What Do We Know About Our Future Selves?
Essays on Sophistication and Prediction.

by

Daniel James Acland

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Economics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Stefano DellaVigna, Chair
Professor Botond Kőszegi
Professor Eugene Smolensky

Fall 2009

What Do We Know About Our Future Selves?
Essays on Sophistication and Prediction.

Abstract

What Do We Know About Our Future Selves?
Essays on Sophistication and Prediction.

by

Daniel James Acland
Doctor of Philosophy in Economics

University of California, Berkeley

Professor Stefano DellaVigna, Chair

What people know about their future preferences and how they take this knowledge into account in their decisions are questions of primary importance in formal models of intertemporal choice and in many domains of public policy. I investigate prediction of changes in state-dependent preferences in the case of habit formation, prediction of future self-control problems, and how agents with self-knowledge with respect to future self-control problems think about the actions and beliefs of their future selves.

In chapter one I and a coauthor extend the gym-attendance study of Charness and Gneezy (2009) by incentivizing subjects to attend the gym for a month and observing their pre- and post-treatment attendance relative to a control group. In addition we elicit subjects' pre- and post-treatment predictions of their post-treatment attendance. We find a habit formation effect similar to that of Charness and Gneezy in the short-run, but with substantial decay caused by winter vacation. We additionally find that subjects seriously over-predict future attendance, which we interpret as evidence of partial naivete with respect to self-control problems. Subjects also appear to have biased beliefs about their future cost of gym attendance. Our design allows us to estimate the monetary value of habit formation—equivalent to a $0.40 per visit subsidy—as well as the welfare cost of naivete.

In chapter two we address whether individuals accurately predict habit-formation, a question of both theoretical and practical interest. Gym-attendance is one domain in which this question is of particular interest to public policy makers. We test for misprediction of habit-formation in gym attendance with a field experiment and find that subjects do form a habit, and do not predict it fully. We develop a simple model that incorporates habit-formation and projection bias in the framework of quasi-hyperbolic discounting and calibrate the parameters of the model.

In chapter three, borrowing from Cognitive Heierarchy Theory, I introduce bounded rationality into the beta–delta model of present-biased preferences. I define a level-two agent—or "k-2–sophisticate"—as one who is aware that her future selves will have present-bias, but believes that they will be naive. The k-2–sophisticate does one round of strategic thinking about her future behavior instead of the unlimited number of rounds of the full sophisticate. In the "doing it once" model of procrastination of O'Donoghue and Rabin (1999) the k-2–sophisticate typically procrastinates and preproperates less than the full sophisti-

cate, and is protected from severe harm from both extreme preproperation and extreme procrastination, though she may suffer from excessive costly preemption due to pessimism about future preemption when costs are immediate.

Professor Stefano DellaVigna
Dissertation Committee Chair

To Geno,
who inspired me at the beginning,
encouraged me in the middle,
and pulled me together at the end.

And to Stefano,
who saw more in me
than I was able to see in myself,
and still does.

Contents

Acknowledgments

Financial support for chapters one and two was provided by the National Institute on Aging through the Center on the Economics and Demography of Aging at UC Berkeley, grant number P30 AG12839. Additional financial support was provided by the UC Berkeley Integrated Graduate Education Research and Training Program in Politics, Economics, Psychology, and Public Policy. Matthew Levy was a wonderful collaborator on chapters one and two. I would also like to thank Gary Charness, Uri Gneezy, Stefano DellaVigna, Matthew Rabin, Botond Koszegi, Teck Hua Ho, Alexander Mas, Ulrike Malmendier, Shachar Kariv, and all of the participants in the UC Berkeley Psychology and Economics Non-Lunch for helpful comments. Special gratitude goes to Brenda Naputi of the Social Science Experimental Laboratory at the Haas School of Business, Brigitte Lossing at the UC Berkeley Recreational Sports Facility, and to Vinci Chow and Michael Urbancic of the UC Berkeley Department of Economics, for extraordinary assistance with implementation.

In addition I would like to thank everyone who supported me though this process, both within the academy and without. In particular, all of my first-year study-group friends, who later became my games-night friends, and later still just my friends. Also my very large, extended spiritual family, and in particular Joe, Kelly, Nat and Nydia, Myrna, Mom, Dad, Bea, Rick, Tom and Durga, and many, many others.

Chapter 1

Habit Formation and Naiveté in Gym Attendance

1.1 Introduction

Incentivizing healthy behaviors, and in particular physical exercise, has received increasing interest in various literatures in the face of growing concern about the cost of health care and the increasing problem of obesity.[1] Of particular interest is the potential to build long-term healthy behaviors with short-term incentive interventions. Charness and Gneezy (2009) provided the first experimental evidence on this possibility in the domain of physical exercise, showing that paying a group of undergraduates to attend the gym for a month raises attendance in the subsequent weeks, despite the removal of the incentive. This effect can be interpreted as habit formation.

Their study raises a number of interesting questions that deserve further investigation. How does the habit decay over time? What is the role of self-control problems in gym attendance? How well do subjects predict various dimensions of their future gym attendance? And is it possible to calibrate the value of the habit? These are key to understanding the welfare effects of the intervention, as well as its policy relevance. In this paper, we present evidence from a field experiment designed to answer these questions.

Charness and Gneezy paid undergraduates to attend the gym for four weeks and found that, after the payment ended, treated subjects had significantly higher gym attendance than did a control group. Their subjects were university undergraduates who were randomized into three groups.[2] A "low incentive" group were offered $25 to attend the gym once during the initial week of the study. A "high incentive" group received the same $25 offer, and were additionally offered $100 to attend the gym another eight times in the subsequent four weeks for a total of nine visits over five weeks. A control group received no offers for gym attendance. Gym-attendance data was collected for all subjects for a period beginning eight weeks before the treatment and ending seven weeks after. By comparing the pre- to post-treatment change in attendance across groups they are able to show that sub-

[1] See Kane, Johnson, Town and Butler (2004) for a review.

[2] We are describing Charness and Gneezy's first study, which our experiment is most similar to. In the same paper they conducted a second study with a slightly different design that yielded similar results.

jects in the high-incentive group continue to have significantly higher gym attendance after the incentive period ends than subjects in the other two groups—an average of 0.67 visits per week more than the control group, and 0.58 visits per week more than the low-incentive group. Furthermore, they found that the increase came from the subset of subjects who had previously attended less than once per week on average, which they refer to as non-regular attenders.

To explore our questions of interest we built on Charness and Gneezy's high-incentive and low-incentive treatments. We recruited 120 subjects who were self-reported non-regular gym attenders. We then collected gym attendance data covering a span of seventeen months, allowing us to investigate habit decay more thoroughly. Further, in addition to the $25 and $100 attendance incentives, we used an incentive-compatible mechanism to elicit subjects' predictions of their post-treatment gym-attendance, conducting the elicitation both immediately before and immediately after the treatment period, allowing us to explore issues of mis-prediction. Finally, the elicitation mechanism involved offering small attendance incentives in some of the post-treatment weeks, which allows us to estimate the costs and benefits associated with the habit.

We find a short-run habit-formation effect among our subjects of 0.256 visits per week, which is smaller than, but statistically indistinguishable from, Charness and Gneezy's result. However, the effect appears to largely decay over the course of winter vacation. Moreover, this treatment effect is highly concentrated in the upper tail of the post-treatment attendance distribution. We also find that subjects substantially over-predict their future gym attendance—even in our simplest elicitation task, subjects over-predicted attendance by roughly a factor of three. Predictions are closer to actual attendance after the treatment period than before. By fixing the delay between the week in which predictions are made and the week about which they are made, we rule out intertemporal discounting as an explanation for this shift, suggesting that subjects also mispredict some other aspect of their gym-attendance decision, such as the opportunity cost of attendance. Finally, we estimate two key parameters of the model: the dollar value of the habit-formation effect, and the value of the unforeseen portion of the foregone long-term gym-attendance benefit lost due to self-control problems. We find that the habit induced in treated subjects is equivalent to a $0.50 per visit subsidy overall, or $4.50 per visit among subjects we identify as habit-formers. The cost of naivete is also large, and indicates that the intervention may be welfare-enhancing.[3] Using these parameters, we set forty-six weeks as an upper bound on how long habituated subjects must retain their gym habit for the intervention to be cost-effective.

The chapter unfolds as follows. Section two presents our model and our parameter-estimation strategy. Section three describes our experimental design. Results are presented in section four. Section five concludes.

[3]By contrast, in a model without time-inconsistency this intervention would increase long-run gym attendance but be inefficient relative to a lump-sum transfer to subjects.

1.2 Model

In this section we develop a simple model of gym attendance that incorporates habit formation and present-biased preferences. Habit—caused by past gym attendance—is modeled as a fixed, additive increase in gym-attendance utility, à la Becker and Murphy (1988) and O'Donoghue and Rabin (1999a). Individuals discount all future periods relative to the present, à la Phelps and Pollak (1968) and Laibson (1997), and are naive or sophisticated with respect to this "quasi-hyperbollic discounting", à la O'Donoghue and Rabin (1999b).

In the spirit of DellaVigna and Malmendier (2004), we consider a finite-horizon, discrete-time model with five unequal periods. Initially all subjects are non-habituated, and are randomly divided into two groups, one of which will be incentivized to attend the gym in period one (treated group), and the other of which will not (control group). In the first period subjects bid, in an incentive compatible auction, on a "p-coupon", a certificate that rewards fourth-period gym attendance, and then predict how many times they will go to the gym that period if they win the coupon.[4] Then, still in the first period, treated subjects attend the gym and develop a habit that will persist through all subsequent periods.

In the second period two things happen. First subjects once again bid on the fourth-period p-coupon and predict their fourth-period attendance. Then, after the auction, all subjects are given a p-coupon.[5] Period three acts as a buffer, ensuring that the target period is considered to be "in the future" when predictions are elicited. In period four, subjects receive p-coupon rewards according to their gym attendance in that period. We explicitly think of periods three and four as weeks, so that subjects decide each day whether to attend the gym that day. Finally, in period five subjects receive the delayed benefit of whatever gym attendance they have engaged in.

Let the immediate utility of gym attendance on day d be $-c + \varepsilon_d$ with $c > 0$, and i.i.d. $\varepsilon_d \sim F$. Let the delayed benefit of gym attendance be $b > 0$. Thus we model gym attendance as an "investment good" in the language of DellaVigna and Malmendier, meaning that costs are immediate while rewards are delayed. Future payoffs are discounted by β, with beliefs about future self-control denoted by $\widehat{\beta}$.[6] Following O'Donoghue and Rabin (1999a), habit formation takes a simple binary form. When subjects are habituated they receive additional, immediate utility for gym attendance of $\eta > 0$, so that the immediate utility of gym attendance for a habituated subject is $\eta - c + \varepsilon_d$. We model utility as quasi-linear in money. Utility from all non-gym sources is normalized to zero.

Let P be the face value of the p-coupon that rewards gym attendance in period four. That is, a p-coupon pays $\$P$, immediately, for each day that the holder attends the gym in period four. Let X_t^g refer to the valuation of a p-coupon in period $t = 1, 2$ of a subject in group $g = 0, 1$ (control=0, treated=1). Let Z^g be the number of days of gym attendance during the target week for a subject in group g.

[4] We refer to period four as the "target-week" as it is the target of the p-coupon.

[5] In the model we are ignoring the fact that the elicitation process requires one or two subjects to wind up with two coupons. In practice, because there were multiple target weeks, most of the auction winners did not end up holding multiple p-coupons for the same week. The two subjects who did wind up with two p-coupons for the same target week simply received double the reward.

[6] Because of the short time horizon, we assume no long-run discounting, i.e. $\delta = 1$.

1.2.1 Attendance decision and the value of a p-coupon.

If a subject attends the gym on a given day during the target week her utility for that day will be $P + \beta b + g\eta - c + \varepsilon_d$. She will attend the gym if this is positive. Thus $Z^g = \sum_{d=1}^{7} \mathbb{1} \cdot \{\varepsilon_d > P + \beta b + g\eta - c\}$. In expectation, total target-week gym-attendance will be,

$$\sum_{d=1}^{7} \Pr(\varepsilon_d > P + \beta b + g\eta - c) = 7 \times \int_{c-\beta b-g\eta-P}^{\infty} dF(\varepsilon). \tag{1.1}$$

However, from the perspective of any previous period, the perceived probability of target-week gym-attendance depends upon the subject's belief about future self-control, $\widehat{\beta}$. She believes she will attend on any given day of the target week if $\varepsilon_d > P + \widehat{\beta} b + g\eta - c$. Thus the subject's ex-ante prediction of her total utility for the target-week, given that she holds a p-coupon, is,

$$7 \times \int_{c-\widehat{\beta} b-g\eta-P}^{\infty} (P + b + g\eta - c + \varepsilon)\, dF(\varepsilon). \tag{1.2}$$

Setting P to zero gives us the predicted utility without a p-coupon. The value of the p-coupon, from the perspective of either period one or period two, is the difference between expected utility with a p-coupon and expected utility without a p-coupon, which is,

$$X_1^g = X_2^g = \left[7 \times \int_{c-\widehat{\beta} b-g\eta-P}^{\infty} P\, dF(\varepsilon) \right] + \left[7 \times \int_{c-\widehat{\beta} b-g\eta-P}^{c-\widehat{\beta} b-g\eta} (b + g\eta - c + \varepsilon)\, dF(\varepsilon) \right]. \tag{1.3}$$

Note that this valuation is the same for pre- and post-treatment elicitations because the target week is in the future (hence "inside β") from the perspective of either elicitation period. The first term in the expression is the expected redemption value of the coupon, which is always weakly positive. The second term is the subject's valuation of the behavioral change that results from holding the coupon, which we will call the incentive value. This is the change in utility caused by those gym-visits that the subject would not have made in the absence of the p-coupon. The sign depends on the subject's ex-ante belief about future self-control problems. If the subject believes that she will not have self-control problems in the target week, the incentive value is negative because the subject believes that the p-coupon will make her attend the gym when the direct utility of doing so is negative. If the subject believes that she will have self-control problems in the target week, then the incentive value may be positive because she may foresee that the p-coupon will make her more likely to attend the gym and gain a long-term benefit that she would otherwise forego due to self-control problems.[7] Note that the net value of the p-coupon is always

[7] Thus, for a sophisticate with self-control problems the incentive value can be thought of as "commitment value" because it is the value of having the p-coupon as a "commitment device" to help her get out the door and down to the gym.

non-negative.

1.2.2 Parameter Identification

We focus our estimation on two parameters that are key to evaluating the welfare effects of the intervention and which can be estimated in a parsimonious two-equation system. The first is the habit-formation effect itself, η, which is the additional, per-visit, gym-attendance utility (measured in dollars) received by a subject in the habituated state. Another way to think of this parameter is that η is the per-visit monetary incentive that would cause a non-habituated subject to attend as often as an unincentivized habituted subject. The second term we are interested in estimating is the per-visit cost of naivete with respect to self-control, $(\widehat{\beta}-\beta)b$. This is the dollar value of the portion of the per-visit future benefit of gym attendance, b, that present bias makes a subject willing to forego, but which a naif fails to foresee.

The first parameter of interest is η, the habit value. Our estimation strategy is essentially equivalent to finding the value of P for which the average target-week attendance in the control group, with a p-coupon, is the same as the average target-week attendance in the treated group, without a p-coupon. Let $\overline{Z}_p^g$ be the average weekly attendance of subjects in group $g \in \{T, C\}$ who are holding a p-coupon, and $\overline{Z}_0^g$ be the same thing for subjects with no p-coupon (i.e. $P = 0$). In terms of our model, we are looking for P^* such that,

$$\overline{Z}_0^T = 7 \times \int_{c-\beta b-\eta}^{\infty} dF(\varepsilon) = 7 \times \int_{c-\beta b-P^*}^{\infty} dF(\varepsilon) = \overline{Z}_p^C. \tag{1.4}$$

Once we know the value of P^*, because $F(\cdot)$ is monotonically increasing, we then have $\eta = P^*$.

The cost of naivete, $(\widehat{\beta}-\beta)b$, is identified by comparing the control group's predicted target-week attendance with their actual attendance. Let $\overline{Y}_p^g$ be the average, unincentivized prediction, in either elicitation session, of gym attendance during a target week with a p-coupon of subjects in group g. The average unincentivized prediction of gym attendance in a target week with a p-coupon with a face value of $\widetilde{P}$, among control subjects, is

$$\overline{Y}_p^C = 7 \times \int_{c-\widehat{\beta}b-\widetilde{P}}^{\infty} dF(\varepsilon) = 7 \times \int_{c-\beta b-(\widehat{\beta}b-\beta b)-\widetilde{P}}^{\infty} dF(\varepsilon). \tag{1.5}$$

We find the value of P^* for which

$$\overline{Y}_p^C = 7 \times \int_{c-\beta b-(\widehat{\beta}b-\beta b)-\widetilde{P}}^{\infty} dF(\varepsilon) = 7 \times \int_{c-\beta b-P^*}^{\infty} dF(\varepsilon) = \overline{Z}_p^C, \tag{1.6}$$

which gives us $(\widehat{\beta}-\beta)b = P^* - \widetilde{P}$. In practice we will evaluate this by setting $\widetilde{P}$ equal to the average value of P among all control subjects. We estimate the moment equations in (2.8) and (1.6) in section .

1.3 Design

We recruited one hundred and twenty subjects from the students and staff of UC Berkeley and randomly assigned them to treated and control groups.[8] Since Charness and Gneezy found the habit-formation effect concentrated among non-attenders we screened for subjects who self-reported that they had not ever regularly attended any fitness facility.[9] Treated and control subjects met in separate sessions on the same day, at the beginning of the second week of the fall semester of 2008. Both treatment and control subjects were asked to complete a questionnaire, and were then given an offer of $25 to attend the gym once during the following week.[10] We call this the "learning week" offer, and it is identical to Charness and Gneezy's low-incentive condition. Our control group is therefore comparable to Charness and Gneezy's low-incentive group. We chose this as our control in order to separate the effect of overcoming the one-time fixed cost of learning about the gym from the actual habit formation that occurs after multiple visits.[11]

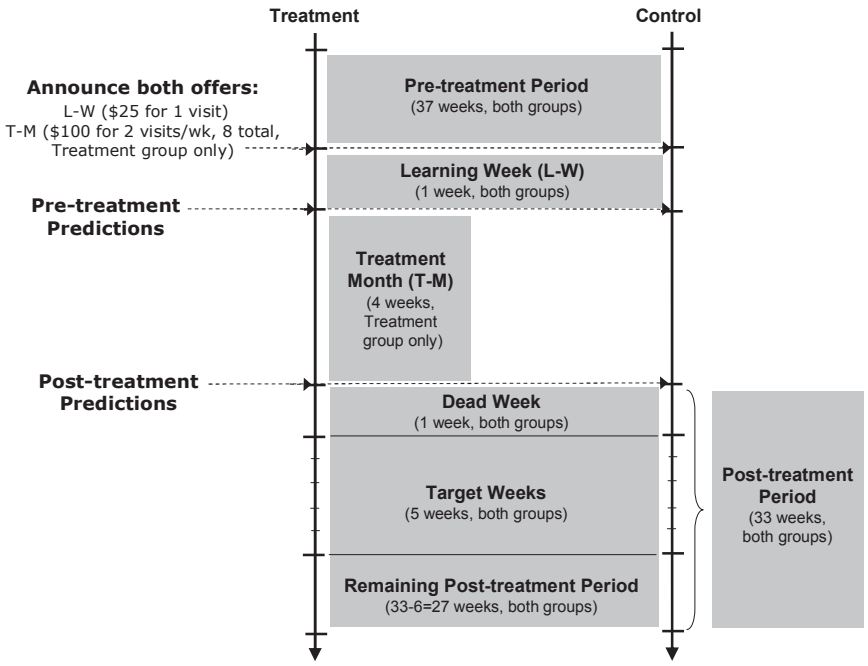

Figure 1.1: Our Experimental Design

[8]Due to attrition and missing covariates, our final sample includes 54 treated subjects and 57 control subjects. Details of the sample appear in appendix B.

[9]Our screening mechanism is described in appendix C.

[10]For this and all subsequent offers, subjects were told that a visit needed to involve at least 30 minutes of some kind of physical activity at the gym. We were not able to observe actual behavior at the gym and did not claim that we would be monitoring activity.

[11]We also paid the $10 gym-membership fee for all students, and filed the necessary membership forms for those who were not already members.

At the same initial meeting, the treatment group received an additional offer of $100 to attend the gym twice a week in each of the four weeks following the learning week. We call this the treatment-month offer, and it is the same as Charness and Gneezy's high-incentive offer, except that they did not require the eight visits to be evenly spaced across the four weeks. The other difference between this offer and Charness and Gneezy's high-incentive offer is that we made our offer at the first meeting, at the same time as the $25 learning-week offer, whereas Charness and Gneezy made their high-incentive offer at their second meeting, a week later. We made our treatment-month offer earlier because we wanted Treatment subjects to have a week to contemplate the idea of going to the gym twice weekly for a month before making predictions. Moreover, if subjects have reference-dependent preferences for money then suddenly announcing a gain of $100 to one group but not the other could introduce systematic bias into the incentive compatible procedure we used to elicit predictions. Waiting a week after treatment subjects learn they will earn $100 will help us overcome a potential "house money effect".

