 of our participants originated in an introductory class and we were given permission to access the (anonymous) grade records for the course, matching student ID numbers for the participants. There were quizzes, a midterm, and a final in the course. One metric would simply be to compare the final-exam grades for winners, non-winners, and non-participants; however, we face an obvious potential selection bias, and it seems best to assess the *change* in relative performance over the course of the

[25] In the weekly-requirements group, of the 23 people who did not complete the study task, 14 were eliminated at the end of the first week, four were eliminated at the end of the second week, one was dropped at the end of the third week, and one was dropped at the end of the fourth week; for the group without weekly requirements, of the 14 people who did not complete the study task, six never studied after the first week, two never studied after the second week, two more dropped out by the end of the third week, and three never studied after the fourth week.

[26] Figure B in Appendix B shows a pattern similar to, but more muted than, Figure 2 for the first two weeks of the study (70% of the people who did not complete the task logged no hours after the second week) for people who did not attain the studying target.

quarter. Our study commenced in the fourth week of the quarter, with two quizzes preceding our study. There is more variability in the quiz grades, with a number of people missing them, particularly after the midterm.[27] We therefore trust the midterm (taken in the second week of the study) and final-exam scores more, but nevertheless include an average for the first two quizzes. Table 5 shows the mean scores by group:

Table 5: Mean scores on tests, by group

Group	N	Quiz	Midterm	Final
Non-participants	403	3.34	8.78	7.56
Participants, non-winners	21	3.21	9.57	7.71
Participants, overall	42	3.43	10.14	8.48
Participants, winners	21	3.67	10.71	9.24

We see that the differences between the non-participants and the non-winners are generally small, although slightly larger for the midterm.[28] In fact, Wilcoxon-Mann-Whitney (two-tailed) ranksum tests confirm that none of these differences are significant (for the midterm comparison, we find that $Z = 1.37$, $p = 0.171$). On the final exam, there was no difference between non-winners and non-participants ($Z = 0.10$); there is also no significant difference between winners and non-winners on the midterm scores ($Z = 1.58$, $p = 0.115$); however, the difference between final scores is in fact significant ($Z = 2.38$, $p = 0.017$). Thus, the data suggest that the difference in test scores increased over the course of the quarter.[29]

[27] The absentee rate on the quizzes after the midterm was more than twice as high as the absentee rate on the quizzes before the midterm.
[28] We note that the midterm took place before many non-winners had stopped logging study hours.
[29] One issue is whether the study hours in the monitored location were simply a substitute for study hours elsewhere. The data suggest that perhaps this is not completely the case. In addition, the results from our pre- and post-experiment questionnaires reveal that only 24% of the eventual winners studied more than 15 hours per week before the experiment started and 64% of winners reported reducing their weekly study hours once the experiment ended. This provides some evidence that the experiment was an exogenous shifter of total hours studied over the five-week period.

Study 2

The task in Study 2 was rather unforgiving, in that participants were required to perform tasks on each of seven consecutive days, with no tolerance for a missed day. As a result, only 15% of the participants (28 of 181 people) completed the task successfully. Sixteen of 69 females (23%) succeeded at the task, compared to 12 of 112 males (11%); the difference in rates is statistically significant ($Z = 2.25$, $p = 0.024$, two-tailed test). It seems that there may have been some confusion (a serious hazard in online experiments), since some people performed all seven sub-tasks, but not on consecutive days.[30] Overall, 60 people who signed up completed no sub-tasks, 69 people completed between one and six subtasks, and 52 people completed all sub-tasks (but only 28 people did so on time). Since so few of the people who registered for the experiment completed the task on time, we report data from various categories.

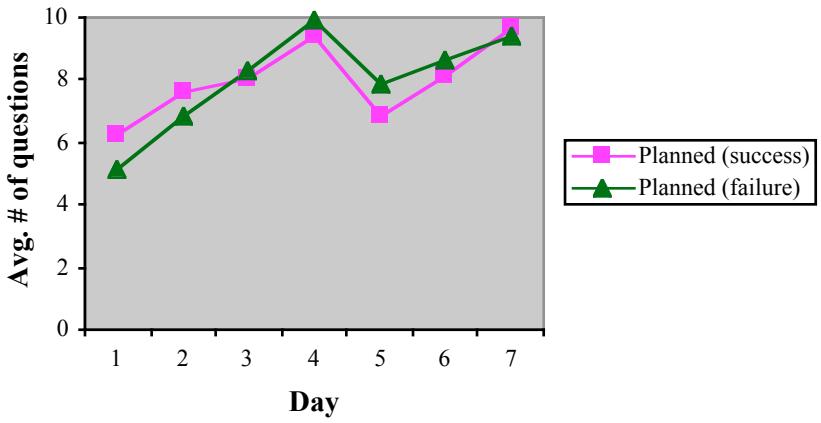
As we asked for the planned task order a week before the sub-tasks could be performed, one natural issue is whether people were more likely to finish on time if they at least started off with their self-imposed plan.[31] We consider the 96 people who completed at least one sub-task on the first day. There were 46 people who stuck to their first-day plan, and 14 of these people finished on time. Of the 50 people who did not stick to their first-day plan, 14 finished on time. There is no significant difference in these proportions ($Z = 0.26$). Thus, having a self-imposed plan doesn't seem to benefit people, much as having an exogenously-imposed plan was unhelpful in Study 1.

