# How are Preferences For Commitment Revealed?

Mariana Carrera, Heather Royer, Mark Stehr, Justin Sydnor, and Dmitry Taubinsky*

June, 2019

**Abstract**

A large literature treats take up of commitment contracts, in the form choice-set restrictions or penalties, as a smoking gun for (awareness of) self-control problems. This paper provides new techniques for examining the validity of this assumption, as well as a new approach for detecting (awareness of) self-control problems. Theoretically, we show that with some uncertainty about the future, demand for commitment contracts is closer to a knife-edge case than to a robust implication of models of limited self-control. In a field experiment with 1292 members of a fitness facility, we find that many participants take up commitment contracts both for going to the gym more and for going to the gym less, and there is a significant positive correlation in demand for these two types of contracts. This suggests that commitment contract take up involved "noisy valuation" or preferences other than the desire to change own future behavior. Moreover, we find that commitment contract take up is negatively related to awareness of self-control problems: a novel information treatment that increased awareness of self-control problems reduced demand for commitment contracts. We address the limitations of using commitment contracts as a measurement tool by showing that a combination of belief forecasts and willingness to pay for linear incentives provide more robust identification of limited self-control and people's awareness of it. We use the methodology to obtain some of the first parameter estimates of partially-sophisticated quasi-hyperbolic discounting in the field.

*We are grateful to seminar and conference participants at Harvard, Wharton, UC San Diego, University of Zurich, Dartmouth, and the Stanford-Berkeley mini conference for helpful comments and suggestions. Paul Fisher, Chang Lee, and Priscila de Oliveira provided excellent research assistance. We are grateful for funding through NIH grant R21AG042051 entitled "Commitment Contracts for Health Behavior Change." Taubinsky also thanks the Sloan Foundation for financial support. This study was approved by the IRB at Case Western Reserve University and the University of California-Santa Barbara.

# 1 Introduction

One of the central insights from economic models of time inconsistency and costly self-control (for short, "present focus") is that people should desire incentives and mechanisms that help them alter their own future behavior (Strotz, 1955; Laibson, 1997; O'Donoghue and Rabin, 1999). A large and growing literature has tested this insight by analyzing the demand for "commitment contracts" that allow people to restrict their own future choice set or to impose penalties for certain (mis)behaviors.[1] This literature has documented that in many settings people take up commitment contracts.

This evidence is typically interpreted as "smoking gun" evidence for awareness of present focus: if people voluntarily restrict their choice sets or agree to penalties, what else could they be revealing other than a desire to change their future selves' behavior? Because of the "smoking gun" status, offers of commitment contracts are typically considered well-targeted policy tools: they would be taken up by those with recognized self-control problems, but would not impose restrictions on others.

However, there are relatively few examples of real-world commitment contracts outside of behavioral economics experiments (Laibson, 2015, 2018; Laibson and Ericson, 2019). As a potential explanation, Laibson (2015) uses numerical simulations in the procrastination model of Carroll et al. (2009) to show that modest uncertainty about the future can quickly erode the desire for commitment contracts—ultimately conjecturing that in theory, take up of commitment contracts is a "hothouse flower that survives only under special parameterizations."

Although it is not uncommon to hear behavioral economists discuss the "puzzle" of why there is not more take up of commitment contracts—even in the existing experiments—numerical exercises such as those of Laibson (2015) suggest that perhaps the puzzle is why we see *so much* take up in our experiments.

Methodologically, our paper builds on this insight in several ways. First, we provide formal and general theoretical results that show that it takes little uncertainty to erode demand for commitment. Second, we develop and implement simple experimental methods for unpacking what take up of commitment contracts really reveals—a genuine desire to change one's future behavior, or perhaps calculation errors or alternative considerations like feeling social pressure to just say "yes" (DellaVigna et al., 2012). Third, we develop a new method for measuring present focus and people's awareness of it, and we show that this method is more robust to the potential problems with using commitment contract take up to reveal present focus.

We begin in Section 2 by modeling the demand for commitment contracts by quasi-hyperbolic decision makers. We show that with at least a moderate amount of uncertainty about the future—formalized as a precise condition on the uncertainty about the costs of engaging in the target activity—no one should desire commitment contracts in the standard qu asi-hyperbolic model.[2] In these situations, people who perceive themselves to be more present focused will actually find

[1] See for example, Ashraf et al. (2006), Gine et al. (2010), Houser et al. (2010), Schwartz et al. (2014), Augenblick et al. (2015), Beshears et al. (2015), Kaur et al. (2015), Royer et al. (2015), Exley and Naecker (2016), Acland and Chow (2018), Toussaert (2018), John (forthcoming), Sadoff et al. (forthcoming), andSchilbach (2019).

[2] See Amador et al. (2006), Heidhues and Kőszegi (2009), Beshears et al. (2015), and John (forthcoming) for qualitative results about the tradeoff between commitment and flexibility.

penalty-based commitment contracts more harmful than those who perceive themselves to be less present focused.

However, even in environments in which standard models do not predict take up of commitment, take up may be observed in practice for several reasons. First, recent work in neuroscience and economics shows that individuals' choices are "noisy" because of imperfect mental representations, even in simple decisions between a pair of lotteries (Woodford, 2012; Wei and Stocker, 2015; Khaw et al., 2017; Frydman and Jin, 2019).[3] Second, commitment contract take up may also reflect "alternative considerations" arising from demand effects, distrust, goal setting, or social pressure. In Section 2.3 we therefore extend the standard quasi-hyperbolic model to allow for both noisy valuation and alternative considerations, modeled in reduced form. The extended model generates predictions and insights that we test and confirm in our field experiment.

We use the model's predictions to generate a set of hypotheses, which are tested using a field experiment conducted with 1,292 members of a fitness facility in three waves. As described in Section 3, the experiment involved a series of online decisions and forecasts, followed by four weeks of incentivized gym attendance. The start of the experiment included an information intervention that randomized half of participants into a treatment aimed at debiasing overoptimistic beliefs about gym attendance. This treatment was intended to serve as an exogenous shock to sophistication about present focus. We then elicited participants' willingness to pay (WTP) for piece-rate incentives for attending the gym over the next month, as well as their beliefs about how often they would attend under the different incentives. They were then asked to make a series of choices about simple commitment contracts. Finally, participants were randomly selected to receive different incentives for attending the gym over the following four weeks, which we track using administrative gym records.

We begin our analysis of the field experiment results in Section 4 by examining the take up of commitment contracts. We elicited demand for commitment contracts tied to attending the gym *at least* 8, 12, or 16 times over the next month. For each of these thresholds, participants chose between an unconditionalpayment of \$80 regardless of how often they used the gym or receiving \$80 conditional on attending at least as many times as that threshold. The conditional options are pure commitment contracts since they feature no financial upside. We find that substantial shares of participants chose these commitment contracts (64% for 8+ visits, 49% for 12+ visits, and 32% for 16+ visits), and that these contracts increased attendance by over 3 visits during the treatment month. The literature would typically interpret this as evidence of a substantial desire to address present-focus.