At the end of the learning week both groups of subjects again met separately and completed pencil-and-paper tasks (described in detail below) designed to elicit their predictions of gym attendance during each of five post-treatment "target weeks". Both groups were reminded of the offers they had received. Four weeks later, at the end of the treatment month, both groups again met separately, completed an additional questionnaire, and completed the same elicitation tasks as in the second session. The target weeks were separated from this second elicitation session by a dead week so that present-biased subjects would see the target weeks as being "in the future" from the perspective of both elicitation sessions. The timeline of the experiment is illustrated in Figure 2.1.

Gym attendance data were collected for a 17-month period stretching from 37 weeks before the learning week to 33 weeks after it. This period includes summer and winter breaks as well as three full semesters.

1.3.1 Elicitation procedures

To elicit predictions of target-week gym attendance we created what we call a "p-coupon", which is a certificate that rewards the holder with $\$P$ for each day that he or she attends the gym during a specified "target week". The value of P, which ranged from $1 to $7, was printed on the coupon, along with the beginning and end dates of the target-week. We used an incentive-compatible mechanism to elicit subjects' valuations for p-coupons of various values with various target weeks.[12] A subject's incentive-compatible bid for a p-coupon is correlated with how many times they think they will attend the gym during the target week of the coupon. A sample p-coupon is included in appendix D, along with the pencil-and-paper task we used to elicit valuations for p-coupons, the instructions we gave them for completing the task, and further description of how the elicitation mechanism worked. Each subject completed this incentive-compatible elicitation task for four of the five target weeks in our design, and for a different value of p-coupon in each of those four

[12]Subjects made a series of choices between a p-coupon and an incrementally increasing fixed amount of money. We infer their valuation from the indifference point between the coupon and the fixed sum. The elicitation mechanism is described in detail in appendix D.

weeks. The values of the p-coupons for the different weeks was randomized among subjects, as was the order in which those weeks were presented.[13]

Subjects' bids for a coupon that pays out as a function of the number of times a certain event occurs in a future target week need not be based entirely on their predictions of how many times that event will occur. Risk-aversion implies we would only observe subjects' certainty equivalents, even for an exogenous event.[14] But for an endogenous event like gym attendance, there is the additional confound that the p-coupon itself incentivizes the subject to go to the gym, thus influencing the very behavior we are asking them to predict. This "incentive effect" may increase or decrease subjects' bids for a p-coupon, and care must therefore be taken not to interpret subjects' bids as directly proportional to their beliefs.

As a check on this mechanism, we also directly asked subjects to state how many times they thought they would go to the gym during the specified target weeks if they had been given the p-coupon they just bid on in the incentive-compatible task. Thus they were making unincentivized *predictions* of hypothetical future attendance under the same set of *attendance* incentives as in the incentivized task.[15] This unincentivized mechanism also allowed us to ask subjects how often they thought they would go to the gym during the one target week for which they were not presented with a p-coupon, the so-called "zero week" (because it is equivalent to a P of zero). The zero week gives us an additional unincentivized prediction of behavior in the absence of any effect of attendance incentives.

Subjects went through exactly the same set of elicitation tasks in both the pre-treatment and post-treatment elicitation sessions. Then, at the end of the second elicitation session, after all of the elicitation tasks had been completed, each subject was given one of the four coupons they had been presented with during the elicitation process. These give-away coupons were in addition to those that had been won earlier in the bidding process. We therefore have two target weeks for each subject in which we can compare their predictions with their actual gym attendance under the same conditions, the first being the zero-week, and the second being the week for which they received a p-coupon in the giveaway. The giveaway was a surprise to the subjects—having been conducted unannounced only after the second elicitation session—and thus did not affect their bids or unincentivized responses during the elicitation tasks.

We discuss compliance with the treatment incentive, attrition, and our randomization procedure in appendix E.

[13] Thus subjects did not all bid on a p-coupon for target-week one, then target-week two, etc, nor did all subjects bid on p-coupons of the same size for each of the target weeks. Among each subject-group/target-week intersection, subgroups of fifteen subjects received \$1, \$2, and \$3 coupons, ten received \$5 coupons, and five received \$7 coupons.

[14] An alternative design which would have allowed us to sidestep assumptions about the linearity of money utility, would have been to have the coupons pay off not with a dollar sum per visit, but with a per-visit increment in the cumulative probability of winning some fixed-sum prize. We believe our design is more intuitive for subjects, and easier for them to understand.

[15] It is important to note that the p-coupons incentivize both target-week attendance and accurate predictions of target-week attendance.

1.4 Results

Of the 54 subjects in our final treatment sample, 43 completed the eight necessary bi-weekly visits in order to earn the $100 incentive–a compliance rate of 80%. In Charness and Gneezy's (2009) high-incentive group the compliance rate was approximately 83%, suggesting that our more restrictive design did not have a significant effect on subjects' ability to make the required number of visits. It is surprising that our sample of non gym-attenders were so easily induced to visit the gym eight times.

1.4.1 Habit formation

Figure 1.2 shows average weekly attendance for the treated and control groups over the duration of the study period.[16] In the pre-treatment period, attendance in the two groups moves together tightly. In the treatment period, treated subjects attend much more than control subjects. In the two months immediately following the treatment period, leading up to, but not including winter vacation, the treatment group consistently attends the gym more than the control group. In the four months after the winter vacation the graph suggests persistence of the increased treatement-group attendance, but the difference is not as striking.

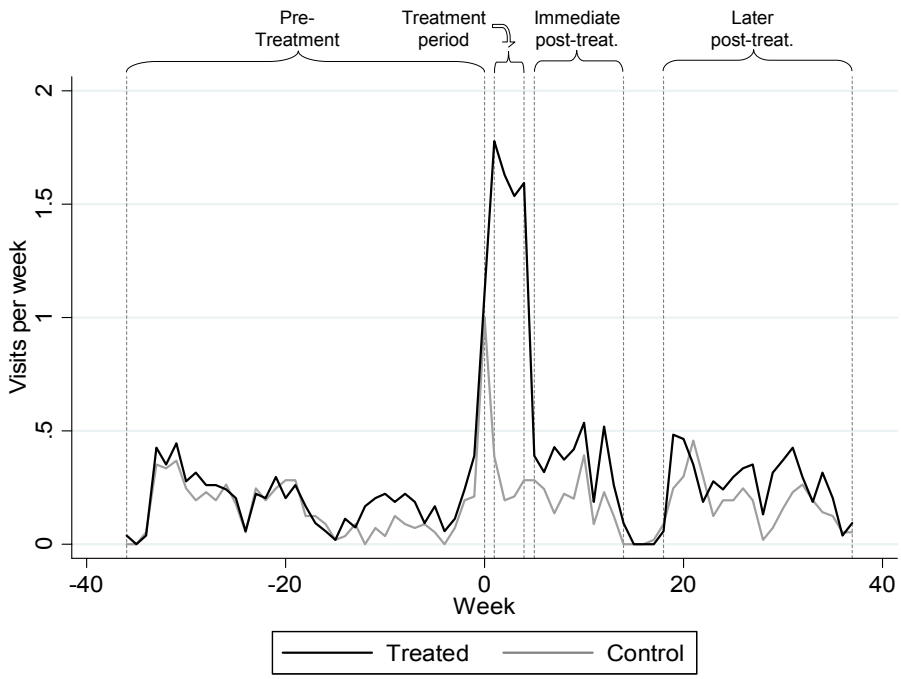

Figure 1.2: Gym Attendance

[16] We have removed observations for target weeks when subjects received p-coupons to make the graph easier to read.

We estimate a linear, difference-in-differences, panel regression model to see if these patterns are statistically significant. Each observation in the panel is a specific individual on a specific week of the study.[17] We regress weekly gym attendance on a treated-group dummy, a set of week-of-study dummies, and the interactions of the treated-group dummy with dummies for the treatment period and each of the two post-treatment periods. The results of this regression appear in the first column of Table 1.1.

Table 1.1: Habit Formation: Regression of average weekly attendance.

	(1)	(2)	(3)	(Charness & Gneezy)
Treated	0.045	0.045		-0.100
	(0.057)	(0.057)		(0.196)
				[0.477][a]
Treatment Period X Treated	1.321***	1.209***		1.275***
	(0.134)	(0.150)		(0.181)
				[0.780][a]
Imm. Post-Trmt X Treated[b]	0.129	0.256**		0.585***
	(0.111)	(0.122)		(0.217)
				[0.186][a]
Later Post-Trmt x Treated[b]	0.050	0.045		–
	(0.095)	(0.098)		
Complied w/ treatment			0.057	
			(0.071)	
Treatment Period X Complied			1.582***	
			(0.180)	
Imm. Post-Trmt X Compliance[b]			0.338**	
			(0.154)	
Later Post-Trmt x Compliance[b]			0.061	
			(0.126)	
Week Efffects	Yes	Yes	Yes	Yes
Controls	–	Yes	Yes	–
IV	–	–	Yes	–
Observations	7433	7433	7433	1520
Num Clusters	111	111	111	80
R-squared	0.15	0.21	0.22	0.13

Notes: [a]Terms in square brackets are p-values from a Chow test of equal coefficients between our sample (column ii) and Charness and Gneezy (2009)'s sample. [b]"Immediate" refers to the 8 weeks following the intervention (excluding the "dead week" for columns (i)-(iii). "Later" refers to the 19 weeks of observations in the following semester (excluding the winter holiday). Robust standard errors in parentheses, clustered by individual. * significant at 10%; ** significant at 5%; *** significant at 1%.

[17]We again exclude observations for the one target week for each subject for which they received an actual p-coupon.

The coefficient on the treated-group dummy tells us that there is no statistically significant difference in gym attendance between treated and control subjects in the pre-treatment period. The coefficient on the interaction of the treated-group and treatment-period dummies reassures us that the treatment-incentive was effective. The coefficient is roughly the product of the twice-weekly incentive target and the 80% compliance rate. The remaining two interaction terms tell us the effect of the treatment on treated-group attendance in the two post-treatment periods. The point-estimate is 0.129 additional visits per week for the immediate post-treatment and 0.050 for the later post-treatment period. Neither of these simple differences-in-differences is statistically significant.

The second column is the same regression with individual-level covariates added.[18] The treatment effect in the immediate post-treatment period is now larger, 0.254, and statistically significant at the 5% level. Thus, when we control for individual characteristics we find an average increase in gym attendance for members of the treated group of a quarter of a visit per week. In the later post-treatment period we still cannot reject that there was no treatment effect. To test whether the coefficient in the immediate post-treatment period is significantly different from the same one in the first column, without controls, we run a Hausman test. Dividing our covariates into four groups–economic, demographic, naivete proxies, and attitudes about gym attendance–we find that the last two explain three-quarters of the change in the coefficient, but none of the groups has a statistically significant effect. The p-value of the test is 0.051, suggesting that we may be correcting for some lumpiness in our randomization.[19]

Because not all subjects in the treatment group made the requisite eight visits to the gym, the results in column two represent the "intention to treat" effect, or ITT. To see the effect on those who complied with the treatment we instrument for compliance with the treated-group dummy, including our vector of individual covariates in the first stage. This gives us the average "treatment effect on the treated", or ATT, controlling for observable differences between compliers and non-compliers. These results are reported in the third column of Table 1.1. Not suprisingly, the ATT is larger than the ITT. We now see an increase in immediate post-treatment gym attendance for the treated-group of a third of a visit per week. In the later post-treatment period we still see no statistically significant increase, despite the apparent difference between treated and control attendance in Figure 1.2. These results suggest that there is habit formation in the immediate post-treatment period, but the habit has decayed when students return from winter break.

To further explore the decay of habit over time we ran a post-estimation Wald test to see whether the immediate post-treatment coefficient is the same as the later post-treatment coefficient. The F-statistic from this test is 2.73 and the probability of seeing a statistic this large is 0.1016. In other words, we cannot reject that the post-winter coefficient is the same as the pre-winter coefficient. This result, together with the results in the table suggest that the habit largely decays over the course of winter break, with perhaps some residual habit remaining into the spring semester.

To compare our results with the results from Charness and Gneezy's first study

[18] These include basic economic and demographic variables, as well as measures of naivete and attitudes towards exercise. The controls and their balance between treatment groups are discussed in Appendix B.

[19] The decomposition of the Hausman test is described in detail in appendix F.

we ran the same regression on their data, the results of which comprise the final column of Table 1.1. The double difference in average weekly attendance between their high-incentive and low-incentive subjects in the immediate post-treatment period was 0.585 visits per week. Stacking their data with ours allows us to conduct a Chow test of the equality of their habit-formation coefficient with the one in our column-two specification. The p-value, reported in square brackets, is 0.186. Thus we cannot reject that the habit-formation effect in our sample was the same as the habit-formation effect in their sample.[20]

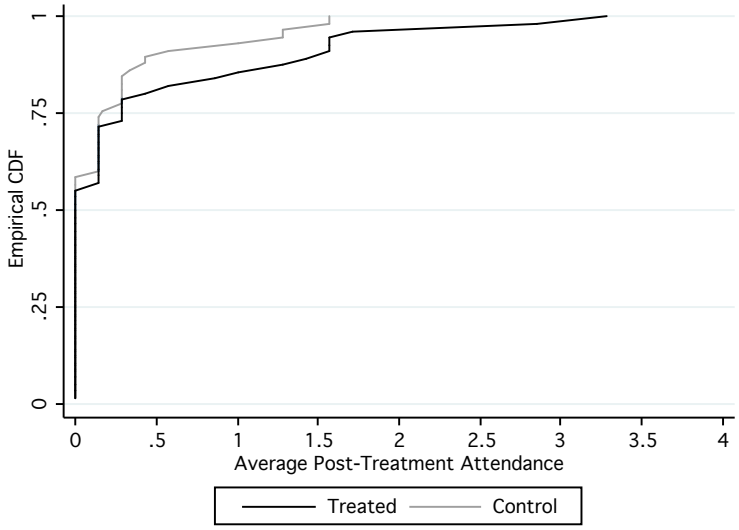

Figure 1.3: Distribution of immediate post-treatment attendance.

To get a better picture of the treatment effect in the immediate post-treatment period, Figure 2.2 plots the empirical CDFs of average post-treatment attendance in the treated and control groups.[21] There is clearly considerable heterogeneity in the treatment effect. The two distributions are similar up to the seventy-fifth percentile—the majority of both treatment and control subjects continue to avoid gym attendance altogether—and then diverge substantially. Thus, though three quarters of our treated subjects complied with the treatement incentive, only about one quarter of them appear to have formed a habit of any size. Similar to Charness and Gneezy, we identify as "habit-formers" those subjects in each group for whom average attendance in the immediate post-treatment period was at least one visit per week greater than an imputed counterfactual based on a regression of attendance on week dummies and covariates using control group data for all weeks and treated group data for the pre-treatment period. This applies to 8 of 54 treated subjects and 3 of 57 control subjects. A test of equal proportions rejects equality at the $p = 0.092$ level, and the one-sided test that there are actually more habit-formers in the control group is rejected at a p-value of 0.046.

[20] The point estimate of the double difference during the treatment period is smaller in the Charness and Gneezy data than in ours. This is largely because baseline attendance was higher in their sample, so that high-incentive subjects needed less of an increase in attendance to earn the $100 incentive.

[21] Attendance in a subject's incentivized week is ommitted from the calculation.

1.4.2 Predictions

We next turn our attention to subjects' predictions. Figure 1.4 shows predicted versus actual gym attendance for the weeks that subjects actually received a p-coupon in the giveaway at the end of the experiment, and for weeks when no p-coupon was offered—so-called "zero-weeks". The two panels break the subjects into control and treated groups. Within each group we separate observations into p-coupon weeks and zero-weeks.[22] Finally, we separate subjects predictions by when they were elicited. We show only subjects' unincentivized predictions for clarity, but Tables 1.2 and 1.3 confirm that incentivized and unincentivized predictions are quite similar.

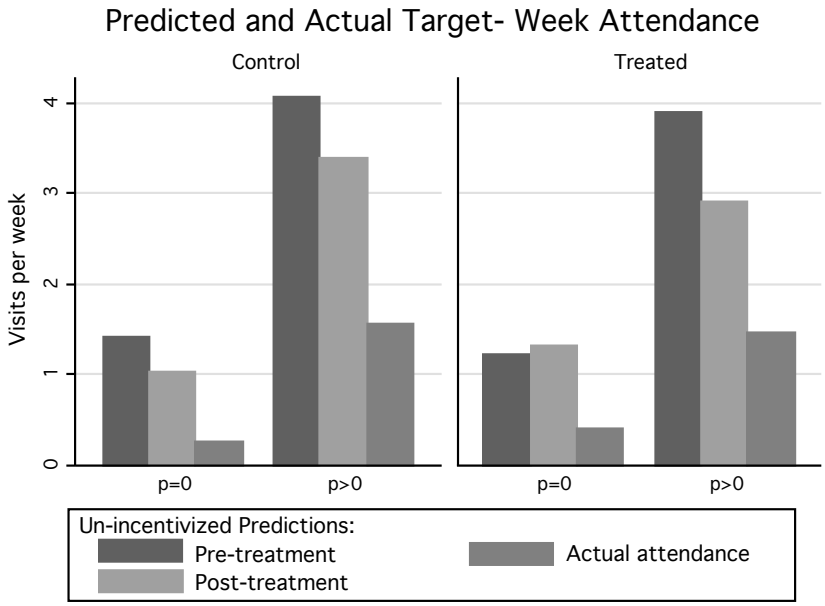

Figure 1.4: Predicted versus Actual Attendance

In both the pre- and post-treatment elicitation sessions, both the treated and control groups predicted future gym attendance that substantially exceeds their actual gym attendance. This pattern holds for both p-coupon weeks and zero-weeks. Furthermore, introducing a p-coupon seems to increase both actual and predicted attendance, as we would expect. Finally, there is a consistent pattern of less over-prediction in the later elicitation session.

Table 1.2 shows the difference between predicted and actual attendance for the different groups and elicitation sessions, pooled over values of the p-coupon. The first column of each panel looks at predictions as captured by subjects' p-coupon bids. The second and third look at their unincentivized predictions, for p-coupon weeks and zero-weeks. In all cases subjects significantly over-predict future gym attendance, by as much as two visits per week. It is particularly striking that subjects substantially over-predict gym

[22] We group all non-zero values of p-coupon together here for simplicity — the effect of each separate p-coupon value is investigated in Table 1.3.

attendance in weeks with no p-coupon, suggesting that the overprediction is not driven by the p-coupon incentives. On the basis of these results we can rule out, in our model, both time consistency ($\beta = 1$) and full sophistication ($\widehat{\beta} = \beta$) if, after the treatment, subjects have rational expectations over their future costs.

Table 1.2: Misprediction of attendance

	Control group			Treatment group		
	Bid $p > 0$	Pred $p > 0$	Pred $p = 0$	Bid $p > 0$	Pred $p > 0$	Pred $p = 0$
Pre-Treatment Predictions						
Predicted attendance	3.868	4.053	1.418	3.63	3.963	1.231
Actual attendance	1.561	1.561	0.255	1.463	1.463	0.365
Difference	2.307	2.491	1.164	2.167	2.500	0.865
St. Error	(0.297)	(0.235)	(0.149)	(0.350)	(0.318)	(0.178)
No. of observations	57	57	55	54	54	52
Post-Treatment Predictions						
Predicted attendance	3.395	3.614	1.058	3.185	3.056	1.313
Actual attendance	1.561	1.561	0.269	1.463	1.463	0.396
Difference	1.833	2.053	0.788	1.722	1.593	0.917
St. Error	(0.321)	(0.299)	(0.144)	(0.315)	(0.299)	(0.171)
No. of observations	57	57	52	54	54	48

Notes: Bid includes only observations for a subject's incentivized week. Pred includes both this week and the unincentivized week for which subjects were asked to make predictions without a p-coupon.

In Table 1.3 we explore the effect of p-coupon value, and the change in predictions over time. The first column regresses actual attendance on dummies for the various values of p-coupon.[23] The point estimates on the p-value dummies indicate a nearly monotonic effect of monetary incentives, and pairwise comparisons of the coefficients do not reject monotonicity. This is reassuring, as it suggests an upward-sloping labor supply curve, as we would expect. The second and third columns regress bids and unincentivized predictions on the same p-coupon dummies, plus a dummy for the post-treatment elicitation session. Subjects appear to predict the slope of their labor-supply curve relatively accurately, despite consistently over-predicting its intercept.

The extent of over-prediction drops for both groups between the first and second elicitation sessions. The session dummy implies that subjects reduce their predictions by roughly two-thirds of a visit per week. These sessions differ in two ways: they are a month apart in time, and the second session is closer to the target weeks than the first. One possibility is that subjects' discount factors decrease smoothly over time rather than abruptly as in the beta–delta model. If so, we would see a change in mispredictions merely because the temporal proximity of the target weeks is greater in the post-treatment elicitation session.

[23] The omitted category is $p = \$7$ throughout this table. This is so that we can compare coefficients across 'Actual' and 'Pred' (for each of which the lowest value is $p = \$0$), and 'Bid' (where the lowest value is $p = \$1$). In addition, all specifications in this table include individual covariates.

