Naiveté, Projection Bias, and Habit Formation in Gym Attendance*

Dan Acland† and Matthew Levy‡

November 11, 2011

Abstract

We develop a model capturing habit-formation, projection-bias, and present-bias in an intertemporal-choice setting, and conduct a field experiment to identify its main parameters. Building on the Charness and Gneezy (2009) paradigm, we incentivize subjects to attend the gym for a month, observe their pre- and post-treatment attendance relative to a control group, and elicit their pre- and post-treatment predictions of post-treatment attendance. Present-biased subjects want their future selves to exercise more than they actually will, and naivete about this bias causes them to over-estimate their future attendance. Projection-biased subjects extend their current state onto their future expectations, and therefore under-estimate any habit-formation effect of our treatment ex-ante. We find subjects form a significant short-run habit, followed by substantial decay caused by the semester break. Importantly, subjects appear not to embed habit formation into their ex-ante predictions. Approximately one-third of subjects formed a habit equivalent to a \$2.60 per-visit subsidy, while their predictions correspond to 90% projection bias over the habit formation. Moreover, subjects greatly over-predict future attendance, which we interpret as evidence of partial naivete with respect to present bias: they expect their future selves to be two-thirds less "present biased" than they currently are.

*The authors would like to thank Stefano DellaVigna, Gary Charness, Uri Gneezy, Teck Hua Ho, Shachar Kariv, Botond Koszegi, Ulrike Malmendier, Matthew Rabin, and seminar participants at UC Berkeley and Harvard for their helpful comments. Financial support 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.

†University of California, Berkeley. acland@econ.berkeley.edu

‡Corresponding author. London School of Economics. m.r.levy@lse.ac.uk

1 Introduction

Individuals routinely make decisions that involve predictions of how their preferences, costs, and beliefs will change in the future. The neoclassical approach assumes that individuals faced with such decisions have rational expectations, and that while exact future preferences, costs, and beliefs may not be known, people know the range of possibilities and make predictions that are correct in expectation. A recent literature has modeled a range of departures from this neoclassical assumption about prediction, and has empirically demonstrated the existence of many such departures in particular contexts. In this paper, we examine the interaction of two of the leading models—partial naiveté with respect to present-biased preferences, and projection bias with respect to state-dependent preferences—when both departures are allowed to affect choices, and we estimate the strength of both biases using a field experiment on physical exercise. These new parameter estimates are of significant theoretical interest, but are also important for efficient design of public policy, not least in the area of public health, and particularly in situations where the appropriate policy response is non-monotonic. For example, as we find here, the effect of a commitment device for present-biased individuals is not straightforward in the case of partially naive present bias—compared to the fully sophisticated or fully naive cases—and varies in important ways when individuals also imperfectly predict habit formation.

We take as given (and then demonstrate) that physical exercise is a domain in which individuals experience both habit formation and self-control issues due to present-biased preferences. We use the O'Donoghue and Rabin (1999a) model of partially-naive present bias, wherein people may not fully predict that they will be just as impatient tomorrow as they are today. To capture predictions of habit formation we follow the "simple projection bias" model of Loewenstein, O'Donoghue and Rabin (2003), in which individuals correctly foresee the direction in which their preferences will change but may under-appreciate the magnitude of the change, using a single parameter α to index the degree of error. Using gym attendance as a proxy for physical exercise, we experimentally test to what extent subjects predict the effect of habit formation, and of their self-control problems, on their future gym attendance. That habit formation plays an important role in physical exercise has long been accepted in the behavioral health literature[1], and has been more recently demonstrated

[1]See Valois, Dersharnais and Godin (1988), Dzewaltowski, Noble and Shaw (1990), Reynolds,

experimentally by Charness and Gneezy (2009), who paid subjects \$100 to attend the gym eight times in one month and found significantly higher gym attendance in the period after the payment ended than in the pre-intervention period. We build on this experimental framework in the current study, but it is important to note that while we are able to generate additional results about habit formation *per se* (including the dollar value of the experimentally induced habit), the exercise paradigm is used primarily to identify the more general parameters of prediction and discounting.

To test for misprediction of future gym preferences, we recruited 120 subjects who were self-reported non-gym attenders and replicated Charness and Gneezy's habit-formation treatment. Building on this framework, we elicited subjects' predictions of their post-treatment gym attendance, conducting elicitations both immediately before and immediately after the treatment period. These attendance predictions consisted of both an incentive-compatible valuation of a contingent-payment contract for future gym attendance, and an un-incentivized direct prediction. 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 pre- and post-treatment predictions should be greater for treated subjects than for control subjects. Moreover, by comparing subjects' immediate post-treatment predictions of attendance with their actual, later post-treatment attendance, we can estimate their degree of naiveté with respect to their self-control problem—that is, how close their prediction of their short-run discount factor for future selves is to their actual short-run discount factor. Finally, by offering small attendance incentives in some of the post-treatment weeks, we are able to estimate the costs and benefits associated with attendance, and hence the cost of subjects' mis-predictions.

We find a significant post-treatment gym-attendance increase of 0.256 visits per week among our subjects, which we interpret as a habit-formation effect. This effect is smaller than, but statistically indistinguishable from, Charness and Gneezy's result, and as they found, the treatment effect is concentrated in the upper tail of the post-treatment attendance distribution.[2] We find no evidence that subjects predicted this

Killen, Bryson, Maron, Taylor, Maccoby and Farquhar (1990), Godin, Valois and Lepage (1993), Godin (1994)

[2]The effect appears to largely decay during the semester break, however, suggesting that this type of habit formation may be short-lived. Indeed, Kane, Johnson, Town and Butler (2004) in a review find that monetary incentives are generally effective at generating short-run behavioral changes, but typically do not have long-run effects that extend even as far as those we identify in this study. Our structural estimates will suggest that our intervention did not induce the steady-state level of

habit-formation effect overall, which we interpret as prima facie evidence of projection bias. We also find that both treated and untreated subjects substantially over-predict their future gym attendance: even in our simplest elicitation task, subjects over-predicted attendance by roughly a factor of three. We interpret this as evidence of partially-naive present-biased preferences. Finally, there is evidence that our student subjects may have begun the semester generally over-optimistic, but became less so over time. By fixing the delay between the week in which predictions are made and the week about which they are made, we rule out any kind of intertemporal discounting as an explanation for this shift.

Having conducted the above reduced-form tests, we estimate a structural model that yields generalizable estimates of prediction-bias parameters, which have previously proven difficult to directly identify. The first parameter of interest is the projection-bias parameter, α. We estimate that whereas the experimentally induced habit has a value equivalent to \$2.60 and takes hold in roughly 1/3 of treated subjects, their gym-attendance predictions prior to being put in the habituated state correspond to a predicted habit value of less than \$0.25. These two estimates together indicate a degree of projection bias $\alpha = 0.9$. This is considerably greater than the $\alpha \in [0.31, 0.50]$ range found by Conlin, O'Donoghue and Vogelsang (2007) for cold-weather clothing catalog sales, although their estimates lie within our 95% confidence interval. This near-total degree of projection bias leads to strong welfare implications. While much attention is given to negative habits such as smoking, equally importantly, if people do not foresee the way that healthy behaviors such as exercise can become more enjoyable after a period of habit formation, they may make suboptimal choices and miss out on important health benefits. The same may be true for a wide range of positive habits.

Our second structural parameter of interest re-parameterizes naiveté over present bias so that it follows a parallel construction to projection bias. If $\beta \in [0, 1]$ captures an individual's actual short-term impatience, and $\hat{\beta} \in [\beta, 1]$ captures their belief about future short-term impatience, we simply re-index her beliefs as a linear combination of the two extremes, $\widehat{\beta} = \omega \cdot 1 + (1-\omega) \cdot \beta$.[3] Thus $\omega = 0$ corresponds to full sophistication,

habituation in subjects.

[3]It is tempting, but not correct, to infer that naiveté in present bias is merely a case of projection bias where a subject's state is given by the current period. Our goal in introducing ω is not to unify these two biases, but rather simply to provide a means of characterizing naiveté in present bias that is independent of the underlying level of time-inconsistency.

and $\omega = 1$ to fully naive beliefs. This re-parameterization allows us to investigate prediction, $\hat{\beta}$, without a separately identified estimate of short-term impatience, β. By using the exercise "commitment value" embedded in subjects' valuations for a contract that rewards future gym attendance, we are able to estimate a value of $\omega = 0.666$: subjects are two-thirds naive about their future self-control problems. If one uses the value $\beta = 0.7$ typically found in other studies (DellaVigna 2009), this corresponds to $\widehat{\beta} = 0.9$. Given the importance of naiveté in the theoretical literature, it is surprising to note the lack of published estimates of $\widehat{\beta}$. Skiba and Tobacman (2008), the only other estimate we could find, use a sample of payday loan borrowers to estimate an almost-identical $\widehat{\beta} = 0.9$; however, more work must be done to confirm the regularity of this result.[4]

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 discusses our findings, and concludes.

2 Model

In this section we develop a simple model that incorporates habit formation, projection bias, and present-biased preferences. The experimental design described in Section 3 will then build on this to identify the key features of this model[5]; Figure 1 may therefore prove useful in understanding the timing of the model. Consider a model with three periods. By design, all subjects are intially non-habituated, and are randomly assigned to a control or treated group.

In the first period, subjects give an incentive-compatible valuation of a contingent-payment contract that provides linear incentives for period-three gym attendance. They also give a direct, but un-incentivized, prediction of how many times they will go to the gym if this contract is implemented. At the end of this period, subjects in

[4] Their estimate unfortunately comes alongside an atypically low $\beta = 0.53$ in addition to an annual long-run discount factor $\delta = 0.45$, suggesting that either their sample put an non-representatively low weight on future consumption, or their model of payday lending is incomplete.

[5] By presenting the model before the experimental design, we hope to underscore the extent to which our empirical strategy is designed to provide identification of key underlying economic parameters. By tightly linking the model to the field experiment, we are following in the "structural behavioral economics" tradition of, among others, DellaVigna, List and Malmendier (2011).

the treated group are endowed with the habit.[6]

In the second period two things happen. First, subjects once again value the third-period incentive contract and predict their third-period attendance. Then, after the elicitation, all subjects are given an incentive contract.[7] We assume there is sufficient time between periods two and three to act as a buffer, ensuring that subjects consider period three to be "in the future" when all predictions are elicited. In period three, subjects receive monetary rewards according to their contemporaneous gym attendance. Finally, at the end of period three, subjects receive the delayed health benefit of whatever gym attendance they have engaged in.