However, because we were sensitive to the role that noise and alternative considerations could play in experimental choices, we also gave participants a set of novel commitment options to go to the gym *less*. Although these commitment choices were not our central focus when we designed the first wave of the experiment—as we did not have strong priors that there would be significant noise—the patterns of choices provide striking new evidence about the role of noise and alternative

[3]See also earlier work on quantal response equilibrium (McKelvey and Palfrey, 1995b).

considerations in commitment contract demand.

We find that 28-34% of participants chose commitment contracts for going to the gym *fewer* than 8, 12, or 16 times over the next month. Taken at face value, these choices could imply that a substantial fraction of participants are either "future focused" or perceive going to the gym to be a "temptation good." However, we find that the take up of "more" and "fewer" contracts at each threshold is significantly positively correlated. This is inconsistent with using commitment contracts as a self-control strategy, but is consistent with noise and alternative considerations, as modeled in Section 2.3. We rule out several alternative explanations for our results, such as participants simply confusing the "fewer visits" contracts for the "more visits" contracts.

Because these empirical results suggest that inferring present focus from commitment take up is problematic, in Section 2.4 we introduce a new method for estimating present-focus models that we show is more robust to noise and alternative considerations. Building on a design feature first introduced by Acland and Levy (2015), as well as theoretical insights introduced in DellaVigna and Malmendier (2004), the key idea is as follows: If a person believes she's time consistent, then the only upside of a piece rate incentive for doing more of an activity is its earnings; the behavior change it induces can only be a negative. Consequently, a person who believes herself to be time-consistent should not value increases in piece-rate incentives above the subjective expected earnings they generate. A person who does value a piece-rate incentive increase above the subjective expected increase in earnings reveals a WTP for behavior change. And because mean-zero errors in valuations translate into mean-zero errors in WTP, finding an average positive WTP for behavior change provides more robust reduced-form evidence of some awareness of present-focus. Moreover, we introduce a sufficient statistics approach for translating such reduced-form measures into parameter estimates of both the perceived and actual short-run discount factor in the partially sophisticated quasi-hyperbolic discounting model.

In Section 5 we utilize this method for assessing participants' desire to change their future selves' behavior. We find a significant average willingness to pay to change behavior in our sample. On average we find that participants value an additional \$1 in the per-day incentive level by between \$0.55 and \$1.40 more than they expect to earn from that additional \$1. Given that on average, a \$1 incentive increases expected attendance by 0.67 visits, this implies that participants were willing to pay \$0.83 to \$2.10 per additional gym visit induced by incentives.

In Section 5.3 we show that the these reduced-form statistics imply a perceived short-run discount factor ($\tilde{\beta}$) in the range of 0.74 to 0.93. However, we find that participants are on average not fully sophisticated, as they overestimate their future attendance. We use these gaps between beliefs and reality to identify the ratio of actual to perceived present focus. Taken together, the evidence strongly implies partial but not full sophistication.

In Section 6 we present the results of the information treatments, which were designed to investigate the malleability of sophistication about present-focus, as well as how exogenous changes in this sophistication affect commitment contract demand. In Wave 1, the treatment simply showed participants information about their past attendance at this gym, but this had no effect on their beliefs or

choices. In Waves 2 and 3 we introduced an enhanced treatment that significantly reduced overestimation of future gym attendance. Unlike the first treatment, this treatment motivated participants to internalize information on past attendance and also informed them about the overestimation by prior participants.

We find that the enhanced (but not basic) information treatment significantly increased our measure of WTP for behavior change. This implies that at least part of the effect of the treatment on beliefs can be attributed to reducing naivete about present focus. We show that the reduced-form impact on WTP for behavior change translates to sizable, though noisy, differences in the estimated level of the perceived level of present focus.

In stark contrast, the enhanced information treatment *reduced* the take up of commitment contracts for more gym attendance by an average of 7 percentage points ($p$-value = 0.02). Although some studies have explored *correlations* between commitment contract demand and proxies for perceived present focus (e.g., Ashraf et al., 2006; Augenblick et al., 2015; Kaur et al., 2015; John, forthcoming), our study is the first to report a causal estimate. The prior evidence on correlations is mixed with some studies finding positive correlations between measured impatience and commitment demand (Augenblick et al., 2015; Kaur et al., 2015) but others finding a negative correlation (Sadoff et al., forthcoming; John, forthcoming). Our result is consistent with theoretical predictions in Section 2.3: because in the presence of uncertainty commitment contracts appear most harmful to individuals with the lower $\tilde{\beta}$, those with lower $\tilde{\beta}$ are less likely to "mistakenly" choose them due to noisy valuation. This suggests that settings with high take up of commitment contracts could be settings in which individuals are particularly naive and noisy, rather than particularly sophisticated.[4]

Our contributions to the literature are threefold. First, while the tradeoff between commitment and flexibility has been acknowledged, existing theoretical results in the literature are of a qualitative nature. Laibson (2015) is the only exception, reporting numerical results from a uniform distribution of task completion costs. We provide general calibration theorems for arbitrary probability distributions about task completion costs, and for a range of economic environments including static, dynamic, and continuous choice.

Second, we provide direct evidence that noisy valuation and alternative considerations could be an important component of the demand for commitment contracts.[5] Prior studies of commitment contract demand have noted that noisy decisions could affect take up.[6] These studies typically support the interpretation of commitment contract demand as a true signal of awareness of limited self-control by appealing to consistency in commitment demand over time or correlations between demand and measured time preferences. We propose a simpler and more direct test: additionally offer participants commitment contracts to do the opposite of the goal activity.

[4] As we discuss in Section 2.2, particular calibrations of future uncertainty could lead to a non-monotonic relationship between $\tilde{\beta}$ and commitment contract demand (as in Heidhues and Kőszegi, 2009; John, forthcoming), which is not inconsistent with the empirical result.

[5] This result is related to Exley and Naecker (2016) who similarly show that considerations other than awareness of present focus can affect take up of commitment contracts, though in a different context with social signaling.

[6] See, for example, Kaur et al. (2015) and Schilbach (2019).

Noisy valuation and alternative considerations may help explain why some prior studies have found that those who show the least indication of present focus are sometimes more likely to take up commitment contracts (Royer et al., 2015; Sadoff et al., forthcoming), or why some studies find that over 90% of participants choose commitment contract thresholds that that they would exceed anyway (Kaur et al., 2015). People with small self-control problems and high motivation should not benefit from commitment contracts, but they also are not likely to be harmed by them, so even small amounts of noise or alternative considerations could lead these people to take up commitment contracts.

