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

Dan Acland† and Matthew Levy‡

October 26, 2012

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. We elicit subjects' pre- and post-treatment predictions of post-treatment gym attendance, using a habit-formation intervention based on Charness and Gneezy (2009) as an exogenous shock to treated subjects' gym preferences. Projection-biased subjects, projecting their current habit state onto their future expectations, will, ex-ante, under-estimate any habit-formation effect of our treatment. Naive present-biased subjects in both groups will overestimate their future attendance. Like Charness and Gneezy, we find subjects do form a significant short-run habit, though we find substantial decay caused by the semester break. Subjects appear not to embed this habit formation into their ex-ante predictions. Approximately one-third of subjects formed a habit equivalent to the effect of a $2.60 per-visit subsidy, while their predictions correspond to 90% projection bias over this habit formation. Moreover, subjects greatly over-predict future attendance, which we interpret as evidence of partial naiveté with respect to present bias: they appear to 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. Special thanks go to Brenda Naputi of the Social Science Experimental Laboratory at the Haas School of Business, Brigitte Lossing at the UC Berkeley Recreational Sports Facility, and to Vinci Chow and Michael Urbancic, for extraordinary assistance with implementation. 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@berkeley.edu

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

1 Introduction

Individuals routinely make intertemporal decisions which require them to predict how their preferences, beliefs, and constraints will change over time. The neoclassical approach typically assumes that individuals faced with such decisions have rational expectations: while they may not know their exact future preferences, beliefs, and constraints, they know the distribution of possibilities, and optimize over expectations. Under this assumption, intertemporal choices can be assumed to maximize individual welfare, in expectation. But there are situations it which it is reasonable to question whether expectations are, indeed, rational, with large welfare consequences if they are not. If someone is deciding today whether to put off until tomorrow an unpleasant task with long-term benefits, it matters whether they correctly predict their ability to resist the same tempation tomorrow. If not, they may procrastinate catastrophically. Similarly, if someone is deciding today whether to invest time in an activity that is habit-forming, it matters whether they predict how habit forming it will be, whether the habit is a desirable but harmful one, such as drug addiction, or an undesirable but beneficial one, such as a life-enhancing health behavior.

Identifying the presence of misprediction in the real world has proven challenging, and estimation of the degree of misprediction has been even more elusive. We take advantage of a field-experimental intervention which has previously been shown to exogenously induce what is interpreted as a gym-attendance habit to explore whether subjects predict the habit formation process, or whether they instead exhibit projection bias with respect to this change in their preferences, as defined by Loewenstein, O'Donoghue and Rabin (2003). Using the same framework, we are also able to explore subjects' predictions of future time preference, i.e. whether they are "naive" or "sophisticated" with respect to self-control problems caused by present bias, as defined by O'Donoghue and Rabin (1999a). In addition, by specifying a structural model of these biases ex-ante, and explicitly designing our experiment with that model in mind, we are able to estimate the degree of both kinds of prediction bias using structural estimation. Finally, the gym-attendance framework we use for studying misprediction allows us to contribute to the understanding of habit formation in gym attendance per se, and the impact of incentives on health behaviors.

We take advantage of the experimental intervention of Charness and Gneezy (2009) (hereafter CG), who conducted a random, controlled experiment in which they

paid treated-group subjects $100 to attend the gym eight times in one month, and found an increase in post-treatment attendance compared to a control group, which they interpret as habit formation.[1] We recruit 120 subjects who were self-reported non-gym attenders, and replicate CG's main gym-attendance intervention almost exactly. We find a significant post-treatment gym-attendance increase of 0.256 visits per week among our subjects, which is smaller than, but statistically indistinguishable from, CG's result. And, like them, we find that the effect is concentrated in the upper tail of the attendance distribution.[2] We utilize this experimental mechanism for inducing an endogenous shift in gym-attendance preferences in a larger experimental framework for exploring whether subjects predict the shift in their preferences. Specifically, we elicit subjects' predictions of their post-treatment attendance, conducting elicitations at two points: first immediately before the intervention, but one week after treated subjects learned of the intervention, and then again immediately after the intervention. Our elicitations consist of both an incentive-compatible valuation of a contingent-payment contract for future gym attendance, and an un-incentivized direct prediction task.

We begin by conducting reduced-form tests of misprediction. For both treated and control groups we test whether subjects correctly predict their actual future gym attendance and find that on average both groups significantly overpredict their actual attendance, by a factor in the range of 2.5–5.5 for pre-treatment predictions, and 2–4 for post-treatment predictions. We interpret this as indicating that subjects are at least partially naive with respect to future self-control problems caused by present bias. That is, they incorrectly predict how their future desire for immediate gratification will affect their gym attendance. Interestingly, we find that both groups revise their predictions downwards after the treatment period. We did not hypothesize

[1] That habit formation plays an important role in physical exercise has long been accepted in the behavioral health literature. See for example, Valois, Dersharnais and Godin (1988), Dzewaltowski, Noble and Shaw (1990), Reynolds, Killen, Bryson, Maron, Taylor, Maccoby and Farquhar (1990), Godin, Valois and Lepage (1993), Godin (1994), but Charness and Gneezy (2009) is the first experimental evidence we are aware of.

[2] In addition, because we track subjects considerably longer than CG, we are able to observe that the effect appears to largely decay during the semester break, 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 one possible reason for this, that our intervention did not induce the steady-state level of habituation in subjects.

this outcome ex ante, but conjecture that it may be the result of initial misprediction of future time constraints, as the initial predictions come at the beginning of the fall semester, while the later predictions are several weeks into the semester. As we will show, we are able to rule out that the downward revision is caused by a gradual improvement over time in subjects' predictions of present bias.

Next we test whether treated subjects correctly predict the increase in their post-treatment attendance that results from what we are calling the habit-formation effect of the intervention. To do this, we compare the change over time in treated subjects' predictions to the change over time for control subjects. Both groups revised their predictions downward over the course of the intervention period. But since treated subjects' actual post-treatment attendance went up relative to control, what we want to know is whether there was a commensurate difference in the amount by which treated subjects revised their attendance predictions. If treated subjects prior to the intervention fully predicted the increase in attendance caused by habit formation, we would expect them to incorporate that increase into both their pre- and post-treatment predictions, and thus the natural downward revision in their predictions over time would be the same as for control subjects, *ceteris paribus*. On the other hand, if treated subjects at least partially failed to predict the habit-formation effect, the downward revision in their predictions would be less than for control groups, because there would be an offseting upward revision in their predictions after the intervention, as they would by then have experienced the habit formation caused by the intervention. This difference-in-differences approach is largely robust to secular trends in prediction that affect treated and control subjects equally.

We find that, for weeks when our measure of un-incentivized predictions is not affected in any way by the subsidy embedded in our incentive-compatible valuation mechanism, treated subjects do indeed revise their predictions downwards by less than control subjects, by 0.394 visits per week, an amount that is compatible with complete failure to predict habit formation. However, we find no statistically significant difference in downard revisions for weeks when attendance is subsidized by the incentive-compatible mechanism. While these results are not conclusive, we feel they are suggestive of projection bias with respect to habit formation. We explore alternative explanations, and discuss the effect of the incentive-compatible mechanism on our measures of predictions.

Having explored these reduced-form results, we turn our attention to the structural

estimation of a formal model. Allowing ourselves to take as given our interpretation of our reduced-form results as evidence of partially-naive present bias, and of projection bias with respect to habit formation, we specify a model of gym attendance that incorporates these behavioral components, and use the generalized method of moments to estimate the parameters of the model, including two prediction-bias parameters that have previously proven difficult to directly identify. The model was specified ex-ante, and the experiment was designed to facilitate structural estimation of the model. 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. In our estimation, the experimentally induced habit increases habituated subjects' average utility valuation of gym attendance by the equivalent of \$2.60 per visit, and takes hold in roughly 1/3 of treated subjects. Meanwhile, 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 highly statistically significant degree of projection bias of $\alpha = 0.9$. This is considerably greater than the $\alpha \in [0.31, 0.50]$ point estimates 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 the welfare loss from misprediction of negative habits such as smoking, equally importantly, if people do not foresee the way that health behaviors such as exercise can become easier to adhere to after a period of habit formation, they may make suboptimal choices about investing in habit formation, and thus miss out on important health benefits.

To capture naiveté with respect to present bias we follow the usual β, δ model of O'Donoghue and Rabin (1999a), in which $\beta \in [0, 1]$ represents an individual's actual short-run discount factor, and $\hat{\beta} \in [\beta, 1]$ represents their belief about what their short-run discount factor will be in the future. We re-parameterize $\hat{\beta}$ as a linear combination of its smallest and largest possible values, $\widehat{\beta} = (1 - \omega) \cdot \beta + \omega \cdot 1$.[3] Thus $\omega = 0$ corresponds to full sophistication, and $\omega = 1$ to fully naive

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

beliefs. This re-parameterization allows us to investigate prediction, $\hat{\beta}$, without a separately identified estimate of the short-run discount factor, 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 statistically significant 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 the design of our experiment. Section three presents our reduced-form results of habit formation and prediction. Section four presents our model of gym attendance, habit formation, and behavioral biases. Section five presents our sturctural estimation of the model. Section six discusses our findings and concludes.