[30] It is of course also possible that students viewed these questions as test preparation for the final exam, since the timing of the study was in late November and early December, but this explanation seems unlikely. A hazard with online experiments is that there is no chance for human feedback if a participant has a question.

[31] Since everyone who stuck to his or her plan finished on time, the best we can do is to analyze whether people stuck to their plans for a smaller number of days, with one day being the cleanest test.

We might also be interested in whether there are differences in the plans of the people who succeed and those who do not. Figure 4 shows the average number of questions planned for each of the seven consecutive days, for people who finished on time and for people who did not:

Figure 4: Avg. number of questions, by day



Visually, there is little difference between the plans of people who completed the study and those who did not. A Chi-squared test confirms that there is virtually no difference at all between the plans of the groups ($\chi^2_6 = 0.126$, $p = 0.999$). How do plans compare to actual completion dates? Table 6 shows how choices of sub-tasks compared to the previous plans of those people who finished on time, on each of the seven days of the task:

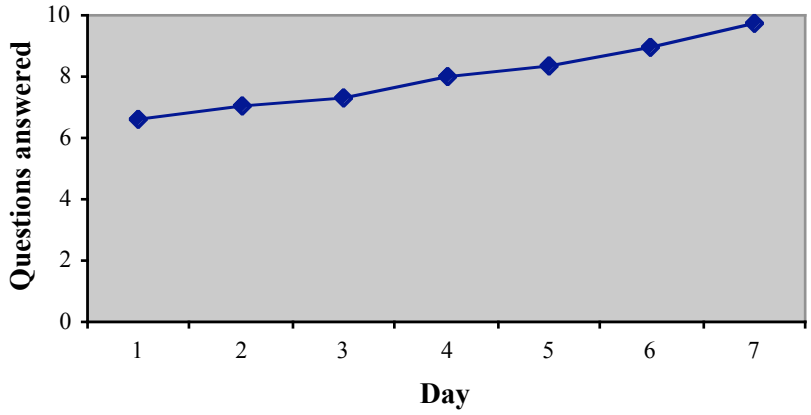
Table 6: Questions answered versus plan, by day

Day	Fewer	Same	More
1	6	14	8
2	11	10	7
3	12	8	8
4	12	8	8
5	5	11	12
6	6	11	11
7	7	11	10

"Easier" ("Same") ["Harder"] means that the sub-task performed had fewer (the same number of) [more] questions than indicated by one's plan.

Overall, during the earlier part of the seven-day period (the first three days), an easier (harder) sub-task was chosen 29 (23) times. By comparison, the easier (harder) task was chosen 18 (33) times during the later part of the seven-day period (the last three days). Thus, there appears to be a tendency towards procrastination even for this disciplined sub-population. Figure 5 bears this out:

Figure 5: Avg. number of questions answered, by task day



We see a steady increase in the number of questions answered per day over the course of the seven-day period. A simple OLS regression of the number of questions answered against the task day gives:

$$\underset{}{\text{Questions answered}} = \underset{(0.130)}{5.949} + \underset{(0.029)}{0.513}*\text{Day},$$

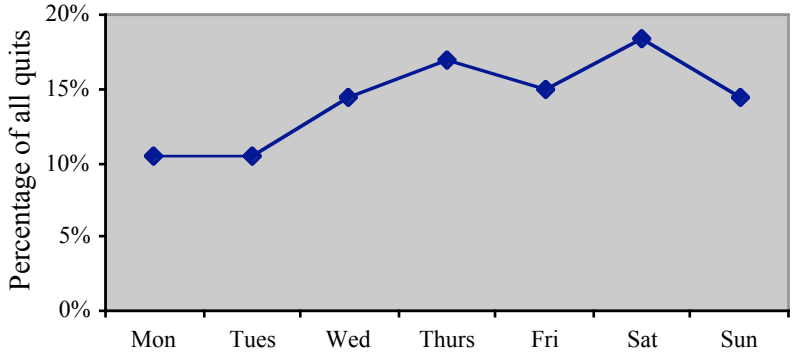
so the number of questions answered increase significantly over the course of the period.[32]

One of the purposes behind our design in Study 2 was to examine whether day-of-the-week effects persist to some degree when the task is online, so that no physical journey is

[32] Further evidence of procrastination is seen in the number of people who completed each sub-task. Out of the 181 participants, 107 completed sub-task 1 (two questions), 89 completed sub-task 2 (four questions) 73 completed sub-task 3, 71 completed sub-task 4, 69 completed sub-task 5, 65 completed sub-task 6, and 76 completed sub-task 7. The likelihood of a task being completed decreases monotonically as the number of questions in the sub-task increases, with the exception of the most difficult task (perhaps some people try to get this out of the way).

required. If we consider the stock of 'studying willpower' over the course of a week, Study 1 suggests that students replenish their supply over the weekend and deplete it once the week is over. To test whether this pattern holds, we examine the day of the week on which people 'quit' (first failed to complete a sub-task); this is illustrated in Figure 6:

Figure 6: Percentage of all quits, by day of the week



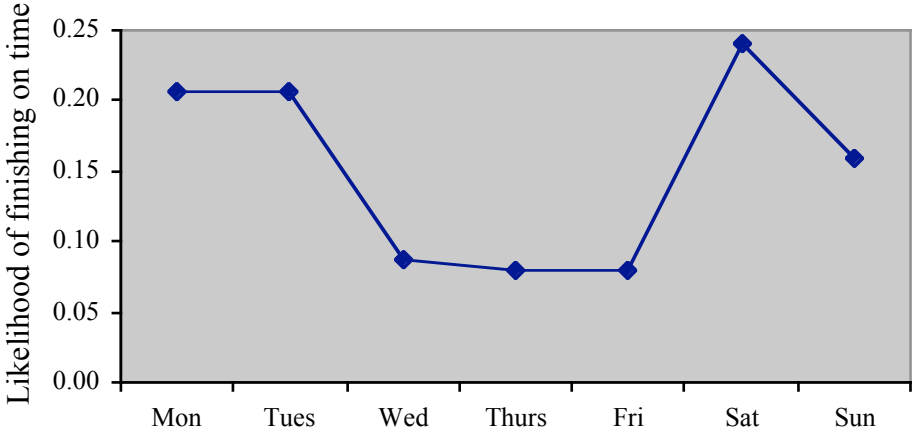
This pattern largely supports the willpower depletion/replenishment story suggested by the weekly cycles in Study 1. Monday and Tuesday have easily the lowest quit rates (10% for each) and, the rate steadily increases through Saturday, with the exception of Friday. A linear regression of the quit rate against the day of the week from Monday (= 0) to Saturday gives:

$$\underset{}{\text{Quit rate}} = \underset{(0.015)}{0.110} + \underset{(0.005)}{0.103}*\text{Day},$$

confirming that there is a significant upward trend over this period.

The complementary issue is that of completion rates. Since everyone was required to answer a chosen number of questions on each day of the week, we might expect that there is no difference in the likelihood of completion according to the starting day of the week. However, Figure 7 suggests otherwise:

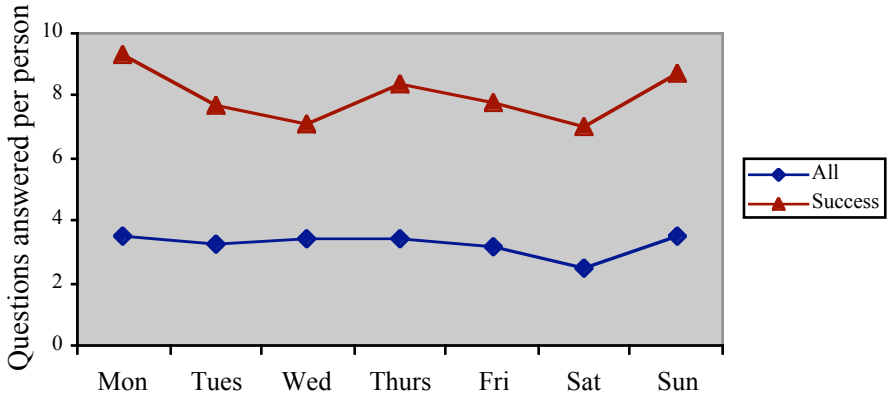
Figure 7: Completion rates, by starting day of the week



We can readily reject the conjecture that the likelihood of finishing on time is independent of the starting day, as it seems that one has the lowest chance of success if the task begins from Wednesday through Friday. The overall success rate is 8% for these starting days and the overall success rate is 20% for Saturday through Tuesday; these rates differ significantly ($Z = 2.22$, $p = 0.026$, two-tailed test). Possible explanations for this pattern include discouragement due to anticipating the weekend ahead and not having developed a regular habit before the weekend arrives.

Finally, if the opportunity cost of doing schoolwork varies substantially over the course of the week, as suggested by the weekly cycles in Study 1, we should expect substantial differences in the number of questions answered on the weekdays or on the weekend. However, there is no dramatic difference in questions answered by day of the week. This is shown for both the entire sample (labeled "All") and for people who finished the task on time (labeled "Success"), as shown in Figure 8:

Figure 8: Average number of questions answered per person, by day of the week



The number of questions answered is slightly higher on Sunday and Monday, while it is lowest on Saturday (and also on Wednesday for people who finished on time). We also see a Saturday effect when we compare planned tasks to actual tasks. While there is no statistical difference between planned tasks and actual tasks comparing across experiment day (i.e. first day of the experiment, second day, etc.), there is one statistical difference between planned and actual tasks comparing across weekdays and this occurs on Saturday. Perhaps not surprisingly, students planned to allocate more effort on Saturday than they actually allocated.

Study 3

Of the 158 people who signed up on line, 100 (63%) completed the task successfully. In a departure from our earlier results, 60 of 85 males (71%) completed this less difficult task, compared to 40 of 73 females (55%); this difference is statistically significant ($Z = 2.05$, $p = 0.040$, two-tailed test). As might be expected, success rates were significantly higher for the Tuesday-Wednesday group, as 57 of 79 people (72%) in this group finished and 43 of the 79 people (54%) in the Friday-Saturday group finished ($Z = 2.31$, $p = 0.021$, two-tailed test).

