Naiveté, Projection Bias, and Habit Formation in Gym Attendance
Author(s): Dan Acland and Matthew R. Levy
Source: *Management Science*, Vol. 61, No. 1 (January 2015), pp. 146-160
Published by: INFORMS
Stable URL: https://www.jstor.org/stable/24551076
Accessed: 05-04-2022 23:16 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

INFORMS is collaborating with JSTOR to digitize, preserve and extend access to *Management Science*

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

MANAGEMENT SCIENCE
Vol. 61, No. 1, January 2015, pp. 146–160
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2014.2091
© 2015 INFORMS

Naiveté, Projection Bias, and Habit Formation in Gym Attendance

Dan Acland
University of California, Berkeley, Berkeley, California 94720, acland@berkeley.edu

Matthew R. Levy
London School of Economics, London WC2A 2AE, United Kingdom, m.r.levy@lse.ac.uk

We implement a gym-attendance incentive intervention and elicit subjects' predictions of their postintervention attendance. We find that subjects greatly overpredict future attendance, which we interpret as evidence of partial naiveté with respect to present bias. We find a significant postintervention attendance increase, which we interpret as habit formation, and which subjects appear not to predict ex ante. These results are consistent with a model of projection bias with respect to habit formation. Neither the intervention incentives, nor the small posttreatment incentives involved in our elicitation mechanism, appear to crowd out existing intrinsic motivation. The combination of naiveté and projection bias in gym attendance can help to explain limited take-up of commitment devices by dynamically inconsistent agents, and points to new forms of contracts. Alternative explanations of our results are discussed.

Data, as supplemental material, are available at http://dx.doi.org/10.1287/mnsc.2014.2091.

Keywords: behavioral economics; experimental economics; habit formation; present bias; projection bias
History: Received May 14, 2013; accepted September 5, 2014, by Uri Gneezy, behavioral economics.

1. Introduction

Individuals routinely make intertemporal decisions that require them to predict how their preferences, beliefs, and constraints will change over time. For example, if one is deciding today whether to put off until tomorrow an unpleasant task with long-term benefits, it matters whether one correctly predicts one's ability to resist the same temptation tomorrow. If not, one may procrastinate at a potentially unbounded welfare cost (O'Donoghue and Rabin 1999). Similarly, if one is deciding today whether to invest time in a habit-forming activity, it matters whether one predicts how habit-forming it will be, either when the habit is a tempting but harmful one, such as drug addiction, or a salutary one, such as a life-enhancing health behavior (Loewenstein et al. 2003). Moreover, a principal designing a behavior-change incentive scheme must take into consideration whether agents' predictions satisfy rational expectations.

Identifying systematic misprediction of future behavior in the real world has proven challenging. We build upon a field-experimental intervention, which has previously been shown to exogenously induce what has been interpreted as a gym-attendance habit, to explore whether subjects predict the apparent habit-formation process or whether they instead exhibit projection bias with respect to this change in their preferences, as defined by Loewenstein et al. (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 (1999). Finally, our experimental design allows us to offer some suggestive evidence about the countervailing phenomena of habit formation and crowding out of intrinsic motivation in the presence of small incentives.

We take advantage of the experimental intervention of Charness and Gneezy (2009) (hereafter, CG), who paid treatment-group subjects $100 to attend the gym eight times in one month and found an increase in posttreatment attendance compared to a control group, which they interpret as habit formation—ruling out long-term crowding out of intrinsic motivation.[1] We recruit 120 subjects and replicate CG's main gym-attendance intervention. Using this intervention as a means to induce an exogenous shift in gym-attendance preferences, we build it into a larger experimental framework to explore whether subjects

[1] That habit formation plays an important role in physical exercise has long been accepted in the behavioral health literature. See, for example, Valois et al. (1988), Dzewaltowski et al. (1990), Reynolds et al. (1990), Godin et al. (1993), and Godin (1994); however, Charness and Gneezy (2009) provide the first experimental evidence of which we are aware.

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

predict such a shift. Specifically, we elicit subjects' predictions of posttreatment attendance, eliciting both immediately before the intervention (but one week after 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 unincentivized direct prediction task.

We find a significant posttreatment gym-attendance increase of 0.188 visits per week among our subjects, which is smaller than, but statistically indistinguishable from, CG's result. Thus, like CG, we find no evidence of long-term crowding out of intrinsic motivation caused by the month-long treatment intervention, and, like them, we interpret the increase in posttreatment attendance as habit formation, although we discuss alternative explanations. Because we track subjects for longer than CG, we can 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 et al. (2004) find that monetary incentives typically do not have long-run effects that extend even as far as those we identify.

We find that the smaller incentives provided by the contingent-payment contracts we use as part of our prediction-elicitation mechanism also result in increased attendance for both treatment and control groups, with attendance monotonically increasing in the size of the incentive. This suggests that smaller incentives (as low as $1 per visit) did not result in short-term crowding out. However, we find also that the difference in posttreatment attendance between treated and control subjects disappears in weeks when subjects held a contingent-payment contract. Given the small number of weeks with these incentives, however, we cannot establish this result with precision.

With respect to predictions, we find that on average treatment and control subjects significantly overpredict their actual attendance, by a factor in the ranges of 2.5–5.5 for predictions made before the intervention, and 2–4 for those made after the intervention. This result is consistent with at least partial naiveté with respect to future self-control problems caused by present bias: subjects fail to predict how their future desire for immediate gratification will affect their gym attendance. Moreover, comparing unincentivized and incentivized predictions, we find that subjects on average did not value being incentivized to go to the gym, providing another test of their beliefs regarding self-control and demand for commitment. Also, we find that both groups revise their predictions downward after the treatment period. This was not one of the hypotheses we set out to test, but because the initial predictions were at the beginning of the semester, it can potentially be interpreted as an initial misprediction of future time constraints. As we will show, our main misprediction results are robust to this time trend in predictions.

Next we test whether treated subjects correctly predict the increase in posttreatment attendance caused by the intervention. We compare the change over time in treated subjects' predictions to the change in control subjects' predictions. If, prior to the intervention, treated subjects fully predicted the increase in their attendance relative to control subjects, we would expect them to incorporate that increase into their predictions both before and after the intervention, and thus the natural downward revision in their predictions over time (mentioned above) would be the same as for control subjects, ceteris paribus. However, if treated subjects at least partially failed to predict the treatment effect, the downward revision in their predictions would be less than that in control subjects' predictions, because there would be an offsetting upward revision in their predictions after the intervention, once they experienced the attendance increase caused by the intervention. We find that treated subjects do revise downward by less than control subjects, by 0.421 visits per week in weeks when they did not receive an additional contingent-payment incentive—a level not significantly different from the observed treatment effect. This result is consistent with projection bias with respect to habit formation. Interestingly, for weeks in which subjects received contingent-payment contracts, treated subjects' predictions do not differ from control subjects' predictions, either before or after the treatment period. We discuss potential explanations of this result, including the possibility that the external incentives simply dominated the treatment effect, making it statistically unobservable.

The remainder of this paper is organized as follows. Section 2 presents the design of our experiment. Section 3 presents our results on attendance and prediction. Section 4 discusses our findings and concludes.