2 Design

We recruited 120 subjects from the students and staff of UC Berkeley and randomly assigned them to treated and control groups. Due to attrition and missing covariates, our final sample includes 54 treated subjects and 57 control subjects.[5] CG found that habit formation was greatest among previous non-attenders, so we screened for subjects who self-reported that they had not ever regularly attended any fitness facility, in hopes that this would give us greater prevalence of habit formation, and thus greater power to investigate predictions of habit formation. In addition, non-attenders constitute a subject pool for which our structural model, which assumes ex-ante non-habituation, is well specified. Our screening mechanism is described 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 a non-representatively low weight on future consumption, or their model of payday lending is incomplete.

[5]Four treated and two control subject dropped out of the study. Three other subjects had missing values for individual covariates. The difference in attrition and missing covariates between treatment and control is not statistically significant. Unfortunately, because demographic variables were collected at the third session, we cannot compare attriters across groups. Details of the sample appear in appendix A.1.

appendix A.2. Treated and control subjects met in four separate sessions, staggered over one afternoon, 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 of $25 to attend the campus gym once during the following week. For this and all subsequent attendance 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. We call this initial offer, made to all subjects, the "learning-week" offer, as it is intended to incentivize subjects to overcome any one-time fixed cost of learning how to access the gym. In this way we hoped to separate true habit formation, resulting from multiple visits, from any increase in post-treatment attendance that might be caused by paying subjects to overcome fixed learning costs. The learning-week offer is identical to the low-incentive condition of CG, so our control group is comparable to their low-incentive group. We also paid the $10 gym-membership fee for all students, and filed the necessary membership forms for those who were not already members.

At the same initial meeting, the treated 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 CG'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.[6]

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". At the time of elicitation, both groups were reminded of the offers they had received the previous week. We announced our treatment-month offer at the initial meeting, one week before the elicitation session, in order to provide treated subjects with time to adjust to the prospect of earning an additional $100, before giving them a multiple price list elicitation task involving a contingent contract. This was to avoid the possibility of treated-subjects' elicitations being influenced by a "house-money" effect that might have led treated subjects to treat a risky prospect differently if

[6]In the event, our compliance rate was not distinguishable from CG's less-restrictive design.

elicitations were done immediately after the announcement of a windfall gain.

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 five target weeks for which attendance predictions were made were separated from this second elicitation session by one week, so that present-biased subjects would see the target weeks as being "in the future" from the perspective of both elicitation sessions. The timeline of the experiment is illustrated in Figure 1.

Gym attendance data were collected for a 17-month period stretching from 37 weeks before the learning week to 33 weeks after the end of the treatment month. This period includes summer and winter academic breaks as well as three full semesters. This attendance data was based on recorded ID-card swipes required for gym entry. Because swipes were necessary to enter the gym but not exit, we cannot determine the length of a visit. Nor were we able to observe subjects' activities while at the gym. We acknowledge that some of the recorded swipes during the treatment month may represent subjects swiping to receive the reward but not exercising. However, there is no reason to continue to engage in such false-visit behavior when incentives are absent during the post-treatment period. To the extent that some subjects may have swiped without exercising during the treatment month, our test of habit formation is biased downwards, i.e. against the finding we report.

2.1 Elicitation procedures

To elicit predictions of target-week gym attendance, and to provide a means by which to calibrate parameters in our structural estimation, 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 was printed on the coupon, along with the beginning and end dates of the target-week. The parameter p took values of 1, 2, 3, 5, and 7.[7] As a reminder, the target weeks were the five weeks immediately following the buffer week, which in turn followed the end of the treatment period.

At the pre-treatment elicitation session, each subject was shown four different p-coupons, corresponding to four of the five target weeks, each with a different value

[7] We conducted a pilot of the elicitation mechanism to determine appropriate values for p.

of p. In addition, for each subject there was one target week for which they were shown no p-coupon, which we refer to as the "zero week", as it corresponds to a week with a coupon with $p = 0$. The week for which they were not shown a coupon, the order in which the coupons were presented, and the value of p for each coupon, were randomized across subjects.[8] For each p-coupon, we asked subjects to complete an incentive-compatible multiple price list task to elicit their valuations for p-coupons of various values with various target weeks.[9] A sample p-coupon is included in Appendix A.3, 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.

The multiple price list task is incentive compatible for subjects' valuations of a contingent contract, but it is clear that could be driven more than just their predictions of future gym attendance. There is an obvious correlation between how many times a subject believes they will go to the gym and the value of a coupon that pays off on a per-visit basis, but the correlation is unlikely to be perfect. In the face of uncertainty about future time and budget constraints, risk-aversion alone implies that we would at best elicit subjects' certainty equivalents for a p-coupon, even for an exogenous event over which subjects have no control.[10] But for an endogenous event like gym attendance, there is the additional effect 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 subjects' future behavior—the "incentive effect" on attendance—and the distortion the incentive introduces into their valuations due to what we refer to as the "commitment value" of the p-coupon, which is the value it offers to subjects as a way to motivate them to go to the gym when they otherwise might not. The incentive effect should be incorporated into subjects' predictions of future attendance,

[8] We used a block-random design such that 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.

[9] 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.3.

[10] 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.

and thus does not confound our inference of beliefs about subsidized attendance. The commitment value, however, drives a wedge between the value a subject places on a p-coupon and the amount of money they predict they will earn from gym attendance. Time-consistent subjects (or sufficiently naive present-biased subjects) will see the p-coupon as an incentive to go to the gym at times when they otherwise would not want to, thus lowering their valuations relative to the expected face value of the coupon. Sufficiently sophisticated present-biased subjects however, who correctly predict their future self-control problems, may value the p-coupon as a commitment device to get their future self to the gym, and will raise their valuation over the expected face value of the coupon. Because this commitment value may increase or decrease subjects' valuation of a p-coupon, care must be taken not to interpret subjects' valuations as exactly proportional to their beliefs, even in the absence of risk-aversion.

In some ways, the commitment value makes these p-coupon valuations a less-than-ideal tool for investigating predictions. We note, however, that as the value of p goes to zero, the distortion induced by the commitment value also goes to zero, and the coupon valuation can be used to directly infer beliefs, a fact we will exploit by focusing particular attention on small values of p. Furthermore, to complement our price-list elicitation mechanism, immediately after the price-list task for each coupon we also asked subjects to directly state how many times they thought they would go to the gym during the specified target weeks, if they were to have been given the p-coupon they had just valued in the price list task. In other words, we asked them to make predictions of future attendance under the hypothetical of facing the same set of attendance incentives as in the price list task. Thus, we have an incentivized elicitation of p-coupon valuations that is correlated with predictions, and an un-incentivized elicitation of predictions themselves. We can therefore use the p-coupon valuations to validate the unincentivized predictions; this will be particularly important when we make use of the unincentivized prediction for the $p = 0$ week.

Our un-incentivized prediction task also allowed us to ask subjects how often they thought they would go to the gym during the zero week, the one target week for which they were not presented with a p-coupon. The zero week gives us an additional un-incentivized prediction of behavior in the absence of any effect of attendance incentives. In addition, under the assumption of quasilinear preferences, by comparing the 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 specifications of our structural model to calibrate the extent of naiveté, and thereby leverage what would otherwise be a confound to provide additional parameter identification.

At the post-treatment elicitation session, each subject was shown exactly the same four p-coupons they had seen in the pre-treatment session, in the same order, and went through the same set of elicitation tasks.[11] 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 shown during the elicitation process.[12] We refer to the target week for which each subject received a coupon as their "coupon week". We therefore have two target weeks for each subject in which we can compare their predictions with their actual gym attendance under the same attendance-incentive conditions, the first being the zero-week, for which there is no attendance incentive, and the second being the coupon week. The giveaway was a surprise to the subjects—having been conducted unannounced only after the second elicitation session—and thus did not affect their p-coupon valuations or un-incentivized predictions during the elicitation tasks. We discuss compliance with the treatment-month offer, attrition, and our randomization procedure in Appendix A.4.

3 Results

Of the 54 subjects in our final treatment sample, 43 completed the eight necessary semi-weekly visits in order to earn the $100 treatment-month incentive: a compliance rate of 80%. In CG's high-incentive group the compliance rate was approximately 83%, suggesting that our more restrictive semi-weekly requirement 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

[11] At the time of the first elicitation, subjects did not know that the second elicitation would take place. Thus, subjects' pre-treatment valuations were not confounded by uncertainty about possible future p-coupons. Furthermore, since the incentivized elicitation mechanism involved randomly selecting only one subject per session-group to receive their choice on the randomly selected line of the multiple price list, in only a very small number of cases did subjects make post-treatment valuations for p-coupons for target weeks for which they already held a p-coupon from the pre-treatment elicitation.