Thirty-nine people answered all 20 questions on the first day, 15 people answered all 20 questions on the second day, and the other 46 people who completed the task answered some questions on both days. Ninety-two people answered some questions on the first day, while 64 people answered questions on the second day.[33] Forty-nine people never answered any questions, while the remaining nine people answered between four and 15 questions. Sixty-one of the 64 people (95%) who answered some questions on the second day completed the task successfully.

In the sample as a whole, people answered an average of 7.62 (5.55) questions on the first (second) day. With respect to the people who completed the task, the average number of questions answered on the first (second) day was 11.53 (8.47); similarly, for the nine people who answered more than zero, but less than 20 questions in total, the average number of questions answered on the first (second) day was 5.56 (3.33).[34] We see some tendency for males to procrastinate more than females, as the number of questions answered on the first day for people who finished was 9.97 for males and 13.88 for females ($Z = -2.25$, $p = 0.025$, two-tailed test).

There were some differences in behavior according to whether one was assigned the easy or hard Stroop test. Twenty-two of the 45 people (49%) who were assigned the easy Stroop and succeeded at the task answered all 20 questions on the first day, compared to 17 of the 55 people (31%) who succeeded at the task and who were assigned the hard Stroop ($Z = 1.83$, $p = 0.034$, one-tailed test). In addition, among people who completed the task, people who were assigned the easy Stroop answered slightly more questions on the first day (12.73 vs. 10.55, $Z = 1.47$, $p =$

[33] While the design of our second and third studies reflected our own learning process, and our intention was never to specifically compare results in one study to another (since many design features were changed simultaneously), we nevertheless note that the proportions of people who never answered any questions are very similar in Study 2 (60 of 181, 33%) and Study 3 (49 of 158, 31%).

[34] Since more questions are answered on the first day, at first glance there is no evidence of procrastination in the aggregate. However, one must keep in mind that each participant was required to answer the Stroop questions on the first day, so that coming back on a second day involved an additional transaction cost.

0.071, one-tailed test); however, this effect is not present in the full sample (7.49 with the easy Stroop vs. 7.73 with the difficult Stroop, $Z = -0.33$).

A rather surprising result is that people who were assigned the hard Stroop were actually slightly more likely to eventually finish the task. Fifty-five of the 79 people (70%) assigned the hard Stroop finished, while 45 of the 79 people (57%) assigned the easy Stroop finished; the test of proportions gives $Z = 1.65$, $p = 0.099$, two-tailed test. The main driver of this difference appears to be the effect on behavior on the second day, as people who are assigned the difficult Stroop answered significantly more questions on the second day than the people assigned the easy Stroop (6.90 vs. 4.20, $Z = 2.45$, $p = 0.014$, two-tailed test).

These results are combined in the regressions in Table 7.[35]

Table 7 – Determinants of success rate and number of questions answered

	Independent variable			
Dependent Variables	(1) Success rate	(2) Questions, 1st Day	(3) Questions, 1st Day	(4) Questions, 2nd Day
Tuesday	.539** (.212)	9.57** (3.97)	1.59 (3.34)	2.42 (3.62)
Hard Stroop	.406* (.212)	.499 (3.85)	-6.92** (3.42)	10.54*** (3.78)
Male	.515** (.213)	.299 (3.87)	-8.04** (3.45)	10.57*** (3.77)
Constant	-.384* (.222)	-1.30 (4.31)	22.11*** (4.17)	-16.29*** (4.67)
Observations	158	158	100	158
Pseudo R^2	0.068	0.010	0.018	0.026

Specification (1) is a Probit regression (Z-statistics) and (2) – (4) are two-sided Tobit regressions (t-statistics). Specification (3) includes only those people who answered all 20 questions. *, **, and *** indicate significance at the 10%, 5%, and 1% level, respectively (two-tailed tests).

[35] We omit interaction terms for the sake of brevity, as these are not significant and including them does not qualitatively change the main results.

Another measure that may reflect willpower depletion, but in any case certainly reflects the quality of work, is the percentage of questions answered correctly. There is no significant difference in the percentage correct across gender (62% for males vs. 60% for females) or type of Stroop test (60% in both cases). In addition, the percentage of correct answers on the first day does not predict success rates (the *t*-statistic on the coefficient for number of questions is 0.21). However we do find that the percentage of correct answers on the first day is significantly higher than on the second day.[36] This is robust to whether we consider the whole sample of 158 people, the 100 people who finished, the 47 people who answered questions on both days, or the 17 people who answered questions on exactly one day.[37] The percentage of answers is slightly higher for the Tuesday-Wednesday group (0.63 vs. 0.56, $Z = 1.78$, $p = 0.075$, two-tailed test).

5. DISCUSSION

In this section, we first focus on the main results and we then comment on how we think they reflect on the standard neo-classical model and the models discussed in section 3. As shall be seen, there is some support for each model, but none of them fit all of the data.