2. Design

We recruited 120 subjects from the students and staff of the University of California, Berkeley and randomly assigned them to treatment and control groups. Due to attrition and missing covariates, our final sample includes 54 treated subjects and 57 control subjects.[2]

[2] Four treated and two control subjects dropped out of the study. An additional two treated subjects and one control subject had missing data. The difference in attrition and missing data between treatment and control is not statistically significant. Unfortunately, because demographic variables were collected at the third session, we cannot compare quitters across groups. Details of the sample and

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

2.1. Timeline, Interventions, and Predictions

Our intervention period was 11 weeks long and began with the second full week of the fall semester of 2008. In the first week of the intervention period—the "learning week"—we paid both groups \$25 if they visited the campus gym one time, to make sure that all subjects overcame any one-time fixed cost of initial gym attendance such as learning how to get to the gym, or how to find the locker room, etc. In this way we hoped to separate true habit formation, resulting from multiple visits, from any increase in posttreatment attendance that might be caused by simply having overcome fixed costs. During the subsequent four weeks—the "treatment month"—subjects in the treatment group received \$100 if they visited the gym twice a week, for a total of eight visits. The learning-week offer is identical to the low-incentive condition of CG, so our control group corresponds to their low-incentive group, and our treatment group is almost identical to their high-incentive group. Our design required two visits per week, theirs eight in a month. We hoped that this change would limit the potential for procrastination so that naive present-biased subjects in the treatment group would be more likely to meet the eight-visit threshold. Our compliance rate was not distinguishable from that under CG's less-restrictive design. The treatment month was followed by a buffer of one week, after which there were five "target weeks," for which subjects made attendance predictions, as described below. Gym attendance was tracked by using card-swipe data.

During the course of the intervention period, subjects were convened for three meetings with researchers: on the first day of the learning week, on the first day of the treatment month, and on the first day of the buffer week. For manageability, each group was broken into two sessions at each meeting, with treatment and control group sessions staggered over the course of an afternoon. To minimize attrition, a participation payment of \$25 was given for attending the first meeting, and another of equal size for attending both of the subsequent meetings. In addition, subjects were reminded of each session by email. At the first session, the learning-week offer was announced to all subjects and the treatment-month offer was announced to treatment-group subjects. In addition, subjects were asked to complete a questionnaire. At the second meeting subjects were asked to complete a series of tasks (described below) intended to elicit gym-attendance predictions for each of the five target weeks. We refer to these as "first-elicitation" predictions. At the time of elicitation, both groups were reminded of the offers they had received the previous week. At the third meeting, subjects were asked to complete the same set of tasks, to elicit their "second-elicitation" predictions for the same five target weeks. Thus, we have predictions of posttreatment target-week gym attendance from immediately before (first elicitation) and immediately after (second elicitation) the treatment month. To the extent that the treatment-month offer resulted in habit formation, these predictions can be thought of as being made before and after the state change from nonhabituated to habituated.[3] Finally, at the end of the third meeting, subjects were asked to complete an additional questionnaire. The timeline of the experiment is illustrated in Figure 1. We discuss compliance with the treatment-month offer, attrition, and our randomization procedure in §A.2 of the appendix.

We announced our treatment-month offer at the initial meeting, one week before the first-elicitation meeting, to provide treated subjects with time to adapt to the prospect of earning an additional \$100, before giving them an elicitation task involving uncertain monetary payoffs (see below). The buffer week was inserted between the treatment month and the target weeks so that present-biased subjects would view the first target week as being in the future from the perspective of the second-elicitation meeting, in an attempt to eliminate any difference between first- and second-elicitation predictions that might have been caused by naive present bias. We are able to confirm that this succeeded in the §3.

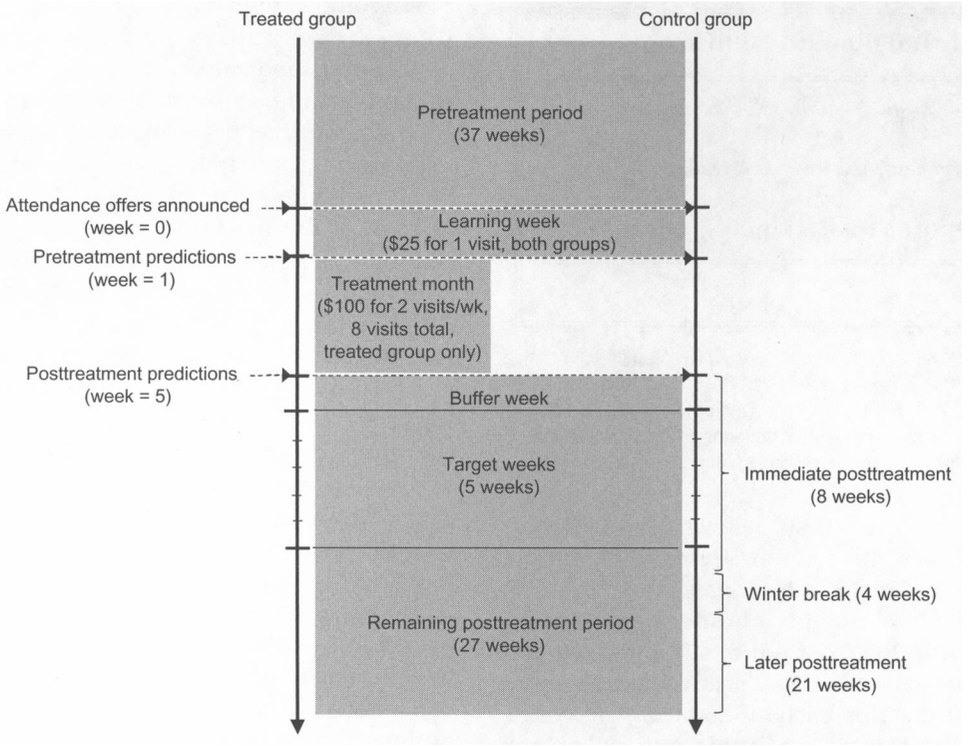
Gym attendance data were collected for a 17-month period stretching from 37 weeks before the learning week to 27 weeks after the end of the fifth target

comparisons of treated and control subjects appear in §A.1 of the appendix. CG found that habit formation was greatest among previous nonattenders, so we screened for subjects who self-reported that they did not regularly attend any fitness facility, with the intent that this would give us greater prevalence of habit formation, and thus greater power to investigate predictions of habit formation. Subjects were unaware that the experiment was related to exercise at the time of recruitment, and there is no evidence that the subjects in our sample manipulated their answers to the screening mechanism to obtain eligibility. Based on the pretreatment pattern of attendance, the screening mechanism appears largely successful at eliminating "regular" gym-goers: 48% never attended at all during the pretreatment period, and 64% did not have gym memberships at the start of the experiment. None of our subjects had average weekly attendance as high as twice a week in the pretreatment period, and only four (3%) had average attendance of once a week. There is no difference in the proportion of treatment and control subjects with nonzero pretreatment attendance. We explore the effect of pretreatment attendance in §B.1 of the online appendix (available at http://dx.doi.org/10.1287/mnsc.2014.2091). We paid the \$10 gym-membership fee for all subjects, filing the necessary membership forms for those who were not already members, and reimbursing the fees of those who were.

[3] We recognize that there are other interpretations for any attendance effect of the treatment-month offer, and we address these alternatives in subsequent sections.

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

Figure 1 Experimental Timeline



week. This period includes summer and winter academic breaks as well as three full semesters. We used administrative data on ID card swipes required for entrance to exercise facilities and locker rooms. Because swipes were necessary to enter facilities but not to exit, we cannot determine the length of a visit. Subjects were told that payments were contingent upon "swiping in" at the gym and engaging in at least 30 minutes of physical activity, but we were not able to observe gym activity, and subjects were aware of this. We acknowledge that some of the recorded swipes during the treatment month, or when holding a contingent-payment contract during the target weeks, may represent subjects swiping to receive the reward but not exercising. However, there would be no reason to continue to engage in such behavior when incentives were absent during the posttreatment period. To the extent that some subjects may have swiped without exercising during the treatment month, our test of habit formation is biased downward, i.e., against the habit-formation finding we report. Furthermore, subjects who used the locker room in addition to exercise facilities, and are thus less likely to have engaged in swiping without exercising, are identifiable in our data, because they had to swipe into the locker room both before and after exercising, in addition to swiping into the exercise facilities. Restricting analysis to this subset does not appear to affect any of our results.