Third, we provide some of the first field estimates of perceived present focus, and the extent of partial sophistication. Augenblick and Rabin (2019) estimate these parameters using a laboratory task. Paserman (2008) and Laibson et al. (2018) estimate present focus using observational data, after *assuming* either naivete or sophistication. Bai et al. (2019) are closest to our work in that they estimate both parameters. Different from our methods, their methodology relies on decisions to take up commitment contracts. Our methods make use of what we think are the most transparent and straightforward moments generated by models of partially sophisticated present focus: how do people's expectations line up with reality, and how much are they willing to pay to change their future selves' behavior.

We conclude the paper in Section 7 with a discussion of implications and best-practice guidelines for future work, as well as some implications for policy design with present focus. We emphasize that the results of our experiment should not be taken to mean that prior work on commitment contracts should be dismissed as confounded. Rather, our results raise the *possibility* of confounds, and our methodology provides some simple experimental techniques researchers can use to detect these issues. We hope that this paper motivates a series of tests that will lead toward a more complete understanding of what demand for commitment contracts reveals.

# 2 Theory

In this section we begin by setting up and characterizing the predictions of the standard model of quasi-hyperbolic discounting (subsections 2.1 and 2.2). We then augment the model to incorporate noisy valuation and alternative considerations (subsection 2.3), and we end by deriving a more robust test of awareness of present focus (subsection 2.4).

## 2.1 Quasi-hyperbolic discounting and commitment demand

To keep exposition as intuitive and concise as possible, we focus here on a simple 3-period model of quasi-hyperbolic discounters who face a binary choice-set. As we discuss in Section 2.5 and show in Appendix B, the results generalize to more dynamic environments, to continuous choice, and to alternative models of limited self-control. All proofs of the results in the body of the paper are contained in Appendix A.

In period 0 agents choose between contracts $(r, p_0, p_1)$ that consist of a fixed (and possibly

negative) reward $r$ and contingent rewards $p_0 \geq 0$ and $p_1 \geq 0$ that they receive in period 2 for taking the actions $a = 0$ and $a = 1$, respectively. In period 1 agents choose the action $a \in \{0, 1\}$. The direct cost of taking the action, $c$, is realized in period 1, and is distributed according to a cumulative density function $F$. The deterministic benefits of taking the action, which are realized in period 2, are $b$. We only require that $c > 0$ with positive probability; we do not preclude the possibility that on some "good days" agents actually find immediate pleasure in activities with delayed benefits such as exercise.

We assume that agents have quasi-hyperbolic preferences given by $U^t(u_t, u_{t+1}, \ldots, u_T) = \delta^t u_t + \beta \sum_{\tau=t+1}^{T} \delta^\tau u_\tau$ , where $u_t$ is the period $t$ utility flow. By assumption, $u_1 = -a \cdot c$, and $u_2 = r + a \cdot (b + p_1) + (1 - a) \cdot p_0$. Following O'Donoghue and Rabin (2001), we allow agents to mispredict their preferences: in period 0, they believe that their period 1 self will have a short-run discount factor $\tilde{\beta} \in [\beta, 1]$. For simplicity, we set $\delta = 1$.

In period 1, agents choose $a = 1$ if $c < \beta(p_1 + b - p_0)$. This decision rule says that for the person to act, the current costs of action have to be less than the discounted future benefits plus contingent rewards from action. In period 0, agents' perceived expected utility given contract $\mathcal{C} = (r, p_0, p_1)$ is

$$V(\mathcal{C}) := \beta \left[ r + \int_{c > \tilde{\beta}(p_1 + b - p_0)} p_0 dF + \int_{c \leq \tilde{\beta}(p_1 + b - p_0)} (p_1 + b - c) dF \right]$$

They choose a contract $\mathcal{C}_A = (r_A, p_{1A}, p_{2A})$ over $\mathcal{C}_B = (r_B, p_{1B}, p_{2B})$ if and only if $V(C_A) > V(C_B)$.

We call a contract $(-p, 0, p)$ a commitment contract for $a = 1$ with penalty $p$, which we denote by $CC(p, 1)$. This contract is perceived as a dominated contract by an agent who believes himself to be time-consistent. We call a contract $(-p, p, 0)$ a commitment contract for $a = 0$ with penalty $p$, which we denote by $CC(p, 0)$. Again, this constitutes a dominated contract for an agent who perceives himself to be time-consistent.

In the exercise context, the action $a = 1$ corresponds to attending the gym, which yields delayed health benefits $b$, but generates immediate costs $c$. The contract $(-p, 0, p)$ specifies that the person pays a fee $p$ if he does not attend the gym. The contract $(-p, p, 0)$ specifies that the person pays a fine $p$ if he *does* attend the gym.

### 2.2 When will commitment contracts be chosen?

Commitment contracts for $a = 1$ will be desired when $\tilde{\beta} < 1$ and there is little need for flexibility. For example, suppose that $Pr(c \leq b) = 1$, so that it is always worthwhile to choose $a = 1$, but that $Pr(c \leq \tilde{\beta} b) < 1$, so that an agent does not believe that he will always choose $a = 1$ in period 1. Then a commitment contract for $a = 1$ with penalty $p \geq \frac{(1-\tilde{\beta})b}{\tilde{\beta}}$ ensures that $a = 1$ is always chosen. This contract is clearly desirable, as it implements desired behavior at no cost to the agent.

More generally, as long as $Pr(c < b)$ is not too high, $\tilde{\beta} < 1$ individuals will be eager to choose fully binding contracts that enforce $a = 1$ with probability 1, and the perceived value of these contracts will be decreasing in $\tilde{\beta}$, for $\tilde{\beta}$ above a certain threshold. Below, we summarize some special cases that formalize this intuition. For a commitment contract $(-p, 0, p)$, we set $\Delta V = V(-p, 0, p) - V(0, 0, 0)$

to denote the perceived value of the contract.