First, one feature common to each study is that behavioral mechanisms that seem attractive often lead to surprising outcomes. Consider the effect of the weekly requirements in Study 1. The Ariely and Wertenbroch (2002) results and our own intuition suggested that this structure would help students to achieve the studying-hours goal by preventing them from falling too far behind. However, the proportion of successful participants was 50% higher when we left the five-week study period unstructured, rather than imposing weekly requirements. The difference appears to be due to the lack of flexibility imposed by weekly constraints. In Study 2,

[36] The order of questions was randomized, so there should be no difference in difficulty levels across days.
[37] The respective comparisons are 0.66 vs. 0.49 ($Z = 4.35$, $p = 0.000$), 0.66 vs. 0.50 ($Z = 4.10$, $p = 0.001$), 0.67 vs. 0.47 ($Z = 3.43$, $p = 0.001$), and 0.65 vs. 0.54 ($Z = 1.93$, $p = 0.053$). All of these tests are two-tailed.

people had to submit a planned sub-task order and we thought that this might be helpful. Yet there is no difference in success rates for people who follow their plans on the first day and for people who ignore their plans. In Study 3, the difficult Stroop test was expected to deplete willpower, and yet the success rates for people who were assigned the hard Stroop were higher (with marginal statistical significance) than the success rates for people who were assigned the easy Stroop test.

We find evidence of procrastination, although it is not as ubiquitous as some might have expected. While there was no aggregate time trend over the five weeks of Study 1, nevertheless 30% of the participants who completed the task deferred logging most of their studying hours to the last two weeks of the five-week period, despite the uncertainty issue. In Study 2, even the people who are able to complete all of the sub-tasks show a clear tendency to delay completing the more difficult sub-tasks until later. In Study 3, 61 of the 100 people (61%) who complete 20 questions don't do so on the first day, even though there is an additional transaction cost in returning for the second day.

We observe strong day-of-the-week effects. In Study 1, there are pronounced weekly cycles, with the most studying done on Monday through Wednesday and the least on Friday and Saturday. In Study 2, the lowest quit rates were on Monday and Tuesday, while completion rates are lowest for people whose starting date was a day or two before the onset of the weekend. In Study 3, people in the weekday group are significantly more likely to complete the task than the weekend group. However, the fact that in Study 2 there was no substantial difference across days of the week in the number of questions people chose to answer would seem to be evidence against the notion that the opportunity cost of working is higher on the weekend (at least when one can do so at home).

There is strong evidence that suggests weekly cycles of willpower depletion and replenishment. In Study 1, logged study hours are highest on Monday-Wednesday, dropping later in the week and during the weekend before returning to the Monday level; in spirit this cycle seems closest to a situation where willpower is depleted over the course of the week and replenished on the weekends; in other words, willpower is renewable. In Study 3, willpower seems considerably higher for the Tuesday–Wednesday group than for the Friday–Saturday group, as the completion rate is significantly higher. While the number of questions answered on the first day is simlar with hard and easy Stroop tests for the group as a whole, people who were assigned the hard Stroop answered significantly more questions on the second day. This is consistent with the notion that succeeding at the hard Stroop serves either as a positive signal to one's self (as in the Bénabou and Tirole model) or increases the will to complete the task.

We find gender effects in each of our studies; however, at least at first glance these are inconsistent. Males are significantly less likely than females to complete the most organizationally-challenging task (Study 2), marginally less likely to complete a lengthy task with some flexibility on a daily basis (Study 1), but are significantly more likely to complete a task of relatively short duration (Study 3). While all the models are essentially agnostic on this point, our interpretation is that female students are generally more successful than males at organizing and completing tasks that require consistent attention over a period of time, males have the willpower and energy to actually do better than females on tasks for which the finish line is never very far from sight, and the task can even be done in one sitting.[38]

We now turn to how our data reflect on the standard model, as well as on the quasi-hyperbolic, self-reputation, dual-self and willpower models described in section 3. In most

[38] We are clearly only speculating on this point. Nevertheless, this would be interesting to study in more detail in future work.

cases, the predictions depend on parameters whose values are unknown to us and would be problematic to estimate. Thus, we make only broad assessments of each model's predictions with respect to the issues discussed in the preceding paragraphs, summarizing these in Table 8:

Table 8 – Consistency of results with standard and behavioral models

	Experimental result		
Model	Restrictions hurt success rates	Tendency to delay work	Day-of-the-week effects/weekly cycles
Standard	(3) Natural	(2) Plausible	(2) Plausible
Quasi-hyperbolic	(2) Plausible	(4) Predicted	(2) Plausible
Self-reputation	(4) Predicted	(3) Natural	(2) Plausible
Dual-self	(2) Plausible	(4) Predicted	(2) Plausible
Willpower	(2) Plausible	(2) Plausible	(3) Natural

We have four broad categories, distinguished as follows: If it is difficult or impossible to reconcile the result with a model, we call this "(1) Inconsistent"; if the data can be rationalized by some choices of parameters, but one must fight the current to do so, we call this "(2) Plausible"; if the result seems to resonate well with a model, we call this "(3) Natural"; if the result is clearly predicted by the model, we call this "(4) Predicted."