We model utility as quasi-linear in money. Without loss of generality, utility from all non-gym sources is normalized to zero. Let the immediate utility of gym attendance on day d be given by $(-c + \varepsilon_d)$ with $c > 0$ and i.i.d. ε_d, and let the present value of the delayed benefits of gym attendance be $b > 0$. Thus we model gym attendance as an "investment good" in the language of DellaVigna and Malmendier (2004), meaning that costs are immediate while rewards are delayed. We abstract from the model of Becker and Murphy (1988) and O'Donoghue and Rabin (1999b) by modeling habituation as a binary state variable rather than a stock variable with geometric decay. While we therefore do not explicitly consider the habit formation process, we assume that a single gym visit or absence is not sufficient to change a subject's state.

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 simply take one of two values. With probability π, a subject has $\eta_i = \overline{\eta}$ strictly greater than zero, and with probability $1 - \pi$, they have $\eta_i = 0$. For simplicity, we present the case in which subjects know their own type ex ante, but neither our modeling nor our empirical strategy distinguishes this from the case where

[6] We are therefore modeling the habit value associated with receiving the treatment offer in our experiment. In practice, the offer will be a $100 reward for attending the gym twice-weekly for one month. Approximately 80% of treated subjects met the 8-visit threshold for earning the reward, while none of the control subjects visited the gym 8 times during this same month.

[7] In the model we abstract from the fact that the elicitation process will require one or two subjects to wind up with two sets of incentives. In practice, because there were multiple target weeks, most of the auction winners did not end up holding multiple rewards for the same week. The two subjects who did wind up with two rewards for the same target week simply received double the reward and are counted in the analysis as such. The analysis is robust to dropping these observations.

subjects merely know $\overline{\eta}$ and π ex ante, but not their own type.

On top of these basic preferences, we allow subjects to exhibit any degree (including none) of two psychological biases. First, individuals may have present-biased time preferences: a time-inconsistent taste for immediate gratification.[8] Following Laibson (1997), the degree of present bias is captured by an extra discount factor $\beta < 1$ applied uniformly to all future periods, in addition to a standard exponential discount factor δ. Because the experimental period is brief, we normalize the long-run discount factor δ to one. Moreover, subjects may be naive about their present bias.[9] Rather than using the correct β, we follow O'Donoghue and Rabin (1999a) and endow subjects with a belief that their future short-run discount factor will be given by $\widehat{\beta} \in [\beta, 1]$. The lower and upper bounds on $\widehat{\beta}$ refer to full sophistication and full naiveté, respectively, while intermediate values correspond to partial naiveté. We note that while the present bias itself may generate an under-investment in exercise relative to one's long-run preferences, naiveté is necessary for subjects to hold systematically biased beliefs (including those that can lead to procrastination). Although the inequality $\widehat{\beta} > \beta$ has been documented in previous research, the literature has far less evidence on its exact magnitude; despite this, it is the magnitude that is critical for policy and welfare analysis.

The second source of bias we consider is "projection bias", whereby subjects do not appreciate the extent to which their future preferences may differ from their current ones as a result of changes to their exercise habit "state". In our setting, this implies that individuals will correctly foresee the direction of the habit-formation process, but may partially or fully "project" their current level of habit onto their future selves.[10] We use the "simple projection bias" defined by Loewenstein, O'Donoghue and Rabin (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

[8]Present bias has been observed in a wide range of contexts, from long-term savings behaviors (Angeletos, Laibson, Repetto, Tobacman and Weinberg 2001) to daily caloric intake (Shapiro 2005). An overview of the literature, with many additional examples, is available in DellaVigna (2009).

[9]Ali (2011) gives a condition under which individuals will learn about their self-control problems over time. That students who have not paid for a gym membership have de facto committed not to attend the gym in the current period places a significant limit on the potential for such learning in our environment.

[10]For example, Read and van Leeuwen (1998) show that people who are currently hungry act as though their future selves will also be relatively hungry, and people who are currently sated act as through their future selves will also be relatively sated.

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.[11]

We define a "p-coupon" as the contingent-payment contract that rewards \$p concurrently with each day the holder attends the gym during an associated period-three week. Let $V_{t,g}(p)$ refer to the valuation of a p-coupon in session $t \in \{pre, post\}$ of a subject in group $g = 0, 1$ (control=0, treated=1). Let $Z^t_{d,g}(p)$ be an indicator for whether a subject in group g actually attends the gym on day $d = 1, ..., 7$ of a week in period t, so that $Z^t_g(p) = \sum_{d=1}^{7} Z^t_{d,g}(p)$ is the number of gym visits during a given week for a subject in group g.

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

If a subject holding a p-coupon 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 $Z^{post}_{d,g}(p) = \mathbb{1} \cdot \{\varepsilon_d > c - p - \beta b - g\eta_i\}$, and $Z^{post}_g(p) = \sum_{d=1}^{7} \mathbb{1} \cdot \{\varepsilon_d > c - p - \beta b - g\eta_i\}$. Denote the unobserved distribution of the daily shock by $F(\varepsilon)$. In expectation, total target-week gym-attendance will be,

$$\sum_{d=1}^{7} \Pr\left(Z^{post}_{d,g}(p) = 1\right) = 7 \times \int_{c-\beta b-g\eta_i-p}^{\infty} dF(\varepsilon) \tag{1}$$

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

$$\sum_{d=1}^{7} \Pr\left(Z^{post}_{d,g}(p) = 1\right) - \sum_{d=1}^{7} \Pr\left(Z^{pre}_{d,g}(p) = 1\right) = 7 \times \int_{c-\beta b-g\eta_i-p}^{c-\beta b-p} dF(\varepsilon). \tag{2}$$

The perceived probability of target-week gym-attendance, from the perspective of

[11]Because of the linearity embedded in our model, the simple projection bias of Loewenstein, O'Donoghue and Rabin (2003) is indistinguishable from subjects systematically holding the incorrect belief that the value of the gym habit is $(1-\alpha)$ times the actual habit value. This ambiguity could in principle be resolved by a modification of our experiment that shocked people out of the habit rather than into it.

any previous period, 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 $\varepsilon_d > c - p - \widehat{\beta}b - g(1-\alpha)\eta_i$. Thus the subject's ex-ante prediction of her total utility for the target-week, given that she holds a p-coupon, is,

$$7 \times \int_{c-\widehat{\beta}b-g(1-\alpha)\eta_i-p}^{\infty} (p + b + g\,(1-\alpha)\,\eta_i - c + \varepsilon)\ dF(\varepsilon). \tag{3}$$

Setting $p = 0$ gives us the perceived utility without a coupon. The perceived value of the p-coupon is simply the difference between expected utility with a p-coupon and expected utility without. In the pre-treatment elicitation session, this is:

$$V_{pre,g}(p) = \left[7 \times \int_{c-\widehat{\beta}b-g(1-\alpha)\eta_i-p}^{\infty} p\,dF(\varepsilon)\right] + \left[7 \times \int_{c-\widehat{\beta}b-g(1-\alpha)\eta_i-p}^{c-\widehat{\beta}b-g(1-\alpha)\eta_i} (b + g(1-\alpha)\eta_i - c + \varepsilon)\,dF(\varepsilon)\right]. \tag{4}$$

And in the post-treatment elicitation session, when the full habit-formation effect is known to the subject, it is:

$$V_{post,g}(p) = \left[7 \times \int_{c-\widehat{\beta}b-g\eta_i-p}^{\infty} p\,dF(\varepsilon)\right] + \left[7 \times \int_{c-\widehat{\beta}b-g\eta_i-p}^{c-\widehat{\beta}b-g\eta_i} (b + g\eta_i - c + \varepsilon)\,dF(\varepsilon)\right]. \tag{5}$$

The first term in both (4) and (5) is the expected redemption value of the coupon, which is always weakly positive. We note that the face value incorporates an estimate of the behavioral response to the p-coupon's implicit subsidy. The second term is the subject's valuation of that behavioral change that results from holding the coupon, which we will call the "commitment 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 commitment value is negative, as the subject believes that the p-coupon will make her attend the gym at times when she would otherwise have preferred not to. If the subject believes that she will have self-control problems in the target week, then the commitment 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 have foregone.

The total ex-ante value of the p-coupon is always non-negative. Intuitively, the endogeneity of the attendance decision gives positive option value to even a p-coupon one does not expect to use. We prove in appendix A.1 that this holds for subjects with both present-biased and projection-biased preferences.

We next turn to our reduced-form test for projection bias. Suppose we compare the difference-in-differences in valuations of a p-coupon for the treated and control subjects, before and after the intervention. For subjects with no projection bias, regardless of their level of present bias, we would expect this difference-in-differences to be zero. Dividing by 7 for simplicity, this double difference is given by:

$$\frac{[\bar{V}_{post,1} - \bar{V}_{pre,1}] - [\bar{V}_{post,0} - \bar{V}_{pre,0}]}{7} = \pi \cdot \int_{c-\widehat{\beta}b-\bar{\eta}-p}^{c-\widehat{\beta}b-(1-\alpha)\bar{\eta}-p} p\, dF(\varepsilon) \tag{6}$$
$$+ \pi \cdot \left[\int_{c-\widehat{\beta}b-\bar{\eta}-p}^{c-\widehat{\beta}b-\bar{\eta}} (b+\bar{\eta}-c+\varepsilon)\, dF(\varepsilon) - \int_{c-\widehat{\beta}b-(1-\alpha)\bar{\eta}-p}^{c-\widehat{\beta}b-(1-\alpha)\bar{\eta}} (b+(1-\alpha)\bar{\eta}-c+\varepsilon)\, dF(\varepsilon) \right]$$

The first term in (6) is the more intuitive one, namely the misprediction of gym attendance caused by projection bias. It is weakly positive for projection-biased agents—strictly when integrating within the support of $F(\varepsilon)$—and identically zero for agents without projection bias, regardless of self-control problems. The latter two terms, however, reflect the change in perceived incentive value from before the treatment to after. While this term is still zero in the absence of projection bias, for a projection-biased subject with a given value of $\widehat{\beta}$, the difference in incentive values depends exclusively on the distribution of ε_d and could take on either positive or negative values.[12] The overall effect is therefore ambiguous, but still provides a test

[12]This result follows from the unknown density $F(\varepsilon)$, which may differ in the regions around the threshold values of ε for the actual habit and the projection-biased predicted habit. If, for example, a subject believed that the coupon would make her much more likely to attend the gym if the habit were $(1-\alpha)\bar{\eta}$ but would be entirely infra-marginal with a habit of $\bar{\eta}$, this latter difference would be negative. The opposite case would yield a positive difference. A noteworthy special case occurs when ε has constant density over the full region traced by (6). In this case, there is no relative change in the perceived commitment value and the term in brackets reduces to zero.

inasmuch as any observed value of the double-difference that is significantly different from zero indicates projection bias. A more powerful test uses only un-incentivized predictions, and therefore reflects only the first term of (6) without incorporating the change in commitment value. We present both in Section 4.