2.2. Elicitation Procedures

Incentive-compatible elicitation of subjects' predictions of their future behavior is complicated by the fact that any scheme to incentivize truthful revelation will also incentivize a change in the behavior being predicted. If subjects value the change in their behavior caused by the incentive—for example, because the incentive helps them to overcome a self-control problem in gym attendance—it may further affect their response to elicitation tasks. There is no accepted solution to this problem. We use two strategies that, together we believe give us a reasonably good claim to have elicited predictions usefully, if not perfectly. The first strategy involves eliciting incentive-compatible valuations for a cash-reward certificate that functions as a contingent-payment contract, the value of which is proportional to the number of gym visits during a future time period. The second involves simply asking subjects how often they believe they would attend the gym during the specified time period if they held such a certificate.

To implement these two strategies, we created a set of "Daily RSF Reward Certificates," each of which entitled the holder to a given number, p, of dollars for each day they attended the campus gym (the "RSF") during a given target week. For conciseness we refer to these certificates hereafter as "p-coupons." Figure 2 shows a sample p-coupon, with $p = 1$. The value that subjects place on a p-coupon is determined, in part,

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

Figure 2 A Sample "*p*-Coupon"

Notes. A *p*-coupon provides a small subsidy, p, for each gym visit during an indicated target week. Subjects were asked to predict their subsidized attendance for $p \in \{1, 2, 3, 5, 7\}$ ("unincentivized predictions") and to give their dollar valuation for the same set of coupons ("coupon valuations"). Most subjects ultimately received exactly one *p*-coupon.

by the number of times they believe they will visit the gym if they own the coupon. There were 5 target weeks, and the parameter p took values of 1, 2, 3, 5, and 7, so that the total number of different *p*-coupons was 35.[4] At both the first- and second-elicitation meetings, we presented each subject with a series of four *p*-coupons, one for each of four target weeks, and asked them to complete a multiple-price-list task to determine their dollar valuation for each of the coupons. Importantly, the task was framed as a choice between the *p*-coupon and a fixed amount of money *on the maturity date of the p-coupon*, which means that time discounting affects both options equally and thus does not affect coupon valuations.[5] By dividing a subject's willingness to pay for a coupon by its face value, p, we computed what we refer to as the normalized coupon valuation, which is *approximately* how many times the subjects believe they will attend the gym during the specified target week under the subsidy provided by the coupon. Presenting each subject with four *p*-coupons left each subject with one target week for which they were not asked to value a coupon. The value of p for each coupon, the week for which there was no coupon, and the order of the target weeks for which the coupons were presented were all randomized across subjects, using a mechanism that ensured a predetermined number of coupons for each value of p, each week, and each possible ordering, and the same distribution in treatment as in control. Immediately after each multiple-price-list task, subjects were asked to indicate the number of days they believed they would attend the gym during the given target week, if they actually held the *p*-coupon for which they had just indicated their dollar valuation. We refer to these as unincentivized predictions.

To ensure that first- and second-elicitation predictions were comparable, we presented each subject with the same series of coupons, in the same order, in each of the meetings. In addition, we selected, at random, only one subject to receive the indicated payoff of only one of their multiple-price-list tasks. Thus, while maintaining incentive compatibility, we minimized the number of instances in which a subject would, at the second-elicitation meeting, be valuing a *second* *p*-coupon for a given target week.[6] At the time of the first elicitation, subjects did not know that the second elicitation would take place. Thus, subjects' first-elicitation valuations were not confounded by uncertainty about possible future *p*-coupons. Also, subjects were instructed that any payment they received as a consequence of their price-list choices—either the payment from the *p*-coupon or the alternative fixed payment—would be issued at the conclusion of the associated target week, so that both options would be subject to the same intertemporal discounting.

Finally, at the end of the second-elicitation meeting, after all of the elicitation tasks had been completed, we gave each subject, to keep, one of the four coupons they had been presented with during the elicitation process, selected at random.[7] The giveaway was a surprise to the subjects—having been conducted unannounced only after the second round of elicitations was completed—and thus did not affect their *p*-coupon valuations or unincentivized predictions during the elicitation tasks. We therefore have two target weeks for each subject for which we can compare their actual gym attendance with their predictions of attendance under the same attendance-incentive conditions: the week for which they actually received a coupon, and the week when no coupon was presented either for predictions or attendance.

The multiple-price-list task is incentive compatible for subjects' valuations of the *p*-coupons, but those valuations are affected by at least three factors in addition to their gym attendance prediction. The first is risk preference: in the face of uncertainty about future preferences and future time and budget constraints, concavity of the utility function for wealth

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

[5] The task asked subjects to make a series of choices between a *p*-coupon and an incrementally increasing fixed amount of money equal to an integer multiple of p. We infer their valuation from the implied indifference point between the coupon and the fixed sum. Subjects were instructed that at most one choice would be implemented, which preserves independent valuations across tasks. The task is described in detail in §B.2 of the online appendix

[6] In these few cases, predictions were simply assumed to be for attendance under the condition of holding a coupon worth $2p$, or subjects were dropped from the analysis.

[7] Here again we randomized in such a way as to ensure predetermined numbers of coupons for each value of p, and for each target week, and the same distribution across groups.

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

(or of the value function in a reference-dependence model) would lower the value of a *p*-coupon. The second factor is the time and effort cost of gym attendance, which we assume would be negative for all subjects and would thus also lower the value of a *p*-coupon. The third factor is the long-term health value of the additional gym attendance incentivized by a *p*-coupon. Present-biased subjects who are at least partially sophisticated about their future self-control problems may value the *p*-coupon as a device to increase the likelihood that their "future self" will attend the gym, which would drive up the value of the coupon. It is possible that this third term could dominate the second for sophisticated present-biased agents. For simplicity, and because we have no way to distinguish them in our data, we refer to the net value of effort cost and long-term health value accruing as a result of *p*-coupon-incentivized attendance as the net "commitment value" of a *p*-coupon. Thus, care must be taken not to interpret subjects' normalized valuations as exactly proportional to their predictions. Our goal in designing the experiment was that additionally asking subjects to directly predict their attendance under the conditions created by holding a *p*-coupon would provide us with a second approach to eliciting predictions, and that the combination of the two approaches would provide us with useful information about predictions.

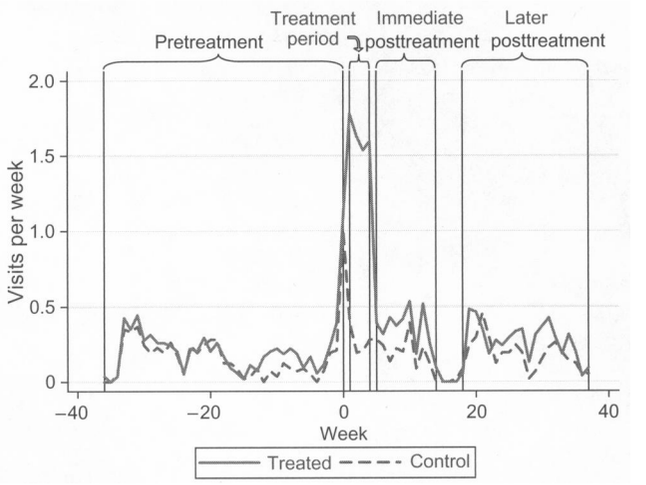
3. Results

Of the 54 subjects in our final treatment sample, 43 completed the eight necessary semiweekly visits to earn the $100 treatment-month incentive, a completion rate of 80%. In CG's high-incentive group the completion rate was approximately 83%, suggesting that our more restrictive semiweekly requirement did not have a substantial effect on subjects' ability to make the required number of visits.

3.1. Posttreatment Attendance

Figure 3 shows average weekly attendance for the treatment and control groups over the duration of the study period. In this and the related regression analysis reported in columns (1)–(3) of Table 1, we have removed observations for each subject for the target week in which they received a *p*-coupon in the giveaway, to isolate the main effect of the treatment-month intervention from the effect of the attendance subsidies provided by the coupons. The broad pattern of results is as follows. In the pretreatment period, attendance in the two groups moves together tightly, as we would expect. In the treatment period, treated subjects attend much more than control subjects, reflecting the effectiveness of our treatment-month incentive. In the immediate posttreatment period—the two months between the end of the treatment

Figure 3 (Color online) 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.

period and the beginning of the semester break—the treated group continues to attend the gym more than the control group. We follow CG in interpreting this as supporting the hypothesis of habit formation, though we discuss alternative explanations later in this section.