[12] Here again, 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.

eight times, which underscores the power of standard economic incentives.

3.1 Habit formation

Figure 2 shows average weekly attendance for the treated and control groups over the duration of the study period. In this and our subsequent regression analysis of habit formation, we have removed observations for each subjects' coupon week—the target week for which they received a p-coupon. This is partly to remove large spikes and make the graph easier to interpret, and partly to rule out the possibility of seeing a spurious habit effect if treated subjects are more likely than control to register a false swipe (without actually exercising) to earn p-coupon rewards. 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, as we would expect. In the immediate post-treatment period—the two months between the end of the treatment period and the beginning of the semester break—the treated group consistently continues to attend the gym more than the control group. During the four weeks of the semester break, there is essentially no attendance in either group. In the later post-treatment period, the four months after the semester break, the difference between groups 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. 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. To increase the precision of our analysis, we control for individual characteristics, including demographics and demand shifters such as travel time to the gym.[13] 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 is 1.209, roughly the product of the twice-weekly incentive target and the 80%

[13] 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.

compliance rate, which reflects the strong effect of the $100 incentive on contemporaneous attendance. 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. In the later post-treatment period, after the winter break, we see no statistically significant difference between the groups, with a point estimate of 0.045 additional visits per week, suggesting that the habit induced by four weeks of exogenous gym attendance largely decayed over a similar period of quasi-exogenous non-attendance. It is worth noting that this post-break decay supports our interpretation of the short-run effect as habit formation rather than alternatives such as learning, for which one would not expect to find decay.

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 (ITT). To see the effect on those who complied with the treatment we instrument for compliance with the treated-group dummy, including our vector of individual covariates in the first stage. This gives us the average treatment effect on the treated (ATT), controlling for observable differences between compliers and non-compliers. This analysis implicitly 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, and the fact that none went exactly 7 times. These results are reported in the second column of Table 1, where the increase in immediate post-treatment gym attendance for the treated-group is now almost exactly a third of a visit per week.

To compare our results with the results from CG's first study we ran the same regression on their data, the results of which constitute the final column of Table 1.[14] 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.[15] 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 do not reject that the habit-

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

[15]Recall that our control group is equivalent to their low-incentive group.

formation effect in our sample was the same as the habit-formation effect in their sample.

Like CG we explore heterogeneity in the habit-formation effect by attempting to identify the increase in attendance at the individual level. CG do this by comparing post-treatment attendance to pre-treatment attendance at the individual level. Our approach is to compare individual treated subjects' post-treatment attendance to a prediction of what their attendance would have been 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 our prediction of their attendance. With this method, we can pick out 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. Relaxing the threshold for habit formation to 0.5 visits/week, which introduces more noise, rejects the null at a p-value of 0.066. Thus, the relatively small average increase in post-treatment attendance in our treated group appears to reflect not a small effect on all subjects, but rather a large mass of unaffected subjects and a smaller mass of subjects for whom the habit formation appears substantial.

It is not surprising that we find heterogeneity in our treatment effect. One possibility, which the data cannot fully address, is that some subjects in the treated group merely swiped their ID cards at the gym during the treatment period, but did not actually exercise. We would not expect such subjects to form any habit, and our estimates of the treatment effect in Table 1 would be biased towards zero by their presence. The fact that any such subjects would have no habit to mispredict would reduce the power of our reduced-form tests of projection bias (and will bias upwards the number of "unaffected" treated subjects in our structural estimates). An alternative explanation is that some subjects would have formed a habit eventually, 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. Our structural model cleanly incorporates all three possibilities with a single parameter.

3.2 Predictions

In Table 2 we turn our attention to our primary focus: subjects' predictions. The two horizontal panels break the subjects into control and treated groups. The columns within each panel correspond to three different measures of predicted target-week attendance. The first is the prediction implied by subjects' valuation of a p-coupon, for the weeks they were given actual p-coupons (their coupon weeks, $p > 0$), pooled over all values of p.[16] The second is subjects' un-incentivized predictions of coupon-week attendance. The third is their un-incentivized predictions of zero-week attendance, $p = 0$. The first two rows of the table show prediction measures from the pre- and post-treatment elicitation sessions. The third row shows actual target-week attendance. The remaining rows test the differences between predicted and actual attendance for the different groups and elicitation sessions, and the change in predictions from pre- to post-treatment elicitations.

In both the pre- and post-treatment elicitation sessions, both groups predicted future gym attendance substantially greater than their actual gym attendance, by as much as two visits per week. This pattern holds for both coupon weeks and zero-weeks. It is particularly noteworthy that subjects substantially over-predict gym attendance in weeks with no p-coupon, suggesting that they are not just overpredicting their response to the p-coupon incentives, but also their underlying gym preferences. Indeed, since overprediction is less pronounced in proportional terms for coupon weeks than for zero weeks, it is not unreasonable to conjecture that subjects predicted the response to the financial incentives more accurately than their underlying gym preferences.

On the basis of this systematic pattern of mis-prediction we can rule out, in the β, δ model of present-biased preferences, both rational time consistency ($\beta = 1$) and fully-sophisticated present bias ($\beta < 1,\ \widehat{\beta} = \beta$), both of which imply rational expectations about future time preference, and hence, correct predictions of attendance, on average.

[16] This is subjects' valuations for a p-coupon divided by the face value, p.

It is possible to brainstorm other biases which could cause the mis-prediction in Table 2, but it is not clear what generalizable, alternative theory would have led us to hypothesize these results ex-ante. Furthermore, while our results do not rule out a role for other models of self control such as the temptation-utility model of Gul and Pesendorfer (2001, 2004), our results suggest that such models cannot fully explain our data, as they too embed rational expectations about choices, which is clearly violated here.

As we would expect, introducing a p-coupon seems to increase both actual and predicted attendance. In Table 3 we present results on the effect of specific values of p on attendance and predictions, controlling once again for individual covariates. 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 $p = \$1$). Not surprisingly, in column (1) we find that attendance is essentially monotonically increasing in p, which is reassuring, as it indicates the upward-sloping supply curve for exercise that one would expect.[17] What is interesting, in columns (2) and (3), is that subjects appear to predict the slope of this supply curve with reasonable accuracy, but utterly fail to predict its vertical intercept, further suggesting that they are better able to make predictions about their response to incentives than about their underlying gym preferences.

Table 2 also reveals a consistent pattern of subjects revising their predictions downward between the pre- and post-treatment elicitation sessions. When we look at predictions by p in Table 3 we see the same thing, reflected in the coefficient on the Post-Treatment dummy in columns (2) and (3), which implies that between the pre- and post-treatment elicitation sessions, subjects reduce their predictions by roughly two-thirds of a visit per week on average. These sessions differ in two ways: they are a month apart in time, and (as a result) the second session is closer to each of the target weeks than the first. One possible explanation of the downward revision between sessions is that the extent to which subjects discount future utility decreases smoothly with distance into the future, rather than abruptly as in the beta-delta model. Another possibility is that subjects' naiveté about future self-control problems increases with distance into the future. Both seem psychologically plausible. In either

[17] Pairwise comparisons of the coefficients do not reject monotonicity, though, interestingly, we also cannot rule out the possibility of crowding-out of intrinsic motivation at low values of p.

case, we would see a decrease in misprediction merely because the target weeks are less far into the future at the time of the post-treatment elicitation session. Because we have multiple target weeks, we can examine this by comparing pre-treatment predictions for the first target week with post-treatment predictions for the fifth target week. This comparison holds temporal proximity constant: in both cases subjects are predicting their attendance five weeks in the future. Columns (4) and (5) of Table 3 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. Thus, we can rule out that temporal proximity alone explains the downward revision in predictions.[18] Rather, there appears to be an effect of the post-treatment session being later in absolute terms. This 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 once they get a few weeks into the semester.[19]

Finally, we find in both Table 2 and Table 3 that in general, subjects' normalized valuations of p-coupons are lower than their un-incentivized predictions. As discussed in the design section, this undervaluation could be caused by risk-aversion, or by a negative commitment value for the coupons. Anticipating this undervaluation, we elicited a measure of risk aversion over small to moderate stakes, using hypothetical lotteries.[20] We find no effect of this measure on undervaluation, suggesting that risk aversion may not be the explanation for the undervaluation. We cannot rule out that our measure of risk aversion is simply uninformative, but this result may suggest that the undervaluation is caused by subjects placing a negative value on the incentive effect of a p-coupon. This would be true for a time-consistent subject, or for a sufficiently naive present-biased subject, both of whom believe that the p-coupon will only incentivize them to go to the gym at times when they would otherwise not want to, inducing a net effort cost that reduces the perceived value of the coupon.

Next we turn to whether treated subjects predict the habit-formation effect. Recall that our approach is to test whether treated subjects revise their predictions downwards between elicitation sessions less than control subjects. If they do, it sug-

[18] Furthermore, the results of our structural estimation are robust to the incorporation of temporal proximity in the model.