Table 1.3: Predictions: Delay versus Session Effects

	(1) Actual	(2) Bid	(3) Pred	(4) Bid	(5) Pred
Session[a]		-0.630***	-0.707***	-0.476**	-0.810***
		(0.132)	(0.112)	(0.226)	(0.187)
p=$0	-2.275***		-3.360***		-3.925***
	(0.611)		(0.498)		(0.598)
p=$1	-1.669**	-0.924	-1.650***	-0.512	-1.618**
	(0.689)	(0.581)	(0.482)	(1.235)	(0.640)
p=$2	-1.304*	-0.760	-1.288***	-1.522	-2.213***
	(0.708)	(0.579)	(0.478)	(1.232)	(0.617)
p=$3	-1.440**	-0.530	-0.924*	-0.489	-1.276**
	(0.714)	(0.580)	(0.472)	(1.233)	(0.634)
p=$5	-0.050	-0.081	-0.272	0.027	-0.698
	(0.808)	(0.623)	(0.523)	(1.241)	(0.648)
Constant	2.600***	3.865***	4.953***	3.988***	5.405***
	(0.609)	(0.613)	(0.497)	(1.233)	(0.590)
Observations	551	875	1088	176	217
R-squared	0.20	0.06	0.27	0.11	0.33
Num Clusters:	111	111	111	110	111
Sample	Full	Full	Full	5-wk delay	5-wk delay

Notes: [a]Pre=0, Post=1. Robust standard errors in parentheses, clustered by individual. * significant at 10%; ** significant at 5%; *** significant at 1%. p = $7 is the omitted category.

We can examine this by comparing first-session predictions for the first target week with second-session predictions for the fifth target week. This comparison holds temporal proximity constant. Columns (4) and (5) report the results of this regression. The coefficients on the session dummy for both bids and unincentivized predictions still show a substantial decrease in over-prediction over time. Apparently something neither we nor the subjects foresaw is happening between the second and sixth weeks of the semester that is causing subjects to lower their predictions of future gym attendance by half to two-thirds of a visit per week. This suggests that there is systematic misprediction along more than one dimension of the gym-attendance decision. One possibility is that subjects begin the semester with overly optimistic beliefs about their amount of free time in the semester, and become more realistic as the semester unfolds.[24]

1.4.3 Structural estimation

Lastly, we estimate two key welfare parameters of the model: the value of the habit, η; and the cost of naivete, $(\widehat{\beta} - \beta)b$. These are identified by a parsimonious system of two equalites described in Section 1.2.2, which we now re-express in terms of regression equation coefficients. Because we varied P in discrete increments, in order to find the precise values of P necessary to estimate our parameters we assume that both unincentivized predictions and attendance are linear in P.[25] Using a seemingly unrelated regressions (SUR) model, we simultaneously estimate

$$\text{ACT}^i_{t,p_{i,t}} = \gamma_{00} + \gamma_{01} \cdot T_i + \gamma_{02} \cdot T_i \cdot p_{i,t} + \gamma_{03} \cdot p_{i,t} \tag{1.7}$$

$$\text{PRED}^i_{t,p_{i,t}} = \gamma_{20} + \gamma_{21} \cdot T_i + \gamma_{22} \cdot T_i \cdot p_{i,t} + \gamma_{23} \cdot p_{i,t}, \tag{1.8}$$

where $\text{ACT}^i_{t,p_{i,t}}$ is the actual attendance of subject i in week t of the immediate post-treatment period, and $\text{PRED}^i_{t,p_{i,t}}$ is subject i's post-treatment, unincentivized prediction of attendance in week t of the same period. T_i is a dummy for whether subject i is in the treated group and $p_{i,t}$ is the value of the p-coupon held by subject i in week t.

To estimate η, we look for P^* such that control subjects holding a $\$P^*$ coupon attend the gym as much as unincentivized treatment subjects. We can now re-express these group means in terms of regression coefficients:

$$\overline{\text{ACT}}^T_{t,0} = \gamma_{00} + \gamma_{01} = \gamma_{00} + \gamma_{03} \cdot P^* = \overline{\text{ACT}}^C_{t,p^*} \tag{1.9}$$

Solving for P^*, and hence for η, we get $\eta = P^* = \gamma_{00}/\gamma_{03}$.

To estimate $(\widehat{\beta} - \beta)b$ we want P^* such that control subjects holding a $\$\widetilde{P}$ coupon

[24] See, e.g. Bénabou and Tirole (2002) for why subjects may begin the semester with overly optimistic beliefs.

[25] We have explored adding curvature to these relationships. It does not change our results significantly. We report the linear approach for tractability.

predict the level of attendance actually achieved by a $\$P^*$ coupon:[26]

$$\overline{\text{PRED}}^C_{t,\widetilde{p}} = \gamma_{20} + \gamma_{23} \cdot \widetilde{P} = \gamma_{00} + \gamma_{03} \cdot P^* = \overline{\text{ACT}}^C_{t,p^*}. \tag{1.10}$$

To implement this we substitute $\overline{P}$, the average value of P in the control group, for $\widetilde{P}$. Solving this for $P^* - \overline{P}$, and hence for $(\widehat{\beta} - \beta)b$, we get $(\widehat{\beta} - \beta)b = P^* - \overline{P} = [\gamma_{20} - \gamma_{00} + (\gamma_{23} - \gamma_{03})\overline{P}]/\gamma_{03}$

Table 2.2 shows the results of the two-equation SUR system, and, beneath these, the estimates of structural parameters of interest. The left-hand panel shows the results when we include the entire treated group. The right-hand panel restricts the sample to include only those treated subjects whose attendance increased by at least one visit per week, our so-called habit formers.

Table 1.4: Parameter Estimation

	All Subjects		Controls and Habit-Formers	
	(1)	(2)	(3)	(4)
	ACT	PRED	ACT	PRED
SUR Results				
Treatment Group	0.180*	0.062	2.020***	1.155**
	(0.106)	(0.245)	(0.205)	(0.499)
Treated X $P	-0.138**	-0.173**	-0.128	0.013
	(0.066)	(0.084)	(0.245)	(0.176)
$P	0.447***	0.558***	0.448***	0.565***
	(0.047)	(0.058)	(0.045)	(0.059)
Constant	0.259***	1.684***	0.258***	1.670***
	(0.074)	(0.170)	(0.071)	(0.170)
Observations	545	545	320	320
Parameter Estimates				
Habit Value	0.403*		4.505***	
	(0.230)		(0.603)	
Cost of Naivete	3.913***		3.906***	
	(0.694)		(0.688)	

Standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

Our estimate of the "cost of naivete" is $3.91. This is the portion of the future benefit of a single gym visit that present bias will cause subjects to forego, and that naivete will cause them to think they will not forego. Put another way, it is the difference, on average, between the dollar value a fully sophisticated subject would put on a 100% effective gym-attendance commitment device, and the dollar value our subjects would put on such a device. It is important to note that this estimate of foregone future benefit does not depend

[26]Note that we are using post-treatment unincentivized predictions, which, given our results in section 1.4.2, we assume are based on correct beliefs about target-week costs. Our model equally allows us to use pre-treatment unincentivized predictions but using post-treatment predictions gives us a more conservative result.

upon any assumptions about the long-term benefits of gym attendance, but is based entirely on subjects' own evaluation of the long-term benefits. Our estimate of the dollar value of the habit-formation effect among the treated group is \$0.40, suggesting that the \$100 per subject treatment incentive increased average gym-attendance utility by the monetary equivalent of forty cents per visit. While this average effect informs the overall cost-effectiveness of the intervention, it masks the heterogeneity of the treatment we observed in Section 1.4.1. If we inflate the habit-value estimate in the full sample by the inverse of the proportion of habit-formers in the treatment group we get a back-of-the-envelope estimate of \$3.11 for the habit value among habit formers.

To address habit-formation heterogeneity in a different way the right-hand panel of Table 2.2 confines the analysis to just those treated subjects identified as habit-formers and estimates the value of their habit. Among treated subjects whose immediate post-treatment attendance increased by at least one visit per week, we find a habit value of \$4.51, much larger than the average for the entire treated group[27], while the cost of naivete remains roughly unchanged. These results depend on the assumption that, after controlling for observables, those in the control group who would have formed a habit respond in the same way to a p-coupon as those who would not have formed a habit. In appendix G we explore the differences in covariates between the habit-formers and non habit-formers in the treatment group, and we are reassured by the fact that their observed behavior responds identically to p-coupons.[28] The only covariate on which they differ significantly is self-reported importance of physical fitness, which is higher among habit-formers. This might help to explain why they formed a habit. But it is hard to see how it would affect their response to the p-coupons, suggesting that this difference may not be a problem for our estimation strategy. However, because we are comparing the habit-formers against all control subjects—rather than only those who would have formed habits had they been treated—these columns should not be treated with the same confidence as our other results.

1.5 Conclusion

We find that incentivizing gym-attendance creates a short-run habit that is smaller than, but statistically indistinguishable from, Charness and Gneezy's (2009) effect, and which decays substantially as the result of an exogenous break in attendance. Although Charness and Gneezy find, at most, very slow decay, a model that incorporates short-term shocks to the cost of gym attendance can rationalize both their findings and ours. Our findings can be explained by the four-week common shock of winter break, while a much slower path of decay would result from a series of smaller, independent shocks over a longer period of time.[29]

Furthermore, we find that subjects have self-control problems of the sort generated by present bias, and that they are at least partially naive with respect to these self-control

[27] But similar to inflating the aggregate habit-value by the inverse of the compliance rate.

[28] In a regression comparing p-responsiveness between habit-formers and non habit-formers (not shown), the coefficients on the p-coupon value differ only by a statistically insignificant -0.039. It is not clear why this comparison should be different between the comparable subjects in the control group.

[29] It seems reasonable that a habit that can be induced by a positive four-week shock can be eliminated by a negative four-week shock.

problems. Even in weeks with no p-coupon to complicate the prediction task, subjects over-predict attendance by about one visit per week—a factor of about three. This is a sufficient degree of mis-prediction to explain the result in DellaVigna and Malmendier (2006) that people purchase monthly health club memberships when their actual attendance only justifies the purchase of single-visit passes.[30]

Because they may be partially, rather than completely, naive about their future self-control problems, we cannot take their predictions as statements of their true preferences, and thus we cannot estimate the full cost of their self-contol problems. However, we are able to estimate the portion of foregone future benefits that they fail to predict—approximately \$4—which serves as a lower bound on the foregone future benefits, and hence on the total future benefits. We also find that for subjects who form a habit, the habit-formation effect almost exactly offsets this cost of naivete. In a population of "procrastinators" who initially believe that, in expectation, they will attend the gym in the future but do not attend in the current period, this term is also the minimum increase in gym-attendance utility necessary to induce attendance (in expectation).

In addition to these results on naive self-control problems, we are able to rule out that the decrease in over-prediction over the course of the treatment month is caused by the increased temporal proximity of outcomes, as would be predicted by a model of true hyperbolic discounting—as opposed to the quasi-hyperbolic discounting captured by the beta–delta model. Instead it appears that subjects' predictions may become more accurate because they are learning something about the distribution of gym-attendance costs as the semester unfolds. We interpret this as being consistent with the literature on overoptimism, but do not propose a specific explanation. Our data also allow us to explore whether subjects predict the habit-formation effect itself, but we do not have the statistical power to effectively answer this question yet.

We found an average habit-formation effect among treated subjects (who complied with the protocol) of approximately one-third of a visit per week, though this effect is heavily concentrated in the upper tail of the distribution. From the standpoint of public policy it is this local average treatment effect that matters because non-compliers do not incur the cost of the treatment incentive. We estimate the unforeseen portion of long-term benefits that treated subjects' self-control problems cause them to forego at roughly four dollars. The overall long-term benefits, therefore, must be at least this much. Adding to this approximately \$0.50 for the average habit value among compliers, we can establish a rough upper bound of sixty-nine weeks on how long the habit would have to persist in order to break even on the cost of the incentive.[31] If the incentive could have been targeted to those we identified as forming a habit, the break-even decay horizon would be just forty-six weeks. In our sample of students, however, we see significant decay after winter break, suggesting that exogenous interruptions in attendance may undermine the intervention. One must also exercise caution in extrapolating these results to other populations, where compliance,

[30] DellaVigna and Malmendier (2006) consider a very different population, of course, so we do not claim that this is driving their result.

[31] This is an upper bound because we do not know the true long-term benefit, which may be substantially higher than just the portion foregone due to self-control problems. We simply divide the expected cost of the intervention (\$100 multiplied by the 80% compliance rate) by the weekly benefits (\$4.50 multiplied by the 0.256 visits/week treatment effect).

habit formation, and habit decay might all be quite different.

Our design also allows us to address the source of gym attendance motivation. Gneezy and Rustichini (2000) argue that introducing small financial incentives may, counterintuitively, reduce a behavior by crowding out intrinsic motivation. We find no evidence that this is the case for gym attendance, either for our main treatment intervention or for our smaller post-treatment incentives. We find that a temporary subsidy increases attendance both while it is in place and in the short run after its removal. We also find that both treated and control subjects respond positively to the incentives provided by our p-coupons. A direct comparison of average attendance during coupon weeks and zero weeks among the treated group strongly rejects the null that unincentivized attendance is higher ($p = 0.0004$). Moreover, we cannot reject that attendance is monotonically increasing in p-coupon value.[32] While intrinsic motivation may still be reduced by our financial incentives, it does not appear to be of first-order significance for our results.

Future research should explore the habit-formation and habit-decay effects in a more policy-relevant population. Subjects might be selected on the basis of health risks such as obesity, and efforts could be made to select true procrastinators. In addition, effort should be made to try to identify the ex-ante determinants of habit formation so that incentives can be more effectively targeted. For example, we find that treatment subjects who ultimately developed a habit had initially expressed stronger beliefs that fitness was important, despite no difference in initial gym attendance. The issue of subjects' predictions also warrants further study, including the critical issue of predicting the habit-formation effect, for which a larger sample is necessary.

[32] That is, for no pair of adjacent coupon values is attendance for recipients of the smaller coupon statistically greater than attendance for recipients of the larger. We do not reject monotonicity in either the full sample or within either experimental group.

Chapter 2

Habit-Formation and Projection-Bias in Gym Attendance

2.1 Introduction

Individuals routinely make decisions that involve predictions about how their preferences, costs, and beliefs will unfold in the future. It is commonly assumed that individuals have rational expectations, which is to say that while exact preferences, costs, and beliefs may not be known, people know the range of possibilities and make accurate predictions based on averages. If, however, people's predictions are wrong then their decisions may fall short of long-run optimality.

Habit-formation is one dimension of future preferences along which misprediction may occur, and for which the welfare costs may by particularly large. If, for example, prospective smokers fail to predict how hooked they will get, they may start smoking at an early age and wind up losing several "quality adjusted life years" worth of utility over the course of a lifetime.[1] But equally, if people do not foresee the way that healthy behaviors can become more enjoyable after a period of habit-formation, they may miss out on a lifetime of health benefits. Becker and Murphy (1988), in their famous "Theory of Rational Addiction", salvage rationality—and hence the welfare theorems—by modeling addicts as perfectly forward-looking with respect to the habit-forming effects of current and future consumption. Loewenstein, O'Donoghue and Rabin (2003) explicitly demonstrate the importance of prediction of preferences for Becker and Murphy's results, and show how misprediction of habit-formation can lead to long-term welfare losses. In particular, they model a form of misprediction, for which they claim support in the psychology literature, in which individuals correctly foresee the direction in which their preferences will change, but underappreciate the magnitude of change. They refer to this kind of misprediction as "projection bias" because people are thought of as projecting their current preferences (in

[1] For example, Gruber (2001) finds that teenage smokers dramatically over-state the probability of quitting within five years, and that heavy teen smokers who believe they will quit are actually *less* likely to quit than those who believe they will not.

this case their current level of habituation) onto their future selves.

One domain in which this kind of misprediction may be important is physical exercise. There is a broad consensus in the health sciences that physical exercise has important physical and psychological health benefits. It is also widely believed in behavioral health that habit-formation plays an important role in physical exercise. Our question as economists is whether, as the theory of rational addiction assumes, people accurately predict the habit-formation process, subject to uncertainty, or whether as the projection-bias model assumes, they systematically mispredict and may thus make suboptimal physical-exercise choices.

In a recent paper, Charness and Gneezy (2009) paid subjects to attend the gym for several weeks and found that they had significantly higher gym attendance than other subjects in the period after the payment ended, suggesting that being paid to attend for a while had led to habit formation. Their subjects were university undergrads who were randomized into three groups. A "low- incentive" group were offered $25 to attend the gym once during the initial week of the study. A "high-incentive" group received the same $25 offer, and were additionally offered $100 to attend the gym another eight times in the subsequent four weeks for a total of nine visits over five weeks. A third group, which received no offers for gym attendance, served as a control group. Gym-attendance data was collected for all students for a period beginning eight weeks before the treatment and ending seven weeks after.[2] By comparing the change in attendance from pre-treatment to post-treatment across groups they are able to show that subjects in the high-incentive group continue to attend the gym significantly more after the incentive period ends than subjects in the other two groups. (An average of 0.67 visits per week more than the control group, and 0.58 visits per week more than the low-incentive group.) They found that the effect was heterogeneous, with most of the increase concentrated in a subset of subjects. Identifying these individuals, they found that they were more likely to be people who had attended less than once per week on average during the pre-treatment period, so-called non-regular attenders.

To test for misprediction of future gym preferences we reran Charness and Gneezy's high-incentive and low-incentive treatments, but with a twist. In addition to the $25 and $100 attendance incentives, we elicited subjects' predictions of their *post*-treatment gym-attendance, conducting the elicitation both immediately before and immediately after the treatment period. If subjects who are paid $100 to attend the gym for a month fail to foresee the way this period of paid gym attendance will change their preferences, then the difference between their pre- and post-treatment predictions should be more positive (or less negative) than for subjects who are paid only the $25 to attend once. Like Charness and Gneezy, we find that subjects who received, and responded to, the $100 incentive do attend the gym more often in the post-treatment period than control subjects, and like them we find heterogeneity in the effect.[3] Furthermore we find that subjects who form a habit do foresee the habit-formation effect, but do *not* correctly predict the magnitude of the increase in their gym-attendance, while subjects who do not form a habit seem to

[2] We are describing Charness and Gneezy's first study. In the same paper they conducted a second study with a slightly different design that yielded largely similar results.

[3] We present these habit-formation results, and explore basic issues of attendance prediction in a separate paper that appears as chapter one of this dissertation.

accurately foresee the lack of habit-formation. We interpret these results as supporting the model of projection bias and discuss how we can distinguish the projection bias model from a rational-expectations model with random habit-value heterogeneity. We estimate the parameters of a structural model of habit formation and projection bias, accounting for the heterogeneity in the habit-formation effect and find that habit formers receive a habit-value of approximately $4 and foresee about two-thirds of it.

The remainder of this paper is organized as follows. Section two presents a simple model of habit formation which nests the rational-addiction model within the projection-bias framework. In section three we describe the experimental design, and in section four we present our results. Section five concludes.

2.2 Model

In this section we develop a simple model of gym attendance that incorporates habit formation, projection bias, and present-biased preferences. Following Becker and Murphy (1988) and O'Donoghue and Rabin (1999a), habituation—resulting from past gym attendance—will be modeled as a binary state variable. The habit-formation effect of being in the habituated state will be modeled as the result of a fixed, additive increase in gym-attendance utility.[4] To explicitly address heterogeneity in habit-formation we will allow the habit-formation effect to vary among habituated individuals. Following Loewenstein et al. (2003), individuals will correctly foresee the direction of this habit-formation process, but may partially or fully "project" their current level of habit onto their future selves. Individuals will discount all future periods relative to the present, à la Phelps and Pollak (1968) and Laibson (1997), and will be naive or sophisticated with respect to this "quasi-hyperbollic discounting", à la O'Donoghue and Rabin (1999b).

In the spirit of DellaVigna and Malmendier (2004), we consider a finite-horizon, discrete-time model with five unequal periods. Initially all subjects are non-habituated, and are randomly assinged into two groups, one of which will be incentivized to attend the gym in period one (treated group), and the other of which will not (control group). In the first period subjects bid, in an incentive compatible auction, on a "p-coupon" that rewards fourth-period gym attendance.[5] Then, still in the first period, treated subjects attend the gym and enter the habituated state, which will persist through all subsequent periods.

In the second period two things happen. First subjects once again bid on the fourth-period attendance-reward coupon that they bid on in the first period. Then, after the auction, all subjects are given a p-coupon.[6] In periods three and four, subjects attend or don't attend the gym according to their preferences, with the only difference between

[4] In Becker and Murphy (1988) the habit-formation effect of being habituated is the effect of increased marginal utility of consumption caused by past consumption. In their model "positive" and "negative" habits are defined by whether past consumption leads to an increase or decrease in total utility. By this definition we model gym-attendance as a neutral habit.

[5] We refer to period four as the "target-week" as it is the target of the p-coupon.

[6] In the model we are ignoring the fact that the elicitation process requires one or two subjects to wind up with two coupons. In practice, because there were multiple target weeks, most of the auction winners did not end up holding multiple p-coupons for the same week. The two subjects who did wind up with two p-coupons for the same target week simply received double the reward.

these periods being that in period four they receive p-coupon rewards for attendance. We explicitly think of periods three and four as weeks, so that subjects decide each day whether to attend the gym that day. Finally, in period five subjects receive the delayed benefit of whatever gym attendance they have engaged in.