**Proposition 1.** *Fix $p$ and suppose that $Pr(c > b)$ is positive for all realizations of $c$. Then there exist $\underline{\beta} > 0$ and $\bar{\beta} < 1$ such that $\Delta V(-p, 0, p) \leq 0$ (i.e., commitment contract with penalty $p$ for action a=1 is undesirable) if $\tilde{\beta} < \underline{\beta}$ or if $\tilde{\beta} > \bar{\beta}$. If $Pr(c > \tilde{\beta}(b + p)) = 0$ then $\Delta V > 0$. When the distribution of $c$ is Bernoulli, there exist thresholds such that $\Delta V > 0$ if and only if $\beta \in (\underline{\beta}, \bar{\beta})$, with $\bar{\beta} > \underline{\beta}$ if $Pr(c > b)$ is sufficiently small.*

Proposition 1 captures the intuition of non-monotonic demand for commitment, analogous to the results of Heidhues and Kőszegi (2009) and John (forthcoming). Those with $\tilde{\beta} = 1$, due to either naivete or actual time-consistency, do not want commitment contracts. Those with very low $\tilde{\beta}$ do not want commitment contracts because they perceive the contracts to be largely ineffective. But those with intermediate levels of $\tilde{\beta}$ do want the contracts. The case in which the distribution of $c$ is binary, $c \in \{\underline{c}, \bar{c}\}$, is analogous to John (forthcoming), who derives this non-monotonicity in the context of consumption and savings. In line with Heidhues and Kőszegi (2009) and John (forthcoming), the results also generalize to the question of whether there exists any commitment contract of size $p \in [0, \bar{p}]$ that is worthwhile. In the proof of the proposition in the appendix, we establish an analogous result for $\max_{p \in [0,\bar{p}]} \Delta V$.[7]

These results about (non-monotonic) demand for commitment depend on strong assumptions about the nature of uncertainty about future costs draws. When the distribution of $c$ is not binary, and when there is even a modest chance that $c > b$, we will show that individuals should not demand commitment at all. In all the results that follow, we assume that $F$ has a density function $f$, and that $f(c) > 0$ if and only if $c \in [\underline{c}, \bar{c}]$ for some constants $\underline{c} < \bar{c}$, with $\bar{c} > 0$. We first ask whether for a fixed penalty $p$ there exists any $\tilde{\beta}$ such that agents will want the contract; we then analyze the question of whether for a given $\tilde{\beta}$ there exists any commitment contract with penalty $p$ that will be desirable.

**Proposition 2.** *Fix $p$ and assume that $f(c_2)/f(c_1) \geq (c_1/c_2)^2$ for all $c_2 > c_1$ in the interval $[\underline{\beta}b, \bar{\beta}(b + p)]$. Then $\Delta V(-p, 0, p)$ is strictly increasing in $\tilde{\beta} \in [\underline{\beta}, \bar{\beta}]$. In particular, if $\underline{\beta} = 0$ and $\bar{\beta} = 1$, then the expected losses are strictly decreasing in $\tilde{\beta}$ for all $\tilde{\beta}$, and thus no individual will want the contract.*

The economic content of the assumption in Proposition 2 is that in the region of cost draws where individuals' decisions can actually be affected by a financial incentive of size $p$, the amount of uncertainty is not "too small"; in particular, that the chances of cost draw $c > b$ do not rapidly vanish to zero. The assumption is satisfied by, for example, a uniform distribution on $[0, \bar{c}]$, where $\bar{c} \geq b + p$. For example, suppose that $c \sim U[0, 1.5b]$, so that time consistent agents do not want to take the action 33% of the time. In this case, there does not exist any $\tilde{\beta}$ for which the the contract looks desirable as long as $p < b/2$.

[7] Heidhues and Koszegi (2009) assume a deterministic cost of commitment. The binary cost shock case analogous to this. The assumption that there is a single shock that makes the commitment contract ineffective is equivalent to having a deterministic cost of $Pr(c = \bar{c})p$.

Proposition 2 fixes the penalty $p$ and provides a result about undesirability of commitment contracts for all $\tilde{\beta}$ (in some interval). We now provide a proposition analogous to that of Proposition 2, but which gives sufficient conditions such that there is no desirable commitment contract for a fixed $\tilde{\beta}$. This includes commitment contracts that simply restrict choice to $a = 1$ with infinite penalties $p = \infty$ for choosing $a = 0$.

**Proposition 3.** *Fix $\tilde{\beta}$ and assume that (i) $f$ is unimodal;*[8] *(ii) $\bar{c} > b + (1 - \tilde{\beta})b$; (iii) $f(c_2)/f(c_1) \geq (c_1/c_2)^2$ for all $c_2 > c_1$ in the interval $[\tilde{\beta}b, \bar{c})$; (iv) $1 - F(b) \geq F(b) - F(\tilde{\beta}b)$ if $f$ does not have a mode in $[\tilde{\beta}b, b + (1 - \tilde{\beta})b]$, and otherwise $1 - F(b) \geq [F(b) - F(\tilde{\beta}b)]/\tilde{\beta}$. Under these four assumptions, there exists no value of $p$, including $p = \infty$, such that a penalty of size $p$ for choosing $a = 0$ is desirable.*

All four of the assumptions of Proposition 3 are satisfied by a uniform distribution with support $[0, \bar{c}]$, where $\bar{c} \geq b + (1 - \tilde{\beta})b$. For example, with $\tilde{\beta} = 0.7$, the assumptions are satisfied by a uniform distribution with support $[0, 1.3b]$; that is, a distribution that generates enough uncertainty that a time consistent individual will not want to take the action 23% of the time.

The economic content of the assumptions of Proposition 3 is again that there is at least some meaningful uncertainty about the desirability of choosing $a = 1$. While assumption (i) is a technical regularity condition, assumptions (ii)-(iv) all provide some bounds on uncertainty. Assumption (ii) ensures that the support is sufficiently wide, assumption (iii) ensures that the density of cost draws does not decrease too quickly, and assumption (iv) ensures that $Pr(c > b)$ is not too low. For $\bar{c}$ high enough, assumption (iv) is implied by assumption (iii).

Figure 1 provides a graphical illustration of commitment contract demand for the case in which $c$ is uniformly distributed on $[0, 1]$. For the uniform case, the bounds of the proposition are sharp: individuals want commitment contracts if and only if $b + (1 - \tilde{\beta})b \leq 1$. Interestingly, for the uniform case, if individuals want a commitment contract at all then they prefer one that is binding. The sharpness of the bounds, and the "all or nothing" nature of demand is specific to the uniform distribution.

**Summary:** In the presence of at least moderate uncertainty, quasi-hyperbolic discounters, however naive or sophisticated, should not choose any kind of commitment contract.

### 2.3 Augmented model with noisy valuation and alternative considerations

In light of the results of the preceding subsection, we reconsider the question of why individuals choose commitment contracts. One possibility is that there is very little uncertainty about the future. Another possibility is that individuals do not behave according to the standard model we have analyzed above. It may be that because evaluating incentives for future behavior is complicated, individuals do so imperfectly, and consequently some individuals overvalue commitment contracts

[8] Formally, there do not exist $c_1 < c_2 < c_3$ such that $f(c_2) < \min(f(c_1), f(c_3))$.

(while others undervalue them). It could also be that some individuals simply enjoy choosing contracts that have clear connotations of setting a "goal," that they feel pressured to enter a contract, or that they assume that anything an authority figure offers them must be worthwhile. To understand if there is empirical content to these alternative explanations for commitment contract demand, in this section we incorporate them into a simple reduced-form model and derive testable predictions.