Finally, we note that this difference-in-differences test is designed to be robust to any un-modeled biases that affect both groups equally. Suppose, for instance, that both control and treated subjects have an additional bias in their ex-ante valuations: $\tilde{V}_{pre,g} = \bar{V}_{pre,g} - \zeta$. For example, all students may begin the academic year with an overly optimistic view about their free time during the semester. Indeed, we will find evidence in Section 4 that is consistent with this. But such a secular drift cancels out in Equation (6). Any difference in how the two groups' beliefs change can only be accounted for by a habit-contingent factor; in other words, by projection bias.[13]

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.[14] To be confident of moving subjects from an unhabituated state to a habituated state, we screened for subjects who self-reported that they had not ever regularly attended any fitness facility.[15] 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 treated and control subjects were asked to complete a questionnaire, and were then given an offer

[13]Because we define the habit in our model as the effect of our experimental treatment, it is still possible that it is not preferences towards exercise that change but some other (possibly unmodeled) dimension of preference or belief. Whatever its precise nature, however, equation (6) and our later structural model will estimate the fraction of the change that treated subjects do not initially account for. Thus somewhat surprisingly, we are more confident in having identified people's biases than their underlying preferences.

[14]Due to attrition and missing covariates, our final sample includes 54 treated subjects and 57 control subjects. Details of the sample appear in appendix A.2.

[15]Our screening mechanism is described in appendix A.3. Using a sample of students who did not regularly attend the gym also increases the power of our habit-formation test, as Charness and Gneezy found the treatment effect to be concentrated among subjects whose average pre-treatment attendance was less than 1 visit per week. While we have some self-reported measures of outside activity, it is possible that this sample may have substituted gym exercise for non-gym exercise. This would certainly affect the welfare implications of the study, but would not affect our estimates of either bias.

of \$25 to attend the gym once during the following week.[16] 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.[17]

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. This difference was intended to limit the potential for procrastination so that naive present-biased subjects in the treated group would be more likely to meet the eight-visit threshold, although our compliance rate was not distinguishable from the less-restrictive design. Our treatment-month offer was announced at the initial meeting in order to provide treated subjects with a week to contemplate the idea of going to the gym twice weekly for a month before making predictions, and to avoid introducing systematic bias into predictions. Waiting a week after treatment subjects learn they will earn \$100 avoids the possibility that subjects would be risk-seeking with this windfall, but does not affect their actual gym attendance decisions.

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 one week, to ensure 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 1.

[16]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.

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

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. This attendance data was based on recorded ID-card swipes required for gym entry.[18]

3.1 Elicitation procedures

To elicit predictions of target-week gym attendance we created what we call a "p-coupon", which is a contingent-payment contract 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 multiple price listing mechanism to elicit subjects' valuations for p-coupons of various values with various target weeks.[19] A sample p-coupon is included in Appendix A.4, 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 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.[20] Because at most one p-coupon would be awarded as part of the elicitation, subjects' valuations are not confounded by uncertainty about how many p-coupons they would receive.

Subjects' willingness to pay for a coupon that pays out as a function of their future behavior will of course not be based entirely on their underlying predictions.

[18]Because swipes were necessary to enter the gym but not exit, we cannot always determine the length of a visit. In some cases where a person swiped multiple times, e.g. to enter the locker room and the pool, we can form a lower bound on the length of the visit. We acknowledge that some of the recorded swipes during the treatment month may be subjects swiping to receive the reward but not exercising. However, there is no reason to continue to engage in such false-visit behavior in the post-treatment period, when the incentives to attend are removed. To the extent that some people did swipe without exercising, our estimates in Section 4 reflect a lower bound on the treatment effect of the program. One may also interpret some proportion of the subjects who do not form a habit (i.e. for whom $\eta_i = 0$) as being in the category of false swipers.

[19]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 A.4.

[20]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.

Risk-aversion alone implies we would only observe subjects' certainty equivalents, even for an exogenous event.[21] 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. There is an important distinction to be made between the effect of the incentive on the underlying behavior—the "incentive effect"—and the distortion the incentive induces for a subject bidding on their own p-coupon. The incentive effect alone would still allow for a direct comparison between p-coupon valuations and attendance, as both incorporate the behavioral response equally. The latter effect, which we call the "commitment value", drives a wedge between the face value of a coupon and a subject's personal value from holding it, and complicates such a comparison. The direction of the distortion depends on the type of present bias a subject experiences. In general, time-consistent and naive subjects will view the incentive as creating a costly distortion and lower their valuations relative to the expected face value. Sophisticated subjects, however, will value the p-coupon as a commitment device for their future selves, and will raise their valuation over the expected face value. We use this later case in naming the effect, but note that this "commitment value" may increase or decrease subjects' demand for a p-coupon, and 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 un-incentivized *predictions* of hypothetical future attendance under the same set of *attendance* incentives as in the incentivized task.[22] This un-incentivized 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 un-incentivized prediction of behavior in the absence of any effect of attendance incentives. More importantly, by comparing the

[21]An alternative design which would have allowed us to sidestep assumptions about 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 was significantly easier for our subjects to understand, and requires us to assume linearity of utility only over very modest stakes.

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

un-incentivized prediction in a week with a p-coupon with the valuation of the corresponding p-coupon, we have an estimate of the commitment value provided by that coupon. We use this difference in one of the structural model specifications to calibrate the extent of naiveté, and thereby leverage what would otherwise be a confound to provide additional identification.

Subjects went through 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.[23] 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 un-incentivized responses during the elicitation tasks. We discuss compliance with the treatment incentive, attrition, and our randomization procedure in Appendix A.5.

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 notable that our sample of non gym-attenders were so easily induced to visit the gym eight times, which underscores the power of standard economic incentives.

[23] We used a block-random design to assign coupons to 12 control and 12 treated subjects in each of the five target weeks. Within each treatment group, 15 subjects received a \$1 coupon, 15 received a \$2 coupon, 15 received a \$3 coupon, 10 received a \$5 coupon, and 5 received a \$7 coupon.

4.1 Habit formation

Figure 2 shows average weekly attendance for the treated and control groups over the duration of the study period.[24] 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 the semester break, the treated group consistently attends the gym more than the control group. In the four months after the semester break the difference is greatly diminished.

We estimate a linear difference-in-differences panel regression model to determine if these patterns are statistically significant. Each observation in the panel is a specific individual on a specific week of the study, and we therefore cluster all standard errors throughout by subject.[25] We regress weekly gym attendance on a treated-group dummy, 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. We also control for individual characteristics, including demographics and demand shifters such as travel time to the gym.[26] The results of this regression appear in the first column of Table 1.

The coefficient on the treated-group dummy indicates no statistically significant difference in pre-treatment gym attendance between treated and control subjects. The coefficient on the interaction of the treated-group and treatment-period dummies, roughly the product of the twice-weekly incentive target and the 80% compliance rate, reiterates that the treatment-incentive was effective. 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 0.256 additional visits per week for the immediate post-treatment period, representing approximately a doubling of average attendance in our sample. While significantly different from zero, it is also misleadingly small as it combines a large mass of unaffected subjects with a smaller

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

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

[26]When we omit the individual characteristics, the main effect is no longer significant at standard levels. A Hausman test between the two specifications obtains a p-value of 0.051, suggesting that we may be correcting for some lumpiness in our randomization. Following Gelbach (2009), we find that demographic covariates, naiveté proxies, and attitudes about gym attendance explain three-quarters of the change in the coefficient, but do not enter significantly.

mass of subjects for whom the habit formation appears substantial.

Because not all subjects in the treatment group made the requisite eight visits to the gym, the results in the first column represent the intention-to-treat effect. 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, controlling for observable differences between compliers and non-compliers. This analysis assumes there is no effect on subjects who did not meet the 8-visit threshold, which is not implausible given the average of only two visits during the treatment period for such subjects. These results are reported in the second column of Table 1, where 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 see no statistically significant difference between the groups, with a later post-treatment ITT effect of 0.045 visits per week, and an ATT effect of 0.061. Neither effect is significant, suggesting that the habit induced by four weeks of exogenous gym attendance was dissipated by a similar period of quasi-exogenous non-attendance. It is worth noting that this long-run decay supports our interpretation of the short-run effect as habit formation over alternatives such as learning, for which one would not expect to find decay. While such a result may disappoint policymakers, this failure to maintain the habit is entirely compatible with the model and, more importantly, suggests what sort of interventions might be suitable to help with long-run maintenance.

To compare our results with the results from Charness and Gneezy's first study we ran the same regression on their data, the results of which constitute the final column of Table 1.[27] 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-one 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.

To get a better picture of the treatment effect in the immediate post-treatment pe-

[27]This specification differs from the one they report, which uses pre- and post-treatment averages rather than the full panel of weeks.

riod, we can also compare the empirical CDFs of average post-treatment attendance in the treated and control groups. Because we had imperfect compliance among treated subjects in our experiment, we can use a nonparametric randomization inference method to estimate the distribution of treatment effects among compliers. Following Frandsen (2010), we can estimate exact finite-sample confidence intervals for the distributions of post-treatment attendance for compliers. Figure 5 shows the 90% confidence intervals for immediate post-treatment attendance, by treatment assignment.[28] We find a large mass of subjects—roughly 60%—did not go to the gym prior to the experiment and did not go to the gym after. The treatment effect is instead concentrated in the upper portion of the distribution, with a significant difference above the 85th percentile. A quantile regression (not shown) confirms a significant treatment effect over this range.

Another straightforward test is to compare individual treated subjects' post-treatment attendance to their predicted attendance had they been in the control group. We imputed this 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, and while it clearly includes noise (for both controls and treated subjects) it is the best prediction of post-treatment attendance in the absence of any intervention. 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 their predicted attendance. Shown in Figure 4, this applies to 8 of 54 treated subjects and 3 of 57 control subjects, the latter serving as an estimate of false positives due to noise. A one-sided test of equal proportions rejects the null that there are more habit formers in the control group at a p-value of 0.046.[29]

It is not surprising that we find heterogeneity in our treatment effect. One possibility, which we cannot fully address, is that some subjects in the treated group merely swiped their ID cards at the gym but did not actually exercise. We would not expect such subjects to form any habit, and our estimates of the treatment effect

[28] That is, we compare the distribution of post-treatment attendance among compliers in the treatment group with that of compliers assigned to the control group. Given that no control subject met the 8-visit threshold during the treatment month, we restrict the proportion of "always-takers" to be zero and therefore estimate a population of "compliers" and "never-takers".