The goal of the model is to develop expressions for expected gym attendance, and for valuations of p-coupons. Let the immediate utility of gym attendance on day d be $-c+\varepsilon_d$ with $c>0$, $\varepsilon_d \sim F$i.i.d., and let the delayed benefit of gym attendance be $b>0$. Thus we model gym attendance as an "investment good" in the language of DellaVigna and Malmendier, meaning that costs are immediate while rewards are delayed. Future payoffs are discounted by β , with beliefs about future self-control denoted by $\widehat{\beta}$. Following O'Donoghue and Rabin (1999a), habituation will take a simple binary form. When subjects are habituated they receive additional, immediate utility for gym attendance of $\eta_i \geq 0$, so that the immediate utility of gym attendance for a habituated subject is $\eta_i - c + \varepsilon_d$. To capture habit-formation heterogeneity parsimoniously, the habit value, η_i, will take one of two values. With probability $\pi, \eta_i = \overline{\eta}$ strictly greater than zero, and with probability $1-\pi, \eta_i = 0$. Subjects have "simple projection bias" as defined by Loewenstein et al. (2003), using $\alpha \in [0,1]$ to index the strength of the bias. That is, when considering future consumption decisions, subjects believe that their future utility function will be an alpha-mixture of their current and future utility functions, with a weight of α on the current utility function and $1-\alpha$ on the future utility function. Thus $\alpha = 0$ refers to the case of no projection bias, in which subjects correctly foresee the actual future instantaneous utility function, and $\alpha = 1$ refers to the case of full projection bias, in which subjects believe that their instantaneous utility function will not change with their state of habituation. We model utility as quasi-linear in money. Without loss of generality, utility from all non-gym sources will be normalized to zero.

We define alpha-sophisticates and alpha-naifs as subjects with $\alpha = 0$ and $\alpha = 1$ respectively, and beta-sophisticates and beta-naifs as subjects with $\widehat{\beta} = \beta$ and $\widehat{\beta} = 1$ respectively, and we can then think in terms of partial naivete with respect to either α or β. In other words, an alpha-sophisticate is a subject with correct beliefs about future habit formation and a beta-sophisticate is a subject with correct beliefs about self-control, etc.

Let P be the face value of the p-coupon that rewards gym attendance in period four. Thus a p-coupon immediately pays $\$P$ for each day that the holder attends the gym in period 4. Let X_t^g refer to the valuation of a p-coupon in period $t = 1, 2$ of a subject in group $g = 0, 1$ (control=0, treated=1). Let $Z_d^g = 0, 1$ be an indicator for whether a subject in group g actually attends the gym on day $d = 1, ..., 7$ of the target week, so that $Z^g = \sum_{d=1}^{7} Z_d^g$ is the number of gym visits during the target week for a subject in group g.

2.2.1 Attendance decision and the value of a p-coupon.

If a subject attends the gym on a given day during the target week her utility for that day will be $P + \beta b + g\eta_i - c + \varepsilon_d$. She will attend the gym if this is greater than zero. Thus $Z_d^g = \mathbb{1} \cdot \{\varepsilon_d > P + \beta b + g\eta_i - c\}$, and $Z^g = \sum_{d=1}^{7} \mathbb{1} \cdot \{\varepsilon_d > P + \beta b + g\eta_i - c\}$. In

expectation, total target-week gym-attendance will be,

$$\sum_{d=1}^{7} \Pr(Z_d^g = 1) = 7 \times \int_{c-\beta b-g\eta_i-P}^{\infty} dF(\varepsilon) \tag{2.1}$$

and the habit-formation effect, the increase in attendance caused by habituation, will be,

$$\sum_{d=1}^{7} \Pr(Z_d^g = 1) = 7 \times \int_{c-\beta b-g\eta_i-P}^{c-\beta b-P} dF(\varepsilon). \tag{2.2}$$

However, from the perspective of any previous period, the perceived probability of target-week gym-attendance depends upon the subject's belief about future self-control, $\widehat{\beta}$ and on her projection bias parameter, α. She believes she will attend on any given day of the target week if $\varepsilon_d > P + \widehat{\beta} b + g(1-\alpha)\eta_i - c$. Thus the subject's ex-ante prediction of her total utility for the target-week, given that she holds a p-coupon, is,

$$7 \times \int_{c-\widehat{\beta} b-g(1-\alpha)\eta_i-P}^{\infty} (P + b + g(1-\alpha)\eta_i - c + \varepsilon)\, dF(\varepsilon). \tag{2.3}$$

Setting $P = 0$ gives us the predicted utility without a p-coupon. The value of the p-coupon is the difference between expected utility with a p-coupon and expected utility without a p-coupon. In period one this is

$$X_1^g = \left[7 \times \int_{c-\widehat{\beta} b-g(1-\alpha)\eta_i-P}^{\infty} P\, dF(\varepsilon) \right] + \left[7 \times \int_{c-\widehat{\beta} b-g(1-\alpha)\eta_i-P}^{c-\widehat{\beta} b-g(1-\alpha)\eta_i} (b + g(1-\alpha)\eta_i - c + \varepsilon)\, dF(\varepsilon) \right]. \tag{2.4}$$

And in period two, when the full habit-formation effect is known to the subject, it is

$$X_2^g = \left[7 \times \int_{c-\widehat{\beta} b-g\eta_i-P}^{\infty} P\, dF(\varepsilon) \right] + \left[7 \times \int_{c-\widehat{\beta} b-g\eta_i-P}^{c-\widehat{\beta} b-g\eta_i} (b + g\eta_i - c + \varepsilon)\, dF(\varepsilon) \right]. \tag{2.5}$$

Note that present-bias does not change these valuations between pre- and post-treatment elicitations because the target week is in the future (hence "inside β") from the perspective of either elicitation period.

The first term in both (2.4) and (2.5) is the expected redemption value of the coupon, which is always weakly positive. The second term is the subject's valuation of the behavioral change that results from holding the coupon, which we will call the incentive value. This is the change in utility caused by those gym-visits that the subject would not have made in the absence of the p-coupon. The sign depends on the subject's ex-ante belief

about future self-control problems. If the subject believes that she will not have self-control problems in the target week then the incentive value is negative because the subject believes that the p-coupon will make her attend the gym at times when she would ex-ante prefer not to. If the subject believes that she will have self-control problems in the target week then the incentive value may be positive because she foresees that the p-coupon will make her more likely to attend the gym and gain a long-term benefit that she would otherwise forego due to self-control problems.[7]

For unhabituated subjects, which is to say control subjects—$g = 0$—and for treated subjects with zero habit value—$\eta_i = 0$—the terms $g(1-\alpha)\eta_i$ and $g\eta_i$, the anticipated and actual habit value, are both zero, so we get $X_1^g = X_2^g$. For treated subjects who form a habit the anticipated habit value is $(1-\alpha)\overline{\eta}$ and the actual habit value is simply $\overline{\eta}$.

The total ex-ante value of the p-coupon is always non-negative. This seems intuitively obvious because the p-coupon is worth money and it helps you to get to the gym, but this intuition is not correct because the incentive value may be negative. The correct intuition is that even with negative incentive value, an individual holding a p-coupon won't go to the gym if the disutility of doing so is greater than the redemption value of the coupon. We prove this in appendix A. In general, the value of any reward or benefit contingent upon gym attendance will be weakly positive, for the same reason.

2.2.2 Reduced-form test for projection bias

Our test of projection bias is simply to compare the average difference in p-coupon valuations from pre- to post-treatment elicitation between treated subjects and control subjects. That is to say, $[\overline{X}_2^1 - \overline{X}_1^1] - [\overline{X}_2^0 - \overline{X}_1^0]$ where the upper-bar denotes a group average. Since $[\overline{X}_2^0 - \overline{X}_1^0] = 0$ the double difference is actually just $[\overline{X}_2^1 - \overline{X}_1^1]$. In the absence of projection bias this difference should be zero. With projection-bias it may be positive or negative depending on the shape of $F(\cdot)$. Consider, dividing by $7 \cdot \pi$ for ease of exposition,[8]

$$\frac{[\overline{X}_2^T - \overline{X}_1^T]}{7 \cdot \pi} = \int_{c-\widehat{\beta}b-\eta-P}^{c-\widehat{\beta}b-(1-\alpha)\eta-P} P\, dF(\varepsilon) \qquad + \tag{2.6}$$

$$\int_{c-\widehat{\beta}b-\eta-P}^{c-\widehat{\beta}b-\eta} (b+\eta-c+\varepsilon)\, dF(\varepsilon) - \int_{c-\widehat{\beta}b-(1-\alpha)\eta-P}^{c-\widehat{\beta}b-(1-\alpha)\eta} (b+(1-\alpha)\eta-c+\varepsilon)\, dF(\varepsilon) \tag{2.7}$$

[7] Thus, for a sophisticate with self-control problems the incentive value can be thought of as "commitment value" because it is the value of having the p-coupon as a "commitment device" to help her get out the door and down to the gym.

[8] Note that for treated subjects the group average is a π mixture of the average for subjects who form a habit and the average for those who don't. And since the average change in valuation of a p-coupon for subjects who do not form a habit is zero, the group average is simply π times the average for habit-formers.

The term in (2.6) is the effect that we are trying to identify, which is the misprediction of gym attendance caused by projection bias. It is weakly positive for alpha-naifs, and zero for alpha-sophisticates, regardless of beliefs about self-control. The difference in (2.7) is the difference in perceived incentive value from before the treatment to after. If $\widehat{\beta}$ is sufficiently close to 1 then the incentive values will both be negative. Conversely, if the subject is sufficiently beta-sophisticated, then the incentive values may both be positive. However, for any given value of $\widehat{\beta}$, the difference in incentive values depends exclusively on the distribution of ε_d. For example, consider a beta-sophisticate, for whom there is positive incentive value for the p-coupon. It could be that before habit formation the subject is just indifferent between going to the gym and staying home, so that the p-coupon has a strong effect, and thus a large incentive value, but that after habit-formation the subject always goes to the gym so the p-coupon no longer has any incentive value. Conversely, it could be that before habit-formation the subject really hated going to the gym and the p-coupon was never enough to get her out the door, but after habit-formation she is on the fence between going and staying home so the p-coupon has a strong incentive value. A similar pair of stories could be told for a beta-naive or time-consistent subject.

Regardless of how the incentive value changes over time, we can still say something definite about projection bias. That is because in the absence of projection bias our test-statistic is always zero. To see this, note that for $\alpha = 0$ both (2.6) and (2.7) collapse to zero. Thus, theoretically, any observed value of the double-difference that is significantly different from zero indicates projection bias.

It is worth noting that β does not appear in this double-difference expression. That is because both the pre- and post-treatment elicitations take place prior to the target week, and observed gym-attendance is not used in the test. We could have designed a test based on the difference in pre-treatment overprediction between treated and control groups, but that test would have been noisier because it would have included a component of misprediction of self-control which we have eliminated in our test.

2.2.3 Structural Estimation

There are three terms that we are interested in estimating. The first is the habit-formation effect itself, η, which is the additional, per-visit, gym-attendance utility (measured in dollars) received by a subject in the habituated state. Another way to think of this parameter, and the key to our estimation strategy, is that η is the per-visit monetary incentive that would cause a non-habituated subject to attend as often as an unincentivized habituted subject. The next term of interest is the portion of the habit-formation effect that subjects foresee, or predict, $(1 - \alpha)\eta$. For subjects with no projection bias, that is $\alpha = 0$, this term is equal to η, because without projection bias subjects foresee the entire habit-formation effect, and vice versa for $\alpha = 1$, the case of complete projection bias. Finally, we are interested in α itself, the projection-bias parameter, which tells us the weight subjects erroneously place on their current preferences when considering choices that will be determined by their future preferences.

Let $\overline{Z}_p^g$ be the average weekly attendance of subjects in group $g \in \{T, C\}$ who are holding a p-coupon, and $\overline{Z}_0^g$ be the same thing for subjects with no p-coupon (i.e. $P = 0$). Let $\overline{Y}_{t,p}^g$ be the average unincentivized prediction, in elicitation session $t \in \{1, 2\}$, of gym

attendance during a target week with a p-coupon of subjects in group g.

Identifying η

Our estimation strategy is essentially equivalent to finding the value of P for which the average target-week attendance in the control group, with a p-coupon, is the same as the average target-week attendance in the treated group, without a p-coupon. In terms of our model, we are looking for P^* such that,

$$\overline{Z}_0^T = 7 \times \int_{c-\beta b-\eta}^{\infty} dF(\varepsilon) = 7 \times \int_{c-\beta b-P^*}^{\infty} dF(\varepsilon) = \overline{Z}_p^C. \tag{2.8}$$

Once we know the value of P^*, because $F(\cdot)$ is monotonically increasing, we then have $\eta = P^*$.

Identifying $(1-\alpha)\eta$ and α

We first need to identify $\left(\widehat{\beta}-\beta\right)b$ as a building block.[9] The average *post*-treatment prediction of gym attendance in a target week with a p-coupon with a face value of $\widetilde{P}$, among control subjects, is

$$\overline{Y}_{2,p}^C = 7 \times \int_{c-\widehat{\beta}b-\widetilde{P}}^{\infty} dF(\varepsilon) = 7 \times \int_{c-\beta b-\left(\widehat{\beta}b-\beta b\right)-\widetilde{P}}^{\infty} dF(\varepsilon). \tag{2.9}$$

We find the value of P^* for which

$$\overline{Y}_{2,p}^C = 7 \times \int_{c-\beta b-\left(\widehat{\beta}b-\beta b\right)-\widetilde{P}}^{\infty} dF(\varepsilon) = 7 \times \int_{c-\beta b-P^*}^{\infty} dF(\varepsilon) = \overline{Z}_p^C, \tag{2.10}$$

which gives us $(\widehat{\beta}-\beta)b = P^* - \widetilde{P}$. In practice we will evaluate this by setting $\widetilde{P}$ equal to the average value of P among all control subjects.

Next we consider the average *pre*-treatment prediction of gym attendance in a target week with a p-coupon with face value of $\widetilde{P}$, among treatment subjects, which is

$$\overline{Y}_{1,p}^T = 7 \times \int_{c-\widehat{\beta}b-(1-\alpha)\eta-\widetilde{P}}^{\infty} dF(\varepsilon) = 7 \times \int_{c-\beta b-\left(\widehat{\beta}b-\beta b\right)-(1-\alpha)\eta-\widetilde{P}}^{\infty} dF(\varepsilon) \tag{2.11}$$

and once again we find the value of P^* in the control group for which

$$\overline{Y}_{1,p}^T = 7 \times \int_{c-\beta b-\left(\widehat{\beta}b-\beta b\right)-(1-\alpha)\eta-\widetilde{P}}^{\infty} dF(\varepsilon) = 7 \times \int_{c-\beta b-P^*}^{\infty} dF(\varepsilon) = \overline{Z}_p^C, \tag{2.12}$$

[9] We discuss this parameter in detail in a separate paper which appears as chapter one of this dissertation.

which gives us $(1-\alpha)\eta = P^* - \widetilde{P} - \left(\widehat{\beta} - \beta\right) b$ and $\alpha = 1 - \frac{P^* - \widetilde{P} - \left(\widehat{\beta} - \beta\right) b}{\eta}$. And again, in practice we will estimate this by replacing $\widetilde{P}$ with the average value of P among treatment subjects.

2.3 Design

We recruited one hundred and twenty subjects from the students and staff of UC Berkeley and randomly assigned them to treated and control groups.[10] Since Charness and Gneezy found the habit-formation effect concentrated among non-attenders we screened for subjects who self-reported that they had not ever regularly attended any fitness facility.[11] Treated and control subjects met in separate sessions on the same day, at the beginning of the second week of the fall semester of 2008. Both treatment and control subjects were asked to complete a questionnaire, and were then given an offer of $25 to attend the gym once during the following week.[12] We call this the "learning week" offer, and it is identical to Charness and Gneezy's low-incentive condition. Our control group is therefore comparable to Charness and Gneezy's low-incentive group. We chose this as our control in order to separate the effect of overcoming the one-time fixed cost of learning about the gym from the actual habit formation that occurs after multiple visits.[13]

At the same initial meeting, the treatment group received an additional offer of $100 to attend the gym twice a week in each of the four weeks following the learning week. We call this the treatment-month offer, and it is the same as Charness and Gneezy's high-incentive offer, except that they did not require the eight visits to be evenly spaced across the four weeks. The other difference between this offer and Charness and Gneezy's high-incentive offer is that we made our offer at the first meeting, at the same time as the $25 learning-week offer, whereas Charness and Gneezy made their high-incentive offer at their second meeting, a week later. We made our treatment-month offer earlier because we wanted Treatment subjects to have a week to contemplate the idea of going to the gym twice weekly for a month before making predictions. Moreover, if subjects have reference-dependent preferences for money then suddenly announcing a gain of $100 to one group but not the other could introduce systematic bias into the incentive compatible procedure we used to elicit predictions. Waiting a week after treatment subjects learn they will earn $100 will help us overcome a potential "house money effect".

At the end of the learning week both groups of subjects again met separately and completed pencil-and-paper tasks (described in detail below) designed to elicit their predictions of gym attendance during each of five post-treatment "target weeks". Both groups were reminded of the offers they had received. Four weeks later, at the end of the

[10] Due to attrition and missing covariates, our final sample includes 54 treated subjects and 57 control subjects. Details of the sample appear in appendix B.

[11] Our screening mechanism is described in appendix C.

[12] For this and all subsequent offers, subjects were told that a visit needed to involve at least 30 minutes of some kind of physical activity at the gym. We were not able to observe actual behavior at the gym and did not claim that we would be monitoring activity.

[13] We also paid the $10 gym-membership fee for all students, and filed the necessary membership forms for those who were not already members.

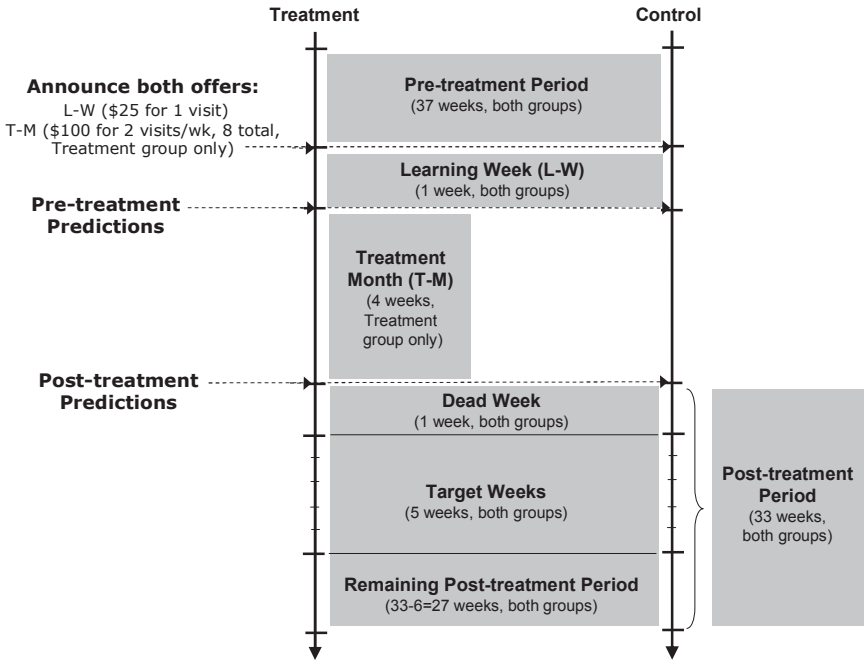

Figure 2.1: Our Experimental Design

treatment month, both groups again met separately, completed an additional questionnaire, and completed the same elicitation tasks as in the second session. The target weeks were separated from this second elicitation session by a dead week so that present-biased subjects would see the target weeks as being "in the future" from the perspective of both elicitation sessions. The timeline of the experiment is illustrated in Figure 2.1.

Gym attendance data were collected for a 17-month period stretching from 37 weeks before the learning week to 33 weeks after it. This period includes summer and winter breaks as well as three full semesters.

2.3.1 Elicitation procedures

To elicit predictions of target-week gym attendance we created what we call a "p-coupon", which is a certificate that rewards the holder with $\$P$ for each day that he or she attends the gym during a specified "target week". The value of P, which ranged from \$1 to \$7, was printed on the coupon, along with the beginning and end dates of the target-week. We used an incentive-compatible mechanism to elicit subjects' valuations for p-coupons of various values with various target weeks.[14] A subject's valuation for a p-coupon is correlated with how many times they think they will attend the gym during the target week of the coupon. A sample p-coupon is included as an appendix, along with the pencil-and-paper task we used to elicit valuations for p-coupons, the instructions we gave them for completing

[14]The elicitation mechanism is described in detail in appendix .

the task, and further description of how the elicitation mechanism worked. Each subject completed this incentive-compatible elicitation task for four out of the five target weeks in our design, and for a different value of p-coupon in each of those four weeks. The values of the p-coupons for the different weeks was randomized among subjects, as was the order in which those weeks were presented.[15]

Subjects' valuation of a coupon that pays out as a function of the number of times a certain event occurs in a future target week need not be based entirely on their prediction of how many times that event will occur. Risk-aversion implies we would only observe subjects' certainty equivalents, even for an exogenous event.[16] But for an endogenous event like gym attendance, there is the additional confound that the p-coupon itself incentivizes the subject to go to the gym, thus influencing the very behavior we are asking them to predict. This "incentive effect" may increase or decrease subjects' valuations for a p-coupon. We ultimately use this endogeneity as a means of estimating the value of subjects' exercise habit, but care must be taken not to interpret subjects' valuations as directly proportional to their beliefs.