We allow for two types of valuation biases or errors in agents' decision making. The first source of bias is due to cognitive limitations, in the spirit of Quantal Response Equilibrium (McKelvey and Palfrey, 1995a) or imperfect perception (Khaw et al., 2017). The second source of bias arises from demand effects, desirability bias, reactance bias (doing the opposite of what one thinks is expected of him/her), or the desire of saying "yes" (DellaVigna et al., 2012). We model the first source as i.i.d. errors that are re-drawn each time an individual encounters a new choice-set. We model the second source of bias as a fixed effect.

Formally, we suppose that for a given decision $j$, individual $i$ behaves as if her expected utility under contract $(r, p_1, p_2)$ is $\widehat{V}(r, p_0, p_1) = [\beta r + V(0, p_0, p_1)\varepsilon_{ij}] + \eta_i \mathbf{1}_{(p_0,p_1)\neq 0}$, where $\mathbf{1}_{(p_0,p_1)\neq 0}$ is an indicator that at least some contingent incentives are involved. We assume that for a fraction $\mu$ of individuals $\varepsilon_{ij} \sim G$ is i.i.d. with $G$ supported on $[0, \infty)$ and $E[\varepsilon] = 1$, while for a fraction $1 - \mu$ of individuals $\varepsilon_{ij} \equiv 1$. That is, we allow for heterogeneity in noisy perception by allowing some individuals to be noiseless. More generally, our results hold as long as there is between-person heterogeneity in the variance of the error term $\varepsilon_{ij}$, and we make this assumption only for notational simplicity.

We model the individual effects $\eta_i$ as additive to reflect the common intuition that social motives such as social desirability bias have a relatively smaller effect at larger stakes . We model noise terms $\varepsilon_{ij}$ as multiplicative to reflect that, setting aside social motives, individuals are likely to have fairly accurate valuations of contracts $(r, p_1, p_2)$ in which $p_1, p_2$ are negligible. But our results hold for any type of mean zero errors around $V$, including errors that are more substantial at smaller stakes.[9] For simplicity, we assume that $\eta_i$ and $\varepsilon$ are unrelated to $\beta_i$ and $\tilde{\beta}_i$.

The reduced-form model is not consistent with all types of errors. In a given choice-set, the agent always prefers contract $(r, p_0, p_1)$ over $(r', p_0', p_1')$ if $r \geq r'$, $p_0 \geq p_0'$ and $p_1 \geq p_1'$ and one of the inequalities is strict.[10] This is a consequence of our assumption that the $\epsilon_{ij}$ does not depend on the contract in a given choice scenario: for choice occasion $j$, the error term is common the different options in choice occasion $j$. As we discuss in Section 3.1.5, this property of the model is in line with our data: almost no individuals choose "obviously dominated" incentive structures like "\$0 for sure" instead of "\$20 for sure."

To characterize the new implications of the model, we begin with the following observation:

*Remark* 1. In the standard model without noise or social motives, no individuals choose commitment contracts for $a = 0$.

However, choice of commitment contracts for $a = 0$ is consistent with our augmented model.

[9] Formally, we just need $E\left[\widehat{V}(r, p_0, p_1)\right] = V(r, p_0, p_1) + \eta_i \mathbf{1}_{(p_0,p_1)\neq 0}$.

[10] For $\mathbf{1}_{(p_0,p_1)\neq 0}$ there is also the technical condition that $\mathbf{1}_{(p_0',p_1')\neq 0}$.

And so is simultaneous choice of commitment contracts for $a = 1$ by the same person, even when the conditions of Proposition 3 are met.

**Proposition 4.** *1. Assume that $\mu > 0$ or $Pr(\eta_i > \beta_i p) > 0$. Then for any $\tilde{\beta}$ including 1, a positive mass of individuals with $\tilde{\beta}_i = \tilde{\beta}$ will choose a $CC(p, 1)$.*

*2. Assume that (i) $\mu > 0$ and (ii) either there are some $\tilde{\beta}_i$ close enough to 1 or $Pr(\eta_i > \beta_i p) > 0$. Then a positive mass of individuals will choose $CC(p, 0)$. In this case, a positive mass of individuals will choose both $CC(p, 1)$ and $CC(p, 0)$.*

Moreover, if the choice of commitment contracts for $a = 1$ is primarily driven by noise rather than a real demand for commitment, then there will be a positive correlation between demand for $CC(p, 1)$ and $CC(p, 0)$.

**Proposition 5.** *Assume that (i) $\mu > 0$ and (ii) either $\mu < 1$ or that the $\eta_i$ are heterogeneous. Then there will be a positive correlation between demand for $CC(p, 1)$ and $CC(p, 0)$ if $E[\tilde{\beta}_i]$ is sufficiently close to 1.*

Intuitively, there are two types of mechanisms that lead to the correlation in Proposition 5. First, if some individuals just like to say "yes" ($\eta_i > 0$) and some do not, then the individuals who like to say "yes" will tend to take up both types of contraction, while the other individuals will tend to not take up any kind of contract. Second, if commitment contracts would generally look unappealing to individuals in the absence of noisy valuation, then only the fraction $\mu \in (0, 1)$ of individuals with noisy valuation will be the ones who choose commitment contracts. But because these individuals choose both types of contracts with positive probability, this induces a positive correlation between choice of contracts. The requirement that $E[\tilde{\beta}_i]$ is sufficiently close to 1 is a requirement that heterogeneity in $\tilde{\beta}_i$ does not overwhelm heterogeneity in noisiness or alternative considerations, since an increase in $\tilde{\beta}_i$ could have opposite effects on the likelihood of choosing the two types of contracts.

And in contrast to common intuitions, and the non-monotonicity results for Bernoulli distributions of cost shocks, the noisy valuation model implies that with at least moderate uncertainty, the likelihood of choosing a penalty-based commitment contract for $a = 1$ will be monotonically increasing in $\tilde{\beta}$. The source of this result is Proposition 2, which shows that under moderate to large uncertainty, the perceived harms of a commitment contract are decreasing in $\tilde{\beta}$. Although in a standard noiseless model these conditions would lead individuals to never choose a commitment contract, in a model with noisy valuation individuals still choose the contract, but with a propensity that is decreasing in the harms that would be perceived by a noiseless individual with the same $\tilde{\beta}$. While the comparative statics about commitment contract choice and $\tilde{\beta}$ are ambiguous in models with noiseless valuation and minimal (or binary) uncertainty,[11] the comparative static is much sharper in a model with noisy valuation and moderate to large uncertainty as shown as below:

**Proposition 6.** *Suppose that $f(c_2)/f(c_1) \geq (c_1/c_2)^2$ for all $c_2 > c_1$ in the interval $[0, b + p]$. Then the likelihood of choosing $CC(p, 1)$ is increasing in $\tilde{\beta}$.*

[11] As shown by our Proposition 1 and by Heidhues and Kőszegi (2009) and John (forthcoming)

Interestingly, the converse of Proposition 6 does not hold for commitment contracts for $a = 0$. Intuitively, this is because a lower $\tilde{\beta}$ dampens the impact of financial incentives in both cases, and thus makes penalty-based contracts potentially more harmful in both cases.