During the four weeks of the semester break there is essentially no attendance in either group. In the later posttreatment period—the four months after the semester break—the difference between groups is greatly diminished, suggesting that the attendance increase induced by four weeks of exogenous gym attendance largely decayed over a similar-length period of quasi-exogenous nonattendance.

We estimate a linear difference-in-differences panel regression model to determine if these patterns are statistically significant. Each observation in the panel is the attendance of 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 the interactions of the treatment-group dummy with dummies for the treatment period and each of the two posttreatment periods. To increase the precision of our analysis we include individual fixed effects.[8] We regress the equation:

$$\begin{aligned} Attendance_{it} = {} & \gamma_1 \cdot During_t + \gamma_2 \cdot During_t \cdot TR_i + \gamma_3 \cdot Post_t \\ & + \gamma_4 \cdot Post_t \cdot TR_i + \gamma_5 \cdot LatePost_t \\ & + \gamma_6 \cdot LatePost_t \cdot TR_i + \mu_i + \varepsilon_{it}, \end{aligned} \quad (1)$$

[8] We thus do not estimate the direct effect of treatment group assignment in the pretreatment period. In Table A.1 of the appendix, we confirm that there are no preexisting differences.

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

Table 1 Effect of Treatment and *p*-Coupons on Average Weekly Attendance

	(1)	(2)	(3)	(4)	(5)	(CG)
During	0.128**	0.128**	0.128**	0.128**	0.128**	0.222**
	(0.064)	(0.064)	(0.064)	(0.064)	(0.064)	(0.103)
Post	0.060	0.060	0.101	0.131*	0.101	0.054
	(0.058)	(0.057)	(0.064)	(0.074)	(0.064)	(0.098)
LatePost	0.047	0.047	0.047	0.047	0.047	
	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	
During × *TR*	1.449***		1.449***	1.449***	1.449***	1.497***
	(0.117)		(0.117)	(0.117)	(0.117)	(0.148)
Post × *TR*	0.188**		0.228**	0.197	0.228**	0.640***
	(0.095)		(0.114)	(0.119)	(0.114)	(0.192)
LatePost × TR	0.096		0.096	0.096	0.096	
	(0.074)		(0.074)	(0.074)	(0.074)	
During × *Complied*		1.787***				
		(0.132)				
Post × *Complied*		0.221*				
		(0.118)				
LatePost × *Complied*		0.109				
		(0.093)				
$p=1$				0.579*	0.407	
				(0.294)	(0.253)	
$p=2$				0.842**	1.369**	
				(0.333)	(0.569)	
$p=3$				0.864**	0.258	
				(0.362)	(0.218)	
$p=5$				2.298***	3.141***	
				(0.531)	(0.655)	
$p=7$				2.433***	2.824***	
				(0.591)	(0.783)	
$TR \times p=1$					0.391	
					(0.626)	
$TR \times p=2$					−1.089*	
					(0.621)	
$TR \times p=3$					1.162*	
					(0.662)	
$TR \times p=5$					−1.687*	
					(1.002)	
$TR \times p=7$					−0.784	
					(1.160)	
R^2	0.14	0.15	0.14	0.19	0.19	0.20
N	7,433	7,433	7,100	7,211	7,211	1,520
Subjects	111	111	111	111	111	80
Sample	All noncoupon weeks	All noncoupon weeks	Noncoupon target weeks	All target weeks	All target weeks	CG
Treatment period Chow test:						0.84
Posttreatment Chow test:						0.06

Notes. Observations of weekly attendance at the subject × week level are reported. Robust standard errors are given in parentheses, clustered by individual. *TR* denotes the treated group. *During*, *Post*, and *LatePost* indicate, respectively, the treatment month, the period between the treatment month and the semester break (8 weeks in column (1), 5 weeks in the remaining columns), and the 19 weeks of observations in the following semester (excluding the semester break). *Complied* is defined as attending 8 times during the treatment month. Chow tests refer to a test of equal coefficients between our sample (column (1)) and Charness and Gneezy (2009)'s sample.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

where $Attendance_{it}$ is subject i's attendance in week t; $During_t$, $Post_t$, and $LatePost_t$ denote the intervention month, immediate posttreatment period, and later posttreatment period, respectively; and TR_i indicates whether subject i is assigned to the treatment group.

The results of this regression appear in column (1) of Table 1. The coefficient on the interaction of the treatment-group and treatment-period dummies is 1.449, roughly the product of the twice-weekly incentive target and the 80% 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 treatment-group attendance in the two posttreatment periods.

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

The point estimate is 0.188 additional visits per week for the immediate posttreatment period, representing approximately a doubling of average attendance in our sample. In the later posttreatment period, after the winter break, we see no statistically significant difference between the groups, with a point estimate of 0.047 additional visits per week. Restricting the immediate posttreatment period to the five target weeks raises the effect on treated subjects to 0.228 visits in column (3), consistent with our interpretation of a decaying habit.

Because not all treatment-group subjects made the requisite eight visits to the gym, column (2) of Table 1 represents the intention-to-treat effect (ITT). We use the treatment-group dummy to instrument for compliance using two-stage least squares to estimate the average treatment effect on the treated (ATT), controlling for baseline differences in attendance. This analysis implicitly assumes that there is no effect on subjects who did not meet the eight-visit threshold, which is not implausible, given that such subjects averaged just two visits during the treatment month, and none went as many as seven times. These results are reported in column (2) of Table 1, where the increase in immediate posttreatment treatment-group attendance is now just under one-quarter of a visit per week on average.

To compare our results with those of CG we show the results of running the ITT regression on their data in the final column of Table 1.[9] The double difference in average weekly attendance between their high-incentive and low-incentive subjects in the immediate posttreatment period was 0.640 visits per week.[10] Stacking their data with ours allows us to conduct a Chow test of the equality of their treatment-effect coefficient with the one in our column (1) specification. The p-value, reported at the bottom of the table, is 0.06. Thus, while we estimate a smaller effect, we do not reject that the treatment effect was the same across studies.

Like CG we explore heterogeneity in the treatment effect. Our approach is to repeat the analysis of Table 1, but to allow for an individual-specific treatment effect during the immediate posttreatment period. We include individual fixed effects as before, and thus our hypothesis is that the distribution of individual treatment effects among treated subjects will first-order stochastically dominate the distribution among controls. We plot each distribution in Figure A.1 of the appendix, and it is clear on inspection that the treatment group includes significantly more subjects with large increases in gym attendance. A Mann–Whitney test rejects equality of the distributions at $p = 0.048$, and a Kolmogorov–Smirnov test rejects equality at a stronger $p = 0.01$. Moreover, the distribution of habits among treated subjects appears to be bimodal, with one group centered around zero change and another centered around an increase of approximately 1–1.5 visits per week. In Table A.1 of the appendix, we find no significant predictors of which subjects experienced a large increase in posttreatment attendance, other than agreement with the statement "fitness is important."

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 treatment group merely swiped their ID cards at the gym during the treatment period and did not actually exercise. Because such subjects would not experience a treatment effect, our estimates of the treatment effect in Table 1 would be biased toward zero by their presence.[11] The fact that any such subjects would have no treatment effect to mispredict would also reduce the power of our test of projection bias. An alternative explanation is that some subjects would have experienced an increase in postintervention attendance if the intervention period had been longer. This interpretation is consistent with recent findings such as those of Lally et al. (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, for some subjects, a period of induced gym attendance simply does not increase their willingness or desire to attend the gym in the future.