[29] Relaxing the threshold for habit formation to 0.5 visits/week, which introduces more noise, also yields a significant result at a p-value of 0.066.

in Table 1 would be biased towards zero by their presence. An alternative explanation would be that some subjects would have formed a habit, but our month-long treatment was too short for them to do so. This interpretation is consistent with recent findings such as Lally, van Jaarsveld, Potts and Wardle (2010), who estimate a range of 18 to 254 days in their subjects' time for habit formation for various tasks. Finally, it is possible that some subjects simply do not find exercising at the gym to be habit-forming. All three interpretations are captured by the fraction $(1 - \pi)$ of treated subjects who receive a habit value of $\eta_i = 0$ from the treatment.

4.2 Predictions

We next turn our attention to subjects' predictions. Figure 5 shows predicted versus actual gym attendance, first for the weeks that subjects actually received a p-coupon in the giveaway at the end of the experiment, and then 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.[30] Finally, we separate subjects' predictions by when they were elicited. We show only subjects' un-incentivized predictions for clarity, but Tables 2 and 3 confirm that incentivized and un-incentivized predictions follow similar patterns. We find in general that un-incentivized predictions are greater than the normalized valuations, indicating that the commitment value of the p-coupons is typically negative for subjects.

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 2 tests differences 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 valuations.[31]

[30] We group all non-zero values of p-coupon together here for simplicity—the effect of each separate p-coupon value is shown in Table 3.

[31] We include subjects' valuations for a p-coupon divided by the subsidy, which clearly does not account for the commitment value of the p-coupon. While we will exploit this in Section 4.3, we present the simple inferences here and caution readers not to interpret the valuations literally as

The second and third refer to their un-incentivized 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 attendance in weeks with no p-coupon, suggesting that the overprediction is not driven by the p-coupon incentives. The systematic pattern of mis-prediction is a central result, and on its basis 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.[32] Furthermore, while it does not rule them out, this result indicates that our data cannot be explained solely by other classes of self control models such as the "temptation utility" of Gul and Pesendorfer (2001, 2004) which embed rational expectations.

In Table 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.[33] 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 indicates the upward-sloping supply curve for exercise that one would expect. The second and third columns regress normalized coupon valuations and un-incentivized predictions on the same p-coupon dummies, plus a dummy for the post-treatment elicitation session. Interestingly, subjects appear to predict the slope of their labor-supply curve relatively accurately, despite consistently over-predicting its intercept.

The session dummy implies that, between the first and second elicitation sessions, 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

predictions. For this section, we prefer the un-incentivized predictions which are directly measuring predicted attendance.

[32]These two cases require rational expectations, which appears to be strongly violated here. It is possible that other un-modeled biases could cause the mis-prediction in Table 2, but we feel that many such alternatives would, unlike naive present bias, be ad hoc for this result.

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

of the target weeks is greater in the post-treatment elicitation session. 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 coupon valuations and un-incentivized predictions, still show a substantial decrease in over-prediction over time. Such a secular drift in misprediction suggests that subjects may begin the semester with overly optimistic beliefs about their amount of free time, and grow more realistic.[34] It is worth noting that typical models of learning would suggest that treated subjects, who have more experience with the gym, should learn about their self control by at least as much as control subjects do. Differences in this learning would therefore bias against our test of projection bias, which uses the relative increase in predicted attendance among treated subjects.

Our test for projection bias is implemented with a difference-in-differences regression of predictions, shown in Table 4. While the GMM estimator in the next section will accommodate the complex effect of projection bias at different values of p-coupons, we restrict our attention here to the simplest cases. We include our standard set of control variables in both columns, as well as a measure of risk aversion.[35] The first column utilizes un-incentivized predictions about unsubsidized weeks, so as not to conflate mistakes about habit formation with mistakes about responsiveness to gym-attendance subsidies. We find that treated subjects revise their predictions upwards by 0.459 visits per week relative to control subjects. This revision is in fact larger than the treatment effect itself, although not significantly. We therefore interpret this as consistent with full projection bias.[36] We replicate this using subjects' valuations of p-coupons in the second column of Table 4. Because no incentives can be obtained for unsubsidized weeks, we instead look at predictions for the smallest subsidy (p=$1), for which any distortion is minimized. We find a similar upward revision among treated subjects of 0.324 visits per week, but the coefficient is no longer significant. Because the change in perceived commitment value may offset learning

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

[35]We obtain subjects' risk aversion from their choices over hypothetical lotteries as in Holt and Laury (2002).

[36]A habituation model with finer gradations than the binary form we use here would in fact be consistent with this double-difference exceeding the actual habit, if post-treatment subjects are temporarily above their steady-state level of habituation.

about one's habit, however, we prefer the first column as a more powerful test of projection bias.

4.3 Structural Estimation

Let $\bar{Z}_{t,g}(p)$ denote the average attendance of group g during period $t \in \{pre, post\}$, when holding coupon p. Let $g \in \{C, T\}$ denote control and treated subjects, respectively. Let $\bar{Y}_{t,g}(p)$ correspond to the analogous average un-incentivized predictions for group g during the pre- and post-treatment elicitation sessions. Let the corresponding observations for an individual i be denoted by $Z_t^i(p)$ and $Y_t^i(p)$. For any individual i, let $p_{i,w}$ denote the w-th p-coupon she was offered in the elicitation session ($w \in \{1, 2, 3, 4\}$) and let $V_{w,i}^t(p_{i,w})$ denote her valuations of these coupons. Then we use the following moments:

$$\begin{aligned}
\bar{Z}_{post,T}(0) &= 7 \cdot \Big[\pi \left(1 - F\left(c - \beta b - \eta; \sigma_\varepsilon\right)\right) \\
&\qquad + (1-\pi)(1 - F(c - \beta b; \sigma_\varepsilon))\Big] && (7) \\
\bar{Z}_{post,C}(0) &= 7 \cdot \Big[1 - F(c - \beta b; \sigma_\varepsilon)\Big] && (8) \\
\bar{Z}_{post,T}(p > 0) &= 7 \cdot \Big[\pi \sum_{i \in \mathcal{T}} \sum_{w=1}^{4} (1 - F(c - \beta b - \eta - p_{iw}; \sigma_\varepsilon)) \\
&\qquad + (1-\pi) \sum_{i=1}^{\mathcal{T}} \sum_{w=1}^{4} (1 - F(c - \beta b - p_{iw}; \sigma_\varepsilon))\Big] && (9) \\
\bar{Z}_{pre,C \cup T}(0) &= 7 \cdot \Big[1 - F(c - \beta b; \sigma_\varepsilon)\Big] && (10) \\
\bar{Y}_{pre,T}(0) &= 7 \cdot \Big[\pi \cdot (1 - F(c - \widehat{\beta} b - (1-\alpha)\eta; \sigma_\varepsilon)) \\
&\qquad + (1-\pi) \cdot (1 - F(c - \widehat{\beta} b; \sigma_\varepsilon))\Big] && (11) \\
\bar{Y}_{pre,C}(0) &= 7 \cdot \Big[1 - F(c - \widehat{\beta} b; \sigma_\varepsilon)\Big] && (12) \\
\bar{Y}_{post,C}(0) &= 7 \cdot \Big[1 - F(c - \widehat{\beta} b; \sigma_\varepsilon)\Big] && (13) \\
\sum_{g_i = T} \mathbb{1}\{\bar{Z}_{pre,i}(0) < \bar{Z}_{post,i}(0)\} &= \sum_{g_i = 1} \Big(\pi + \frac{1}{2}(1-\pi)\Big) && (14)
\end{aligned}$$

The habit-formation effect itself is identified by equations (7) and (8), relying on the assumption that the habit is the only systematic driver of post-treatment differences in attendance among control and treated subjects in un-incentivized weeks. The rest of the attendance parameters are identified by (9) and (10), which establish the responsiveness to coupons and the pre-treatment baseline. Equation (11) mirrors (7), but uses pre-treatment predictions of attendance rather than actual attendance in order to identify the ex-ante expectations of the habit value. Finally, we use control subjects' pre- and post-treatment predictions to identify the general over-confidence driven by naiveté about self-control. For these three sets of expectations, we use un-incentivized weeks to avoid embedding an assumption that subjects correctly predict their response to small monetary incentives.

We introduce (14) to estimate the fraction of subjects developing a strictly positive habit. As the number of pre- and post-treatment periods grows large, the probability that a subject with a positive habit will have higher average attendance in the post-treatment period converges to 1. The corresponding probability for a subject who did not form a habit converges to 0.5. We use this limit as our moment for estimation, but note that (14) gives a conservative estimate of π due to the finite pre- and post-period samples. Calibrations at the estimated coefficients suggest the approximation is good.

In addition, in Model 2 we use the difference between coupon valuations and the valuation implicit in the un-incentivized predictions to gain an additional moment, which will allow us to identify naiveté. Letting $\mathbb{G}_i = \mathbb{1}\cdot\{g_i = T\}$ and $\mathbb{T} = \mathbb{1}\cdot\{t = post\}$ for convenience, we write:

$$\sum_{t\in\{pre,post\}}\sum_i \frac{1}{4}\sum_{w=1}^{4}\left(V^t_{w,i}(p_{i,w}) - Y^i_t(p_{i,w})\cdot p_{i,w}\right) =$$

$$\sum_{t\in\{1,2\}}\sum_i \frac{1}{4}\sum_{w=1}^{4} 7\cdot\left[\pi\int_{c-\widehat{\beta}b-\mathbb{G}_i(1-\alpha(1-\mathbb{T}))\eta-p_{i,w}}^{c-\widehat{\beta}b-\mathbb{G}_i(1-\alpha(1-\mathbb{T}))\eta}(b+\mathbb{G}_i(1-\alpha(1-\mathbb{T}))\eta - c+\varepsilon)dF(\varepsilon) + (1-\pi)\int_{c-\widehat{\beta}b-p_{i,w}}^{c-\widehat{\beta}b}(b-c+\varepsilon)dF(\varepsilon)\right] \tag{15}$$

While Equation (15) may appear complex, it has a straightforward interpretation. The left-hand side is the average difference between incentivized and un-incentivized predictions, both pre- and post-treatment, across both groups. The right-hand side is the commitment value described in (4) and (5), taking into consideration the different coupons offered to subjects and their different beliefs at the time of each prediction. The linearity that simplified the previous moments does not extend into the distribution of ε, and so we must write out the averages using summations. Because this moment relies strongly on the assumption that the only difference between the un-incentivized and incentivized predictions comes through the commitment value of a p-coupon, we present the results from estimating a model both without (Model 1) and with (Model 2) making use of Equation (15).