Finally, it is also helpful to note that even if individuals are observed to be more likely to choose $CC(p, 1)$ than $CC(p, 0)$, that does not imply that there must be some individuals with $\tilde{\beta}_i < 1$. Such an implication arises only if individuals think they are *unlikely* to choose $a = 1$, so that choosing $CC(p, 1)$ involves a higher financial loss than choosing $CC(p, 0)$. If $CC(p, 1)$ is more attractive to $\tilde{\beta}_i = 1$ individuals than $CC(p, 0)$, in the absence of noise, then $CC(p, 1)$ will be chosen more often.

This shows that under noisy valuation and "alternative considerations" for contract choice, commitment contract demand can be a problematic indicator of demand for behavior change. However, we show in the next subsection that more robust tests are available.

**Summary:** Noisy valuations and alternative considerations lead individuals to choose commitment contracts even in the presence of significant uncertainty, and even when they *discourage* actions with delayed benefits and immediate costs. These tendencies can generate a positive correlation in the take up of both types of commitment contracts. And in the presence of at least moderate uncertainty, these considerations predict that take up of contracts that encourage actions with delayed benefits and immediate costs will be decreasing in sophistication.

### 2.4 Robust tests of demand for behavior change

Although our augmented model implies that commitment contract demand could be a poor indicator of actual awareness of self-control problems, we now show that identification of the quasi-hyperbolic model is still possible using alternative techniques.

For a contract $(0, 0, p)$, which provides incentives $p$ for choosing $a = 1$, define an individual's willingness to pay for the contract, $w(p)$, to be the smallest $r$ such that the individual prefers $(r, 0, 0)$ over $(0, 0, p)$. Define $\alpha_i(p)$ to be be individual $i$'s expected probability of taking action $a = 1$ under contract $(0, 0, p)$.

**Proposition 7.** *If $p > 0$ or $E[\eta_i] = 0$, then*

$$E[w'(p)] = E[\alpha_i(p)] + E[(b_i + p)(1 - \tilde{\beta}_i)\alpha_i'(p)].$$

*If the terms $\left\{(\Delta p)^n \frac{d^m}{dp^m}\alpha(0, 0, p)_{p=p_1}\right\}_{\{n\geq 1, m\geq 2\}}$ are negligible, then*

$$E[w(p + \Delta p) - w(p)] \approx (\Delta p)\frac{E[\alpha_i(p + \Delta p)] + E[\alpha_i(p)]}{2} + E\left[(b_i + p)(1 - \tilde{\beta}_i)(\alpha_i(p + \Delta p) - \alpha_i(p))\right] \quad (1)$$

To obtain intuition for the proposition, consider first the case in which $\tilde{\beta}_i = 1$ for all $i$. Then, Proposition 7 states that on average, time-consistent (and fully naive) individuals should value a marginal \$1 increase in stakes by approximately the average of their expectations of choosing $a = 1$

with and without incentives. The value of a \$1 increase in the contingent incentive $p$ should be approximately valued at the perceived likelihood of receiving it. E.g., if this individual perceives a 50% chance of choosing $a = 1$, then this marginal \$1 increase should be worth approximately \$0.50 to this individual.

Valuations that are in excess of the $\tilde{\beta}_i \equiv 1$ benchmark are due to a demand for behavior change that results from $\tilde{\beta}_i < 1$. The degree of the demand for behavior change is a function of two components. The first component is $(b_i + p)(1 - \tilde{\beta}_i)$, a measure of the degree to which the future self under-appreciates the benefits of choosing $a = 1$. This depends on how large the delayed benefits $b_i + p$ are and on how time-inconsistent the individual perceives himself to be. The second force is $\alpha_i(p + \Delta p) - \alpha_i(p)$, the perceived change in behavior induced by the variation in the incentive.

Note that this formula allows individuals' valuations to be noisy, as long as the noise in valuations of *increases* (not necessarily levels) in incentives is mean-zero. Although this mean-zero assumption is not innocuous, it does illustrate that the alternative approach here is a more robust measure of demand for behavior change than take up of commitment contracts. As we have already shown, mean-zero noise in valuations translates to non-mean-zero noise in binary decisions, such as commitment contract choices. Intuitively, this is because if everyone's true valuation for a commitment contract is negative, some may still choose it due to noise. However, direct elicitations of individuals' valuations using a money metric implies that the money-metric estimates will have mean-zero noise, and thus the average of the elicitations is an unbiased estimate of the true valuation.

By this same logic, the formula continues to hold if in place of individuals' true beliefs $\alpha_i(p)$ we use elicited beliefs $\hat{\alpha}_i(p)$, which may be noisy or even systematically upward or downward biased biased, as long $\hat{\alpha}_i(p + \Delta p) - \hat{\alpha}_i(p)$ is an unbiased estimate of $\alpha(p + \Delta p) - \alpha(p)$ for $p > 0$.

It is also important to note that core to this result is that WTP is completely unrestricted, ranging from positive to negative. Restricting WTP to be non-negative, as in Milkman et al. (2014), would naturally lead to an upward bias in valuations, since negative draws of errors in valuation would be censored at 0. Similarly, presenting experimental participants with a continuous commitment contract range of many possible penalties or targets, as in, e.g., Kaur et al. (2015), would lead to bias if the range only allows participants to commit to doing more of something, but not less of something.

The assumptions leading to (1) are essentially the same as the assumptions that are utilized in the canonical Harberger (1964) formula of the deadweight loss of taxation: that the change in incentives is not too large and that curvature of the behavior response is negligible in the region of incentive change. The economics of the result can be easily intuited through the following public finance analogy: the period-0 self's desired behavior is the "social marginal benefit curve" from the supplied activity, the period-1 self's behavior is the "actual market supply," and the piece-rate incentives are subsidies that bring the two closer together. The WTP for behavior change corresponds to the reduction in deadweight loss generated by a corrective subsidy.

**Summary:** In the presence of noisy valuation and alternative considerations, a more robust test of partial awareness of time inconsistency is testing whether the willingness to pay for a marginal increase in piece-rate incentives is greater than the expected earnings from that marginal increase.