Our interpretation of the treatment effect as habit formation is useful for fitting the results into the framework of existing economic theory, but it obscures important distinctions. In the standard economic model of habit formation developed by Becker and Murphy (1988), any increase in the marginal utility of a good or activity that results from past consumption (i.e., any "adjacent complementarity" in consumption across periods) regardless of the psychological, physiological, or even social cause, is labeled as habit formation. One possible explanation of our treatment effect is that subjects who have not previously exercised at a gym may have systematically downward-biased beliefs about the net benefits of doing so, and a month of attendance may cause them to learn the true values. Another possibility is that our treatment effect could be the result of treated subjects forming social connections with people they interact

[9] This specification differs from the one they report, which uses pre- and posttreatment averages rather than the full panel of weeks. We thank CG for making their data available.

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

[11] As an additional robustness check, we repeat the analysis of Table 1 using card swipes at the gym locker room rather than the front entrance. The short-run effect in this specification increases to 0.214 (s.e. 0.075).

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

with at the gym, leading to an increased desire to attend the gym for social reasons. Both possibilities could generate a positive treatment effect, but would not generally be considered habit formation, as it is commonly understood.

These distinctions matter, both for our understanding of behavior change, and for the policy implications of our findings. If the learning story is correct, there are alternative approaches to correcting false beliefs that might be more cost-effective than incentives, and more feasible at the population level. Meanwhile, the social connection story would imply that incentives need to be tied to exercise in a social context per se, as opposed to in isolation, and would suggest a range of alternative interventions to bolster the social rewards of exercise. It is possible that incentives might be an effective and/or cost-effective way to increase long-term gym attendance, regardless of the underlying mechanism, but for both academic and policy reasons, it would be valuable to understand the mechanism more fully. One might expect that learning the true utility of gym attendance or building social ties would persist longer than the four weeks of winter break, but we are a long way from drawing firm conclusions on this matter.

3.2. Short-Term Attendance Under p-Coupons

There is a growing literature suggesting that small incentives can lead to "crowding-out" of intrinsic motivation.[12] Our relatively large treatment-month incentive increased treated subjects' attendance both during and after the intervention, which, like CG, we interpret as ruling out crowding-out both while incentives are in place (short term) and after removal of incentives (long term). However, it is possible that the smaller incentives of the p-coupons could lead to crowding out. In column (4) of Table 1, weeks in which p-coupons were actually received have been added back into the sample, and the immediate post-treatment dummy has been interacted with dummies for the different values of p. The coefficients for these interactions show the increase in attendance for each level of p relative to weeks without p-coupon subsidies, pooling treated and control subjects, and controlling for the difference between treatment and control. The effect is positive, and statistically significant at the 10% level or above, for all values of p, despite the small cell size when cutting the data so finely. The effect is monotonically increasing in p, as predicted by standard theory.

An interesting result appears when we look at the differential effect of p-coupons between treated and control subjects. The regression reported in column (5) of Table 1 shows the difference in p-coupon effect between treatment and control groups for each value of p. The uninteracted values of p reflect the effect of p-coupons on control-group attendance. The pattern is qualitatively similar to the effect in the pooled sample. The interaction terms show the difference in attendance under p-coupons for treated subjects relative to control subjects. These terms simultaneously capture the short-term effect of p-coupons on treated subjects and the long-term effect of the treatment-month incentive. The coefficients are variously positive and negative, and none is significant above the 10% level. Only for $p = 3$ can we reject that the difference between treated and control subjects in a week with coupons is zero (i.e., we reject that $Post \times TR + TR \times p = 3$ is zero). Pooling all nonzero coupons together, we do not reject the null that there is no relative increase for treated subjects in coupon weeks as a whole.

It is not obvious how these results should be interpreted. Just as there is no evidence of crowding out from the larger treatment-month incentive among treated subjects, so there is no evidence of crowding out from p-coupon incentives among the control group. However, the relative increase in attendance between treatment and control is not observed among treated subjects on weeks with a p-coupon. One possibility is that the small incentives from the p-coupons crowd out the intrinsic motivation recently generated by the treatment effect.[13] However, we find no evidence of persistent crowding out on subsequent weeks, as might be expected under this hypothesis. Our study does not have power to explore this with any precision. Alternatively, since the treatment effect occurred in a smaller subset of treated subjects than the effect of p-coupon incentives, it may be that the combined effect of the two in those for whom there was a treatment effect is simply drowned out in our data.

3.3. Predictions

Next we explore the relationship between predictions and actual attendance, as well as the relationship between first-elicitation and second-elicitation predictions. In columns (1) and (4) of Table 2, we regress our two measures of target-week attendance predictions—subjects' unincentivized predictions of attendance under p-coupon subsidies, and the predictions implied by their normalized valuations of

[12] See Gneezy et al. (2011) for one review.

[13] Another possibility is that the treatment effect and the effect of the p-coupon incentives may simply be nonadditive, in a way that shrouds their combined effect.

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

p-coupons[14]—on the posttreatment dummy and dummies for values of p. We estimate

$$Prediction_{iwt} = \gamma_1 \times 2nd_t + \theta' \mathbf{p}_{iw} + \mu_i + \varepsilon_{iwt}, \quad (2)$$

where $Prediction_{iwt}$ is subject i's predicted attendance for week w at time t.

Comparing either column to column (4) of Table 1 reveals an unambiguous pattern: Subjects in both groups dramatically overpredict their future gym attendance, in both the first and second elicitations, by as much as two visits per week, for all values of p and for weeks without p-coupons. For every level of p-coupon subsidy, postestimation tests reject at $p < 0.01$ that either prediction measure equals actual attendance.

These results allow us to rule out, in the β, δ model of present-biased preferences, both rational time consistency ($\beta = 1$) and fully sophisticated present bias ($\beta < 1, \hat{\beta} = \beta$), both of which imply correct attendance predictions on average. Under this interpretation, the results provide direct evidence of naiveté in prospective gym-attendance plans, which complements naiveté findings of DellaVigna and Malmendier (2006) based on procrastination in gym-contract cancelation.[15]

It is also worth noting in Table 2 that unincentivized predictions and normalized coupon valuations are quite close for all values of p, suggesting that our two elicitation methods are similarly valid as measures of beliefs. However, in general, normalized valuations are lower than unincentivized predictions. As discussed in §2, this undervaluation could either be caused by risk attitudes or by the commitment value of p-coupons—the value of the combined short-term cost and the long-term health benefit of the additional attendance incentivized by a p-coupon. To help to distinguish between risk attitudes and commitment value, we elicited a measure of risk aversion over small-to-moderate stakes using hypothetical lotteries.[16] We find no effect of this measure on undervaluation, suggesting it is not driven by risk preferences. Although our measure of risk aversion may simply be noisy, this test supports the interpretation that the commitment value of a p-coupon is negative for subjects. This would be the case for a time-consistent subject or for a sufficiently naive present-biased subject, both of whom believe that the p-coupon will increase their attendance only when the net marginal utility of attendance is negative. Given that time-consistency is incompatible with systematic misprediction, these results lend additional support to the conclusion that our subjects are naive about their self-control problems.

In addition to the dramatic overprediction reported above, there is a striking pattern of subjects revising their predictions downward over the course of the treatment month. This is reflected in the coefficient on the *2nd elicitation* dummy in Table 2. Predictions on average decreased by approximately three-quarters of a visit per week between the first- and second-elicitation meetings. These meetings differ in two ways: first, they are separated by a month, during which subjects' lives may have changed in behaviorally relevant ways; and second, the second-elicitation meeting is closer in time to each of the target weeks than the first-elicitation meeting. One possible explanation of the downward revision in predictions between meetings is that the extent to which subjects discount future utility (their per-period discount rate) increases smoothly with temporal proximity, rather than abruptly as in the quasi-hyperbolic model. Another is that subjects' naiveté about future self-control problems decreases with temporal proximity. Both seem psychologically plausible. We can test for whether either of these explanations is sufficient because the temporal proximity of the first elicitation to the first target week is the same as that of the second elicitation to the fifth target week. Thus, when we restrict the regressions reported in columns (2) and (5) to these weeks, the coefficient on the posttreatment dummy reflects only the effect of the second elicitation being later in absolute terms than the first. The results are virtually identical to those using the full sample of target weeks. Thus we can rule out that temporal proximity alone explains the downward revision in predictions. Rather, there appears to be an effect of the second elicitation being later than the first elicitation in absolute terms. Running the regression using the *difference* between prediction and attendance confirms that predictions are becoming more accurate over time. This secular decrease in misprediction suggests that subjects may begin the semester with incorrect beliefs about some aspect of the gym-attendance decision. One possibility is that they are initially overly optimistic about their future free time and grow more realistic once they get a few weeks into the semester.[17] Any number of alternative explanations could explain the results.