The results of estimating the structural models using GMM are presented in table 5. We assume a Type 1 extreme value distribution for ε with a zero mean and scale parameter σ_ε.[37] Panel A presents those parameters estimated directly, while additional parameters derived from these are presented in Panel B. The structural parameters confirm the reduced-form results in Sections 4.1 and 4.2. On average, the immediate utility cost of gym attendance exceeds the discounted future benefits by \$4.71. Naiveté about their future self-control problems causes people to under-estimate their future impatience about gym attendance by \$3.10, however. This corresponds to the significant over-prediction of future attendance relative to actual attendance found in the previous results. This naiveté also explains why un-incentivized predictions lie above the normalized valuations of p-coupons, as subjects underestimate their need for commitment and view the p-coupons as including a costly distortion.

Turning to the habit formation, we find that 32% of treated subjects formed a habit equivalent to a \$2.60 daily gym-attendance subsidy.[38] This is a significant habit—our \$100 treatment would be recouped after only 50 visits. It is still substantially smaller than the net daily cost, however, so that even a habituated subject would not on average enjoy going to the gym. This helps explain why we estimate one-third of subjects forming the habit but far fewer actually regularly going to the gym—many of the habituated subjects still did not receive sufficiently good shocks to push them over the edge. In contrast to the not-inconsequential actual habit, subjects' predicted habit value was trivial. We estimate that subjects expected a habit worth only \$0.16, and the coefficient is not significantly different from zero. By combining the predicted habit with the actual habit, we can estimate a large and highly

[37] As a robustness check, we estimate the same models in Table A.4 using a normally distributed error term.

[38] It is also possible to estimate a model with a homogeneous treatment effect assumption by restricting $\pi = 1$. In this case the overall habit value is \$1.33 (s.e. 0.45), although the overidentifying restrictions are now rejected at p = 0.031.

significant degree of projection bias, $\alpha = 0.94$. That is, while we can strongly reject the null of "no projection bias" ($\alpha = 0$), we cannot reject "complete projection bias" ($\alpha = 1$).

The second column of Table 5 presents the results from incorporating the additional moment restriction on the commitment value of our p-coupons. While the effect on most of the parameters is small, this additional restriction allows us to estimate an additional parameter, $(1 - \widehat{\beta})b$. Because this reduces to zero in the case of complete beta-naiveté, our significant estimate of 1.500 allows us to reject this null. We can combine this estimate with the cost of naiveté (which similarly lets us reject "complete beta-sophistication") in order to estimate the extent of beta-naiveté. We show this in Panel B, finding a tightly estimated $\omega = 0.666$. This means that subjects believe their future discount factor $\widehat{\beta}$ will be a weighted average between 1 and their actual discount factor β, with 2/3 of the weight on 1. We cannot separately estimate the underlying true discount factor, unfortunately, but for most values our estimate would imply an economically meaningful degree of naiveté. If we use the typical value of $\beta = 0.7$ found in other studies, this would imply that subjects hold the partially-naive $\widehat{\beta} = 0.9$. One can easily model a setting in which such beliefs and preferences would generate considerable welfare loss.

Under the same assumptions of $\beta = 0.7$, our estimate for the demand for commitment, $(1 - \hat{\beta})b = \$1.50$, also implies a value of the perceived value of the long-term benefits of gym attendance, $b = \$15$ per day. This is not insubstantial for our student sample, and we expect we would find a higher value among older subjects or subjects with higher incomes. Given the average increase of 0.256 visits per week and the 80% compliance rate, the increased attendance would have to last approximately 20 weeks for the program to break even. This break-even duration threshold was not met in our sample of students, at least in part because of the significant decay over winter break, which came just eight weeks after the intervention. It is possible that a smaller, or differently designed incentive would have been sufficient to obtain compliance for a significant subset of our subjects, which would lower the break-even duration.[39]

Finally, Figure 6 uses the estimated structural parameters to plot the probability of nonzero weekly gym attendance as a function of an individual's habit value. This figure makes clear the link between utility shifters and behavioral changes: in partic-

[39]For example, Volpp et al. (2008) shows that incentivizing healthy behaviors with lottery-style rewards can be particularly cost-effective.

ular, our estimated habit value of \$2.60 increases the probability that a subject will attend the gym at least once in a given week by 47 percentage points. On the one hand, this can be viewed as a large effect. On the other hand, such a habit value is still only expected to lead to gym attendance in 63% of weeks. This is another way to understand why we estimate a larger proportion of "habit formers" than we find evidence of using behavior changes in the data. It also prompts us to expect that the habit effect may have decayed on its own in the absence of the quasi-exogenous break imposed by the semester break, as weeks with insufficient utility shocks gradually cause subjects to de-habituate.[40] More optimistically, however, we note that by our estimation a habit value of \$4.30 would generate a 95% probability of weekly gym attendance, which we view as the steady-state level of habituation. This leaves as an open question for future work whether such a habit can be achieved by varying the treatment, or by complementing the habit with long-term small subsidies.

5 Discussion

Using gym exercise as a field setting, we find evidence of two simultaneous dimensions of misprediction: projection bias, with respect to habit formation; and naiveté, with respect to present bias. We use a large financial incentive to induce a gym-attendance habit in subjects, and find significant heterogeneity in the habit: two-thirds of treated subjects appear unaffected, while one-third of subjects developed a gym-attendance habit equivalent in utility to a \$2.60 per-visit subsidy. Despite this significant habit, subjects did not appear to predict any habit formation ex ante—an effect we interpret through the framework of projection bias. Furthermore, we find that subjects are greatly overoptimistic about their subsequent gym attendance in general, which we interpret as over-optimism about their own self-control. Even in weeks with no p-coupon to complicate the prediction task, subjects over-predict attendance by 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, given their actual attendance, per-visit passes would be a cheaper alternative.[41] Thus, somewhat counterintuitively, we find that projection

[40]Although we have examined this point in our data, we have insufficient power to detect a downward trend in post-treatment attendance in addition to the the semester break effect.

[41]DellaVigna and Malmendier (2006) consider a population purchasing commercial health club memberships, so we cannot claim that our evidence applies to their result.

bias leads subjects to under-value regular exercise, while at the same time naiveté about present bias causes them to over-predict their willingness to engage in it. Using our parameter estimates, including subjects' perceived long-term health benefits, we find that the habit would need to last about 2.5 times longer than it does for our intervention to break even. One should exercise caution in extrapolating results from college students to other populations. The compliance rate, habit-formation effect, decay rate, opportunity cost of time and perceived long-term health benefits might be quite different in a different, perhaps more policy-relevant population.

However, while there is reason to suspect that some of the context-specific parameters of the model may differ across populations, it does not follow that the more fundamental parameters of time preference and biased beliefs would necessarily change, and these parameters have an importance to economic theory and public policy beyond the application of physical-exercise adherence. In particular, that subjects predict essentially none of the changes in their state-dependent preferences can be seen as a refutation of the rational-expectations assumption in the standard "rational addiction" model.[42] Without rational expectations it does not follow that pursuing harmfully addictive substances or behaviors is expected-utility maximizing. Indeed, there is evidence that policies that increase the price of addictive goods may increase welfare.[43]. As we have shown, however, the potential failure of utility maximization is not limited to harmful habit formation. Stigler and Becker (1977) point out, using appreciation of classical music as an example, there are many activities that become less costly or more rewarding over time. More consequential examples may deserve more serious attention. The failure of rational utility maximization as a priori grounds for policy intervention is controversial, but follows in the same way from other failures in the assumptions of the fundamental welfare theorems, such as market power or externalities.[44]

We have identified mistakes which people would themselves recognize as such and which thus lower the normative bar for intervention,[45] but it is critical to note that

[42]Subsequent extensions of the Becker and Murphy (1988) model have allowed imperfect foresight, but maintained rational expectations for ex ante choice optimality.

[43]e.g. Gruber and Mullainathan (2005)

[44]Following Executive Order 13563, Federal agencies have in fact begun using such justifications for regulatory intervention when no standard market failure is present. For example, the FDA recently proposed new food labeling regulations on the basis of "systematic biases in how consumers weigh current or immediate benefits... against future or long-term costs." (76 FR 19221).

[45]See Bernheim and Rangel (2007) for a discussion of the difficulties in applying standard welfare

they would only recognize their errors after the fact. Ex ante, subjects are unaware of the full extent of their biases. This may be seen as paternalism, and attempts to make people better off, even according to their own preferences, may be viewed as unnecessary or even unwelcome. It is therefore particularly important to understand belief parameters such as the degree of beta-naiveté and the degree of projection bias more fully, and across a greater range of subject heterogeneity. Public policy that is only partly informed could easily do more harm than good. For example, our estimate of partial naiveté presents challenges for policies offering "commitment", as the partial awareness may cause consumers to seek out this commitment but then purchase too little of it. Through the direct cost of such commitment and mis-judging the cost of complementary behaviors, the availability of commitment devices can in fact make such consumers worse off. This fact may help explain the limited take-up and limited success rates of some commitment contracts in the field[46], and suggests caution in designing an intervention. It may also explain the limited availability of commercial commitment devices, especially where other biases such as projection bias reinforce the effect of naiveté on limiting demand. More importantly, perhaps, there are many market settings where firms would instead choose to design contracts which exploit partially naive consumers a la Gabaix and Laibson (2006). Indeed, commercial gym contracts are likely to be one such area: the archetypal unlimited monthly attendance plan lowers a consumers' daily cost in a manner that provides weak commitment, which may be valued by consumers who are partly aware of their self-control problems. We find in this paper, however, that the socially optimal gym contract would instead increase the monthly membership fee and provide a negative daily rate for attendance. That firms do not offer this contract suggests that the market finds it more profitable to make use of biases in this setting than to correct them.

A final lesson to be learned from this paper is the importance of considering the interaction of multiple biases, particularly when designing optimal policy. In the case of an "investment good" with positive habit formation such as exercise, the effects of projection bias and naive present bias are superadditive in reducing consumption from its optimal level. The importance of their interaction in designing an optimal

criteria to nonstandard preferences.

[46]For instance, Gine, Karlan and Zinman (2010) find that while a smoking cessation commitment device was overall more effective than typical interventions, fully two-thirds of those taking up the device lost their deposits after two months.

intervention, however, is more subtle. In the case of our experiment, moving subjects into the habituated state was found to lower the perceived "commitment value" of our p-coupons. Although the sign of this effect in general is ambiguous, a policy-maker with this information can engineer the details of timing to improve take-up and willingness to pay for an intervention that provides commitment. In our case, having subjects evaluate an exercise commitment contract prior to habit-formation generates a prediction that the effect of projection bias would in some measure offset the effect of partial naiveté. We hope that our results will encourage researchers and policy-makers to examine other such predictions in designing future mechanisms.