### 2.5 Robustness and extensions of the theoretical results

**Dynamics** For the sake of concreteness and ease of exposition, we have developed our theoretical results under the assumption that there is only one period in which an action can be taken. In our empirical application, however, we consider a dynamic setting in which individuals choose a contract in period $t = 0$, and take actions $a_t \in \{0, 1\}$ (go to the gym or not) in periods $t = 1, \ldots T$ . We show that the main conclusions from the static setting carry over to the dynamic setting in Appendix B.1. The commitment contracts we consider involve a penalty $p$ that must be paid whenever the total number of times that $a_t = 1$ is chosen is below some threshold $X$.

In the dynamic setting, the key condition for commitment contracts to be unattractive is that the density of cost shocks in period $t$, conditional on any period $t$ history of actions, does not diminish too quickly toward zero, in the sense of Proposition 2. Under this condition, backwards induction using repeated application of Proposition 2 establishes a result analogous to Proposition 2. Although one possible intuition, in the spirit of the Central Limit Theorem, is that uncertainty becomes less of an issue when there are more opportunities to act, this is counteracted by the fact that future selves' misbehavior is also more of an issue in dynamic settings in which payoffs are not separable in actions—and this non-separability is generated by commitment contracts to meet a certain threshold.

Having established analogous results about the (un)desirability of commitment contracts in the dynamic setting, we then show that the results of Sections 2.3 and 2.4 carry through essentially verbatim. Consequently, the testable predictions that we summarize in 2.6 are valid for the dynamic setting we analyze in our field experiment.

**Continuous choice** We continue to explore the robustness of our Section 2.2 results about the undesirability of commitment contracts in Appendix B.2. Another natural question is whether the spirit of our results carries over to continuous choice, such as costly effort provision to generate future earnings, or such as saving for the future. In Appendix B.2 we verify that the spirit of our results carries over to these contexts as well. For "continuous penalty" contracts that involve a penalty of $p(X-x)$ for all choices of $x$ (effort, savings) below some threshold $X$ (as in, e.g., penalties on early withdrawl from a savings account), we derive the following striking result both for models of effort provision and savings for the future: If there is a positive probability of states of the world in which the period 0 self would desire a choice of $x < X$ under the commitment contract, then the contract is unappealing for any $\tilde{\beta} \in [0, 1]$, and its perceived damages are decreasing in $\tilde{\beta}$.

For "discontinuous penalty" contracts that consist of a fixed penalty $p$ that is paid whenever $x < X$ (as in, e.g., a stickk.com contract), we derive a condition similar to the one in our binary model: If the density of cost shocks does not decrease "too quickly" in a region of cost shocks at

which agents with $\tilde{\beta} \in [\underline{\tilde{\beta}}, 1]$ are on the margin for choosing $x = X$, then the commitment contract is unappealing to all agents with $\tilde{\beta}$ in that region, and its perceived damages are decreasing in $\tilde{\beta}$.[12]

**Other models** Finally, in Appendix B.3, we consider the robustness of our results about lack of demand for commitment contracts to alternative models of individual behavior that might generate demand for commitment. We show that in models such as those of Fudenberg and Levine (2006) and Gul and Pesendorfer (2001), penalty-based commitment contracts such as the ones we consider can never be desired, and their expected damages are increasing in the (perceived) cost of self-control, as in the quasi-hyperbolic model. On the other hand, choice-set restrictions are more desirable in the costly self-control models than in the quasi-hyperbolic model,[13] though uncertainty about future costs erodes the benefits of those contracts as well.

## 2.6 Testable predictions of the augmented model, applied to gym attendance

If individuals behave according to the augmented model in Section 2.3, and if there is sufficient uncertainty in the environment, our theoretical results generate four testable predictions. For the sake of concreteness, and to emphasize connection to our experimental setting, we state these predictions for the case of exercise. We assume that exercise has delayed benefits and immediate costs, and that the costs distribution satisfies the conditions assumed in our theoretical results (i.e., there is sufficient uncertainty in the realized costs).

Our first three predictions are direct consequences of Propositions 4, 5 and 6.

**Prediction 1.** *Individuals will demand commitment contracts to both exercise more and to exercise less.*

**Prediction 2.** *There will be a positive correlation between take up of commitment contracts to exercise more and take up of commitment contracts to exercise less.*

**Prediction 3.** *Increasing individuals' sophistication about their time-inconsistency will decrease their demand for commitment contracts to exercise more.*[14]

Our fourth prediction is derived from Proposition 7. We refer to $\omega(p, \Delta p) := \frac{E[w(p+\Delta p)-w(p)]}{\delta} - \frac{E[\alpha_i(p+\Delta p)]+E[\alpha_i(p)]}{2}$ as the per-dollar willingness to pay for behavior change.

**Prediction 4.** *Increasing individuals' sophistication about their time-inconsistency will increase their per-dollar willingness to pay for behavior change.*

[12] We recognize that with continuous choice, the space of possible commitment contracts is very large. A general penalty based commitment contract is a function $\pi$, $\pi(x) \geq 0$, that prescribes a penalty for any possible choice $x$. Analyzing this fully general space of contracts is beyond the scope of this paper. But we doubt that the spirit of results would be different for a more exotic choice of penalties than the one we analyze.

[13] Intuitively, this is because a choice-set restriction eliminates a costly temptation even in states of the world in which it would not have changed choice. See Toussaert, 2018 for further discussion.

[14] As we discuss in Section (4.1), the converse of this prediction does not hold for commitment contracts to exercise less.

# 3 Experimental design and sample for analysis

## 3.1 Design

Our study recruited members of a private fitness facility in a large city in the Midwest U.S. The facility is affiliated with a private university, offering subsidized memberships to graduate students, faculty, and staff, but is also open to the public.

Members of the facility were recruited to participate in a study that elicited, via an online survey, information related to their beliefs about gym attendance and preferences over contingent incentives for using the gym during the 4 weeks following the survey. Members were randomly assigned to various incentives or a control group at the end of the survey. The study was open for three recruitment periods starting in October 2015 and ending in March 2016. During each recruitment period, the study was advertised through email invitations and flyers posted near the gym. The study was open to any gym member for a two week period in each enrollment wave.[15]

In each wave, enrollment was limited to people over the age of 18 who had held memberships over the past eight weeks and who had not participated in the study during any prior wave. Over the three waves, 4,953 members were emailed invitations to participate and 1,292 participated. Waves 1, 2, and 3 had 350, 528, and 414 participants, respectively.

To verify their eligibility, and to enable the match with their prior visits data, members entered the barcode number of their gym ID on the first screen of the online survey. Then, after agreeing to participate via an informed consent form, members began the online component of the study.