[14] As explained in §2, this is simply subjects' valuations for a p-coupon divided by the per-visit subsidy, p.

[15] Furthermore, although 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), they do suggest that such models cannot fully explain our data, as they too embed rational expectations about choices, which are clearly violated here.

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

[17] See, for example, Bénabou and Tirole (2002) for why subjects may begin the semester with overly optimistic beliefs. Another alternative would be that subjects begin the semester with high intrinsic motivation to attend the gym, and project this enthusiasm in their

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

Table 2 Predictions of Target-Week Attendance

	Prediction			Cpn value		
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	1.573***	1.440***	1.333***	3.440***	3.450***	3.466***
	(0.120)	(0.292)	(0.121)	(0.101)	(0.321)	(0.105)
$p=1$	1.742***	1.854***	2.012***			
	(0.150)	(0.388)	(0.240)			
$p=2$	2.095***	2.305***	2.533***	0.159*	0.291	0.136
	(0.149)	(0.407)	(0.227)	(0.083)	(0.506)	(0.135)
$p=3$	2.467***	2.735***	2.854***	0.391***	0.622	0.364**
	(0.152)	(0.442)	(0.232)	(0.103)	(0.409)	(0.183)
$p=5$	3.101***	3.105***	3.412***	0.822***	0.862*	0.844***
	(0.172)	(0.409)	(0.191)	(0.130)	(0.446)	(0.199)
$p=7$	3.743***	3.927***	4.657***	1.300***	1.583***	0.919**
	(0.291)	(0.738)	(0.418)	(0.325)	(0.226)	(0.434)
2nd elicitation	−0.720***	−0.783***	−0.423***	−0.631***	−0.583**	−0.858***
	(0.111)	(0.183)	(0.160)	(0.133)	(0.226)	(0.187)
$TR \times p=1$			0.095			
			(0.357)			
$TR \times p=2$			−0.277			−0.015
			(0.345)			(0.187)
$TR \times p=3$			−0.117			−0.031
			(0.352)			(0.237)
$TR \times p=5$			−0.233			−0.170
			(0.365)			(0.296)
$TR \times p=7$			−0.469			1.181
			(0.872)			(1.033)
$TR \times$ 2nd			0.421**			0.358
			(0.209)			(0.298)
$TR \times 2nd \times p=1$			−0.667*			
			(0.347)			
$TR \times 2nd \times p=2$			−0.731**			−0.316
			(0.323)			(0.225)
$TR \times 2nd \times p=3$			−0.650**			0.018
			(0.323)			(0.259)
$TR \times 2nd \times p=5$			−0.763**			−0.232
			(0.341)			(0.307)
$TR \times 2nd \times p=7$			−1.221			−2.158*
			(1.278)			(1.211)
R^2	0.49	0.51	0.51	0.17	0.20	0.18
N	1,088	217	1,088	875	176	875
Subjects	111	111	111	111	110	111
Sample	Full	5-wk delay	Full	Full	5-wk delay	Full

Notes. Predictions and coupon valuations at the subject × target-week × elicitation-session level. "Prediction" refers to subjects' unincentivized predictions. "Cpn value" refers to subjects' valuations of *p*-coupons divided by the face value of those coupons. *TR* is an indicator for treated subjects; *2nd elicitation* indicates the posttreatment elicitation session. The "5-wk delay" sample includes only predictions for target weeks five weeks after each elicitation session. All specifications include individual fixed effects and cluster standard errors by subject.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Finally, we turn our attention to whether subjects predict the postintervention attendance increase reported in §3.1. To do so we begin with the assumption that second-elicitation predictions reflect correct beliefs about the treatment effect, because by the time of the second-elicitation meeting, treatment subjects have experienced the treatment effect. If treated subjects correctly foresee the ways in which the treatment month will affect their posttreatment attendance, their predictions should change over time in the same way as those of control subjects. Both their first- and second-elicitation predictions will be higher than those of control subjects, so the downward revision will be the same. If treated subjects fail to predict the treatment effect, we would expect to see a smaller downward revision in their predictions, because there would be an offsetting upward revision caused by the treatment effect becoming known. We test for this

beliefs about late-semester attendance. Both would correspond to an upward shift in subjects' first set of predictions, but our data cannot distinguish these and other similar mechanisms.

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

with a series of difference-in-differences regressions of unincentivized predictions and normalized coupon valuations.

In columns (3) and (6) of Table 2 we regress predictions on dummies for values of p; a dummy for predictions being posttreatment, rather than pretreatment; and the full set of interactions of these terms with each other and with the treatment dummy. We estimate the equation

$$\begin{aligned} Prediction_{iwt} = {} & \gamma_1 \times 2nd_t + \gamma_2 \times TR_i \cdot 2nd_t + \theta_1' \mathbf{p}_{iw} \\ & + \theta_2' \mathbf{p}_{iw} \times TR_i + \theta_3' \mathbf{p}_{iw} \times TR_i \times 2nd_t \\ & + \mu_i + \varepsilon_{iwt}. \end{aligned} \tag{3}$$

The main effect of treatment in the pretreatment period is absorbed by the fixed effects. We also note that there are of course no p-coupon valuations on weeks for which subjects were not presented with a p-coupon to value, which means that the omitted category for unincentivized predictions in column (3) is $p = 0$, but for normalized coupon valuations in column (6) it is $p = 1$.

The results in the two columns are quite similar. We begin with the unincentivized predictions in column (3) because they provide results for weeks with and without p-coupon subsidies, and thus they allow us to investigate whether the differences in treatment effect between these weeks are mirrored in predictions. The coefficient for weeks without p-coupon subsidies ($TR \times 2nd$) is positive and significant, and we cannot reject that it is identical to the estimated treatment effect for these weeks reported in column (2) of Table 1, suggesting that treatment-group subjects largely failed to predict the increased attendance caused by the treatment month. Meanwhile, the coefficients on $TR \times 2nd \times p$ for all positive values of p are negative and, except for $p = 7$, are significant at least at the 10% level. Adding each of these coefficients to the coefficient on the $TR \times 2nd$ dummy gives us the differential revision in predictions between treatment and control for each value of p on weeks where p-coupon subsidies are present. Thus, we do not reject the hypothesis that there is no differential revision in predictions between treatment groups when p-coupon incentives are present, unlike during weeks without the extra p-coupon subsidies. The same pattern can be seen in the coefficients on the triple-interaction terms in the regression of normalized coupon valuations in column (6).

Our attempt to test misprediction of the treatment effect using an incentivized prediction elicitation mechanism is undermined by the fact that the treatment effect is only evidenced when there are no p-coupon incentives. However, we are encouraged by the fact that our two prediction measures move together closely across weeks with and without p-coupons. Taken together, our results for attendance and unincentivized prediction appear to be consistent with the idea that treated subjects failed to foresee the increased attendance caused by the treatment month; although there was no increase in attendance on weeks for which they held a p-coupon, their treatment-effect predictions appear to be correct because there was essentially no treatment effect for them to mispredict. This is consistent with a model of projection bias over habit formation, in the manner of Loewenstein et al. (2003), but there are certainly other possibilities. For example, subjects may have had correct beliefs about habit formation per se but failed to predict that they would attend the gym enough to form a habit during the treatment month. Another possibility is that the treatment effect was caused not by habit formation but by one of the alternative causes mentioned in §3.1, i.e., treated subjects either learning the true utility of gym attendance or generating utility from social contacts at the gym. In either case our results suggest that they did not predict the effect.