References

Ali, S. Nageeb, "Learning Self-Control," *The Quarterly Journal of Economics*, 2011, *126*, 857–893.

Angeletos, George-Marios, David Laibson, Andrea Repetto, Jeremy Tobacman, and Stephen Weinberg, "The Hyperbolic Consumption Model: Calibration, Simulation, and Empirical Evaluation," *The Journal of Economic Perspectives*, Summer 2001, *15* (3), 47–68.

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.

Bernheim, B. Douglas and Antonio Rangel, "Behavioral Public Economics," in Peter Diamond and Hannu Vartiainen, eds., *Behavioral Economics and Its Applications*, Princeton University Press, 2007.

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

Conlin, Michael, Ted O'Donoghue, and Timothy J. Vogelsang, "Projection Bias in Catalog Orders," *The American Economic Review*, September 2007, *97* (4), 1217–1249.

DellaVigna, Stefano, "Psychology and Economics: Evidence from the Field," *Journal of Economic Literature*, 2009, *47* (2), 315–372.

___ **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.

—— **, John List, and Ulrike Malmendier**, "Testing for Altruism and Social Pressure in Charitable Giving," *The Quarterly Journal of Economics*, Forthcoming 2011.

Dzewaltowski, David, John Noble, and Jeff Shaw, "Physical activity participation: social cognitive theory versus the theories of reasoned action and planned behavior," *Sport Psychology*, December 1990, *12* (4), 388–405.

Frandsen, Brigham, "Randomization Inference on Quantiles Under Imperfect Compliance," *working paper*, 2010.

Gabaix, Xavier and David Laibson, "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets," *The Quarterly Journal of Economics*, 2006, *121* (2), 505–540.

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

Gine, Xavier, Dean Karlan, and Jonathan Zinman, "Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation," *American Economic Journal: Applied Economics*, 2010, *2* (4), 213–235.

Godin, Gaston, "Theories of reasoned action and planned behavior: usefulness for exercise promotion," *Medicine and Science in Sports and Exercise*, November 1994, *26* (11), 1391–1394.

—— **, Pierre Valois, and Linda Lepage**, "The pattern of influence of perceived behavioral control upon exercising behavior: An application of Ajzen's theory of planned behavior," *Journal of Behavioural Medicine*, 1993, *16* (1), 81–102.

Gruber, Jonathan and Sendhil Mullainathan, "Do Cigarette Taxes Make Smokers Happier," *Advances in Economic Analysis and Policy*, 2005, *5* (1).

Gul, Faruk and Wolfgang Pesendorfer, "Temptation and Self-Control," *Econometrica*, November 2001, *69* (6), 1403–1435.

—— **and** —— , "Self-Control and the Theory of Consumption," *Econometrica*, January 2004, *72* (1), 119–158.

Holt, Charles A. and Susan K. Laury, "Risk Aversion and Incentive Effects," *The American Economic Review*, 2002, *92* (5), 1644–1655.

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.

Lally, Phillppa, Cornelia H. M. van Jaarsveld, Henry W. W. Potts, and Jane Wardle, "How are habits formed: Modelling habit formation in the real world," *European Journal of Social Psychology*, 2010, *40* (6), 998–1109.

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

____ **and** ____ , "Addiction and Self Control," in Jon Elster, ed., *Addiction: Entries and Exits*, Russel Sage Foundation, 1999.

Read, Daniel and Barbara van Leeuwen, "Predicting Hunger: The Effects of Appetite and Delay on Choice," *Organizational Behavior and Human Decision Processes*, November 1998, *76* (2), 189–205.

Reynolds, Kim, Joel Killen, Susan Bryson, David Maron, C. Barr Taylor, Nathan Maccoby, and John Farquhar, "Psychosocial predictors of physical activity in adolescents," *Preventive Medicine*, September 1990, *19* (5), 541–551.

Shapiro, Jesse, "Is There a Daily Discount Rate? Evidence From the Food Stamp Nutrition Cycle," *Journal of Public Economics*, 2005, *89* (2), 303–325.

Skiba, Paige and Jeremy Tobacman, "Payday Loans, Uncertainty and Discounting: Explaining Patterns of Borrowing, Repayment, and Default," *Vanderbilt Law and Economics Research Paper No. 08-33*, August 2008.

Valois, Pierre, Raymond Dersharnais, and Gaston Godin, "A comparison of the Fishbein and Ajzen and tbe Triandis attitudinal models for the prediction of exercise intention and behavior," *Journal of Behavioural Medicine*, 1988, *11* (5), 459–472.

Volpp, Kevin, Leslie John, Andrea Troxel, Laurie Norton, Jennifer Fassbener, and George Loewenstein, "Financial IncentiveBased Approaches for Weight Loss," *Journal of the American Medical Association*, 2008, *300* (22), 2631–2637.

Figure 1: Our Experimental Design

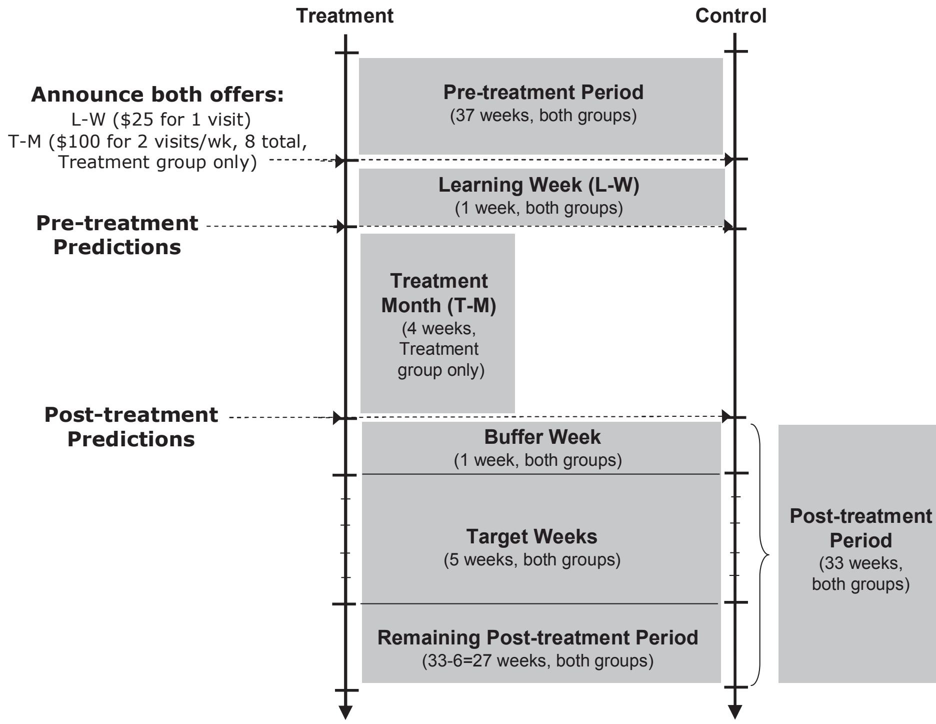
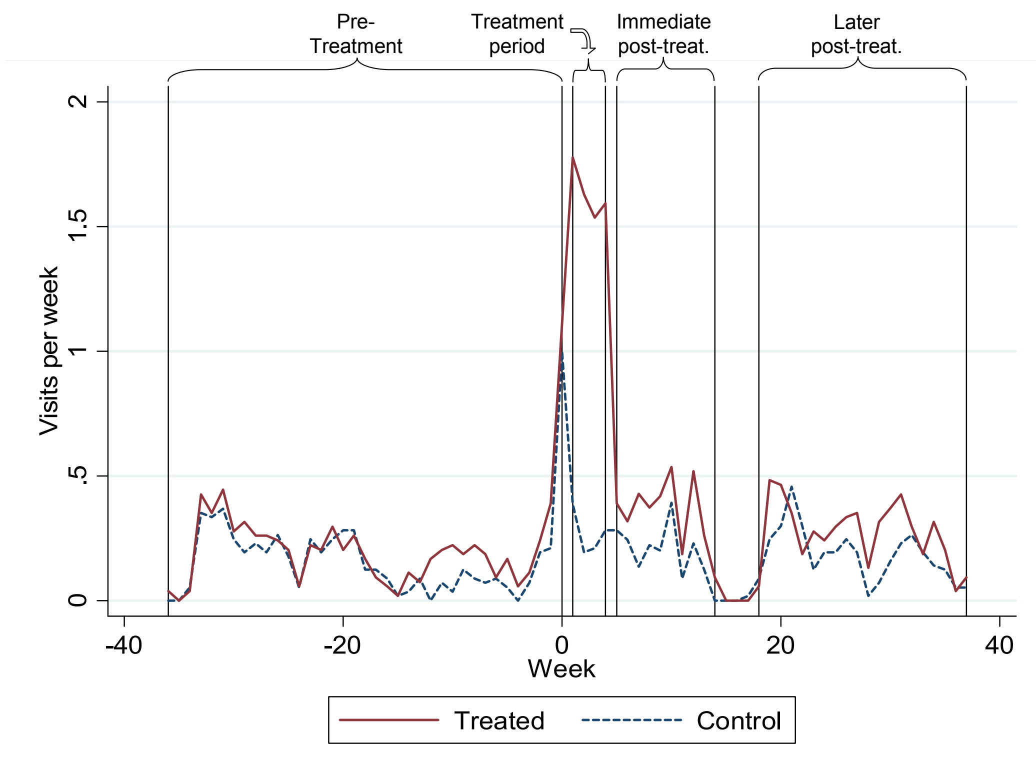


Figure 2: Gym Attendance



Notes: Average weekly gym attendance, by treatment group status. Weeks in which a subject received a p-coupon for attendance are omitted from this figure.

Table 1: Habit Formation: Regression of average weekly attendance.