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

[20] We use a hypothetical-stakes version of the mechanism outlined in Holt and Laury (2002).

gests that they initially failed to fully predict the habit effect of the treatment month, and then incorporated an upward revision in their post-treatment predictions after the habit was established.[21] We implement this test using a difference-in-differences regression of predictions on dummies for treatment group, post-treatment session, and the interaction of the two, as well as individual covariates. The first three columns of Table 4 show the results of this regression for the same three measures of predictions we used earlier: coupon valuations, and un-incentivized predictions, for target weeks for which subjects were shown a p-coupon, pooling values of p; and un-incentivized predictions for zero weeks.[22] We see a positive double difference for coupon values and un-incentivized zero-week predictions, though the former is not statistically significant. Meanwhile, for un-incentivized predictions for weeks when a coupon was shown, the point estimate is negative, but also not statistically significant.

One reason we might see a significant double-difference for zero-week predictions, but not for weeks with a p-coupon is that, for large values of p, the presence of the coupon might dominate the effect of the gym habit. For example, if some subjects planned to swipe their IDs without actually exercising during target weeks, then for these subjects, both coupon-week prediction measures would be capturing their predictions of this behavior, which might drown out our ability to identify other subjects' predictions of actual gym attendance. To explore this possibility, in column four we restrict coupon valuations to the smallest coupon value, $p = 1$, for which the false-swipe motivation is smallest. This result is not statistically significant, which partly reflects the smaller number of observations due to restricting the value of p, but the point estimate, 0.314, is similar to that for un-incentivized zero-week predictions, 0.394. Thus, when the incentive effect of p-coupons is minimized or eliminated, we get results which appear to support the hypothesis that treated subjects revised their predictions downwards less than control subjects, though we do not have sufficient power for a conclusive test. The actual increase in attendance for treated subjects was 0.256 visits per week. The double difference in predictions is larger than this, though not significantly larger, so if we take the point estimates at face value, these

[21] Alternatively, if they learned that gym-attendance was costlier than initially believed, treated subjects would have a greater downward revision.

[22] Note that in table two we restricted the first two columns to weeks when subjects received an actual coupon (so-called coupon weeks) so that we could compare predictions with actual attendance. Here we are comparing predictions to predictions, so we include all weeks for which subjects were shown a coupon in the elicitation process.

results are consistent with complete projection bias with respect to habit formation.

4 Model

4.1 Setup

In this section we develop a simple model of gym attendance that incorporates habit formation, projection bias, and present-biased preferences, and then estimate the parameters of that model using the generalized method of moments. Our model embeds rationality with respect to intertemporal preferences and predictions, and as such allows us to test for the presence of biases. But it is important to note that our main objective in this section is not test one theory against another, but rather to develop a model that is rich enough to capture what we believe to be going on in our experiment, yet simple enough to be feasibly estimated.

Our model consists of three periods. Period 1 corresponds to the pre-treatment period, including a period of pre-treatment attendance, the announcement of learning-week and treatment-month offers, and the pre-treatment elicitation session. Period 2 corresponds to the post-treatment elicitation session. Period 3 corresponds to the post-treatment target weeks. These periods are not of equal length, but our approach to intertemporal discounting renders this choice non-problematic, as we will explain. In the period 1 all subjects are assumed to be non-habituated—since we screened for non-attenders—and are randomly assigned to control or treated groups. Gym-attendance offers are made, and eliciations of predictions of period-three target-week attendance take place, according to our experimental design. We assume that habit formation occurs instantaneously for treated subjects betweeen periods 1 and 2, though we allow for heterogeneity in the habit effect of the treatment. Elicitation is reiterated in period 2, and subjects receive an unanounced p-coupon.[23] Period 3 constitutes the target weeks during which p-coupons pay off according to subjects' attendance. We assume there is sufficient time between periods 2 and 3 to act as a time buffer, ensuring that from the perspective of period-2 elicitations, subjects consider

[23] 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.

period 3 target weeks to be "in the future" with respect to present bias. This reflects the buffer week we included between the treatment month and the first target week. In periods 1 and 3, when pre- and post-treatment gym attendance takes place, we will model the predicted and actual weekly attendance of an individual by defining gym-attendance utility on a daily basis and aggregating to the weekly level. This allows us to capture the fact that the p reward is earned on a day-by-day basis, while predictions and valuations are made on a weekly basis. Finally, following DellaVigna and Malmendier (2004), subjects recieve long-term health benefits from gym attendance after period three, which we model as a single exponentially discounted sum which is experienced as being in the future from the perspective of all three periods.

We model utility as quasi-linear in money. Without loss of generality, utility from all non-gym sources is normalized to zero. Because our subjects are non-attenders, let the immediate non-habituated utility of gym attendance on day d of individual i be given by $(-c + \varepsilon_{d,i})$ with $\varepsilon_{d,i}$ i.i.d. with cdf $F(\varepsilon)$, and let the exponentially discounted sum of long-term benefits of gym attendance be $b > 0$.[24] 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 models 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 growth and decay. Thus, we implicitly assume that a single gym visit is not sufficient to change a subject's state, but that the exogenous month of attendance induced by our treatment intervention is enough. Furthermore, for simplicity, we do not model habit decay.[25] When subjects are habituated they receive additional, immediate utility for gym attendance of $\eta_i \geq 0$, where, to capture the heterogeneity across subjects in habit-formation parsimoniously, with probability π, a subject has $\eta_i = \overline{\eta}$ strictly greater than zero, and with probability $1 - \pi$, they have $\eta_i = 0$, corresponding to no habit effect.[26] Thus, for subject i in group $g \in 0, 1$ (control = 0, treated=1) the immediate utility of gym attendance is $g\eta_i - c + \varepsilon_d$, which collapses to non-habituated gym utility for control subjects. For simplicity, we present this in terms of subjects knowing their own type ex ante, but the law of iterated expectations

[24] Informally we think of c as being positive, reflecting the fact that the majority of our subjects were non-attenders, but this restriction is not formally imposed.

[25] As an empirical matter, our data do not allow us to identify decay over the immediate post-treatment period we are using to estimate our model.

[26] In our structural estimation we do not restrict $\bar{\eta}$ to be positive, but we find that it is.

means that neither our modeling nor our empirical strategy distinguishes this from the case where subjects merely know $\overline{\eta}$ and π ex ante, but not their own type.

We model time preferences using the β, δ model of Laibson (1997) and O'Donoghue and Rabin (1999a).[27] Subjects discount future periods relative to one another with the standard exponential discount factor, $\delta \in (0, 1)$, and additionally discount all future periods relative to the present with short-run discount factor, $\beta \in (0, 1]$, which captures preferences for immediate gratification, with $\beta = 1$ embedding pure exponential discounting, and $\beta < 1$ leading to potential self-control problems. Subjects' belief about their future short-run discount factor is represented by $\widehat{\beta} \in [\beta, 1]$, with $\widehat{\beta} = \beta$ capturing full "sophistication" or correct prediction of future self-control problems, and $\widehat{\beta} = 1$ capturing complete "naiveté", or total failure to predict future self-control problems, while $\widehat{\beta} \in (\beta, 1)$ allows for partial naiveté.[28] We note that while present bias itself may generate an under-investment in exercise relative to one's long-run preferences, some degree of naiveté is necessary for subjects to hold systematically biased beliefs about future behavior (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 actual magnitude of $\widehat{\beta}$ relative to β, i.e. the degree of naiveté.

Finally, we incorporate the phenomenon of "projection bias," whereby subjects fail to appreciate the extent to which their future preferences may differ from their current ones as a result of changes in state, such as moving from non-habituated to habituated. In our setting, projection bias implies that individuals correctly foresee the direction of the habit-formation process, but may mispredict the strength of the effect by partially or fully "projecting" their preferences under their current level of habituation onto their future selves. We use the "simple projection bias" model of Loewenstein, O'Donoghue and Rabin (2003), in which an individual correctly predicts their future state, but their belief about their future utility function is an alpha mixture of their current utility function and their true future utility function after the state change, where $\alpha = 0$ captures fully correct beliefs about future preferences

[27] 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).

[28] 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.

and $\alpha = 1$ captures complete projection bias, or totally failure to predict state-dependent preference changes. As with naiveté, above, there is a growing literature that suggests $\alpha > 0$,[29], but very little evidence on the actual magnitude of α. For both of these misprediction parameters, estimates of actual magnitudes are necessary if models are to be successfully used for concrete policy and welfare analysis. Because of the difficulty of estimating these parameters using observational data, there is a role for a controlled field experiment such as ours, which is designed explicitly to permit the estimation of structural parameters.