4. Discussion

Using a field-experimental intervention to exogenously shift preferences toward gym attendance in a student sample, we find systematic evidence consistent with two simultaneous dimensions of misprediction: projection bias with respect to habit formation, and naiveté with respect to present bias. We develop a novel tool for eliciting predictions that serves both to incentivize subjects' predictions and to shift their actual future behavior, allowing us to explore both their attendance predictions and the value they place on incentives for future attendance, as well as the effect of small incentives. We find that our 8-visit intervention raised posttreatment attendance among treated subjects by approximately 0.2 visits per week in the immediate posttreatment period. This relative increase in attendance did not persist after the hiatus in attendance created by the winter break. The difference in differences in treated and control subjects' first- and second-elicitation predictions reveals a significant relative increase in predicted attendance of 0.4 visits per week in those weeks without p-coupon subsidies, which is not distinguishable from the actual change in attendance. Finally, subjects in general predicted approximately one to two visits more than they ultimately made.

Under the strong assumptions needed to model these results formally as habit formation, projection bias, and partially naive present bias, it is possible to derive and estimate a structural model to recover the parameters associated with the relevant models. We follow this approach in a previous working version of this paper, Acland and Levy (2013). Under

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

the assumption that agents have quasi-linear utility in money, disutility from gym attendance, and the long-run health benefits of gym attendance, we find that the habit reported in this paper corresponds to a monetary value of \$2.60.[18] In contrast, subjects only predicted a habit of \$0.16, corresponding to a degree of projection bias of $\alpha = 0.94$ in the model of Loewenstein et al. (2003). Subjects' general overoptimism regarding gym attendance corresponds to a degree of naiveté regarding present bias of 0.67—that is, for example, if subjects have a short-run discount factor $\beta = 0.7$ in the Laibson (1997) model, they believe all future selves will instead have a short-run discount factor $\hat{\beta} = 0.9$.

We believe that our results can help to explain several features of contracts typical in gym markets. Our results provide one explanation for the prevalence of unlimited prepaid gym contracts rather than pay-per-visit memberships or the "commitment contracts" often suggested as a remedy for present bias. Following DellaVigna and Malmendier (2004), firms will provide contracts for investment goods such as exercise that feature below-marginal cost-per-use prices, either because of sophisticated consumers' demand for commitment or to obtain the rents from naive consumers' "fictitious surplus" caused by overoptimistic beliefs. We provide direct evidence that it is the latter effect that appears to drive consumer behavior, albeit in our highly selected sample.[19]

Moreover, our findings suggest an explanation for the dearth of contracts replicating the habit-formation intervention used here and by Charness and Gneezy (2009). Both papers find that such an intervention can help people to develop positive exercise habits, at least in the short run, and in principle this could be marketed by firms. That our subjects failed to predict the increase in gym attendance caused by the intervention means that there is likely to be little demand for such a contract.[20] The effect is magnified by subjects' naiveté, because they do not value the effect an exercise habit would have in balancing their future selves' self-control problems. Employers or policymakers seeking to provide incentives for increased physical activity are thus constrained in the offerings that will be deemed valuable, and this problem plausibly extends beyond the gym. Contracts that use the overestimation of future gym attendance to raise the cost of immediate misbehavior(e.g., allowing individuals to choose a target level of attendance with rewards for matching it and penalties for missing)are likely to be particularly effective.

We acknowledge other potential interpretations for these two prediction failures. For example, the failure of treated subjects to predict the treatment effect could be explained if subjects entered the experiment with systematically biased beliefs about the desirability of gym attendance, and subsequently only treated subjects learned that it was more pleasant (or less unpleasant) than their prior belief. Alternatively, treated subjects may have formed rewarding social contacts at the gym during the treatment month, making future attendance more desirable for social reasons. However, any alternative must require that treated subjects initially held incorrect beliefs, and will have largely similar welfare implications: individuals will underinvest in beneficial activities for which an initial period of attendance leads to a persistent increase in future attendance.

A model of projection bias over habit state predicts that habituated subjects will also mispredict the decay of the treatment effect. If dehabituation is sufficiently rapid (as our data suggest may be true in our setting),projection bias with respect to habit decay might cause individuals to underinvest in maintaining a habit once it has been formed. It is possible that learning about the true utility of gym attendance, or the value of social contacts at the gym, could decay rapidly over a span of a few weeks, but this seems less plausible than the decay of habit formation. It is possible that further research could help to fully disentangle our projection-bias interpretation from alternatives by eliciting predictions of attendance over a period of habit decay.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mnsc.2014.2091.

Acknowledgments

The authors thank Stefano DellaVigna, Gary Charness, Uri Gneezy, Teck-Hua Ho, Shachar Kariv, Botond Koszegi, Ulrike Malmendier, Matthew Rabin, and seminar participants at the University of California, Berkeley (UC Berkeley) and Harvard University, as well as the editors and anonymous referees 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 P30 AG12839].

[18] That is, the extra consumption utility of gym attendance when in a habituated state is equivalent to a \$2.60 external subsidy.

[19] DellaVigna and Malmendier (2006) show that infrequent gym-users choose contracts with prices that are apparently dominated ex post but point out that plan choice is consistent with either sophisticated or naive consumers and instead establish their naiveté result using procrastination in plan cancellation.

[20] It is telling that the increasingly popular "fitness boot camps," which offer participants several weeks of intense training, tend to advertise their product as providing immediate benefits such as weight loss rather than as creating a lasting exercise habit.

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

Appendix

A.1. Sample

Our initial sample consisted of 120 subjects, randomly assigned to treatment and control groups of 60 subjects each. Table A.1 provides a comparison of the treatment and control groups. Because of attrition and missing data, in our analysis the final number of treated subjects is 54 and of control subjects 57. In addition to basic demographic variables, we included discretionary budget and the time and money cost of getting to campus to control for differences in the cost of gym attendance and the relative value of monetary incentives. The 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