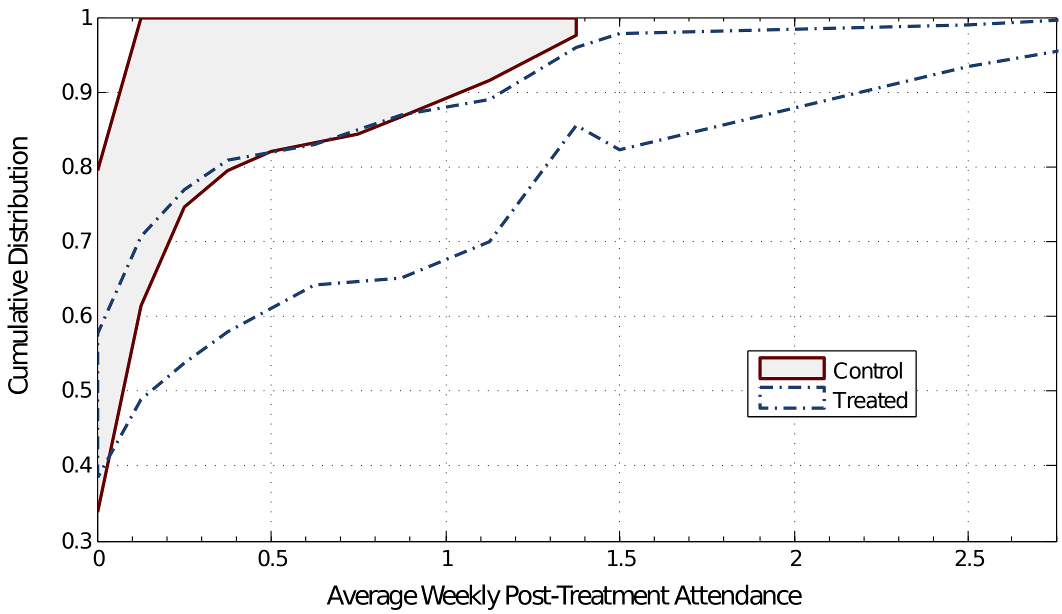
	(1)	(2)	(Charness & Gneezy)
Treated	0.045		-0.100
	(0.057)		(0.196)
			[0.477][a]
Treatment Period X Treated	1.209***		1.275***
	(0.150)		(0.181)
			[0.780][a]
Imm. Post-Treatment X Treated[b]	0.256**		0.585***
	(0.122)		(0.217)
			[0.186][a]
Later Post-Treatment x Treated[b]	0.045		–
	(0.098)		
Complied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

Notes: Observations of weekly attendance at the subject-week level. Robust standard errors in parentheses, clustered by individual. * significant at 10%; ** significant at 5%; *** significant at 1%.

[a]Terms in square brackets are p-values from a Chow test of equal coefficients between our sample (column 1) and Charness and Gneezy (2009)'s sample.

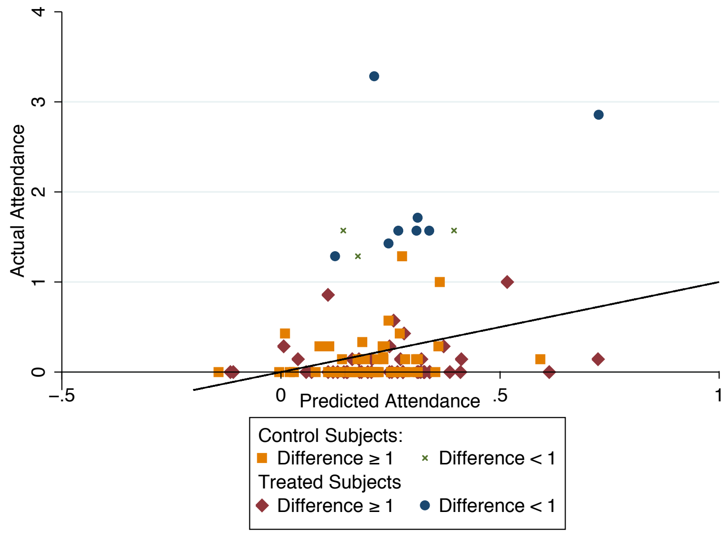
[b]"Immediate" refers to the 8 weeks following the intervention (excluding the "dead week" for columns (1) and (2). "Later" refers to the 19 weeks of observations in the following semester (excluding the semester break).

Figure 3: Cumulative Distribution of Immediate Post-Treatment Attendance



Notes: Cumulative distribution functions of post-treatment attendance among compliers in control and treated groups, based on finite-sample randomization inference with imperfect compliance. 90% confidence intervals shown; corresponding interval for fraction of compliers is [0.7083, 0.8083]; fraction of always-takers is constrained to zero. Weeks in which subjects received a p-coupon are omitted.

Figure 4: Actual vs. Predicted Post-Treatment Attendance



Notes: Average weekly gym attendance in the 8 weeks following the treatment month plotted against predicted attendance conditional on receiving no treatment. Weeks in which a subject received a p-coupon for attendance are omitted from this calculation. Predicted attendance is based on an OLS regression of attendance on week dummies and covariates using control group data for all weeks and treated group data for the pre-treatment period.

Figure 5: Predicted versus Actual Attendance

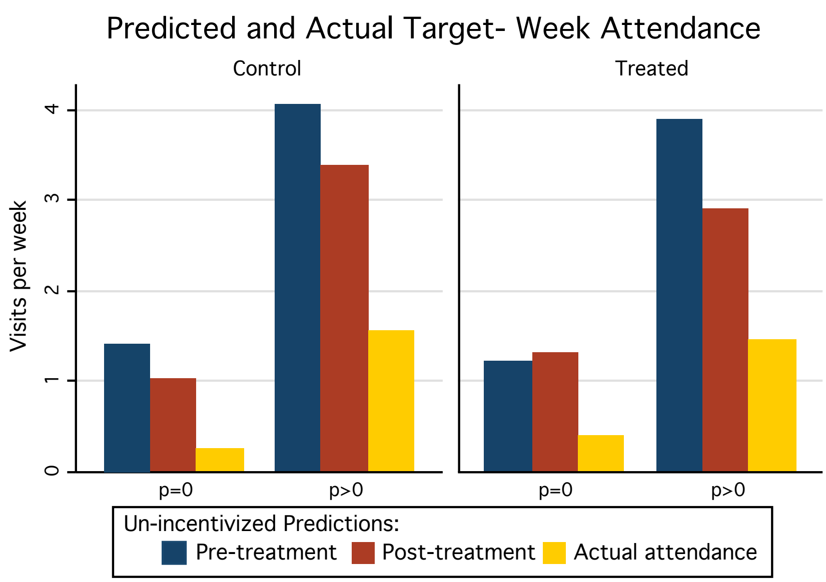


Table 2: Misprediction of attendance

	Control group			Treatment group		
	Coupon Value $p > 0$	Un-Incent'd $p > 0$	Un-Incent'd $p = 0$	Coupon Value $p > 0$	Un-Incent'd $p > 0$	Un-Incent'd $p = 0$
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

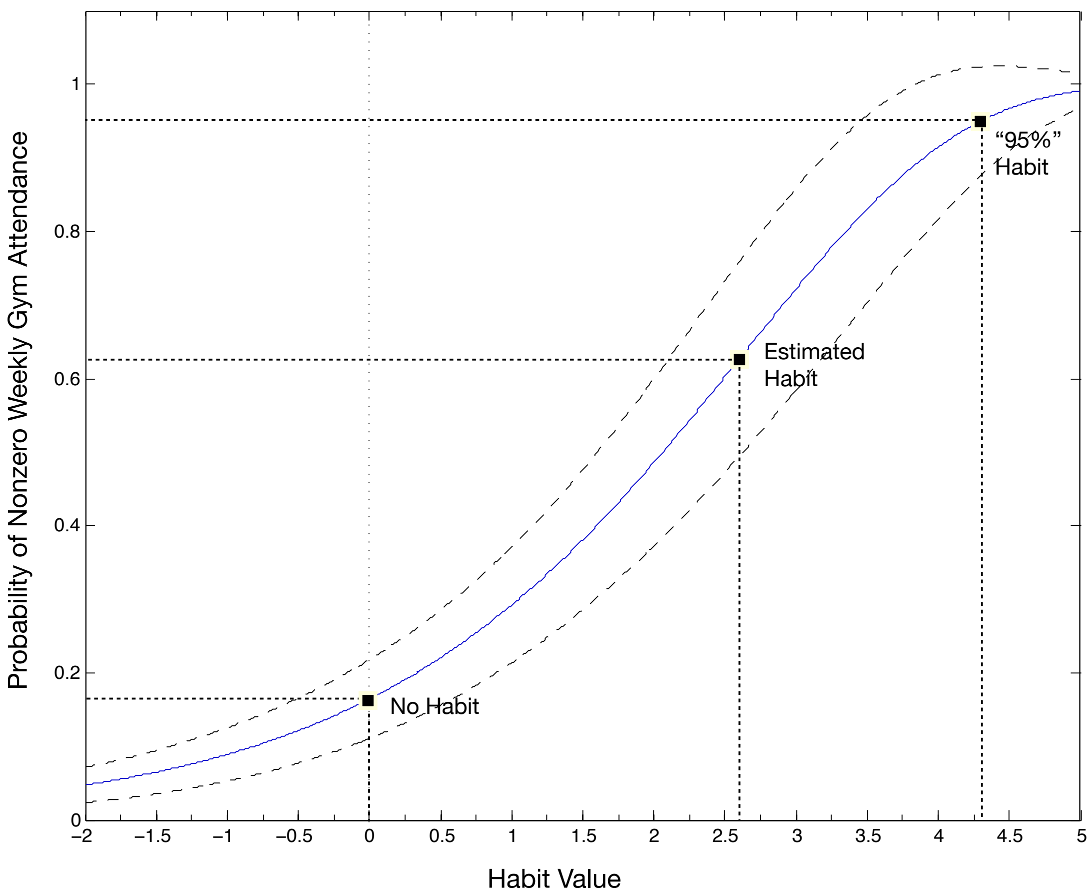
Notes: Coupon value refers to the average valuation of a p-coupon normalized by its subsidy, and includes only observations for the week a subject actually received a p-coupon. Un-incentivized refers to subjects' direct predictions, and is separated into this subsidized week and the unincentivized week for which subjects were only asked to make predictions without a p-coupon.

Table 3: Predictions: Delay versus Session Effects

	(1) Actual	(2) Coupon Value	(3) Un-incentivized	(4) Coupon Value	(5) Un-incentivized
Post-Treatment		-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: Observations are at the subject-week level. Coupon value refers to the average valuation of a p-coupon normalized by its subsidy, and includes only target weeks associated with a non-zero subsidy. Un-incentivized refers to subjects' direct predictions, and includes all target week predictions. Robust standard errors in parentheses, clustered by individual. * significant at 10%; ** significant at 5%; *** significant at 1%. p = $7 is the omitted category.

Figure 6: Simulated Weekly Attendance



Notes: Simulation of the probability of observing a non-zero weekly attendance as a function of the habit value, based on Model 1 in Table 5. Dashed lines indicate the 95% confidence interval.

Table 4: Difference-in-differences in Predictions

	Unincentivized ($0)	Incentivized ($1)
Post-Trmt X Treated	0.459**	0.324
	(0.228)	(0.317)
Post-Trmt	-0.397**	-0.824***
	(0.171)	(0.194)
Treated	-0.280	-0.043
	(0.233)	(0.351)
R-squared	0.255	0.3753
Num clusters	107	111

Notes: Robust standard errors in parentheses, clustered by individual. * significant at 10%; ** significant at 5%, *** significant at 1%. The dependent variable in column 1 is subjects' subjective predictions for weeks in which they received no p-coupon, and in column 2 is subjects' valuations of p-coupons for p=$1. Both columns include our standard vector of controls, as well as a measure of subjects' risk-aversion.