As a check on this mechanism, we also directly asked subjects to state how many times they thought they would go to the gym during the specified target weeks if they had been given the p-coupon they just bid on in the incentive-compatible task. Thus they were making unincentivized *predictions* of hypothetical future attendance under the same set of *attendance* incentives as in the incentivized task.[17] This unincentivized mechanism also allowed us to ask subjects how often they thought they would go to the gym during the one target week for which they were not presented with a p-coupon, the so-called "zero week" (because it is equivalent to a P of zero). The zero week gives us an additional unincentivized prediction of behavior in the absence of any effect of attendance incentives.

Subjects went through exactly the same set of elicitation tasks in both the pre-treatment and post-treatment elicitation sessions. Then, at the end of the second elicitation session, after all of the elicitation tasks had been completed, each subject was given one of the four coupons they had been presented with during the elicitation process. These give-away coupons were in addition to those that had been won earlier in the bidding process. We therefore have two target weeks for each subject in which we can compare their predictions with their actual gym attendance under the same conditions, the first being the zero-week, and the second being the week for which they received a p-coupon in the give-away. The give away was a complete surprise to the subjects—having been conducted unannounced only after the second elicitation session—and cannot have affected their bids or unincentivized responses during the elicitation tasks.

We discuss compliance with the treatment incentive, attrition, and our random-

[15] Thus subjects did not all bid on a p-coupon for target-week one, then target-week two, etc, nor did all subjects bid on p-coupons of the same size for each of the target weeks. Among each subject-group/target-week intersection, subgroups of fifteen subjects received \$1, \$2, and \$3 coupons, ten received \$5 coupons, and five received \$7 coupons.

[16] An alternative design which would have allowed us to sidestep assumptions about the linearity of money utility, would have been to have the coupons pay off not with a dollar sum per visit, but with a per-visit increment in the cumulative probability of winning some fixed-sum prize. However, this would not have allowed us to take advantage of variation in p-coupon value for parameter estimation purposes.

[17] It is important to note that the p-coupons incentivize both target-week attendance and accurate predictions of target-week attendance.

ization procedure in appendix E.

2.4 Results

The immediate post-treatment habit-formation effect among our treated subjects who complied with the treatment was 0.338 visits per week.[18] However, this aggregate result masks heterogeneity in the effect. Figure 2.2 plots the empirical CDF's of average post-treatment attendance in the treated and control groups.[19] The two distributions move together up to the seventy-fifth percentile–with the majority of both treatment and control subjects continuing to avoid gym attendance altogether–and then diverge substantially. Thus, though three-quarters of our treated subjects complied with the treatement incentive, only about one-quarter of them appear to have formed a habit of any size.

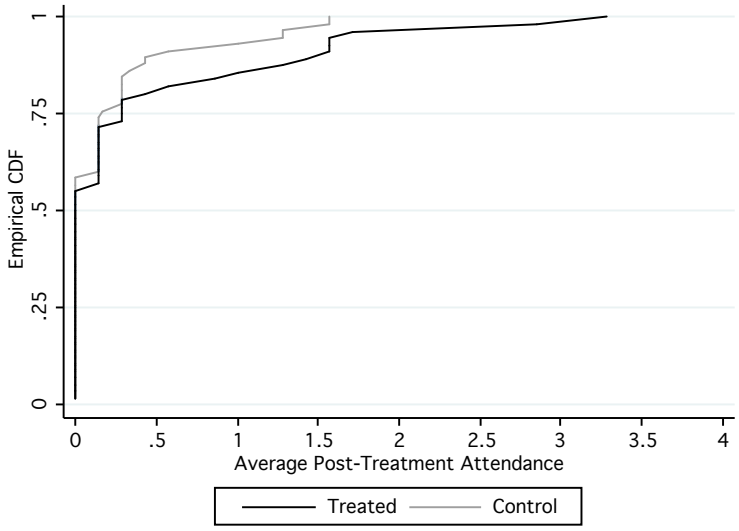

Figure 2.2: Distribution of immediate post-treatment attendance.

To explore this heterogeneity further we look at the change in attendance among treated subjects at the individual level. Figure 2.3 plots average attendance for treated subjects in the immediate post-treatment period against an imputed counterfactual based on a regression of attendance on week dummies and covariates using control-group data for all weeks and treated-group data for the pre-treatment period. Following Charness and Gneezy we designate as "strong habit-formers" those subjects whose actual attendance was at least one visit per week greater than the counterfactual.[20] They are marked with blue crosses in the figure. We designate those below the forty-five degree line as "non–habit-

[18] With a standard error of 0.154. These results are discussed in detail in (cite our other paper).

[19] Attendance in a subject's incentivized week is ommitted from the calculation.

[20] This applies to 8 of 54 treated subjects and 3 of 57 control subjects. A test of equal proportions rejects equality at the p = 0.092 level, and the one-sided test that there are actually more habit-formers in the control group is rejected at a p-value of 0.046.

formers", marked with green circles. In between lie those individuals marked with red diamonds, whom we have designated as "weak habit-formers". It is not clear whether these are people who have actually formed a weak habit, or simply people for whom the random component of our counterfactual was slightly negative. To avoid this ambiguity we will focus our attention on the comparison between our strong habit-formers and non–habit-formers.[21]

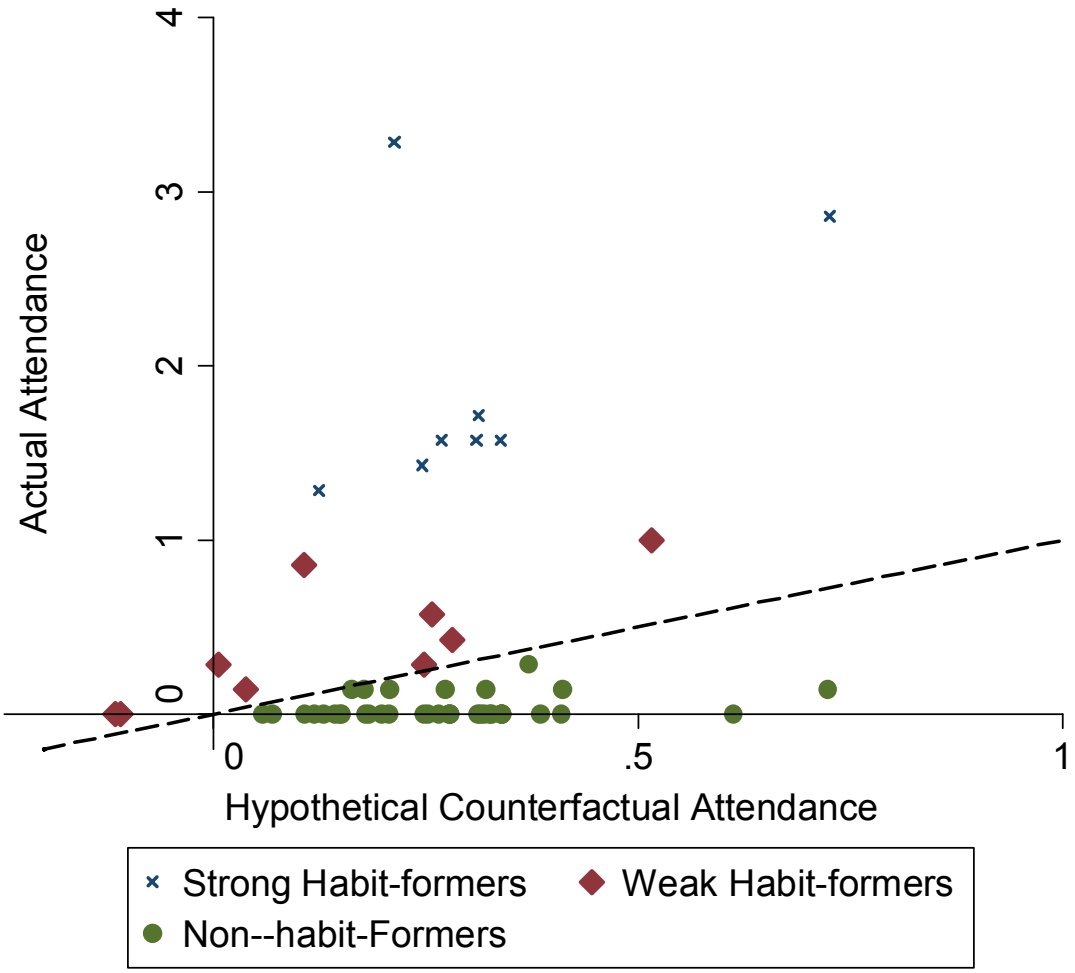

Figure 2.3: Actual versus Counterfactual Attendance

The heterogeneity in habit formation needs to be taken into consideration in testing for projection bias. The test described in section 2.2 is to compare the average pre- to post-treatment change in p-coupon valuation between treated and control groups. In Loewenstein, O'Donoghue and Rabin's model of projection bias, habit-formation heterogeneity would simply downward bias this aggregate double difference. This is because in their model individuals correctly foresee the direction in which their preferences will change,

[21] We have run all of our regressions on the weak habit-formers and the results are qualitatively similar to the results for non–habit-formers.

but under-appreciate the degree of change. Thus, for an individual whose preferences do not change—i.e. a treated subject with habit-value of zero—there can be no evidence of projection bias. Treated subjects who do not form a habit may or may not suffer from projection bias, but since their preferences don't change, projection bias cannot affect their predictions. In a population of subjects with projection bias our model would predict that habit-formers would have a non-zero double difference and non–habit-formers would have a zero double difference.

If, instead, habit-value heterogeneity is the result of random variation around a common mean, and all subjects accurately foresee the average habit-value ex-ante but then realize either a high or low habit-value ex-post, then we would expect habit-formers to revise their predictions upward after the treatment and non–habit-formers to revise theirs downward, so that the aggregate double difference would no longer be an informative test.[22] Looking at habit-formers and non–habit-formers separately provides insight into which of these stories is more likely to be true. To do this we need to make an additional assumption, which is that habit-formers and non–habit-formers would have the same unobservable gym-attendance proclivities in the untreated condition. We realize that this is an assumption that takes us away from the clean exogeneity of randomization. We explore the differences between habit-formers and non–habit-formers on observables in appendix G.

We regress individual weekly attendance in the pre- and immediate post-treatment periods on a dummy for being in the treated group, a dummy for being in the post-treatment period, and the interaction of the two.[23] The double difference in predictions is the coefficient on the interaction term. Table **??** presents the results. The first two columns include all treated subjects, looking first at the incentivized predictions implied by the "BDM" p-coupon valuations, and then at the unincentivized "Self" predictions. The double difference is statistically insignificant for both measures, suggesting no projection bias. In columns three and four we exclude all treated subjects except strong habit-formers. Now we see a double difference in the BDM measure of 0.640, significant at the 5% level. As the coefficient on the post-treatment dummy shows, all subjects' predictions went down over time, but for these habit-forming treated subjects, predictions went down by two-thirds of a visit less than for control subjects, suggesting that before the treatment they failed to predict the habit-formation effect they later experienced. For the Self measure the double difference is still statistically insignificant, thought the point estimate has gone from negative to positive. Columns five and six tell the story for the non–habit-formers. For the BDM measure the double difference is a loosely estimated zero, and for the Self measure it is a statistically insignificant negative.

We feel that these results support the Loewenstein, O'Donoghue and Rabin projection-bias model. Particularly for the incentivized BDM measure what we seem to be seeing is that habit-formers foresee that they will form a habit, but not how much. Their actual habit-formation is at least one visit per week. They appear to have foreseen at most thirty-five percent of that. Meanwhile, it appears that non–habit-formers foresee that they will

[22] We are mindful that the incentive value of the p-coupon could change in either direction for either habit-formers or non–habit-formers. It appears in the data that the incentive value was not a major confound, but as an attempt to address this concern we conduct all our tests on both the incentivized and unincentivized elicitations.

[23] We include our vector of individual covariates in all of our regressions.

Table 2.1: Changes in Predicted Attendance

	(1)	(2)	(3)	(4)	(5)	(6)
	All Treated		Habit-formers		Non Habit-formers	
	BDM	Self	BDM	Self	BDM	Self
Post-Trmt X Treated	0.194	-0.162	0.640**	0.226	0.005	-0.278
	(0.271)	(0.227)	(0.259)	(0.336)	(0.347)	(0.267)
Post-Trmt	-0.733***	-0.636***	-0.733***	-0.636***	-0.732***	-0.637***
	(0.159)	(0.148)	(0.162)	(0.150)	(0.160)	(0.149)
Treated	0.001	0.025	-0.045	0.277	-0.107	-0.210
	(0.301)	(0.288)	(0.586)	(0.449)	(0.331)	(0.306)
Constant	6.858***	3.457**	5.474**	0.276	6.165***	3.480**
	(1.420)	(1.339)	(2.196)	(2.127)	(1.485)	(1.486)
Observations	875	1087	511	635	741	919
R-squared	0.33	0.42	0.35	0.48	0.34	0.43
Num Clusters:	111	111	65	65	94	94

Notes: "Robust standard errors in parentheses," clustered by individual. * significant at 10%; ** significant at 5%; *** significant at 1%

not form a habit. The story is not as clear for the unincentivized Self measure. Taking the point estimates at face value we might interpret these results as supporting the uncertainty story, with habit-formers revising their predictions upward from a rational expectation, and non–habit-formers revising downward from the same baseline. However, these results are statistically insignificant, suggesting that more data may be necessary to fully answer the question.

2.4.1 Estimation strategy

We now estimate the parameters of our model, which we identified in section 2.2. Because we varied P in discrete increments, in order to find the precise values of P necessary to estimate our parameters we assume that both predictions and attendance are linear in P.[24] Using Zellners' Seemingly Unrelated Regressions we estimate

$$\text{ATT}^i_{t_p} = \gamma_{00} + \gamma_{01} \cdot T_i + \gamma_{02} \cdot T_i \cdot p_i + \gamma_{03} \cdot p_i \tag{2.13}$$

$$\text{PRED1}^i_{t_p} = \gamma_{10} + \gamma_{11} \cdot T_i + \gamma_{12} \cdot T_i \cdot p_i + \gamma_{13} \cdot p_i \tag{2.14}$$

$$\text{PRED2}^i_{t_p} = \gamma_{20} + \gamma_{21} \cdot T_i + \gamma_{22} \cdot T_i \cdot p_i + \gamma_{23} \cdot p_i \tag{2.15}$$

We can now express each of the identifying equalities above in terms of the coefficients of this regression equation, and then solve for the parameters of interest. Thus, to

[24] We have explored adding curvature to these relationships. It does not change our results significantly. We report the linear approach for tractability.

estimate η we are looking for P^* such that

$$\text{ATT}_0^T = \gamma_{00} + \gamma_{01} = \gamma_{00} + \gamma_{03} \cdot P^* = \text{ATT}_p^C. \tag{2.16}$$

Solving this for P^*, and hence for η we get $\eta = P^* = \gamma_{00}/\gamma_{03}$.

To estimate $\left(\widehat{\beta} - \beta\right) b$ we want P^* such that

$$\text{PRED2}_p^C = \gamma_{20} + \gamma_{23} \cdot \widetilde{P} = \gamma_{00} + \gamma_{03} \cdot P^* = \text{ATT}_p^C. \tag{2.17}$$

To implement this we substitute $\overline{P}$, the average value of P in the control group, for $\widetilde{P}$. Solving this for $P^* - \overline{P}$ and hence for $\left(\widehat{\beta} - \beta\right) b$ we get $\left(\widehat{\beta} - \beta\right) b = P^* - \overline{P} = (\gamma_{20} - \gamma_{00} + (\gamma_{23} - \gamma_{03})\overline{P})/\gamma_{03}$.

To estimate $(1 - \alpha)\eta$ and α we want P^* such that

$$\text{PRED1}_p^T = (\gamma_{10} + \gamma_{11}) + (\gamma_{12} + \gamma_{13})\widetilde{P} = \gamma_{00} + \gamma_{03} \cdot P^* = \text{ATT}_p^C. \tag{2.18}$$

Once again we implement this by substituting $\overline{P}$, the average value of P, this time in the treatment group, for $\widetilde{P}$. Solving this we get $P^* - \overline{P} = ((\gamma_{10} + \gamma_{11} - \gamma_{00}) + (\gamma_{12} + \gamma_{13} - \gamma_{03})\overline{P})/\gamma_{03}$. Subtracting off the expression we derived above for $\left(\widehat{\beta} - \beta\right) b$ we can easily derive the appropriate expressions for $(1 - \alpha)\eta$ and α.

2.4.2 Estimation results

Table 2.2 shows the results of the three-equation SUR system, and, beneath these, the estimates of structural parameters of interest.[25] The left-hand panel shows the results when we include the entire treated group. The middle panel restricts the sample to strong habit-formers in the treated group, and the left-hand panel to non–habit-formers.

For the full treated group the estimated habit value is about forty cents, significant at the 10% level, and the estimated habit-value is about eighty cents, significant at 5%. This would seem to suggest that in the aggregate subjects overpredicted the degree of habit-formation, resulting in the statistically insignificant habit-formation parameter point estimate of -1.043, well out of the range allowed in the model. From the standpoint of projection bias, however, what matters is not whether the predicted habit is different from zero, but whether it is different from the actual habit. The difference between the two parameters is 0.422 with a standard error of 0.380 so we cannot reject that the predicted habit is the same as, or smaller than, the actual habit. Our α parameter is one minus the ratio of the two, and as such, because both terms are small, it is highly sensitive to small differences between the two, in this case spuriously pushing it into negative territory. To get a cleaner test of projection bias in the aggregate we would need more statistical power.

Looking at strong habit-formers in the middle panel of table **??** we find a habit value of about \$4.50, of which subjects seem to have predicted about \$3.00. Both of these estimates are significant at the 1% level. For these subjects α is 0.324. They foresee

[25] Standard errors generated by the delta-method for non-linear combinations of regression coefficients are not invariant to the algebraic form of the parameters being estimated. We have used the most obvious algebraic forms, as shown above.

Table 2.2: Parameter Estimation

	All Subjects		Controls and Habit-Formers	
	(1)	(2)	(3)	(4)
	ACT	PRED	ACT	PRED
SUR Results				
Treatment Group	0.180*	0.062	2.020***	1.155**
	(0.106)	(0.245)	(0.205)	(0.499)
Treated X $P	-0.138**	-0.173**	-0.128	0.013
	(0.066)	(0.084)	(0.245)	(0.176)
$P	0.447***	0.558***	0.448***	0.565***
	(0.047)	(0.058)	(0.045)	(0.059)
Constant	0.259***	1.684***	0.258***	1.670***
	(0.074)	(0.170)	(0.071)	(0.170)
Observations	545	545	320	320
Parameter Estimates				
Habit Value	0.403*		4.505***	
	(0.230)		(0.603)	
Cost of Naivete	3.913***		3.906***	
	(0.694)		(0.688)	

Standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%

about two-thirds of the habit value, which is to say they "project" about one-third of their current habit state into the future. For non–habit-formers, in the right-hand panel, the habit value is estimated as −$0.520, significant at the 5% level. Our procedure for imputing counterfactual post-treatment attendance forces these subjects to have a small negative change in attendance, so we take this result with a grain of salt. The predicted habit value is not significantly different from zero, suggesting that these non–habit-formers are predicting no change in attendance. The estimate of α is 2.013 which is above the range in the model, but this is almost certainly because of the spurious negative habit value. If the uncertainty story were valid we would expect to see these subjects predicting the same habit value as habit-formers and then revising downward. Instead, we interpret these results as supporting the projection-bias model.

2.5 Conclusion

Summarize the heterogeneity story saying that we feel it supports projection bias, not uncertainty but it isn't robust because we don't have enough statistical power. Make it clear our result is very robust for habit formers, but the full story is not clear overall, so we do know that habit-formers mispredict but we don't know if projection bias is the right model.

As explained in section 2.4, habit-value heterogeneity is a key factor in our exploration of misprediction of habit-formation. If we ignore heterogeneity then our result is either downward biased, or confounded, depending on what model of predictions is cor-

rect, and we can't distinguish between models. If we take heterogeneity into account we can differentiate between a model of projection bias and a model of rational expectations with random habit-value heterogeneity. This is because the two models have different implications for attendance predictions among non–habit-formers. Projection bias assumes that individuals predict the direction in which their preferences will change, but underappreciate the magnitude of change. Non–habit-formers thus accurately foresee that their preferences will not change, so that there is no scope for changes in prediction. A rational-expectations model with random habit-value assumes that subjects correctly predict the average habit-value and then revise based on actual realizations, implying that non–habit-formers predictions will go down from the pre- to post-treatment elicitations. Because we are not able to distinguish habit-formers from non–habit-formers in the control group, our test of projection bias, and our parameter estimation strategy, rely upon the assumption that the two types would respond similarly to p-coupons, and make similar attendance predictions, under the control-group conditions. Allowing ourselves that assumption, we find that habit-formers do foresee the direction in which their preferences change, and do underappreciate the magnitude of change by a factor of about two-thirds, and we find evidence to suggest that non–habit-formers do foresee that they will not form a habit. Our results for habit-formers are quite statistically significant, but for non–habit-formers we do not have adequate statistical power to draw strong conclusions. Taken at face value, these results support the model of projection bias. However, we cannot rule out the rational expectations model. Furthermore, though habit-formers and non–habit-formers appear largely similar on observables, we do not know what effect our assumption of control-condition homogeneity is having on our results.