4.2 Observables

We now define our experimentally observable variables, and derive expressions for them from our model. This will allow us to explore some of the features of our prediction measures and our reduced-form test of projection bias, and to define a set of moment equations for GMM estimation. To make notation easier to follow, let $t \in \{Pre, Post\}$ indicate whether attendance or elicitation occurs before or after the treatment intervention. Thus, for attendance variables, t will index pre- versus post-treatment attendance, which occur in periods 1 and 3 respectively (there is no modeled attendance in period 2). Meanwhile, for prediction variables, t will index pre- versus post-treatment elicitation sessions, which occur in periods 1 and 2 respectively (there are no predictions in period 3). Let $Z_{i,d}^{t,g}(p)$ be an indicator for whether subject i in group g holding a coupon of value p (with $p = 0$ for no coupon) attends the gym on day $d = 1, ..., 7$ of a week during period t, so that $Z_i^{t,g}(p) = \sum_{d=1}^{7} Z_{i,d}^{g,t}(p)$ is the number of gym visits during a given week for said subject, and $\bar{Z}^{t,g}(p)$ is the average attendance of subects in group g for a week in that period, at that coupon value. Let $V_i^{t,g}(p)$ be a subject's valuation at elicitation session t of a p-coupon for a post-treatment target week, conditional on their group, and the face value of the coupon, and let $Y_i^{t,g}(p)$ be the subject's un-incentivized attendance prediction for the same coupon. Thus, $\bar{V}^{t,g}$ and $\bar{Y}^{t,g}$ are the group averages for those measures.

If a subject in group g holding a p-coupon attends the gym on a given day during a target week in the post-treatment period, her utility for that day will be $p + \beta b +$

[29]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.

$g\eta_i - c + \varepsilon_{i,d}$, and she will attend the gym if this is greater than zero. Thus $Z_{i,d}^{Post,g}(p) = \mathbb{1} \cdot \{\varepsilon_{i,d} > c - p - \beta b - g\eta_i\}$, and $Z_i^{Post,g}(p) = \sum_{d=1}^{7} \mathbb{1} \cdot \{\varepsilon_{i,d} > c - p - \beta b - g\eta_i\}$. In expectation, total target-week gym-attendance will be,

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

In the absence of habit formation, expected attendance would be the same integral without the $g\eta_i$ term, so the habit-formation effect, the expected increase in attendance caused by habituation, will be the difference between these two integrals:

$$7 \times \int_{c-\beta b - g\eta_i - p}^{\infty} dF(\varepsilon) - 7 \times \int_{c-\beta b - p}^{\infty} dF(\varepsilon) = 7 \times \int_{c-\beta b - g\eta_i - p}^{c-\beta b - p} dF(\varepsilon). \tag{2}$$

The *perceived* probability of target-week gym attendance, from the perspective of the pre-treatment elicitation, 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 $\varepsilon_{i,d} > c - p - \widehat{\beta} b - g(1-\alpha)\eta_i$.

Thus, her un-incentivized pre- and post-treatment predictions of target week attendance are,

$$Y_i^{Pre,g}(p) = 7 \times \int_{c-\widehat{\beta} b - g(1-\alpha)\eta_i - p}^{\infty} dF(\varepsilon) \quad \text{and,} \quad Y_i^{Post,g}(p) = 7 \times \int_{c-\widehat{\beta} b - g(\eta_i - p}^{\infty} dF(\varepsilon). \tag{3}$$

Note that the only difference between these terms is the presence or absence of projection bias. Because of the buffer week, misprediction of self-control, $\widehat{\beta}$, is implicated in both predictions. And since we assume ε is i.i.d. across individuals and days, the double difference in average predictions, normalized by 7 for clarity is,

$$\frac{[\bar{Y}^{Post,1} - \bar{Y}^{Pre,1}] - [\bar{Y}^{Post,0} - \bar{Y}^{Pre,0}]}{7} = \pi \cdot \int_{c-\widehat{\beta} b - \bar{\eta} - p}^{c-\widehat{\beta} b - (1-\alpha)\bar{\eta} - p} dF(\varepsilon). \tag{4}$$

This object corresponds to the coefficients on the interaction terms in columns 2 and 3 of Table 4. The integral tells us, for a subject who will develop a habit ($\eta_i = \bar{\eta}$), the probability on any given day that ε will lie between the value necessary to get her to the gym and the value that projection bias tells her will be necessary. This is multiplied by π to reflect that fact that not all subjects will develop a habit as a result of the treatment. Thus, it is portion of the habit-driven attendance increase that a subject fails to predict due to projection bias, scaled by the proportion of subjects who form a habit. In the absence of projection bias, $\alpha = 0$, this double difference is zero. Assuming positive habit value, $\bar{\eta} > 0$, any degree of projection bias makes the double difference positive.

To develop an expresion for the double difference in the average valuation for a p-coupon, consider that a subject's pre-treatment prediction of her total utility for the target-week, given that she holds a p-coupon, is,

$$7 \times \int_{c-\widehat{\beta}b-g(1-\alpha)\eta_i-p}^{\infty} (p + b + g\,(1-\alpha)\,\eta_i - c + \varepsilon)\; dF(\varepsilon). \tag{5}$$

This is just 7 times the value she gets from going to the gym, integrated over the range of ϵ that she beleives will get her to go. Setting $p = 0$ gives us the perceived utility without a coupon. Since we assume quasi-linear money utility, the perceived dollar value of the p-coupon is simply the difference between expected utility with a p-coupon and expected utility without. From the perspective of the pre-treatment elicitation session, this is:

$$V_i^{Pre,g}(p) = \left[7 \times \int_{c-\widehat{\beta}b-g(1-\alpha)\eta_i-p}^{\infty} p\, dF(\varepsilon)\right] + \left[7 \times \int_{c-\widehat{\beta}b-g(1-\alpha)\eta_i-p}^{c-\widehat{\beta}b-g(1-\alpha)\eta_i} (b + g(1-\alpha)\eta_i - c + \varepsilon)\, dF(\varepsilon)\right]. \tag{6}$$

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

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

The first term in both (6) and (7) is the expected redemption value of the coupon, which is always weakly positive. We note that the redemption value incorporates the subject's prediction of their behavioral response to the p-coupon. The second term is the subject's valuation of that behavioral response, which we will call the "commitment value". This is the change in perceived *utility* caused by the additional gym-visits that the subject foresees she will make as a result of the p-coupon. The sign depends on the subject's ex-ante belief about their 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 the long-run net benefit of doing so is negative. 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 at times when the long-term benefit outweighs the effort cost, but self-control problems would have kept her from going.

We can now derive the double difference in coupon valuations, which corresponds to the coefficient on the interaction term in columns 1 and 4 of Table 4, by comparing the average change in valuations of a p-coupon, from pre- to post-treatment elicitations, between the treated and control subjects. Once again dividing by 7 for simplicity, this double difference is,

$$\frac{[\bar{V}^{Post,1} - \bar{V}^{Pre,1}] - [\bar{V}^{Post,0} - \bar{V}^{Pre,0}]}{7} = \pi \cdot \int_{c-\widehat{\beta}b-\bar{\eta}-p}^{c-\widehat{\beta}b-(1-\alpha)\bar{\eta}-p} p\, dF(\varepsilon) \tag{8}$$
$$+ \pi \cdot \left[\int_{c-\widehat{\beta}b-\bar{\eta}-p}^{c-\widehat{\beta}b-\bar{\eta}} (b + \bar{\eta} - c + \varepsilon)\, dF(\varepsilon) - \int_{c-\widehat{\beta}b-(1-\alpha)\bar{\eta}-p}^{c-\widehat{\beta}b-(1-\alpha)\bar{\eta}} (b + (1-\alpha)\bar{\eta} - c + \varepsilon)\, dF(\varepsilon) \right]$$

For subjects with no projection bias, regardless of their level of present bias, we would expect this difference-in-differences to be zero. The first term in (8) is just the misprediction of gym attendance, which we saw in equation 4. It is weakly positive for projection-biased agents—strictly for $p > 0$ 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 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 on the sign of commitment value, and the distribution of ε and could take on either positive or negative values.[30] The overall effect is therefore ambiguous, but still provides a test inasmuch as any observed value of the double-difference that is significantly different from zero indicates projection bias.

Finally, we note that this difference-in-differences test is designed to be robust to many 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 3 that is consistent with this. But such a secular drift cancels out in Equation (8). 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.[31]

5 Structural Estimation

In the preceding section we derived expressions for the basic observable variables in our experiment—attendance (Z), un-incentivized predictions (Y), and p-coupon valuations (V), for treated and control subjects, before and after the treatment-month, and for various levels of coupon value, p—in terms of the parameters of our structural model. Next, we use those expressions to derive a set of moment equations that identify all of the parameters in the model, and we use the generalized method of moments (GMM) to find the parameter values that minimize the difference between

[30] 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 with positive commitment value 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 (8). In this case, there is no relative change in the perceived commitment value and the term in brackets reduces to zero.