Table 5: GMM Parameters

Name	Parameter	Model 1	Model 2
Panel A: Directly Estimated Parameters			
Net daily cost	$C - \beta b$	4.713***	4.582***
		(0.417)	(0.421)
Cost of naivete	$(\widehat{\beta} - \beta)b$	3.099***	2.993***
		(0.289)	(0.290)
Habit value	η	2.602***	2.618***
		(0.733)	(0.738)
Predicted habit value	$(1 - \alpha)\eta$	0.160	0.226
		(0.747)	(0.757)
Probability of habituation	π	0.320**	0.306**
		(0.133)	(0.136)
Demand for commitment	$(1 - \widehat{\beta})b$	–	1.500***
			(0.250)
Scale parameter, daily shock	σ	1.528***	1.490***
		(0.148)	(0.150)
Panel B: Extended Parameters			
Degree of projection bias	α	0.939***	0.914***
		(0.285)	(0.286)
Degree of beta-naivete	ω	–	0.666***
			(0.037)

Notes: The daily shock, ϵ, is drawn from a mean-zero type-1 extreme value distribution. Standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. Parameters in Panel B are calculated by transformations of parameters in Panel A, with standard errors implied by the delta rule. Model 2 includes an additional moment restriction on the difference between unincentivized predictions and p-coupon valuations.

A For Online Publication

A.1 Value of a p-coupon

The ex-ante value of a p-coupon is

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

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

$$\int\limits_{c-\widehat{\beta}b-g\cdot\eta-P}^{c-\widehat{\beta}b-g\cdot\eta} dF(\varepsilon)\cdot\left[(1-\widehat{\beta})b - g\cdot\eta - P\right].$$

Thus, dividing by 7 for notational convenience, and noting that by assumption $b \geq 0$:

$$\begin{aligned}
\frac{X_2^C}{7} &= \frac{X_6^C}{7} > \int\limits_{c-\widehat{\beta}b-g\cdot\eta-P}^{\infty} P\, dF(\varepsilon) + \int\limits_{c-\widehat{\beta}b-g\cdot\eta-P}^{c-\widehat{\beta}b-g\cdot\eta} dF(\varepsilon)\cdot\left[(1-\widehat{\beta})b - g\cdot\eta - P\right] \\
&= \int\limits_{c-\widehat{\beta}b-g\cdot\eta-P}^{c-\widehat{\beta}b-g\cdot\eta} P\, dF(\varepsilon) + \int\limits_{c-\widehat{\beta}b-g\cdot\eta}^{\infty} P\, dF(\varepsilon) - \int\limits_{c-\widehat{\beta}b-g\cdot\eta-P}^{c-\widehat{\beta}b-g\cdot\eta} P\, dF(\varepsilon) + \int\limits_{c-\widehat{\beta}b-g\cdot\eta-P}^{c-\widehat{\beta}b-g\cdot\eta} (1-\widehat{\beta})b\, dF(\varepsilon) \\
&= \int\limits_{c-\widehat{\beta}b-g\cdot\eta}^{\infty} P\, dF(\varepsilon) + \int\limits_{c-\widehat{\beta}b-g\cdot\eta-P}^{c-\widehat{\beta}b-g\cdot\eta} (1-\widehat{\beta})b\, dF(\varepsilon) \geq 0
\end{aligned}$$

A.2 Sample

Our initial sample consisted of 120 subjects, randomly assigned to treated and control groups of 60 subjects each. Table A.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 get at 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.

A.3 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 did return to the screening site and modify their answers until they hit upon the correct answer to join the experiment. (Which was a "no" on

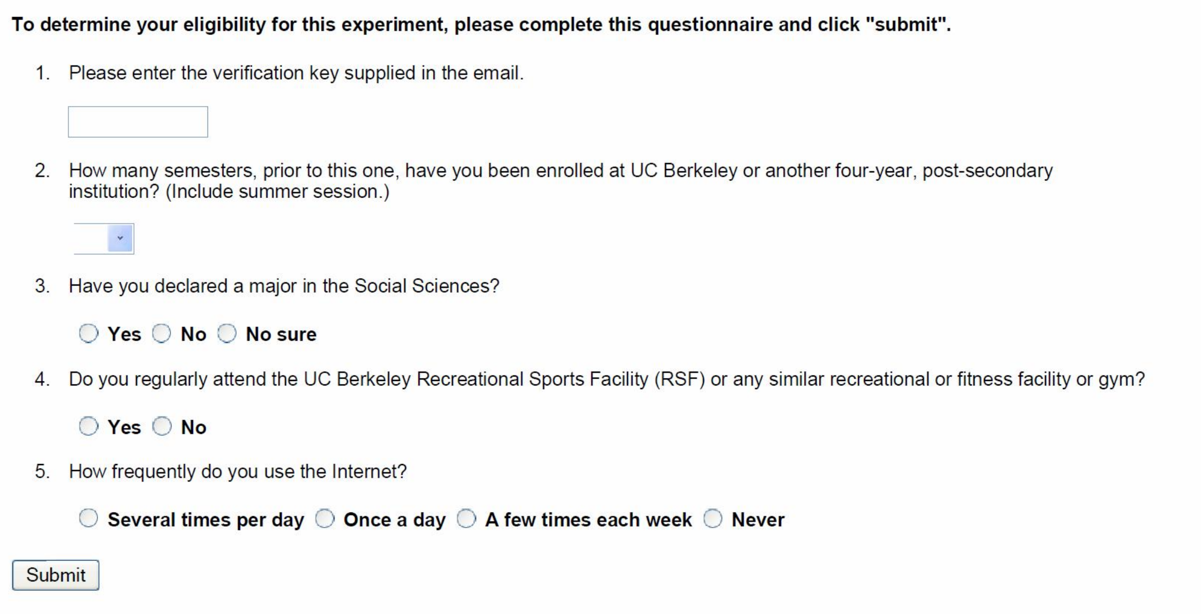
Table A.1: Comparison of Treated and Control groups.

	(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.

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 A.1: Screening Site



To determine your eligibility for this experiment, please complete this questionnaire and click "submit".

1. Please enter the verification key supplied in the email.

2. How many semesters, prior to this one, have you been enrolled at UC Berkeley or another four-year, post-secondary institution? (Include summer session.)

3. Have you declared a major in the Social Sciences?

   Yes **No** **No sure**

4. Do you regularly attend the UC Berkeley Recreational Sports Facility (RSF) or any similar recreational or fitness facility or gym?

   Yes **No**

5. How frequently do you use the Internet?

   Several times per day **Once a day** **A few times each week** **Never**

Submit

A.4 Elicitation mechanisms

Figure A.2 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 A.3.

A.5 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.[47] 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

[47]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.

Figure A.2: Sample p-coupon and incentive-compatible elicitation task

[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 A.3: Unincentivized and other elicitation tasks

[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.

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 left the study at this point.[48] 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.[49]

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.

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

[49] 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 A.2: 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.

Table A.3: Comparison of Habit-Formers and Non Habit-Formers

	(1) Treated Group	(2) "Habit-Formers"	(3) Non "Habit-Formers"	(4) T-test p-value
Demographic covariates				
Age	22.204	19.750	22.630	0.306
	(0.990)	(0.453)	(1.150)	
Gender (1=female)	0.648	0.625	0.652	0.885
	(0.066)	(0.183)	(0.071)	
Proportion white	0.333	0.250	0.348	0.596
	(0.065)	(0.164)	(0.071)	
Proportion Asian	0.630	0.750	0.609	0.454
	(0.066)	(0.164)	(0.073)	
Proportion other race	0.037	0.000	0.043	0.557
	(0.026)	(0.000)	(0.030)	
Economic covariates				
Discretionary budget	208.333	181.250	213.043	0.699
	(28.830)	(92.068)	(30.274)	
Travel cost to campus	0.648	0.000	0.761	0.424
	(0.334)	(0.000)	(0.391)	
Travel time to campus (min)	14.398	9.688	15.217	0.252
	(1.703)	(1.666)	(1.958)	
Naivete proxy covariates				
"Forget[a,b]"	1.556	1.500	1.565	0.800
	(0.090)	(0.327)	(0.091)	
"Spontaneous[a,b]"	2.574	2.250	2.630	0.198
	(0.104)	(0.164)	(0.118)	
"Things come up[a,b]"	2.611	2.375	2.652	0.363
	(0.107)	(0.263)	(0.117)	
"Think ahead[a,b]"	2.944	3.000	2.935	0.778
	(0.081)	(0.189)	(0.090)	
"Procrastinate[a,b]"	3.056	2.875	3.087	0.473
	(0.104)	(0.295)	(0.111)	
Exercise experience andattitude covariates				
Pre-trt Godin Activity Scale	36.500	41.688	35.598	0.474
	(2.983)	(3.823)	(3.434)	
"Fitness is important[a,b]"	2.981	3.500	2.891	0.010
	(0.086)	(0.189)	(0.089)	
"Appearance is important[a,b]"	3.259	3.375	3.239	0.620
	(0.096)	(0.183)	(0.109)	
N obs.	54	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.

A.6 GMM Robustness Checks

Table A.4: GMM Parameters - Alternative Specification

Name	Parameter	Model 1	Model 2
Panel A: Directly Estimated Parameters			
Net daily cost	$C - \beta b$	5.219***	5.057***
		(0.504)	(0.491)
Cost of naivete	$(\widehat{\beta} - \beta)b$	2.746***	2.641***
		(0.282)	(0.274)
Habit value	η	2.229***	2.250***
		(0.731)	(0.740)
Predicted habit value	$(1 - \alpha)\eta$	0.175	0.246
		(0.825)	(0.839)
Probability of habituation	π	0.320**	0.306**
		(0.134)	(0.136)
Demand for commitment	$(1 - \widehat{\beta})b$	–	1.506***
			(0.290)
Standard deviation, daily shock	σ	2.670***	2.594***
		(0.264)	(0.258)
Panel B: Extended Parameters			
Degree of projection bias	α	0.922***	0.891***
		(0.367)	(0.369)
Degree of beta-naivete	ω	–	0.637***
			(0.045)

Notes: The daily shock, ϵ, is drawn from a mean-zero normal distribution. Standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. Parameters in Panel B are calculated by transformations of parameters in Panel A, with standard errors implied by the delta rule. Model 2 includes an additional moment restriction on the difference between unincentivized predictions and p-coupon valuations.