To to make further headway on this issue, future research needs to overcome to weaknesses of our study, inadequate statistical power and inability to control for heterogeneity in the control group. Three approaches to addressing these two shortcomings suggest themselves. First, obviously, a larger sample would increase statistical significance. Second, knowing ex-ante which subjects are most likely to form a habit, and in particular a strong habit, and which ones will not, would simultaneously increase statistical power—by increasing the variation in habit-value and attendance predictions—and help to address control-group heterogeneity—by allowing us to identify potential habit-formers and non–habit-formers in the control group. A broader, and more informed battery of pre-treatment survey questions, and more extensive collection of pre-treatment observable measures, could help to pin down the heterogeneity in the control group. It might also be helpful to start by recruiting subjects from a pool that is more prone to habit-formation, for example, those who express a desire to establish an exercise routine but have not done so.[26]

Third, to properly pin down the attendance and prediction behavior of habit-formers and non–habit-formers under control-group conditions it will be necessary to conduct an experiment that renders within-subject variation in the treatment condition. Pre-

[26]It is worth noting that this is a somewhat more challenging task than we originally hoped. Charness and Gneezy observed that habit-formation was concentrated among those who did not attend the gym regularly in the several weeks before treatment. On the basis of this observation we thought we would get stronger habit-formation results, and thus greater statistical power for our prediction tests, by selecting non-attenders. In this we were disappointed. Future inquiry into the determinants of habit-formation will need to be much more extensive and systematic.

liminary consideration suggests a design in which subjects would first be subjected to the control condition, and then to the treatment condition. To eliminate calendar effects it would be necessary to run the experiment in waves with subjects in the control condition from one wave coinciding with subjects in the treatment condition in the previous wave. And to control for learning about prediction-making through repeated exposure to the task it would be necessary to subject a group of subjects to the control condition twice.[27] We feel further research with these changes is warranted given the importance of the theoretical issues at stake, and the public policy value of better understanding habit-formation and prediction in gym attendance.

[27]Treating this concern as an order-effect would not work because implementing the treatment condition first would make it impossible to properly implement the control condition, for which subjects must be ex-ante unhabituated.

Chapter 3

A Bounded Rationality Approach to Beta–Delta Preferences.

3.1 Introduction

Behavioral economists have converged on the quasibyperbolic, or beta–delta model of Phelps and Pollak (1968) and Laibson (1997) to represent the psychological phenomenon of present-biased preferences, and explore issues of self control that may arise in the presence of such preferences. The predictions of the model frequently depend crucially on what assumptions are made about individuals' beliefs about their future preferences. O'Donoghue and Rabin (1999b) worked out what has become the standard way of incorporating beliefs by introducing the $\widehat{\beta}$ parameter to capture naivete ($\widehat{\beta} = 1$), sophistication ($\widehat{\beta} = \beta$), and partial naivete ($\beta < \widehat{\beta} < 1$) with respect to future preferences.

There are times when this approach leads to results that seem counterintuitive or less than fully satisfactory. For example, one might hope that self knowledge would protect agents from severe harm, but in the "doing it once" setting of O'Donoghue and Rabin (1999b) (hereafter O'D–R), complete sophistication can cause a mildly present-biased individual to experience severe welfare loss when benefits are immediate, where a naif with the same preferences would experience only mild harm. This is because a sophisticate is modeled as "unboundedly rational", in the sense that she is able to foresee an unlimited number of iterations of future behavior, and future foresight, right up to the terminal period. Put another way, in the O'D–R model the sophisticate's action in each period is determined through backward induction all the way from the terminal period, so that her pessimism about future self control can be compounded many times over. This observation leads to an obvious query: could more natural results be obtained by modeling foresight in a more natural way?

In this paper I borrow an idea from Cognitive Heierarchy Theory (CHT), which is to restrict the number of iterations of foresight that a sophisticated agent engages in. In CHT each player in a strategic game believes that the other players are less sophisticated, and therefore doing fewer rounds of strategic thinking, than themselves.[1] If we think of a

[1] See Camerer, Ho and Chong (2004). I depart from their distributional assumption in modeling agents of level k as believing that all their future selves are level $k-1$.

discrete-time intertemporal model as a strategic game between a current self and a series of future selves then this kind of heierarchical approach can be applied quite naturally. In particular, in this paper I model time-consistent agents as level zero and naive agents—who believe their future selves will be time consistent—as level one. Then I introduce a new concept in intertemporal decision-making, the "k-2–sophisticate" who believes that all her future selves will be naive, or level one, and will thus be modeled as level two.[2] This approach allows for sophisticated beliefs about future preferences, while limiting the number of iterations of strategic thinking the sophisticated agent engages in. Extensive backward induction is no longer necessary, and the baleful pheneomenon of repeatedly compounded pessimism about future self control is mitigated.

I explore behavioral and welfare results for the k-2–sophisticate in the "doing it once" setting of O'D–R. I find that the k-2–sophisticate's behavior is qualitatively similar to O'D–R's full sophisticate, though the k-2–sophisticate procrastinates less than the full sophisticate when costs are immediate, and under natural restrictions on the evolution of delayed costs, preproperates less when rewards are immediate. In addition I find that, like the full sophisticate, the k-2–sophisticate with mild present bias is protected from disasterous procrastination when costs are immediate, but unlike the full sophisticate, is protected from disasterous preproperation when rewards are immediate. However, when costs are immediate she may engage in highly costly pre-emptive behavior due to excessive pessimism about future pre-emptive behavior, with the upper bound on harm, counter-intuitively, positively correlated with β. Section two reviews the O'D–R model and the behavioral and welfare results from that paper. Section three introduces k-2–sophistication and presents behavioral results. Section four presents welfare results for the k-2–sophisticate. Section five concludes.

3.2 Doing it Once: Setup and Results from O'D–R

Agents have periods $t = 1, 2, \ldots, T$ to do an action one time. Doing the action in period t renders reward v_t and cost c_t, one of which will be immediate and the other delayed. The vectors $\boldsymbol{v} = (v_1, v_2, \ldots, v_T)$ and $\boldsymbol{c} = (c_1, c_2, \ldots, c_T)$ fully define the setting. If rewards are immediate then $U^t(t)$, the agent's period-t instantaneous utility for doing it in period t, is $v_t - \beta c_t$ and if costs are immediate it is $\beta v_t - c_t$, while in either case $U^t(\tau)$, the period-t instantaneous utility of doing it in any period $\tau > t$, is $\beta(v_\tau - c_\tau)$, with $\beta \in [0, 1]$ capturing present bias.[3] Beliefs about future present bias are captured by $\widehat{\beta} \in [\beta, 1]$. Agents are of three types, $a \in \{TC, N, S\}$, for Time-Consistent ($\beta = 1$), Naive ($0 < \beta < 1$ and $\widehat{\beta} = 1$), and Sophisticated ($0 < \beta < 1$ and $\widehat{\beta} = \beta$). An agent's strategy, $\boldsymbol{s} = (s_1, s_2, \ldots, s_T)$, with $s_t \in \{Y, N\}$, describes whether she will do it in each period conditional on not having done it already.

[2] One could model levels above two, but they are less obviously natural in this setting than in game theory, and I do not explore them in this paper. It is worth noting, however, that in the limit as the level approaches infinity the CHT approach renders the O'Donoghue–Rabin full sophisticate. It is also worth noting that my CHT-based approach still allows for partial naivete as it includes the $\widehat{\beta}$ parameter to capture beliefs about future preferences.

[3] For simplicity O'D–R let $\delta = 1$.

Solution concepts for the three types are based on the principle that each period's choice must be optimal with respect to what the agent believes she will do in the future. O'D–R define "perception perfect" strategies for the three types. Actual behavior for an agent of type a in any given setting is to do it in the first period for which $s_t^a = Y$. That period is refered to as τ_a.

Definition 1 (O'D–R 2) *A perception perfect strategy for TCs is a strategy* $\boldsymbol{s}^{tc} \equiv \left(s_1^{tc}, s_2^{tc}, \ldots s_T^{tc}\right)$ *that satisfies for all* $t < T \;\; s_t^{tc} = Y$ *if and only if* $U^t(t) \geq U^t(\tau)$ *for all* $\tau > T$.

Definition 2 (O'D–R 3) *A perception perfect strategy for naifs is a strategy* $\boldsymbol{s}^{n} \equiv (s_1^{n}, s_2^{n}, \ldots s_T^{n})$ *that satisfies for all* $t < T \;\; s_t^{n} = Y$ *if and only if* $U^t(t) \geq U^t(\tau)$ *for all* $\tau > T$.

Definition 3 (O'D–R 4) *A perception perfect strategy for sophisticates is a strategy* $\boldsymbol{s}^{s} \equiv (s_1^{s}, s_2^{s}, \ldots s_T^{s})$ *that satisfies for all* $t < T \;\; s_t^{s} = Y$ *if and only if* $U^t(t) \geq U^t(\tau')$, *where* $\tau' \equiv \min_{\tau > t} \{\tau \mid s_\tau^s = Y\}$.

A time-consistent agent does it in the period with the highest net benefit. A naif does it in the first period which his taste for immediate gratification tells him is better than all future periods. A sophisticate does it in the first period that her taste for immediate gratification tells her is better than all future periods in which her future self would do it, given what she foresees about what her future selves will foresee about what subsequently future selves will foresee about... You get the point. The solution concept for sophisticates requires $T - t$ iterations of "strategic" thinking in every period.

The examples in O'D–R elucidate these solution concepts. A cinema shows one film each Saturday for four weeks with ascending values of 3, 5, 8, and 13. In the first example, of immediate costs, agents must miss a film to complete a report on one of the four Saturdays, rendering delayed reward of $\bar{\nu}$. In the second example, of immediate rewards, agents have a coupon good for one film and cannot see more than one, and delayed cost is normalized to zero. In both examples we explore the behavior of TCs, and of naifs and sophisticates with $\beta = \frac{1}{2}$.

Example 4 (O'D–R 1) *Immediate costs:* $\boldsymbol{v} = (\bar{\nu}, \bar{\nu}, \bar{\nu}, \bar{\nu})$ $\boldsymbol{c} = (3, 5, 8, 13)$
$\boldsymbol{s}^{tc} = (Y, Y, Y, Y), \;\; \tau_{tc} = 1$
$\boldsymbol{s}^{n} = (N, N, N, Y), \;\; \tau_{n} = 4$
$\boldsymbol{s}^{s} = (N, Y, N, Y), \;\; \tau_{s} = 2.$

The TC does the report promptly, the naif procrastinates disasterously, the sophisticate procrastinates less.

Example 5 (O'D–R 2) *Immediate rewards:* $\boldsymbol{v} = (3, 5, 8, 13)$ $\boldsymbol{c} = (0, 0, 0, 0)$
$\boldsymbol{s}^{tc} = (N, N, N, Y), \;\; \tau_{tc} = 4$
$\boldsymbol{s}^{n} = (N, N, Y, Y), \;\; \tau_{n} = 3$
$\boldsymbol{s}^{s} = (Y, Y, Y, Y), \;\; \tau_{s} = 1.$

The TC exercises full restraint, the naif preproperates a bit, the sophisticate preproperates disasterously.

Why does the sophisticate fare so badly in example 5? To decide whether to see the first movie she has to figure out which future movies she will go to if she skips the first. This involves putting herself into the shoes of her period two self, but to figure out what she'll do next week she has to put herself into the shoes of her period three self. In each case she foresees a future of one-period-at-a-time preproperation and in despair mopes off to see the worst film.

O'D–R next demonstrate that this pattern of behavior is quite general.[4]

Proposition 6 (O'D–R 1) *(1) If costs are immediate, then $\tau_n \geq \tau_{tc}$. (2) If rewards are immediate, then $\tau_n \leq \tau_{tc}$.*

The naif always does the wrong thing relative to the TC, which O'D–R call the present-bias effect.

Proposition 7 (O'D–R 2) *For all cases, $\tau_s \leq \tau_n$.*

The sophisticate foresees the trouble her present bias will cause her in the future and either procrastinates less or preproperates more—which O'D–R call the sophistication effect—in both cases because she realizes that some prefered future period is not a real option.

Furthermore, O'D–R show that the pattern of potential harm implied by the examples is also quite general. Restricting attention to settings in which there is an upper bound, $\bar{X}$, to the reward and/or cost of any given period they work out the worst-case scenarios for naive and sophisticated agents with arbitrarily mild present bias. Their welfare comparisons are based on a long-term view of utility, which is mathematically the same as utility for a time-consistent agent. Notationally, the long-term utility of period t is $U^0(t) \equiv v_t - c_t$

Proposition 8 (O'D–R 3) *Suppose costs are immediate and consider all $\boldsymbol{v}$ and $\boldsymbol{c}$ such that $v_t \leq \bar{X}$ and $c_t \leq \bar{X}$ for all t:*
(1) $\lim_{\beta\to 1}(\sup_{(\boldsymbol{v},\boldsymbol{c})}[U^0(\tau_{tc}) - U^0(\tau_s)]) = 0$, *and*
(2) For any $\beta < 1$, $\sup_{(\boldsymbol{v},\boldsymbol{c})}[U^0(\tau_{tc}) - U^0(\tau_n)] = 2\bar{X}$.

In certain settings even a minutely present-biased naif may put off the task repeatedly, always thinking he will do it in the next most prefered period, incurring only a small welfare cost each time, but eventually losing all. A sophisticate with the same preferences will always accurately foresee her entire strategy. If she doesn't do it in τ_{tc} it can only be because her present bias convinces her τ_{tc} is less desirable than some other period when she actually does do it, and because her present bias is tiny, the difference between that period and τ_{tc} must also be tiny.

Proposition 9 (O'D–R 4) *Suppose rewards are immediate and consider all $\boldsymbol{v}$ and $\boldsymbol{c}$ such that $v_t \leq \bar{X}$ and $c_t \leq \bar{X}$ for all t:*
(1) $\lim_{\beta\to 1}(\sup_{(\boldsymbol{v},\boldsymbol{c})}[U^0(\tau_{tc}) - U^0(\tau_n)]) = 0$, *and*
(2) For any $\beta < 1$, $\sup_{(\boldsymbol{v},\boldsymbol{c})}[U^0(\tau_{tc}) - U^0(\tau_n)] = 2\bar{X}$.

[4] Proofs of O'D–R's results can be found in the appendix to their paper.

A naif always thinks he will do it in τ_{tc} and thus compares each period to that most prefered period. Thus, if he is only minutely present biased then he will do it in a period that is only minutely less prefered than τ_{tc}. In certain settings a sophisticate with the same preferences will foresee an unwinding backward sequence of future selves foreseeing that their future selves will preproperate, and therefore do it in the worst period because her present bias makes her think it is just marginally better than her next-best realistic alternative.

In the same way that some results in game theory which involve agents doing many rounds of strategic thinking are unsatisfactory, this catastrophic outcome for a minutely present biased sophisticate, relying as it does upon many rounds of pessimistic foresight, leaves something to be desired. It seems intuitively reasonable that a drastically present-biased sophisticate could second-guess herself and do the task in a drastically sub-optimal period. But for a minutely present-biased to do so seems counter-intuitive. And it is the assumption of unbounded rationality that is driving the odd result.

3.3 K-2–sophistication: Definition and Behavior

The crucial step in applying Cognitive Heierarchy Theory to a novel setting is to define level-zero behavior, as all other levels are defined in terms of this single building block. The natural starting place in the β, δ is time consistency, which involves no consideration of future selves preferences. Careful inspection of definition 2 reveals that a naif thinks that all his future selves will be time consistent, or level zero, so in the framework of CHT a naif is level one. Taking things to the next level, a level two agent thinks all of her future selves will be level one, or naive. This allows for foresight with respect to present-biased preferences, while introducing bounded rationality with respect to the number of iterations of foresight a sophisticate engages in. For this reason I refer to a level two agent as a "k-2–sophisticate".[5] Thus, a k-2–sophisticate does it in any period that appears better than the next period in which a naif would do it in. And since a naif's behavior can always be determined prospectively, so can that of a k-2–sophisticate. To formalize these concepts:

Definition 10 *A perception perfect strategy for k-2–sophisticates is a strategy* $\boldsymbol{s}^k \equiv (s_1^k, s_2^k, \dots s_T^k)$ *that satisfies for all* $t < T \quad s_t^k = Y$ *if and only if* $U^t(t) \geq U^t(\tau')$*, where* $\tau' \equiv \min_{\tau > t}\{\tau \mid s_\tau^n = Y\}$.

If she has not done it already, a k-2–sophisticate will do it in period t if and only if the utility of doing so is greater than the perceived (beta-discounted) utility of doing it in the next period in which a naif would do it.

The goal of the exercise is to preserve the qualitative behavioral results of sophistication while improving the welfare results. I begin by exploring behavioral results.[6]

Proposition 11 *For all cases,* $\tau^k \leq \tau^n$.

[5] The "k" comes from the terminology of CHT, in which k refers to the level of an agent.

[6] All propositions are proved in appendix I.

The k-sophisticate always does it as soon or sooner than the naif. Thus the sophistication effect of O'D–R is preserved under k-2–sophistication.

The behavioral comparison between the k-2–sophisticate and the full sophisticate is slightly more complicated. The following proposition addresses results for a limited but interesting set of cases.

Proposition 12 *(1) If rewards are immediate and $c_t \geq c_{t+1}$ for all t, then $\tau^s \leq \tau^k$. (2) If costs are immediate, then $\tau^k \leq \tau^s$*

What proposition 2b says is that when delayed costs are constant or decreasing, and for any sequence of delayed benefits, the k-2–sophisticate does less of the bad thing than the full sophisticate. She preproperates less because she does only one round of strategic thinking and thus avoids the tragedy of endless second guessing that causes the full sophisticate to abandon any hope of exerting self-control.[7] She procrastinates less because, once again doing only one round of strategic thinking, she compares each period to a worst-case scenario that the full sophisticate knows she won't actually have to face.

We can see proposition 12 in action by looking at what a k-2–sophisticate would do in the cinema examples of O'D–R.

Example 13 *A k-2–sophisticate goes to the cinema.*
(1) In the immediate costs setting of example 4 we have $\boldsymbol{s}^k = (Y, Y, Y, Y),\ \tau_k = 1$.
(2) In the immediate rewards setting of example 5 we have $\boldsymbol{s}^k = (N, Y, Y, Y),\ \tau_k = 2$.

In keeping with 12, when costs are immediate the k-2–sophisticate procrastinates less than the full sophisticate because she is more pessimistic about her future self-control. In period one she says, "I know myself. I'll put this off until the last moment and miss the best film. I need to get it out of the way now or all hope will be lost." It is true that she knows herself, in the sense that she knows she has a persistent problem with self-control, but it is also true that she applies that self-knowledge to the consideration of her future behavior in a limited way. In this case it works in her favor. When rewards are immediate she sees the film in period two because she foresees herself preproperating in period three. But in period one she does not foresee her period-two preproperation because she is only thinking of what a naif would do, which is to see the film in the third period. She knows herself, but not fully. In this case, once again, bounded rationality works in her favor.

3.4 Welfare

The O'D–R cinema examples are ideal for the k-2–sophisticate, giving her a better welfare outcome than the full sophisticate whether costs or rewards are immediate. As O'D–R point out in their paper, fully general welfare comparisons are prohibitively complicated. However, a couple of examples will show how things can backfire on the k-2–sophisticate, relative to the full sophisticate. First, imagine adding to example 4 an additional week, at the beginning, when the cinema is playing quite a good film, worth 6.

[7] I consider the limited set of cases in which a k-2–sophisticate preproperates more than a full sophisticate in appendix H.

Example 14 *Immediate costs:* $\boldsymbol{v} = (\bar{\nu}, \bar{\nu}, \bar{\nu}, \bar{\nu}, \bar{\nu})$ $\boldsymbol{c} = (6, 3, 5, 8, 13)$
$\boldsymbol{s}^{tc} = (N, Y, Y, Y, Y),\ \tau_{tc} = 2$
$\boldsymbol{s}^{n} = (N, N, N, N, Y),\ \tau_{n} = 5$
$\boldsymbol{s}^{s} = (N, N, Y, N, Y),\ \tau_{s} = 3.$
$\boldsymbol{s}^{k} = (Y, Y, Y, Y, Y),\ \tau_{k} = 1$

The addition of the quite good film doesn't change the behavior of the time consistent agent, the naif, or the full sophisiticate. But the k-2–sophisticate, in the first period, because she does not think through what she will do in periods four or three, thinks her only chance to get the better of her impulsive future self is to get the report out of the way immediately. One way to think of this is that, though she is less pessimistic about her future self control problems than the full sophisticate, she is more pessimistic about her future preemtive behavior. As we will see, this kind of excessive preemption of procrastination is the only way that a k-2–sophisticate with mild present bias can get hurt badly.