[31] 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 (8) 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.

the expressions derived from the model and their sample analogs in our experimental data. For ease of presentation, we need a few slight refinements of notation. For individual i, let $p_{i,w}$ denote the face value of the w-th p-coupon she was offered during the elicitation sessions ($w \in \{1,2,3,4\}$) and let $V_{w,i}^{t,g}(p_{i,w})$ denote her valuation for that coupon. Also, for clarity of exposition, we will use $g \in \{C,T\}$ (for control and treated) to denote groups, with $g = C \cup T$ denoting the union of treated and control groups. Finally, let σ_ε be the scale parameter of the distribution of ε.

In our main model (Model 1), the moment equations are as follows:

$$\bar{Z}^{Post,C}(p>0) = 7 \cdot \sum_{i \in \mathcal{C}} \sum_{w=1}^{4} (1 - F(c - \beta b - p_{i,w}; \sigma_\varepsilon)) \tag{9}$$

$$\bar{Z}^{Pre,C \cup T}(0) = 7 \cdot \Big[1 - F(c - \beta b; \sigma_\varepsilon)\Big] \tag{10}$$

$$\begin{aligned} \bar{Z}^{Post,T}(0) = 7 \cdot \Big[& \pi \left(1 - F\left(c - \beta b - \eta; \sigma_\varepsilon\right)\right) \\ & + (1-\pi)(1 - F(c - \beta b; \sigma_\varepsilon))\Big] \end{aligned} \tag{11}$$

$$\bar{Z}^{Post,C}(0) = 7 \cdot \Big[1 - F(c - \beta b; \sigma_\varepsilon)\Big] \tag{12}$$

$$\bar{Y}^{Pre,C}(0) = 7 \cdot \Big[1 - F(c - \widehat{\beta} b; \sigma_\varepsilon)\Big] \tag{13}$$

$$\bar{Y}^{Post,C}(0) = 7 \cdot \Big[1 - F(c - \widehat{\beta} b; \sigma_\varepsilon)\Big] \tag{14}$$

$$\begin{aligned} \bar{Y}^{Pre,T}(0) = 7 \cdot \Big[& \pi \cdot (1 - F(c - \widehat{\beta} b - (1-\alpha)\eta; \sigma_\varepsilon)) \\ & + (1-\pi) \cdot (1 - F(c - \widehat{\beta} b; \sigma_\varepsilon))\Big] \end{aligned} \tag{15}$$

$$\sum_{g_i = T} \mathbb{1}\{\bar{Z}_i^{Pre,g_i}(0) < \bar{Z}_i^{Post,g_i}(0)\} = \sum_{g_i = T} \left(\pi + \frac{1}{2}(1-\pi)\right) \tag{16}$$

Equations (9) and (10) identify the baseline, discounted, net cost of gym attendance, $c - \beta b$, and the scale parameter of the distribution of attendance cost shocks, σ_ε, by comparing the post-treatment attendance of control subjects in response to p-coupon incentives with the pre-treatment (i.e. un-incentivized) attendance of all subjects. Because we assume preferences are quasi-linear in money, control subjects' responsiveness to p allows us to calibrate these parameters in dollar terms. The habit-formation effect is identified by equations (11) and (12), which capture the difference in post-treatment attendance between treated and control subjects, in the absence of p-coupons. Here we rely on the assumption that the habit induced by the

treatment is the only systematic driver of post-treatment differences in attendance between control and treated subjects in un-incentivized weeks. Together, the first four equations allow us to put a dollar value on the habit effect. In other words, our estimate of habit utility η is effectively the size of p-coupon subsidy necessary to increase control-group attendance in the post-treatment period by as much as the habit effect increases treated-group attendance.

Equations 13 and 14 use control subjects' pre- and post-treatment predictions to identify over-confidence about future self-control, driven by naiveté with respect to present bias. Specifically, these equations allow us to estimate what we call the "cost of naiveté", $(\widehat{\beta} - \beta)b$ which is that part of the long-run benefit of attendance which subjects mistakenly believe they will value in the future. Equation (15) parallels (11), but captures treated subjects' pre-treatment predictions of their post-treatment attendance rather than their actual post-treatment attendance, in order to identify their ex-ante expectations of the habit value, $(1 - \alpha)\eta$. The ratio of this parameter and the actual habit value, η, allows us to identify the degree of projection bias, α. In all of these prediction moment equations, we use weeks with no p-coupons to avoid embedding an assumption that subjects correctly predict their response to small monetary incentives.

Finally, we introduce (16) to estimate the fraction of subjects developing a strictly positive habit, π. This equation takes advantage of the fact that 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 than in the pre-treatment period converges to 1. The corresponding probability for a subject who did not form a habit converges to 0.5. We use these limits to derive this moment equation, but we note that (16) 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.

Though Model 1 allows us to identify the cost of subjects' naiveté with respect to present bias, it does not allow us to identify the actual degree of naiveté. To do this, in Model 2 we include an additional moment equation based on the difference between subjects' valuations of p-coupons and the expected face value of those coupons implicit in their un-incentivized predictions of coupon-week attendance. Letting, for convenience, $\mathbb{G}_i = \mathbb{1}\cdot\{g_i = T\}$ and $\mathbb{T} = \mathbb{1}\cdot\{t = post\}$, we get:

$$\sum_{t\in\{Pre,Post\}}\sum_{i}\frac{1}{4}\sum_{w=1}^{4}\left(V^{t}_{w,i}(p_{i,w})-Y^{i}_{t}(p_{i,w})\cdot p_{i,w}\right)=$$

$$\sum_{t\in\{Pre,Post\}}\sum_{i}\frac{1}{4}\sum_{w=1}^{4}7\cdot\left[\pi\int_{c-\widehat{\beta}b-\mathbb{G}_i(1-\alpha(1-\mathbb{T}))\eta-p_{i,w}}^{c-\widehat{\beta}b-\mathbb{G}_i(1-\alpha(1-\mathbb{T}))\eta}(b+\mathbb{G}_i(1-\alpha(1-\mathbb{T}))\eta-c+\varepsilon)dF(\varepsilon)+(1-\pi)\int_{c-\widehat{\beta}b-p_{i,w}}^{c-\widehat{\beta}b}(b-c+\varepsilon)dF(\varepsilon)\right] \tag{17}$$

While Equation (17) may appear complex, it has a straightforward interpretation. The left-hand side is the average difference between subjects' valuations of p-coupons and the expected face value implicit in their un-incentivized predictions, pooling pre- and post-treatment elicitations, across both groups. The right-hand side is the expression for commitment value described in equations (4) and (5) in section 4, 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, so for this equation we must write out the averages using summations.

Under the assumption that the difference between valuations and predicted face values is driven by the commitment value of the p-coupons, this allows us to identify what we call the "demand for comitment", $(1-\widehat{\beta})b$, which is the part of the long-run benefit of attendance that present bias causes subjects to forego. Recalling that $\widehat{\beta}$ lies in the interval $[\beta, 1]$, we reparameterize $\widehat{\beta}$ as a linear combination of it's smallest and largest possible values, $\widehat{\beta}=(1-\omega)\cdot\beta+\omega\cdot 1$, so that $\omega=0$ corresponds to full sophistication, and $\omega=1$ corresonds to complete naiveté. We can then compute ω by taking the ratio of the cost of naiveté and the demand for commitment. Because this additional moment equation relies strongly on the assumption that the only difference between the un-incentivized and incentivized predictions is caused by the commitment value of a p-coupon, we present the results from estimating a model both without it (Model 1) and with it (Model 2).

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.[32] Panel A presents the parameters that are estimated directly

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

by GMM, while additional parameters derived from these are presented in Panel B. The structural parameters confirm the reduced-form results in Sections 3.1 and 3.2. On average, the immediate utility cost of gym attendance exceeds the discounted future benefits by \$4.71, causing most of our subjects to not attend most of the time, in the absence of incentives. Naiveté about their future self-control problems causes subjects to under-estimate their future gym-attendance self-control problems, resulting in an estimate of the cost of naiveté of \$3.10. This result corresponds to the significant over-prediction of future attendance relative to actual attendance found in our reduced-form results, and since this term is zero when $\widehat{\beta} = \beta$, we can rule out the hypothesis of full sophistication. This finding of naiveté can also explain why un-incentivized predictions lie above the normalized valuations of p-coupons, as subjects underestimate their need for commitment and thus have a negative commitment value for the p-coupons.

Turning to the habit formation, we find that 32% of treated subjects formed a habit equivalent to a \$2.60 daily gym-attendance subsidy.[33] This is a significant habit—if it persisted, our \$100 treatment would be recouped after only 50 visits on the basis of immediate gym utility alone. It is still substantially smaller than the net daily cost, however, so that even a habituated subject does not on average enjoy going to the gym. This helps explain why we estimate one-third of subjects forming the habit but find far fewer actually regularly going to the gym—apparently, 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 significant degree of projection bias, $\alpha = 0.94$, which appears in panel B of table 5. Indeed, 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. We find a statistically significant demand for commitment of \$1.5. The effect of the additional

error term.