While our classification scheme is arbitrary, we nevertheless can, to some extent, justify our choices and we suspect that we are at least close in all cases. First, regarding the result that interim weekly requirements decreased the success rate, this seems a fairly clear prediction of the Benabou and Tirole (2004) model. The restrictions may or may not bind in the standard model, so that we should get either no effect or the observed decrease; it seems natural that adding restrictions lowers willpower. The predictions of the quasi-hyperbolic and dual-self models depend on the severity of the procrastination problem and the degree of uncertainty about the future, and it's not clear that a five-week time horizon contains that much uncertainty.

Second, the tendency to delay work is a major feature of both the quasi-hyperbolic and dual-self models. To some extent, an increasing work profile is natural with a self-reputation

model, as incremental successes serve as positive signals. The willpower model primarily predicts a declining work profile, but this really depends on the willpower renewal/depletion rate. The standard model predicts (with a daily discount rate) a tendency to delay work, but this may be reversed in the presence of uncertainty.

Finally, none of the models explicitly predict the day-of-the-week effects or the strong weekly cycles, although all of them can rationalize these with the argument that opportunity costs for working are higher on the weekend. Of course, the relatively constant number of questions answered per day in Study 2 makes this argument a bit less tenable. The willpower model also considers the notion of depletion and replenishment periods, so a cycle according to the portion of the week seems fairly natural to the model.

Overall, each behavioral model resonates well with at least some of our results, but has problems with some other result. At the same time, the standard model does not do badly, especially when one also considers its relative success at predicting the effects of the behavioral interventions, whereas the behavioral models have more mixed success in this regard.

6. CONCLUSION

Our study is one of the first to provide empirical evidence concerning procrastination, willpower, and monetary incentives. Charness and Gneezy (2006) find that it may be possible to 'crowd in' a taste for exercise by paying people to go to the gym; on the other hand, DellaVigna and Malmendier (2006) find the perverse effect that people who choose to pay a flat monthly fee for membership in a gym pay more than if they would have chosen to pay a fixed cost per visit.[39] In a certain sense, this result from the latter study is similar to the result in Ariely and

[39] A possible explanation of the finding is that, as a self-control mechanism, people choose the more expensive plan because it reduces the marginal cost of attending to zero, and they (mistakenly) believe that this will encourage them to attend the gym.

Wertenbroch (2002), who find that people do not do as well with managing their coursework without an imposed schedule; in both cases, people do not anticipate their own limitations when devising a plan of attack. Yet, we do not find that imposing a structure is beneficial in our studies. In fact, none of the outcomes of the behavioral interventions in our study matched our predictions, although these results would not surprise a neo-classical theorist.

We find clear evidence of procrastination and willpower depletion, as well as behavior suggestive of self-reputation considerations. In this respect, each of the behavioral models identifies an important aspect of how people deal with tasks. It is possible that some elements of these models can be combined to produce a more descriptive model; for example, it seems entirely conceivable that the stock of willpower could affect important parameter values in all of the other models.

We consider the area of time structuring, procrastination, and incentives to be just coming into its first full flowering. One clear message from our results is that people may not treat different portions of the week the same, so that this should be embodied in models of discounting or willpower; another is that behavioral interventions often have unintended consequences. We hope our results help spark more research and theory in this nascent area.

REFERENCES

Ainslie, G. (1992), *Picoeconomics: The Interaction of Successive Motivational States within the Individual*, Cambridge: Cambridge University Press.

Ainslie, G. (2001), *Breakdown of Will*, Cambridge: Cambridge University Press.

Akerlof, G., (1991), "Procrastination and Obedience," *American Economic Review*, **81**, 1-19.

Angrist, J. and V. Lavy (2002), "The Effect of High School Matriculation Awards: Evidence from Randomized Trials," mimeo.

Angrist, J., D. Lang, and P. Oreopoulos (2007), "Incentives and Services for College Achievement: Evidence from a Randomized Trial," mimeo.

Ariely, D. and K. Wertenbroch (2002), "Procrastination, Deadlines, and Performance: Self-Control by Precommitment," *Psychological Science*, 13, 219-224.