When rewards are immediate there are also cases where the k-2–sophisticate fares worse than the full sophisticate. Consider the film-coupon setup of example 5 and imagine that a large conference has been planned at a nearby hotel for the third week. The cinema has decided to maximize the take from conference goers by reducing the value of the coupons they give out to locals, requiring them to pay a portion of the ticket price that week worth 4. In addition, to make the example work, imagine that the first film is worth 2.25 and the last, 11.

Example 15 *Immediate rewards:* $\boldsymbol{v} = (2.25, 5, 8, 11)$ $\boldsymbol{c} = (0, 0, 4, 0)$
$\boldsymbol{s}^{tc} = (N, N, N, Y),\ \tau_{tc} = 4$
$\boldsymbol{s}^{n} = (N, N, Y, Y),\ \tau_{n} = 3$
$\boldsymbol{s}^{s} = (N, Y, Y, Y),\ \tau_{s} = 2$
$\boldsymbol{s}^{k} = (Y, Y, Y, Y),\ \tau_{k} = 1.$

Both the naif and the sophisticate go to the film in week three, which means that in week two both the k-2–sophisticate and the full sophisticate go to the film. However, in week one the sophisticate foresees that she'll get the better deal of week two while the k-2–sophisticate still has her eyes on week three because she hasn't worked out that the added cost that week will make her want to go in week two. Again, what hurts the k-2–sophisticate in this case is her excessive pessimism about future self control. She consistently fails to predict the positive steps her future selves will be willing to take to manage her self control problem. However, as we will see, in the case of immediate rewards this kind of mistake cannot cause greivious harm to a k-2–sophisticate with only mild present bias.

Following O'D–R I next consider worst-case welfare scenarios when present bias is mild. The essence of their welfare results is the number of rounds of self-destructive decision making or strategic thinking that agents engage in. When costs are immediate the naif is capable of procrastinating over and over again, hurting himself each time by an amount that is bounded by a diminishing function of β, but potentially accumulating a large welfare loss over many periods of iterative decision making. The full sophisticate avoids all of this iteration by accurately foreseeing all of the periods she might do it and choosing the one she likes best. Only one round of self-destructive decision making, the

cost of which is bounded by that same diminishing function of β, so that serious harm can only come to an agent with a serious self-control problem. Meanwhile, when benefits are immediate the naif does it in the first period that looks better than τ_{TC}, one round of decision making and again, the amount of his welfare loss from that single round of decision making is bounded by a diminishing function of β, so he can't get that badly hurt unless he has an overwhelming self-control problem. The full sophisticate, however, is capable of engaging in an unlimited number of rounds of pessimistic backward induction about her future behavior, concluding, with each round of strategic thinking, that her preproperation will cause her to do it earlier and earlier, and leading, potentially, to extreme preproperation and large welfare loss.

By contrast, the mildly present-biased k-2–sophisticate is protected from the naif's many rounds of procrastination by her foresight, and from the full sophisticates many iterations of pessimistic foresight by her bounded rationality. The only serious harm she can come to is excessive preemption of procrastination. First we consider the procrastination result.

Proposition 16 *Suppose costs are immediate and consider all* $\boldsymbol{v}$ *and* $\boldsymbol{c}$ *such that* $v_t \leq \bar{X}$ *and* $c_t \leq \bar{X}$ *for all* t*:*

(1) $[\tau_k \geq \tau_{tc}]$: $\lim_{\beta\to 1}(\sup_{(\boldsymbol{v},\boldsymbol{c}\ |\ \tau_k\geq\tau_{tc})}[U^0(\tau_{tc}) - U^0(\tau_k)]) = 0$

(2) $[\tau_k < \tau_{tc}]$: *For any* $\beta < 1$, $\sup_{(\boldsymbol{v},\boldsymbol{c}\ |\ \tau_k<\tau_{tc})}[U^0(\tau_{tc}) - U^0(\tau_k)] = (1+\beta)\bar{X}$

Whenever a mildly present-biased k-2–sophisticate hasn't done it before τ_{tc}, if she doesn't do it in τ_{tc} it must be because τ_n is not that much worse than τ_{tc} and since τ_k has to be weakly better than τ_n the welfare loss is bounded and vanishes as β approaches one.[8]

However, in cases where a time-consistent agent doesn't do it in the first period, difficulty may arise for the k-2–sophisticate, even when present bias is mild. The k-2–sophisticate looks at the horrendous outcome that she believes lies in wait for her and, believing that she won't do it at, or after, τ_{tc}, she does it in the first period that feels better in the short-term than her discounted assesment of τ_n, which may be a much less desirable period than τ_{tc}. However, she is protected by her present bias. If she has very mild present bias then she is realistic about how painful τ_n is going to be, and will be willing to do it in an almost as painful early period. If, instead, she has substantial present bias then she erroneously believes that τ_n will not be so bad, and thus passes over very painful early periods and only does it in an early period if the short term cost is relatively low. Thus, ironically, as β approaches one the k-2–sophisticate may lose everything.

Next I consider the case of immediate rewards.

Proposition 17 *Suppose rewards are immediate and consider all* $\boldsymbol{v}$ *and* $\boldsymbol{c}$ *such that* $v_t \leq \bar{X}$ *and* $c_t \leq \bar{X}$ *for all* t*:*

$\lim_{\beta\to 1}(\sup_{(\boldsymbol{v},\boldsymbol{c})}[U^0(\tau_{tc}) - U^0(\tau_k)]) = 0$

When rewards are immediate the k-2–sophisticate with mild present bias cannot be severly harmed by extreme preproperation. The naif does one round of preproperation, and foreseeing this, the k-sophisticate does one more round of preproperation. In each of these rounds

[8] It may be worth noting that in the case of constant or diminishing (check this) delayed rewards we get $\tau_k = \tau_n$ because in this case $\beta v_t - c_t \leq \beta v_{\tau_{tc}} - c_{\tau_{tc}}$, $\forall t$, and in particular, for $\tau_{tc} < t < \tau_n$, $\beta v_t - c_t \leq \beta v_{\tau_{tc}} - c_{\tau_{tc}} < \beta U^0(\tau_n)$.

the welfare loss is limited as a function of β. The thing that can lead the full sophisticate to ruin is that she is capable of foreseeing an unlimited number of iterations of preproperation and the accumulation of small welfare losses can become severe.

3.5 Conclusion

Introducing bounded rationality into a model of present-biased preferences by borrowing from Cognitive Heierarchy Theory appears to render more natural results for procrastination and preproperation in a setting where an agent must do a task with either immediate costs or immediate rewards one time in a fixed number of periods. A "boundedly rational" k-2–sophisticate typically preproperates less, and always procrastinates less, than an "unboundedly rational" full sophisticate, while still exhibiting the sophistication effect of always doing the task before a naif. When present-bias is mild, like the naif, the k-2–sophisticate is protected from extreme preproperation, and like the full sophisticate, is protected from extreme procrastination, except in cases of excessive preemptive behavior.

This is a very preliminary exploration of the role of bounded rationality in models of present bias. An important step would be to review existing results for full sophistication in various models and see whether k-2–sophistication preserves and/or improves those results. In particular, it would be very helpful to know whether limiting the number of rounds of prospective thinking sophisticated agents engage in, and thus largely obviating backward induction, could lead to unique solutions in infinite-horizon settings where full sophistication often leads to multiple solutions. It may also be worth exploring levels of cognitive heierarchy higher than two.

One of the interesting features of the CHT approach I have developed in this paper is that it separates agents' beliefs about their future preferences from their beliefs about their future beliefs. In the O'D–R model the $\widehat{\beta}$ parameter does double duty by simultaneously capturing beliefs about future preferences and beliefs about future beliefs. If a decision maker has a preference parameter β, the model tells us not only that she believes her future selves will have a preference parameter of $\widehat{\beta}$ but also that she believes her future selves will have the same belief about their respective future-selves' preferences. By contrast, a k-2–sophisticate believes that her future selves will have preference parameter $\beta = \widehat{beta}$ but belief parameter $\widehat{beta} = 1$. It could be useful to explore other approaches to separting beliefs about preferences from beliefs about beliefs.

Bibliography

Becker, Gary and Kevin Murphy, "A Theory of Rational Addiction," *Journal of Political Economy*, August 1988, *96* (4), 675–700.

Bénabou, Roland and Jean Tirole, "Self-Confidence and Personal Motivation," *The Quarterly Journal of Economics*, August 2002, *117* (3), 871–915.

Camerer, Colin, Teck-Hua Ho, and Juin-Kuan Chong, "A Cognitive Hierarchy Model of Games," *The Quarterly Journal of Economics*, August 2004, pp. 861–898.

Charness, Gary and Uri Gneezy, "Incentives to Exercise," *Econometrica*, May 2009, *77* (3), 909–931.

DellaVigna, Stefano and Ulrike Malmendier, "Contract Design and Self-Control: Theory and Evidence," *The Quarterly Journal of Economics*, May 2004, *119* (2), 353–402.

___ **and** ___ , "Paying Not To Go To The Gym," *The American Economic Review*, June 2006, *96* (3), 694–719.

Gelbach, Jonah, "When Do Covariates Matter? And Which Ones, and How Much?," *Working Paper*, June 2009.

Gneezy, Uri and Aldo Rustichini, "Pay Enough, or Don't Pay at All," *The Quarterly Journal of Economics*, August 2000, *115* (3), 791–810.

Gruber, Jonathan, "Youth Smoking in the 1990's: Why Did It Rise and What Are the Long-Run Implications?," *The American Economic Review*, May 2001, *91* (2), 85–90.

Kane, Robert, Paul Johnson, Robert Town, and Mary Butler, "A Structured Review of the Effect of Economic Incentives on Consumers' Preventive Behavior," *American Journal of Preventive Medicine*, 2004, *27* (4).

Laibson, David, "Golden Eggs and Hyperbolic Discounting," *The Quarterly Journal of Economics*, May 1997, *112* (2), 443–477.

Loewenstein, George, Ted O'Donoghue, and Matthew Rabin, "Projection Bias in Predicting Future Utility," *The Quarterly Journal of Economics*, March 2003, *118* (4), 1209–1248.

O'Donoghue, Ted and Matthew Rabin, "Addiction and Self Control," in Jon Elster, ed., *Addiction: Entries and Exits*, Russel Sage Foundation, 1999.

___ **and** ___ , "Doing It Now or Later," *The American Economic Review*, March 1999, *89* (1), 103–124.

Phelps, Edmund and Robert Pollak, "On Second-Best National Savings and Game-Equilibrium Growth," *The Review of Economic Studies*, April 1968, *35* (2), 185–199.

Zellner, Arnold, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, June 1962, *57* (298), 348–368.

Appendix A

Value of a p-coupon

The ex-ante value of a p-coupon is

$$X_2^g = X_6^g = 7 \times \int_{c-\widehat{\beta}b-g\cdot\eta-P}^{\infty} P\, dF(\varepsilon) + 7 \times \int_{c-\widehat{\beta}b-g\cdot\eta-P}^{c-\widehat{\beta}b-g\cdot\eta} (b + g\cdot\eta - c + \varepsilon)\, dF(\varepsilon).$$

To see that this is weakly positive, note that the first integral is always non-negative, and the second integral is bounded below by

$$\int_{c-\widehat{\beta}b-g\cdot\eta-P}^{c-\widehat{\beta}b-g\cdot\eta} dF(\varepsilon)\cdot\Big[(1-\widehat{\beta})b - P\Big]$$

This would be the case if all of the mass in the integral were at the lower limit. Thus,

$$\begin{aligned} X_2^C = X_6^C &\geq 7 \times \int_{c-\widehat{\beta}b-g\cdot\eta-P}^{\infty} P\, dF(\varepsilon) + \int_{c-\widehat{\beta}b-g\cdot\eta-P}^{c-\widehat{\beta}b-g\cdot\eta} dF(\varepsilon)\cdot\Big[(1-\widehat{\beta})b - P\Big] \\ &= \int_{c-\widehat{\beta}b-g\cdot\eta}^{\infty} P dF(\varepsilon) + \int_{c-\widehat{\beta}b-g\cdot\eta-P}^{c-\widehat{\beta}b-g\cdot\eta} (1-\widehat{\beta})b\, dF(\varepsilon) \geq 0 \end{aligned}$$

Appendix B

Sample

Our initial sample consisted of 120 subjects, randomly assigned to treated and control groups of 60 subjects each. Subjects were solicited by email through the Xlab at the Haas School of Business at UC Berkeley, and via supplementary email sent through the undergraduate advisors of several of the larger academic departments on UC Berkeley campus. Table B.1 provides a comparison of the treated and control groups. Due to attrition and missing covariates the final number of treated subjects in our analysis is 54 and of control subjects 57. Comparing the two groups on the covariates that we used in all of our analysis we find no significant differences in means, and the F-test of joint significance of the covariates in a linear regression of the treatment-group dummy on covariates is 0.387. In addition to basic demographic variables we included discretionary budget and the time and money cost of getting to campus in order to control for differences in the cost of gym attendance and the relative value of monetary incentives. The pre-treatment Godin Activity Scale is a self-reported measure of physical activity in a typical week prior to the treatment. The self-reported importance of physical fitness and physical appearance were included as a proxy for subjects' taste for the outcomes typically associated with gym-attendance. The naivete proxy covariates are subjects answers to a series of questions that we asked in order to assess their level of sophistication about self-control problems. Answers were given on a four-point scale from "Disagree Strongly" to "Agree Strongly". The exact wording of these questions is as follows:

Variable	Question
Forget	I often forget appointments or plans that I've made, so that I either miss them, or else have to rearrange my plans at the last minute.
Spontaneous	I often do things spontaneously without planning.
Things come up	I often have things come up in my life that cause me to change my plans.
Think ahead	I typically think ahead carefully, so I have a pretty good idea what I'll be doing in a week or a month.
Procrastinate	I usually want to do things I like right away, but put off things that I don't like.

Table B.1: Comparison of Treated and Control groups.

	(1) Full sample	(2) Treated group	(3) Control group	(4) T-test p-value
Original sample	120	60	60	
No. of attriters	6	4	2	
No. w/ incomplete controls	3	2	1	
Final sample size	111	54	57	
$25 learning-week incentive		Yes	Yes	
$100 treatment-month incentive		Yes	–	
Demographic covariates				
Age	21.919	22.204	21.649	0.639
	(0.586)	(0.990)	(0.658)	
Gender (1=female)	0.685	0.648	0.719	0.425
	(0.044)	(0.066)	(0.060)	
Proportion white	0.36	0.333	0.386	0.568
	(0.046)	(0.065)	(0.065)	
Proportion Asian	0.559	0.63	0.491	0.145
	(0.047)	(0.066)	(0.067)	
Proportion other race	0.081	0.037	0.123	0.01
	(0.026)	(0.026)	(0.044)	
Economic covariates				
Discretionary budget	192.342	208.333	177.193	0.404
	(18.560)	(28.830)	(23.749)	
Travel cost to campus	0.901	0.648	1.14	0.37
	(0.273)	(0.334)	(0.428)	
Travel time to campus (min)	14.662	14.398	14.912	0.811
	(1.071)	(1.703)	(1.335)	
Naivete proxy covariates				
Forget[a,b]	1.595	1.556	1.632	0.573
	(0.067)	(0.090)	(0.099)	
Spontaneous[a,b]	2.486	2.574	2.404	0.281
	(0.079)	(0.104)	(0.117)	
Things come up[a,b]	2.586	2.611	2.561	0.731
	(0.072)	(0.107)	(0.097)	
Think ahead[a,b]	2.874	2.944	2.807	0.338
	(0.071)	(0.081)	(0.116)	
Procrastinate[a,b]	3.036	3.056	3.018	0.8
	(0.075)	(0.104)	(0.108)	
Exercise experience and attitude covariates				
Pre-trt Godin Activity Scale	36.05	36.5	35.623	0.855
	(2.376)	(2.983)	(3.689)	
Fitness is important[a,b]	3.081	2.981	3.175	0.092
	(0.057)	(0.086)	(0.076)	
Appearance is important[a,b]	3.252	3.259	3.246	0.917
	(0.065)	(0.096)	(0.088)	
F-test of joint significance				0.387

Notes: [a] 1= Disagree Strongly, 2=Disagree Somewhat; 3=Agree Somewhat; 4=Agree Strongly. [b] Wording of questions in appendix. Standard errors in parentheses.

Appendix C

Screening mechanism

The webpage we used to screen for non-attenders is shown below. We included three "dummy" questions to make it harder for subjects to return to the site and change their answers in order to be able to join the experiment. Despite this precaution, a handful of subjects may have returned to the screening site and modified their answers until they hit upon the correct answer to join the experiment. (Which was a "no" on question four.) Out of a total of 497 unique IP addresses in our screening log, we found 5 instances of subjects possibly gaming the system to gain access to the study. We have no way to determine if these subjects wound up in our subject pool.

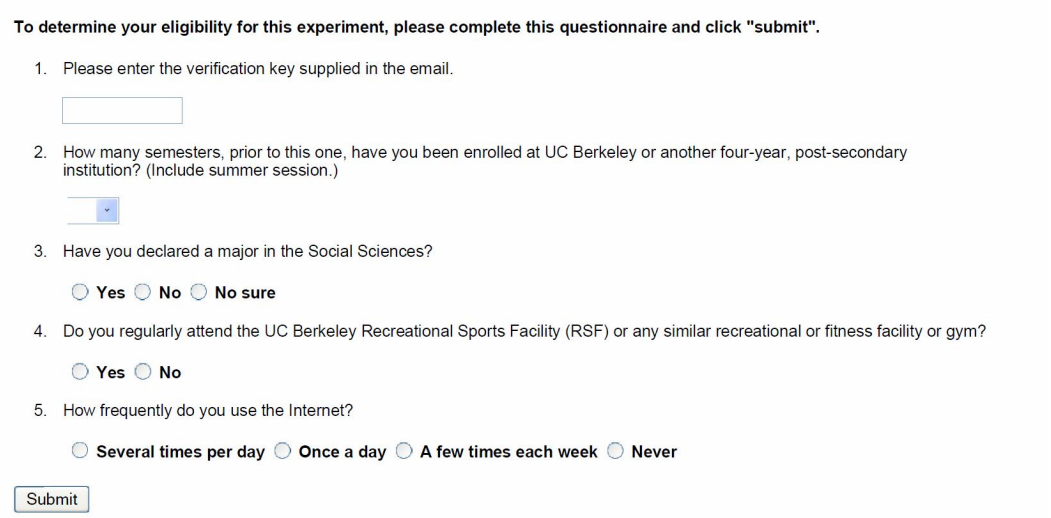

Figure C.1: Screening Site

Appendix D

Elicitation mechanisms

Figure D.1 depicts the sample p-coupon and instructions that subjects saw to prepare them for the incentive-compatible elicitation task. Verbal instructions given at this time further clarified exactly what we were asking subjects to do. Note that the sure-thing values in column A are increments of $\$P$. The line number where subjects cross over from choosing column B to choosing column A bounds their valuation for the p-coupon. We used a linear interpolation between these bounds to create our "BDM" variable. Thus, for example, if a subject chose B at and below line four, and then chose A at and above line five we assigned them a p-coupon valuation of $\$P \times 4.5$ In general subjects appear to have understood this task clearly. There were only three subjects who failed to display a single crossing on every task, and all of them appear to have realized what they were doing before the end of the first elicitation session. The observations for which these three subjects did not display a single crossing have been dropped from our analysis.

By randomly choosing only one target week for only one subject we maintain incentive compatibility while leaving all but one subject per session actually holding a p-coupon, and for only one target week. This is important because what we care about is the change in their valuation of a p-coupon from pre- to post-treatment elicitation sessions. Subjects who are already holding a coupon from the first session would be valuing a second coupon in the second session, making their valuations potentially incomparable, rather like comparing willingness-to-pay for a first candy bar to willingness-to-pay for a second candy bar.

The instructions and example for the unincentivized prediction task and the task for prediction of other people's attendance appear as figure D.2.

[PRACTICE]

This exercise involves nine questions, relating to the Daily RSF-Reward Certificate shown at the top of the page. Each question gives you two options, A or B. For each question check the option you prefer.

You will be asked to complete this exercise four times, once each for four of the five target weeks. The daily value of the certificate will be different for each of these four target weeks. For one of the five weeks you will not be asked to complete this exercise.

At the end of the session I'll choose one of the five target weeks at random. Then I'll choose one of the nine questions at random. Then I'll choose one subject at random. The randomly chosen subject will receive whichever option they checked on the randomly chosen question for the randomly chosen target week. Thus, for each question it is in your interest to check the option you prefer.