[33] 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 rejected at p = 0.031.

moment on most of the other parameters is small. Because the demand for commitment is zero when $\widehat{\beta} = 1$, we can reject the hypothesis of complete naiveté. Panel B of table 5 reports a tightly estimated $\omega = 0.666$. This means that subjects are essentially $\frac{2}{3}$ naive. We cannot separately estimate the underlying true discount factor, *beta*, 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 that subjects place a value on the long-term benefits of gym attendance of $b = \$15$ per daily visit. Adding this benefit to the increase in immediate gym utility caused by the habit, and 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 treatment intervention as a whole 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.[34]

Finally, Figure 3 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 particular, 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 could identify based only on large 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 could be expected to gradually cause subjects to de-habituate.[35] More

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

[35]Although we have examined this point in our data, we have insufficient power to detect a

optimistically, however, this simulation suggests that 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 for subjects who do not consider the strategic benefit of current attendance on their future selves' likelihood of gym attendance. 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.

6 Discussion

Using a field intervention to exogenously shift preferences toward gym-attendance in a student sample, we find systematic evidence of two simultaneous dimensions of misprediction: projection bias with respect to habit formation; and naiveté with respect to present bias. We develop a novel estimation tool which serves both to incentivize subjects' predictions and to shift their actual future behavior, allowing us to monetize the value of the exercise habit and subject's predictions. We estimate that 32% of our treated subjects formed an exercise habit equivalent to a $2.60 per-visit subsidy in response to receiving a $100 incentive to visit the gym 8 times during the course of the treatment month. While subjects were generally well-calibrated about their responsiveness to additional gym-attendance subsidies, treated 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 the level of subsequent gym attendance, which our structural model interprets as subjects being two-thirds "naive" about their present bias.

We acknowledge there are, of course, other potential interpretations for these two prediction failures. For example, the failure of treated subjects to predict the habit formation could be explained if subjects entered the experiment with systematically biased beliefs about the desirability of gym attendance, and only treated subjects learned that it was more pleasant than their prior. Any alternative must rule out that treated subjects initially hold correct beliefs, though, and will have similar welfare implications: individuals will underinvest in beneficial, potentially habit-forming activities. Projection bias, however, further predicts that habituated subjects will continue to make errors that simple learning models rule out, and if de-habituation is sufficiently rapid—as our data suggest applies to our setting—projection-biased indi-

downward trend in post-treatment attendance in addition to the the semester break effect.

viduals will fail to invest in maintaining a habit. This suggests how further research may help to fully disentangle our model from alternatives.

Furthermore, our use of a student sample clearly raises the question of the external validity of our parameter estimates. While some parameters—such as the opportunity cost of exercise—are certainly context-specific, it is simply not yet known how the estimates of projection bias and naivete regarding present bias will compare to other contexts and populations. Indeed, one of the main contributions of this current study is in providing an easy blueprint for researchers to use in identifying the role of biased beliefs in other settings. Both the overall experimental design and the structural estimation technique can be straightforwardly applied to many interventions involving an exogenous shift of treated subjects' "state", and we hope that future research will help to address this important question.

However, we believe that our estimates can help explain several features of gym contracts. For example, we provide direct evidence that subjects over-predict their gym attendance sufficiently to explain through naiveté the finding in DellaVigna and Malmendier (2006) that infrequent gym-users purchase monthly memberships when, given their actual attendance, per-visit passes would be a cheaper alternative. Futhermore, our estimate that subjects are two-thirds naive about their present bias can help explain the lack of stronger commitment in membership contracts. We estimate that the socially optimal gym contract would increase the monthly membership fee and provide a negative rate for attendance, but because naive agents insufficiently value this commitment, firms instead design contracts which exploit partially naive consumers a la Gabaix and Laibson (2006). That agents are also significantly projection biased places an additional constraint on commitment contracts for behavior change, and points to new contract designs which may help improve welfare.

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.

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.

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.

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.

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.

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.

Figure 1: Our Experimental Design

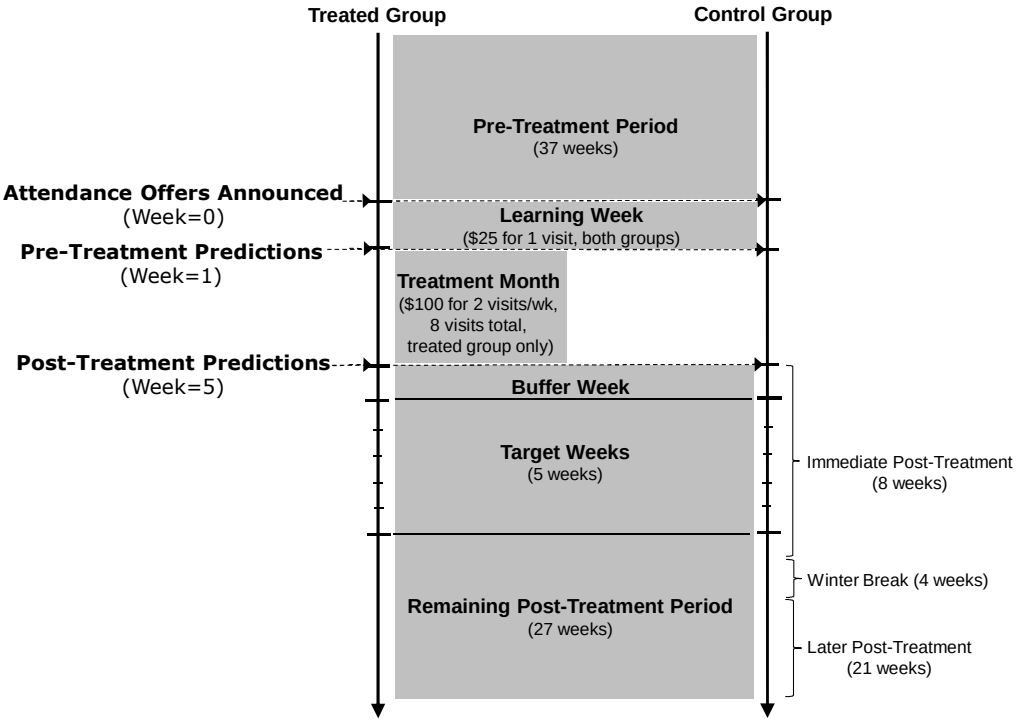
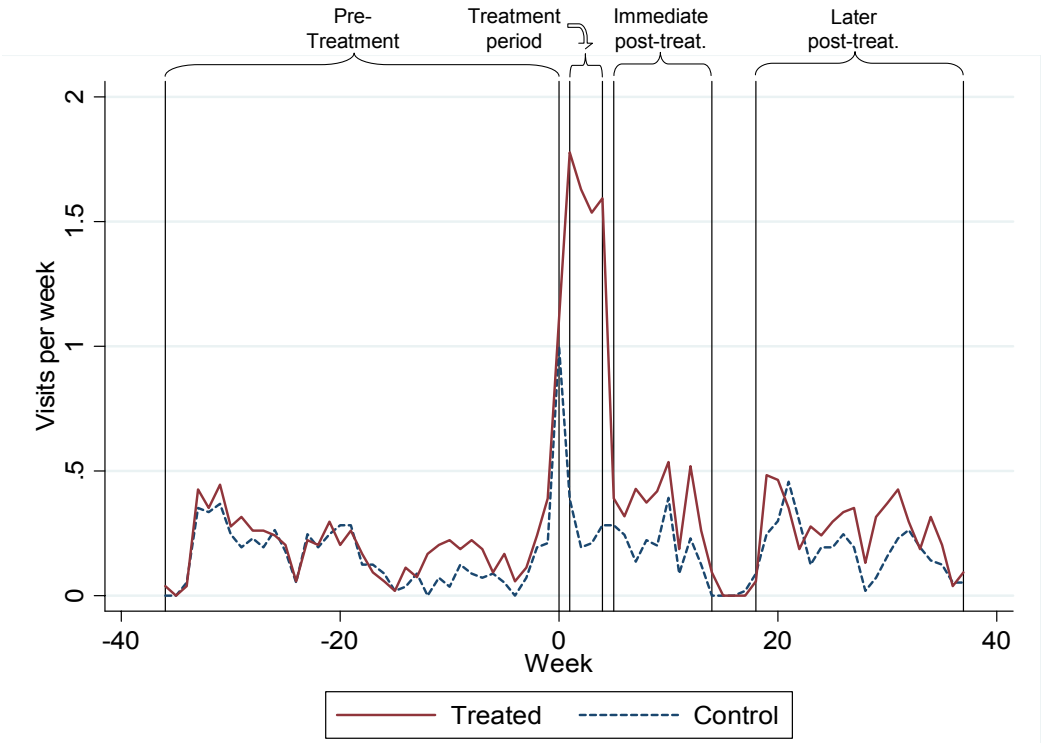