Benabou, R. and J. Tirole (2004) "Willpower and personal rules," *Journal of Political Economy*, **112**, 848-886.
Bisin, A. and K. Hyndman (2008), "Procrastination, Self-imposed Deadlines and Other Commitment Devices: Theory and Experiment," mimeo.
Burger, N. and J. Lynham (2007), "Betting on Weight Loss… and Losing: Personal Gambles as Commitment Mechanisms," mimeo.
Charness, G. and U. Gneezy (2006), "Incentives to Exercise," mimeo.
Charness, G., E. Haruvy, and D. Sonsino (2007), "Social distance and reciprocity: An Internet experiment," *Journal of Economic Behavior and Organization*, 63, 88-103.
Cook, T. and D. Campbell (1979), *Quasi-experimentation: design & analysis issues for field settings*, Boston: Houghton Mifflin.
DellaVigna, S. and U. Malmendier (2006), "Paying Not to Go to the Gym," *American Economic Review*, **96**, 694-719.
Ellis, A. and W. Knaus (1977), *Overcoming Procrastination*, Institute for Rational Living: New York.
Fischer, C. (2001), "Read This Paper Later: Procrastination with Time-consistent Preferences," *Journal of Economic Behavior and Organization*, **46**, 249-269.
Frederick, S., G. Loewenstein, and T. O'Donoghue (2002), "Time Discounting and Time Preference: A Critical Review," *Journal of Economic Literature*, **40**, 351-401.
Fudenberg, D. and D. Levine (2006), "A Dual-Self Model of Impulse Control," *American Economic Review*, **96**, 1449-1476.
Glasnapp, D. and J. Poggio (1985), *Essentials of Statistical Analysis for the Behavioral Sciences*, Columbus: Merrill.
Laibson, D. (1997), "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, **112**, 443-477.
O'Donoghue, T. and Rabin, M. (1999), "Doing It Now or Later," *American Economic Review*, **89**, 103-124.
Ozdenoren, E., S. Salant, and D. Silverman (2007), "Willpower and the Optimal Control of Visceral Urges," mimeo.
Roth, Alvin (1995), "Introduction," in *The Handbook of Experimental Economics*, J. Kagel and A. Roth, eds., Princeton University Press: Princeton, NJ, 3-109.
Rubinstein, A. (2003), "Economics and Psychology? The Case of Hyperbolic Discounting," *International Economic Review*, **44**, 1207-1216.
Schelling, T. (2002), "Self-command: A New Discipline," in J. Elster and G. Loewenstein (Eds.), *Choice over Time*, New York: Russell Sage Foundation, 167-176.
Solomon, L. and Rothblum, E. (1984), "Academic Procrastination: Frequency and Cognitive-behavioral Correlates," *Journal of Counseling Psychology*, **31**, 503-509.
Strotz, R. (1956), "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, **23**, 165-180.
Suvorov, A. and J. van de Ven (2008), "Goal-setting as a Self-regulation Mechanism," mimeo.
Wertenbroch, K. (1998), "Consumption Self-control by Rationing Purchase Quantities of Virtue and Vice," *Marketing Science*, **17**, 317-337.

APPENDIX A – INDIVIDUAL STUDY HOURS

WEEKLY STUDY HOURS IN NO-WEEKLY GROUP, BY INDIVIDUAL

ID	Female	Week 1	Week 2	Week 3	Week 4	Week 5	Total
1	1	0	0	0	0	0	0.00
2	1	12.31	20.15	13.82	14.44	14.40	75.12
3	0	9.18	14.78	4.12	1.53	0	29.61
4	0	11.16	15.16	16.89	17.92	14.30	75.42
5	1	8.17	9.97	2.46	9.26	0	29.85
6	0	16.70	13.52	6.66	17.47	20.90	75.25
7	1	1.83	0	0	0	0	1.83
8	1	2.15	18.75	4.35	31.19	18.63	75.07
9	1	9.34	6.77	0	2.74	14.31	33.17
10	1	10.27	0.67	3.43	0	0	14.36
11	1	17.72	17.74	20.98	10.01	8.74	75.18
12	1	8.69	18.11	17.43	10.24	20.64	75.11
13	1	12.13	12.43	5.62	5.03	39.93	75.13
14	1	2.82	0	0	0	0	2.82
15	0	18.11	11.99	12.42	18.20	14.49	75.21
16	1	12.54	13.74	10.16	19.61	19.26	75.30
17	0	14.14	22.03	18.05	14.30	7.27	75.80
18	1	0	12.87	13.76	2.74	0	29.38
19	0	16.37	6.58	27.47	8.33	16.63	75.37
44	1	13.69	9.77	13.50	13.05	25.07	75.08
45	1	14.98	12.14	16.97	22.90	8.21	75.19
46	1	24.56	28.81	16.97	5.46	0	75.80
47	1	17.68	13.94	11.94	15.77	16.56	75.89
48	1	1.33	0	0	0	0	1.33
49	1	3.81	14.67	12.08	27.67	17.48	75.72
50	1	16.32	15.49	9.53	14.44	19.38	75.16
51	1	15.19	17.00	11.51	14.79	16.79	75.27
52	1	6.20	11.79	13.74	23.09	20.56	75.38
53	1	5.71	15.80	6.40	29.69	17.58	75.17
54	1	28.93	24.16	19.03	3.33	0	75.46
55	0	1.07	3.79	0	0	0	4.87
56	1	0	0	0	0	0	0.00
57	0	12.93	0.99	0	0	0	13.92
58	0	10.64	0	3.53	0	0	14.17
59	0	7.60	0	0	0	0	7.60
60	1	16.93	19.94	17.73	20.99	0	75.60