$1 **Daily RSF-Reward Certificate** **$1**

This certificate entitles the holder to

$1

for every day that he or she attends the RSF during the week of

Monday, Oct 13 *through* **Sunday, Oct 19.**

$1 **$1**

	S	M	T	W	T	F	S
SEPT		1	2	3	4	5	6
	7	**8**	9	10	11	12	13
	14	**15**	16	17	18	19	20
	21	22	23	24	25	26	27
OCT	28	29	30	1	2	3	4
	5	6	7	8	9	10	11
	12	**13**	14	15	16	17	18
	19	20	21	22	23	24	25
NOV	26	27	28	29	30	31	1
	2	3	4	5	6	7	8
	9	10	11	12	13	14	15
	16	17	18	19	20	21	22
	23	24	25	26	27	28	29

For each question, check which option you prefer, A or B.

	Option A			**Option B**	
1. Would you prefer	☐	$1 for certain, paid Monday, Oct 20.	or	☐	The Daily RSF-Reward Certificate shown above.
2. Would you prefer	☐	$2 for certain, paid Monday, Oct 20.	or	☐	The Daily RSF-Reward Certificate shown above.
3. Would you prefer	☐	$3 for certain, paid Monday, Oct 20.	or	☐	The Daily RSF-Reward Certificate shown above.
4. Would you prefer	☐	$4 for certain, paid Monday, Oct 20.	or	☐	The Daily RSF-Reward Certificate shown above.
5. Would you prefer	☐	$5 for certain, paid Monday, Oct 20.	or	☐	The Daily RSF-Reward Certificate shown above.
6. Would you prefer	☐	$6 for certain, paid Monday, Oct 20.	or	☐	The Daily RSF-Reward Certificate shown above.
7. Would you prefer	☐	$7 for certain, paid Monday, Oct 20.	or	☐	The Daily RSF-Reward Certificate shown above.
8. Would you prefer	☐	$8 for certain, paid Monday, Oct 20.	or	☐	The Daily RSF-Reward Certificate shown above.
9. Would you prefer	☐	$9 for certain, paid Monday, Oct 20.	or	☐	The Daily RSF-Reward Certificate shown above.

Figure D.1: Sample p-coupon and incentive-compatible elicitation task

[PRACTICE]

For each target week you will also be asked to complete the following two exercises. Both of these exercises relate to the Daily RSF-Reward Certificate shown at the top of the page, which is the same as the one shown at the top of the preceding page. In addition, there will be one target week for which you will be shown no certificate, and you will be asked to complete only these last two exercises.

$1 **Daily RSF-Reward Certificate** **$1**

This certificate entitles the holder to
$1
for every day that he or she attends the RSF during the week of
Monday, Oct 13 *through* **Sunday, Oct 19.**

$1 **$1**

	S	M	T	W	T	F	S
SEPT		1	2	3	4	5	6
	7	**8**	9	10	11	12	13
	14	**15**	16	17	18	19	20
	21	22	23	24	25	26	27
OCT	28	29	30	1	2	3	4
	5	6	7	8	9	10	11
	12	**13**	14	15	16	17	18
	19	20	21	22	23	24	25
NOV	26	27	28	29	30	31	1
	2	3	4	5	6	7	8
	9	10	11	12	13	14	15
	16	17	18	19	20	21	22
	23	24	25	26	27	28	29

Imagine that you have just been given the Daily RSF-Reward Certificate shown above, and that this is the only certificate you are going to receive from this experiment.

How many days would you attend the RSF that week if you had been given that certificate? _____

Now imagine that everyone in the room *except you* has just been given the Daily RSF-Reward Certificate shown above, and that this is the only certificate they are going to receive from this experiment.

What do you think would be the average number of days the other people in the room *(not including you)* would go to the RSF that week? _____

(Your answer does not have to be a round number. It can be a fraction or decimal.)

Notes: As part of this experiment some subjects will receive real certificates.

I will give a $10 prize to the subject whose answer to this exercise is closest to the correct, average RSF-attendance for subjects *(other than themselves)* who receive the certificate shown above. The prize money will be paid by check, mailed on Monday,.Oct 20.

Figure D.2: Unincentivized and other elicitation tasks

Appendix E

Compliance, attrition, and randomization.

About 80% of Charness and Gneezy's high-incentive subjects complied with the $100 treatment incentive by attending the gym eight times during the treatment month. A similar percentage, 75%, of our treatment subjects complied with our treatment incentive by attending the gym twice a week during the treatment month. In our data, a direct comparison of means between treatment and control will only allow us to estimate an "intention to treat" effect (ITT). If compliance were random we could simply inflate this by the inverse of the compliance rate to estimate the average treatment effect. Since compliance is almost certainly not random, we will do our best to estimate an "average treatment effect on the treated" (ATT) by using our rich set of individual covariates to help us control for differences between compliers and non-compliers.

To mitigate attrition over our three sessions we gave subjects two participation payments of $25 each, in addition to the various gym-attendance offers. The first payment was for attendance at the first session. The second payment required attendance at both the second and third sessions.[1] Despite this titration of rewards, six of the 120 subjects did not complete the study. Two control subjects and two treatment subjects left the study between the first and second sessions, and two more treatment subjects left between the second and third. In order to include an additional handful of subjects who were not able to make the third session, and otherwise would have left the study, we held make-up sessions the following day. Four control subjects and two treatment subjects attended these sessions and we have treated them as having completed the study.

Randomizing subjects into treatment and control presented some challenges. Our design required that treatment and control subjects meet separately. For each of the three sessions we scheduled four timeslots, back-to-back, and staggered them between Control and Treatment. When subjects responded to the online solicitation, and after they had completed the screening questionnaire, they were randomly assigned to either treatment or control and were then asked to choose between the two timeslots allocated to their assigned group. Subjects who could not find a timeslot that fit their schedule voluntarily

[1] Gym-attendance offers were not tied to attendance because this would have created a differential between the treatment and control groups in the incentive to complete the study.

left the study at this point.[2] As it turned out, subjects assigned to the treatment group were substantially less likely to find a timeslot that worked for them, and as a result the desired number of subjects were successfully enrolled in the control group well before the treatment group was filled. Wanting to preserve the balanced number of Treatment and Control subjects, maintain power to identify heterogeneity within the Treatment group, and stay within the budget for the study, we capped the control group and continued to solicit participants in order to fill the treatment group. Subjects who responded to the solicitation after the Control group was filled were randomly assigned to treatment or control, and those assigned to control were then thanked and told that the study was full. Our treatment group therefore includes subjects who were either solicited later, or responded to the solicitation later than any of the subjects in the control group.[3]

To the extent that these temporal differences are correlated with any of the behaviors we are studying, simple comparisons of group averages may be biased. It appears, however, that the two groups are not substantially different along any of the dimensions we observed in our dataset, as a joint F-test does reject that the two groups were randomly selected from the same population based on observables. A comparison of the two groups appears in a separate appendix. To address the possibility that they may have differed significantly on unobservables we use observable controls in our hypothesis tests.

[2] Technically they were considered to have never joined the study, and received no payment.

[3] Additionally, the two groups of subjects were available at different times of day. To the extent that what made it hard for Treatment subjects to find a timeslot that fit the schedule may have been correlated with gym-attendance behavior (if, for example, the Treatment timeslots happen to have coincided with the most prefered times for non-gym exercise), then the group averages for some outcome variables may be biased.

Table E.1: Comparison of Compliers and Non-Compliers

	(1) Treated Group	(2) Compliers	(3) Non-Compliers	(4) T-test p-value
Demographic covariates				
Age	22.204	22.605	20.636	0.429
	(0.990)	(1.234)	(0.472)	
Gender (1=female)	0.648	0.651	0.636	0.929
	(0.066)	(0.074)	(0.152)	
Proportion white	0.333	0.349	0.273	0.640
	(0.065)	(0.074)	(0.141)	
Proportion Asian	0.630	0.651	0.545	0.526
	(0.066)	(0.074)	(0.157)	
Proportion other race	0.037	0.000	0.182	0.004
	(0.026)	(0.000)	(0.122)	
Economic covariates				
Discretionary budget	208.333	222.093	154.545	0.350
	(28.830)	(34.475)	(41.808)	
Travel cost to campus	0.648	0.616	0.773	0.853
	(0.334)	(0.386)	(0.679)	
Travel time to campus (min)	14.398	13.372	18.409	0.237
	(1.703)	(1.790)	(4.564)	
Naivete proxy covariates				
"Forget[a,b]"	1.556	1.465	1.909	0.047
	(0.090)	(0.096)	(0.211)	
"Spontaneous[a,b]"	2.574	2.442	3.091	0.011
	(0.104)	(0.101)	(0.285)	
"Things come up[a,b]"	2.611	2.558	2.818	0.333
	(0.107)	(0.101)	(0.352)	
"Think ahead[a,b]"	2.944	2.977	2.818	0.436
	(0.081)	(0.091)	(0.182)	
"Procrastinate[a,b]"	3.056	2.977	3.364	0.135
	(0.104)	(0.118)	(0.203)	
Exercise experience andattitude covariates				
Pre-trt Godin Activity Scale	36.500	38.360	29.227	0.221
	(2.983)	(3.137)	(7.961)	
"Fitness is important[a,b]"	2.981	2.977	3.000	0.914
	(0.086)	(0.097)	(0.191)	
"Appearance is important[a,b]"	3.259	3.256	3.273	0.944
	(0.096)	(0.095)	(0.304)	
N obs.	54	43	11	
F-test of joint significance				0.635

Notes: [a] 1= "Disagree Strongly, 2=Disagree Somewhat; 3=Agree Somewhat;" 4=Agree Strongly. [b] Wording of questions in appendix. Standard errors in parentheses.

Appendix F

Hausman Test

Following Gelbach (2009), if we decompose the change in the treatment effect caused by the addition of covariates into the contributions of our four categories of covariates, we get:

Table F.1: Hausman Decomposition

	Change in coef	p-value
Total	0.127	0.051
Demographics	0.031	0.358
Economic	0.007	0.883
Naivete	0.048	0.233
Exercise	0.041	0.213

Appendix G

Habit Formers

Table G.1: Comparison of Habit-Formers and Non Habit-Formers

	(1) "Habit-Formers"	(2) Non Habit-Formers	(3) T-test p-value
Demographic covariates			
Age	19.750	22.630	0.306
	(0.453)	(1.150)	
Gender (1=female)	0.625	0.652	0.885
	(0.183)	(0.071)	
Proportion white	0.250	0.348	0.596
	(0.164)	(0.071)	
Proportion Asian	0.750	0.609	0.454
	(0.164)	(0.073)	
Proportion other race	0.000	0.043	0.557
	(0.000)	(0.030)	
Economic covariates			
Discretionary budget	181.250	213.043	0.699
	(92.068)	(30.274)	
Travel cost to campus	0.000	0.761	0.424
	(0.000)	(0.391)	
Travel time to campus (min)	9.688	15.217	0.252
	(1.666)	(1.958)	
Naivete proxy covariates			
Forget[a,b]	1.500	1.565	0.800
	(0.327)	(0.091)	
Spontaneous[a,b]	2.250	2.630	0.198
	(0.164)	(0.118)	
Things come up[a,b]	2.375	2.652	0.363
	(0.263)	(0.117)	
Think ahead[a,b]	3.000	2.935	0.778
	(0.189)	(0.090)	
Procrastinate[a,b]	2.875	3.087	0.473
	(0.295)	(0.111)	
Exercise experience and attitude covariates			
Pre-trt Godin Activity Scale	41.688	35.598	0.474
	(3.823)	(3.434)	
Fitness is important[a,b]	3.500	2.891	0.010
	(0.189)	(0.089)	
Appearance is important[a,b]	3.375	3.239	0.620
	(0.183)	(0.109)	
N obs.	8	46	
F-test of joint significance			0.663

Notes: [a] 1= Disagree Strongly, 2=Disagree Somewhat; 3=Agree Somewhat; 4=Agree Strongly. [b] Wording of questions in appendix. Standard errors in parentheses.

Appendix H

When does k preproperate more than s?

$\tau_k < \tau_s$ requires that there be some period when k does it and s doesn't, which means $U^{\tau_k}(\tau'_k) \leq U^{\tau_k}(\tau_k) < U^{\tau_k}(\tau'_s)$. Rewriting the ends of this double inequality gives us $v_{\tau'_s} - c_{\tau'_s} > v_{\tau'_k} - c_{\tau'_k}$, which we can rearrange to get $v_{\tau'_s} - v_{\tau'_k} > c_{\tau'_s} - c_{\tau'_k}$ Next, notice that $\tau_k < \tau_s$ requires that in τ_k we have τ'_s strictly before τ'_k, which means that a naif would not do it in τ'_s. Now, the only reason this can be true is if there is some period, say t', after τ'_k in which the naif, in τ'_s thinks she will do it.[1] This requires $U^{\tau'_s}(t') > U^{\tau'_s}(\tau'_s)$ which, by the definition of τ'_k requires $U^{\tau'_k}(\tau'_k) > U^{\tau'_s}(\tau'_s)$ which gives us $v_{\tau'_s} - \beta c_{\tau'_s} > v_{\tau'_k} - \beta c_{\tau'_k}$. Rearranging this and combining with the earlier result we get the full condition:

$$c_{\tau'_s} - c_{\tau'_k} < v_{\tau'_s} - v_{\tau'_k} < \beta(c_{\tau'_s} - c_{\tau'_k})$$

Notice that this double inequality can only hold when costs are increasing between τ'_s and τ'_k, and in particular, increasing more than rewards, but not too much more.

[1] Need to check this assertion. Basically it has to be the case that if n doesn't do it at tauprime s it must be because there's some period she thinks will be better, so I just need to show that that period cannot come before tauprime k without violating the definition of tauprime k.

Appendix I

Proofs

Proof of Proposition 11.

Recall that the naif does it in period t if and only if $U^t(t) \geq U^t(\tau)$ for all $\tau > t$, while the the k-2–sophisticate does it in period t if and only if $U^t(t) \geq U^t(\tau')$. Since $\{\tau'\} \subseteq \{\tau \mid \tau > t\}$ and the maximum of a subset is weakly less than the maximum of the superset, the k-2–sophisticate does it whenever the naif does, and in particular may do it when the naif doesn't, i.e. sooner.

Proof of Proposition 12.

Let $t < T$ be an arbitrary, non-terminal period. Relative to t we refer to the τ' in definition 3 as τ'_s and the τ' in definition 10 as τ'_k.

(1) By proposition 7 $\tau'_s \leq \tau'_k$. The proof consists of showing that $U^t(\tau'_s) \geq U^t(\tau'_k)$ so that if k does it in period t, s does too, and may do it when k does not. Now $v_{\tau'_k-1} - \beta c_{\tau'_k-1} < v_{\tau'_k} - \beta c_{\tau'_k}$ because if not the definition of τ'_k is contradicted. To see this, notice that by the definition of τ'_k we have $v_{\tau'_k} - \beta c_{\tau'_k} \geq \max_{\tau > \tau'_k}\{\beta(v_\tau - c_\tau)\}$, and since $v_{\tau'_k} - \beta c_{\tau'_k} > \beta(v_{\tau'_k} - c_{\tau'_k})$, if $v_{\tau'_k-1} - \beta c_{\tau'_k-1} \geq v_{\tau'_k} - \beta c_{\tau'_k}$ then $s^n_{\tau'_k-1} = Y$ which contradicts the definition of τ'_k. By iteration, $v_\tau - \beta c_\tau < v_{\tau'_k} - \beta c_{\tau'_k}$, for all $t < \tau < \tau'_k$ and since for all t $c_t \geq c_{t+1}$ we get $v_\tau - c_\tau < v_{\tau'_k} - c_{\tau'_k}$ for all $t < \tau < \tau'_k$ which means $v_{\tau'_s} - c_{\tau'_s} \leq v_{\tau'_k} - c_{\tau'_k}$. Thus $U^t(t) \geq U^t(\tau'_k) \Longrightarrow U^t(t) \geq U^t(\tau'_s)$, which means s does it whenever k does it.

(2) The proof consists of showing that $U^t(\tau'_k) \geq U^t(\tau'_s)$ so that if s does it in period t, k does too, and may do it when s does not. By proposition 2 we have $\tau'_s \leq \tau'_k$ and because s does it whenever n does it, $s^s_{\tau'_k} = Y$. By the definition of τ'_s we have $\beta v_{\tau'_s} - c_{\tau'_s} \geq \beta(v_{\tau'_k} - c_{\tau'_k})$, and since $\beta c_{\tau'_s} < c_{\tau'_s}$ we have $v_{\tau'_s} - c_{\tau'_s} \geq v_{\tau'_k} - c_{\tau'_k}$. Thus $U^t(t) \geq U^t(\tau'_s) \Longrightarrow U^t(t) \geq U^t(\tau'_k)$, which means k does it whenever s does it.

Proof of proposition 16

(1) If $\tau_k = \tau_{tc}$ then $U^0(\tau_{tc}) - U^0(\tau_k) = 0$. If $\tau_k > \tau_{tc}$ we know from proposition 11 that $\tau_k \geq \tau_n$ and by the definition of τ_k we have $\beta v_{\tau_k} - c_{\tau_k} \geq \beta v_{\tau_n} - \beta c_{\tau_n}$, and since $\beta v_{\tau_k} - \beta c_{\tau_k} \geq \beta v_{\tau_k} - c_{\tau_k}$ we get $U^0(\tau_k) \geq U^0(\tau_n)$. Now $\tau_k > \tau_{tc} \implies s^k_{\tau_{tc}} = N \implies \beta v_{\tau tc} - c_{\tau_{tc}} < U^0(\tau_n) \leq U^0(\tau_k)$. Rearranging we get $\beta U^0(\tau_{tc}) - (1-\beta)c_{\tau_{tc}} < \beta U^0(\tau_k)$ and rearranging again we get $0 \leq U^0(\tau_{tc}) - U^0(\tau_k) < \frac{1-\beta}{\beta} c_{\tau_{tc}} \leq \frac{1-\beta}{\beta}\bar{X}$, where the first inequality arises from the definition of τ_{tc} as the period with the highest ex-ante utility. Hence $0 \leq \sup_{(\boldsymbol{v},\boldsymbol{c} \mid \tau_k \geq \tau_{tc})}[U^0(\tau_{tc}) - U^0(\tau_k)] < \frac{1-\beta}{\beta}\bar{X}$ and the result follows from the squeeze theorem.

(2) $U^0(\tau_{tc}) - U^0(\tau_k) = [U^0(\tau_{tc}) - U^0(\tau_n)] - [U^0(\tau_k) - U^0(\tau_n)]$ By proposition 8 we know that $\sup_{(\boldsymbol{v},\boldsymbol{c})}[U^0(\tau_{tc}) - U^0(\tau_n)] = 2\bar{X}$ and from the proof of that proposition in O'D–R we know that the welfare loss converges to this supremum when $(v_{\tau_{tc}}, c_{\tau_{tc}}, v_{\tau_n}, c_{\tau_n}) = (\bar{X}, \varepsilon, 0, \bar{X})$, where $\varepsilon \in (0, \bar{X})$ is some arbitrarily small positive number. Now let us add a period before $v_{\tau_{tc}}$ and call this period 1, and let $v_1 = 0$, and $c_1 = \beta\bar{X}$ so that $s^k_1 = Y$, $\tau_k = 1$, and $U^0(\tau_k) = -\beta\bar{X}$. Thus $U^0(\tau_k) - U^0(\tau_n) = -(\beta\bar{X}) - (-\bar{X}) = (1-\beta)\bar{X}$. As this is the smallest value of $U^0(\tau_k) - U^0(\tau_n)$ for which $\tau_k < \tau_{tc}$ we have $\sup_{(\boldsymbol{v},\boldsymbol{c} \mid \tau_k < \tau_{tc})} -[U^0(\tau_k) - U^0(\tau_n)] = (1-\beta)\bar{X}$ and since this supremum and the one above are both approached by the same $(\boldsymbol{v}, \boldsymbol{c})$ vector we get $\sup_{(\boldsymbol{v},\boldsymbol{c} \mid \tau_k < \tau_{tc})}[U^0(\tau_{tc}) - U^0(\tau_k)] = 2\bar{X} - (1-\beta)\bar{X} = (1+\beta)\bar{X}$.

Proof of Proposition 17.

$U^0(\tau_{tc}) - U^0(\tau_k) = [U^0(\tau_{tc}) - U^0(\tau_n)] + [U^0(\tau_n) - U^0(\tau_k)]$ By the definition of τ_n we know that $U^0(\tau_{tc}) - U^0(\tau_n) \leq \frac{1-\beta}{\beta} v_{\tau_n} \leq \frac{1-\beta}{\beta}\bar{X}$. (This is derived in the proof of prosposition 4.1 in O'D–R.) If $\tau_k = \tau_n$ then $U^0(\tau_n) - U^0(\tau_k) = 0$. Otherwise, by the definition of τ_k, we have $v_{\tau_k} - \beta c_{\tau_k} > \beta U^0(\tau_n)$ which by rearranging gets us $U^0(\tau_n) - U^0(\tau_k) < \frac{1-\beta}{\beta} v_{\tau_k} \leq \frac{1-\beta}{\beta}\bar{X}$. Thus we get that $0 \leq U^0(\tau_{tc}) - U^0(\tau_k) \leq 2\frac{1-\beta}{\beta}\bar{X}$ which implies $0 \leq \sup_{(\boldsymbol{v},\boldsymbol{c})}[U^0(\tau_{tc}) - U^0(\tau_k)] \leq 2\frac{1-\beta}{\beta}\bar{X}$, and the result follows from the squeeze theorem.