Table A.1 Comparison by Treatment, Compliance, and Habit Formation

	(1) Full sample	(2) Treated group	(3) Control group	(4) *t*-test, *p*-value (2) = (3)	(5) Compliers	(6) Non-compliers	(7) *t*-test, *p*-value (5) = (6)	(8) "Habit formers"	(9) Non-"habit formers"	(10) *t*-test, *p*-value (8) = (9)
Original sample	120	60	60							
No. of quitters	6	4	2							
No. w/incomplete data	3	2	1							
Final sample size	111	54	57		43	11		8	46	
					Demographic covariates					
Age	21.919	22.204	21.649	0.639	22.605	20.636	0.429	19.750	22.630	0.306
	(0.586)	(0.990)	(0.658)		(1.234)	(0.472)		(0.453)	(1.150)	
Gender (1 = female)	0.685	0.648	0.719	0.425	0.651	0.636	0.929	0.625	0.652	0.885
	(0.044)	(0.660)	(0.060)		(0.074)	(0.152)		(0.183)	(0.071)	
Proportion white	0.360	0.333	0.386	0.568	0.349	0.273	0.640	0.250	0.348	0.596
	(0.046)	(0.065)	(0.065)		(0.074)	(0.141)		(0.164)	(0.071)	
Proportion Asian	0.559	0.630	0.491	0.145	0.651	0.545	0.526	0.750	0.609	0.454
	(0.047)	(0.066)	(0.067)		(0.074)	(0.157)		(0.164)	(0.073)	
Proportion other race	0.081	0.037	0.123	0.010	0.000	0.182	0.004	0.000	0.043	0.557
	(0.026)	(0.026)	(0.044)		(0.000)	(0.122)		(0.000)	(0.030)	
					Economic covariates					
Discretionary budget	192.342	208.333	177.193	0.404	222.093	154.545	0.350	181.250	213.043	0.699
	(18.560)	(28.830)	(23.749)		(34.475)	(41.808)		(92.068)	(30.274)	
Travel cost to campus	0.901	0.648	1.140	0.370	0.616	0.773	0.853	0.000	0.761	0.424
	(0.273)	(0.334)	(0.428)		(0.386)	(0.679)		(0.000)	(0.391)	
Travel time to campus (min)	14.662	14.398	14.912	0.811	13.372	18.409	0.237	9.688	15.217	0.252
	(1.071)	(1.703)	(1.335)		(1.790)	(4.564)		(1.666)	(1.958)	
					Naiveté proxy covariates					
Forget[a,b]	1.595	1.556	1.632	0.573	1.465	1.909	0.047	1.500	1.565	0.800
	(0.067)	(0.090)	(0.099)		(0.096)	(0.211)		(0.327)	(0.091)	
Spontaneous[a,b]	2.486	2.574	2.404	0.281	2.442	3.091	0.011	2.250	2.630	0.198
	(0.079)	(0.104)	(0.117)		(0.101)	(0.285)		(0.164)	(0.118)	
Things come up[a,b]	2.586	2.611	2.561	0.731	2.558	2.818	0.333	2.375	2.652	0.363
	(0.072)	(0.107)	(0.097)		(0.101)	(0.352)		(0.263)	(0.117)	
Think ahead[a,b]	2.874	2.944	2.807	0.338	2.977	2.818	0.436	3.000	2.935	0.778
	(0.071)	(0.081)	(0.116)		(0.091)	(0.182)		(0.189)	(0.090)	
Procrastinate[a,b]	3.036	3.056	3.018	0.800	2.977	3.364	0.135	2.875	3.087	0.473
	(0.075)	(0.104)	(0.108)		(0.118)	(0.203)		(0.295)	(0.111)	
					Exercise experience and attitude covariates					
Pretreatment Godin activity scale	36.050	36.500	35.623	0.855	38.360	29.227	0.221	41.688	35.598	0.474
	(2.376)	(2.983)	(3.689)		(3.137)	(7.961)		(3.823)	(3.434)	
Fitness is important[a,b]	3.081	2.981	3.175	0.092	2.977	3.000	0.914	3.500	2.891	0.010
	(0.057)	(0.086)	(0.076)		(0.097)	(0.191)		(0.189)	(0.089)	
Appearance is important[a,b]	3.252	3.259	3.246	0.917	3.256	3.273	0.944	3.375	3.239	0.620
	(0.065)	(0.096)	(0.880)		(0.095)	(0.304)		(0.183)	(0.109)	
F-test of joint significance				0.387			0.635			0.663

Note. Standard errors in parentheses.

[a]1 = Disagree strongly; 2 = disagree somewhat; 3 = agree somewhat; 4 = agree strongly.

[b]Wording of questions in the appendix.

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms

for subjects' taste for the outcomes typically associated with gym attendance. The naiveté proxy covariates are subjects' answers to a series of unincentivized questions as follows.

Variable	Description
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.

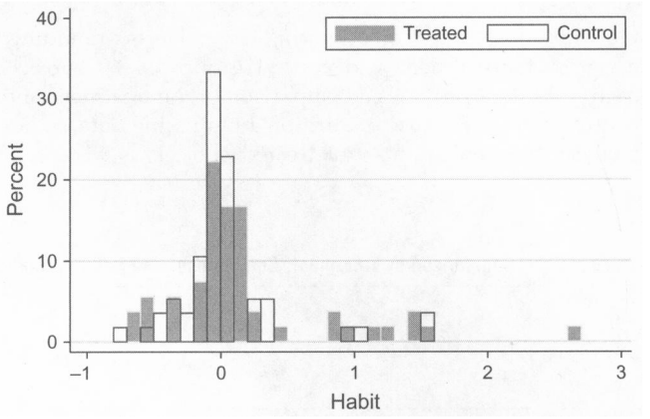
A.2. Compliance, Attrition, and Randomization

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.[21] Despite this titration of rewards, 6 of the 120 subjects did not complete the study. Two control subjects and two treatment subjects left the study between the first and second sessions, and two more treatment subjects left between the second and third. 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 makeup sessions the following day. Four control subjects and two treatment subjects attended these sessions, and we have treated them as having completed the study.

A comparison of compliers and noncompliers with the treatment-month incentive among treated subjects appears in columns (5) and (6) of Table A.1.

A.3. Individual Effects

Figure A.1 (Color online) Individual-Level Habits



Notes. Individual level habit effects, by treatment assignment. Habits are individual changes in attendance during the immediate posttreatment period and are analogous to the group-level habits estimated in Table 1.

References

Acland D, Levy MR (2013) Naiveté, projection bias, and habit formation in gym attendance. Working paper, University of California, Berkeley, Berkeley.

Becker G, Murphy K (1988) A theory of rational addiction. *J. Political Econom.* 96(4):675–700.

Bénabou R, Tirole J (2002) Self-confidence and personal motivation. *Quart. J. Econom.* 117(3):871–915.

Charness G, Gneezy U (2009) Incentives to exercise. *Econometrica* 77(3):909–931.

DellaVigna S, Malmendier U (2004) Contract design and self-control: Theory and evidence. *Quart. J. Econom.* 119(2): 353–402.

DellaVigna S, Malmendier U (2006) Paying not to go to the gym. *Amer. Econom. Rev.* 96(3):694–719.

Dzewaltowski D, Noble J, Shaw J (1990) Physical activity participation: Social cognitive theory versus the theories of reasoned action and planned behavior. *Sport Psych.* 12(4):388–405.

Gneezy U, Meier S, Rey-Biel P (2011) When and why incentives (don't) work to modify behavior. *J. Econom. Perspect.* 25(4): 1–21.

Godin G (1994) Theories of reasoned action and planned behavior: Usefulness for exercise promotion. *Medicine Sci. Sports Exercise* 26(11):1391–1394.

Godin G, Valois P, Lepage L (1993) The pattern of influence of perceived behavioral control upon exercising behavior: An application of Ajzen's theory of planned behavior. *J. Behav. Medicine* 16(1):81–102.

Gul F, Pesendorfer W (2001) Temptation and self-control. *Econometrica* 69(6):1403–1435.

Gul F, Pesendorfer W (2004) Self-control and the theory of consumption. *Econometrica* 72(1):119–158.

Holt CA, Laury SK (2002) Risk aversion and incentive effects. *Amer. Econom. Rev.* 92(5):1644–1655.

Kane R, Johnson P, Town R, Butler M (2004) A structured review of the effect of economic incentives on consumers' preventive behavior. *Amer. J. Preventive Medicine* 27(4):327–352.

Laibson D (1997) Golden eggs and hyperbolic discounting. *Quart. J. Econom.* 112(2):443–477.

Lally P, van Jaarsveld CHM, Potts HWW, Wardle J (2010) How are habits formed: Modelling habit formation in the real world. *Eur. J. Soc. Psych.* 40(6):998–1109.

Loewenstein G, O'Donoghue T, Rabin M (2003) Projection bias in predicting future utility. *Quart. J. Econom.* 118(4): 1209–1248.

O'Donoghue T, Rabin M (1999) Doing it now or later. *Amer. Econom. Rev.* 89(1):103–124.

Reynolds K, Killen J, Bryson S, Maron D, Taylor CB, Maccoby N, Farquhar J (1990) Psychosocial predictors of physical activity in adolescents. *Preventive Medicine* 19(5):541–551.

Valois P, Dersharnais R, Godin G (1988) A comparison of the Fishbein and Ajzen and the Triandis attitudinal models for the prediction of exercise intention and behavior. *J. Behav. Medicine* 11(5):459–472.

[21] 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.

This content downloaded from 128.32.10.230 on Tue, 05 Apr 2022 23:16:51 UTC
All use subject to https://about.jstor.org/terms