WEEKLY STUDY HOURS IN WEEKLY GROUP, BY INDIVIDUAL

ID	Female	Week 1	Week 2	Week 3	Week 4	Week 5	Total
20	0	0	0	0	0	0	0.00
21	0	0	0	0	0	0	0.00
22	0	12.34	11.62	0.02	0	0	23.98
23	0	19.07	15.41	15.86	13.10	13.29	76.74
24	0	14.97	17.59	8.66	12.72	21.11	75.05
26	1	19.70	14.12	12.74	7.08	21.47	75.11
27	1	13.84	11.67	11.12	7.40	0	44.03
28	0	14.70	13.85	8.35	27.88	11.19	75.97
29	0	12.22	3.21	0	0	0	15.42
30	0	0	0	0	0	0	0.00
31	1	18.02	8.23	6.65	15.43	23.89	72.21
32	1	0	0	0	0	0	0.00
33	0	30.06	16.99	12.86	15.11	0	75.02
34	0	12.02	12.07	12.21	22.59	16.24	75.13
35	1	0	0	0	0	0	0.00
36	1	13.53	13.85	19.62	10.30	17.79	75.09
37	0	0	0	0	0	0	0.00
38	1	0	0	0	0	0	0.00
39	0	0	0	0	0	0	1.73
40	0	0	0	0	0	0	0.00
41	0	11.81	4.52	0	0	0	16.33
42	1	5.31	0	0	0	0	5.31
43	1	15.55	13.10	10.23	16.69	20.50	76.06
61	0	12.43	3.82	0	0	0	16.25
62	1	13.02	15.41	10.49	12.42	23.80	75.14
63	1	15.44	28.28	4.82	9.00	17.59	75.13
64	1	14.81	13.60	9.98	13.20	23.67	75.26
65	1	0	0	0	0	0	0.00
66	0	1.41	0	0	0	0	1.41
67	1	0.95	0	0	0	0	0.95
68	1	12.53	12.32	11.93	16.05	2.34	55.18
69	0	14.40	1.61	0	0	0	16.01
70	1	15.60	12.30	7.70	16.07	6.28	57.95
71	0	17.40	7.24	20.59	24.37	5.44	75.04
72	1	19.58	15.91	6.25	21.57	12.07	75.37
73	1	18.36	11.05	16.21	7.83	21.57	75.02
74	1	13.22	11.87	1.77	22.24	26.02	75.11

WEEKLY STUDY HOURS IN NO-WEEKLY GROUP, BY INDIVIDUAL (WINNERS)

ID	Female	Week 1	Week 2	Week 3	Week 4	Week 5	Total
2	1	12.31	20.15	13.82	14.44	14.40	75.12
4	0	11.16	15.16	16.89	17.92	14.30	75.42
6	0	16.70	13.52	6.66	**17.47**	**20.90**	75.25
8	1	2.15	18.75	4.35	**31.19**	**18.63**	75.07
11	1	17.72	17.74	20.98	10.01	8.74	75.18
12	1	8.69	18.11	17.43	10.24	20.64	75.11
13	1	12.13	12.43	5.62	**5.03**	**39.93**	75.13
15	0	18.11	11.99	12.42	18.20	14.49	75.21
16	1	12.54	13.74	10.16	**19.61**	**19.26**	75.30
17	0	14.14	22.03	18.05	14.30	7.27	75.80
19	0	16.37	6.58	27.47	8.33	16.63	75.37
44	1	13.69	9.77	13.50	**13.05**	**25.07**	75.08
45	1	14.98	12.14	16.97	22.90	8.21	75.19
46	1	**24.56**	**28.81**	16.97	5.46	0	75.80
47	1	17.68	13.94	11.94	15.77	16.56	75.89
49	1	3.81	14.67	12.08	**27.67**	**17.48**	75.72
50	1	16.32	15.49	9.53	14.44	19.38	75.16
51	1	15.19	17.00	11.51	14.79	16.79	75.27
52	1	6.20	11.79	13.74	**23.09**	**20.56**	75.38
53	1	5.71	15.80	6.40	**29.69**	**17.58**	75.17
54	1	**28.93**	**24.16**	19.03	3.33	0	75.46
60	1	16.93	19.94	17.73	20.99	0	75.60

WEEKLY STUDY HOURS IN WEEKLY GROUP, BY INDIVIDUAL (WINNERS)

ID	Female	Week 1	Week 2	Week 3	Week 4	Week 5	Total
23	0	19.07	15.41	15.86	13.10	13.29	76.74
24	0	14.97	17.59	8.66	12.72	21.11	75.05
26	1	19.70	14.12	12.74	7.08	21.47	75.11
28	0	14.70	13.85	8.35	**27.88**	**11.19**	75.97
33	0	**30.06**	**16.99**	12.86	15.11	0	75.02
34	0	12.02	12.07	12.21	**22.59**	**16.24**	75.13
36	1	13.53	13.85	19.62	10.30	17.79	75.09
43	1	15.55	13.10	10.23	*16.69*	*20.50*	76.06
62	1	13.02	15.41	10.49	*12.42*	*23.80*	75.14
63	1	**15.44**	**28.28**	4.82	9.00	17.59	75.13
64	1	14.81	13.60	9.98	*13.20*	*23.67*	75.26
71	0	17.40	7.24	20.59	24.37	5.44	75.04
72	1	19.58	15.91	6.25	21.57	12.07	75.37
73	1	18.36	11.05	16.21	7.83	21.57	75.02
74	1	13.22	11.87	1.77	**22.24**	**26.02**	75.11