Figure 2: Gym Attendance



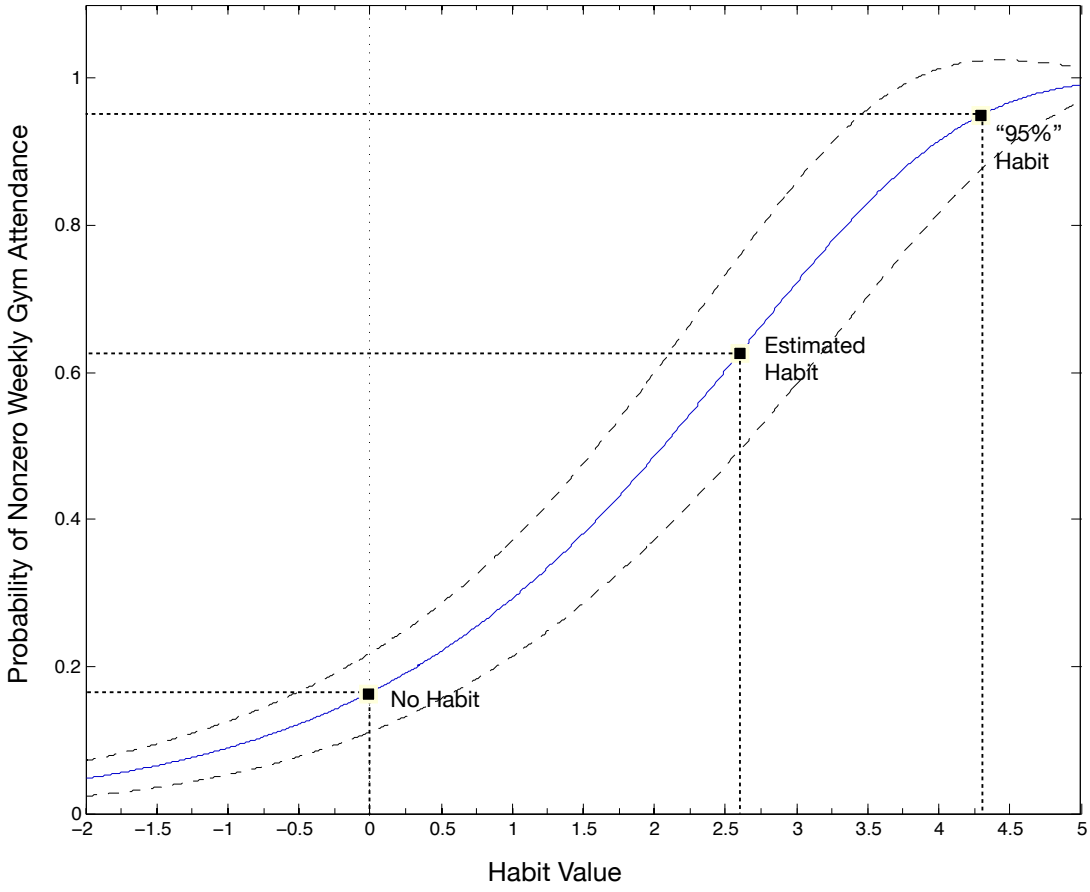
Notes: Average weekly gym attendance, by treatment group status. Weeks in which a subject received a p-coupon for attendance are omitted from this figure.

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]Imm. Post-Treatment refers to the 8 weeks following the intervention (excluding the buffer week for columns (1) and (2). Later Post-Treatment refers to the 19 weeks of observations in the following semester (excluding the semester break).

Figure 3: 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 2: Misprediction of attendance — Predicted and actual target-week 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-trmt prediction (Pre)	3.868	4.053	1.453	3.63	3.963	1.333
Post-trmt prediction (Post)	3.395	3.614	1.058	3.185	3.056	1.313
Actual attendance	1.561	1.561	0.264	1.463	1.463	0.396
Pre minus Actual	2.307	2.491	1.189	2.167	2.500	0.934
(St. Error)	(0.297)	(0.235)	(0.153)	(0.350)	(0.318)	(0.189)
Post minus Actual	1.833	2.053	0.774	1.722	1.593	0.917
(St. Error)	(0.321)	(0.299)	(0.142)	(0.315)	(0.299)	(0.171)
Pre minus Post	0.474	0.439	0.415	0.444	0.907	0.021
(St. Error)	(0.159)	(0.179)	(0.160)	(0.243)	(0.220)	(0.135)
No. of observations	57	57	53	54	54	48

Notes: Coupon value refers to subjects' valuations of coupon-week p-coupons divided by the face value of those coupons. Un-incent'd refers to subjects' un-incentivized predictions of target-week attendance, and is separated into coupon weeks ($p > 0$) and zero weeks ($p = 0$).

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***	4.365***	4.953***	4.488***	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. All specifications in this table include individual covariates.

Table 4: Misprediction of treatment effect — Difference-in-differences in predictions.

	Coupon Value $P > 0$	Un-Incent'd $P > 0$	Un-Incent'd $P = 0$	Coupon Value $P = 1$
Post-Trmt X Treated	0.194	-0.288	0.394*	0.327
	(0.269)	(0.261)	(0.225)	(0.314)
Post-Trmt	-0.728***	-0.684***	-0.415**	-0.827***
	(0.157)	(0.165)	(0.173)	(0.193)
Treated	0.003	0.092	-0.187	-0.012
	(0.301)	(0.336)	(0.231)	(0.363)
R-squared	0.292	0.240	0.233	0.327
Num clusters	111	111	101	111

Notes: Robust standard errors in parentheses, clustered by individual. * significant at 10%; ** significant at 5%, *** significant at 1%. The dependent variables in the different columns are identical to those in table 2.

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_ε	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 Additional Materials

A.1 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.2 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

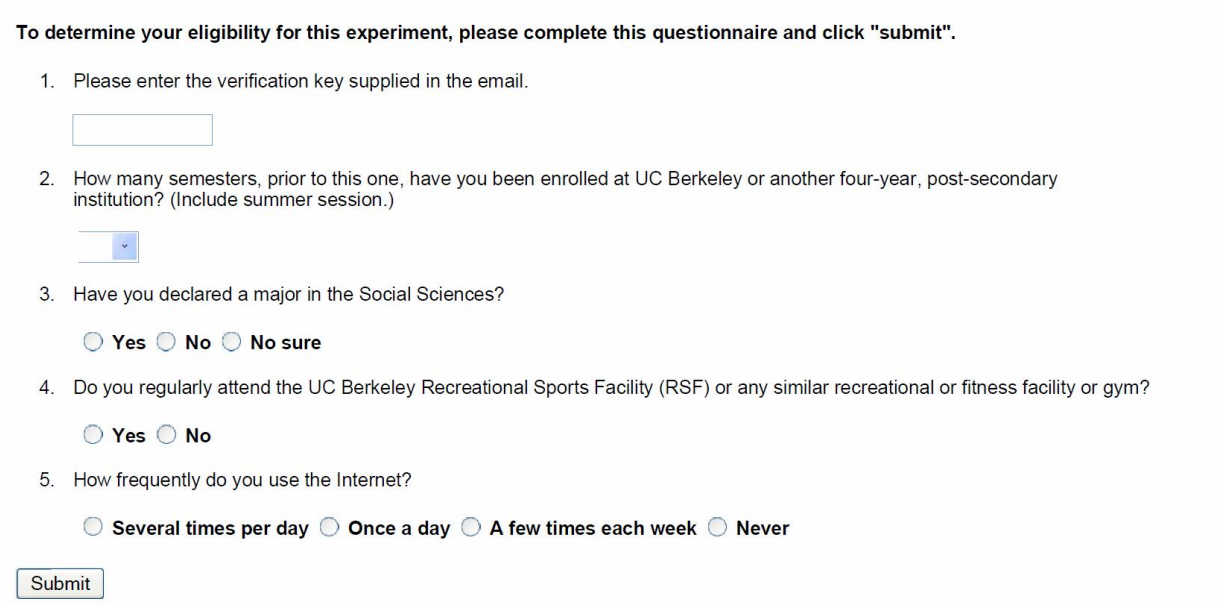
Table A.1: Comparison of Treated and Control groups.

	(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.

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 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



A.3 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.4 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.[36] Despite this titration of rewards, six of the 120 subjects did not complete the study. Two control subjects and two treatment

[36] 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.

subjects left the study between the first and second sessions, and two more treatment subjects left between the second and third. In order to include an additional handful of subjects who were not able to make the third session, and otherwise would have left the study, we held make-up sessions the following day. Four control subjects and two treatment subjects attended these sessions and we have treated them as having completed the study.

Randomizing subjects into treatment and control presented some challenges. Our design required that treatment and control subjects meet separately. For each of the three sessions we scheduled four timeslots, back-to-back, and staggered them between Control and Treatment. When subjects responded to the online solicitation, and after they had completed the screening questionnaire, they were randomly assigned to either treatment or control and were then asked to choose between the two timeslots allocated to their assigned group. Subjects who could not find a timeslot that fit their schedule voluntarily left the study at this point.[37] 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.[38]

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.

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

[38] 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.5 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.