The online component contained six sections. The first section elicited consent. The second through fourth sections included information provision, piece rate incentives, and commitment contracts respectively, which we summarize in detail below. The fifth section collected demographic information, and the sixth section administered a randomly selected incentive package to each participant. Appendix Figure A.1 shows the ordering of all parts of the online component of the study.

### 3.1.1 Past attendance and information treatment

The first section of the online component asked participants about their past gym attendance and elicited participants' beliefs and goals regarding their future gym attendance. First, all participants were asked to estimate how many visits they had made in the past 100 days and how many days they thought they should have gone, but did not. Next, participants were assigned with 50% chance to receive an information treatment.

In Wave 1 of the study, the information treatment consisted of a graph showing the number of visits made by the participant in each of the past twenty weeks (Figure 2a). Participants were required to confirm whether they could see the graph in order to proceed to the next page of the

[15] Because many gym members are university students or employees, we scheduled the four-week incentive periods so as to avoid long breaks in the academic calendar. Thus, the first wave of the survey was in the fall semester, the second wave was in the spring semester preceding spring break, and the third wave was in the spring semester following spring break.

survey. In Waves 2 and 3, we enhanced the information treatment in two ways. First, participants were asked to enter their best estimate for the average number of weekly visits they had made, while viewing the graph of their past visits. We anticipated that this would prompt them to pay more attention and better process the information shown. Second, participants were informed that participants from the prior wave of the study had on average overestimated their future attendance by 1 visit per week (Figure 2c).

Participants randomized into the control group (no information) did not see the graph of their own past visits or information about the overestimates of other participants. Instead, they proceeded directly to the elicitation of beliefs and goals regarding gym attendance over the next four weeks. All participants were asked to give their "best guess" of the number of days they would visit over the next 4 weeks (starting the Monday following the date of the survey), their goal number of visits over that period, and their perceived probability of meeting their goal.

#### 3.1.2 Piece-rate incentives

In the next section of the online component, participants were asked to consider six distinct piece-rate (i.e., per day) incentives for going to the gym. These incentives applied over the same period for which they had reported beliefs and goals: the four weeks starting the Monday after they completed the survey. The incentives were \$1/day, \$2/day, \$3/day, \$5/day, \$7/day, and \$12/day. Each incentive was presented on a separate page of the survey, and the order of these pages was randomized.

Participants were asked to estimate how many days (0-28) they expected they would visit the gym over the next four weeks under each incentive. On the same page, they used a slider to indicate their willingness to pay (WTP) for this incentive; i.e., the largest possible fixed payment over which they would prefer to receive the piece-rate incentive. All payments, both fixed and contingent, were to be paid out after the four-week period.

The WTP elicitation was incentive-compatible: at the end of the online component, participants would learn which of the questions had been randomly chosen to apply to them, and which randomly chosen fixed payment would be compared to their WTP to determine their outcome. The online component devoted several pages to developing participants' understanding of how to use a slider to indicate willingness to pay and to explain its incentive compatibility. It also included two questions testing participants' comprehension of the slider. Participants who answered one or both of these questions incorrectly were given another chance to answer correctly before moving to the next section of the survey. See Appendix F for the survey prompts and comprehension test details.

#### 3.1.3 Commitment contracts

In the next section, participants were presented with commitment contract options targeting both more and fewer visits over the same four-week period. For example, in all three waves, participants were given the "more visits" commitment choice shown in Figure 3(a) and the "fewer visits" commitment choice shown in Figure 3 (b).

In Waves 1 and 2, participants made a series of binary choices between an unconditional $80 payment and $80 conditional on making "8 or more," "12 or more," "16 or more," "7 or fewer," "11 or fewer," and "15 or fewer" visits to the gym (i.e., a series of 6 choices). In Wave 3, this section of the survey was modified. Participants were only asked to consider commitments to visit "12 or more" and "11 or fewer" days, but they were also asked for their beliefs about their probabilities of meeting these commitments.

### 3.1.4 Assignment of incentives

To encourage participants to think about these decisions carefully, one question was randomly chosen to count for each participant. When the selected question involved a piece-rate incentive, the participant's WTP for that incentive was compared against a randomly drawn fixed payment. Fixed payments were drawn from a mixture distribution with two components: a uniform distribution from $0-$7 (mixture weight = 0.99), and a uniform distribution from the full range of slider values (mixture weight = 0.01). The rationale for this distribution was to avoid the endogenous assignment of incentives to participants with higher WTPs for those incentives. If the WTP was at least as large as the fixed payment, the participant was assigned the incentives.[16] For the majority of participants with WTPs above $7 for the incentives (i.e., those for whomthe randomly-drawn fixed payment always lay below their WTP and hence, were always assigned the incentives)the incentives they were assigned corresponded to the question we randomly selectedt; e.g., WTP for $5 or WTP for $7.

Given this design, piece-rate incentives were exogenously assigned, with the exception of two rare cases. The first case is when the fixed payment draw exceeded $7 (n=9), and the second case is when a participant indicated a WTP value within the $0-$7 range from which our fixed payments were heavily drawn (n=32). In these two cases, participants with higher WTP values are more likely to receive an incentive, which would bias our estimation of incentive effects on gym visits due to selection. These 41 observations are excluded from the analyses in Sections 6.4 and 8 because they require estimates of the response of exercise to incentives whereas the elicited preference measures do not.

We targeted a small number of questions with high probabilities in order to power our comparisons of the incentive effects. In wave 1, for example, the questions about the $2 and $7 piece rate incentives were each assigned a 0.33 probability of being chosen. To create a control group that did not face any incentive to visit the gym, the survey also included a choice between a $0 incentive and a $20 fixed payment, and this question was also chosen with 0.33 probability. The remaining 1% was a random draw from all six piece-rate incentives and commitment contract questions. In our sample, there were six participants whose randomly selected questions differed from the questions targeted in that wave, resulting in three participants receiving the $5 piece rate incentive, two

[16] To minimize the chance of disappointment, we told participants the following: "To keep within our grant budget, incentives and fixed payments with lower amounts are more likely to be randomly selected, but every incentive and fixed amount we ask you about has some chance of being selected."

receiving the $3 incentive, and one receiving the $1 incentive.

While the 0.33 probability of receiving the $0 incentive remained fixed across the three waves, the targeted incentives were varied in order to document the effects of different incentive sizes.[17] In Wave 2, we shifted half of the probability mass at the $7 piece-rate incentive to the $5 piece-rate incentive, to better understand the curvature of attendance as a function of the linear incentives: 33% $0 incentive; 33% $2 incentive; 16.5% $5 incentive; 16.5% $7 incentive.

In Wave 3, we added a group that would receive $80 conditional on making 12 or more visits, an incentive equivalent to receiving one of the commitment contracts. Participants in this group would receive the $80 conditional payment as long as they had chosen option (a) for the question: "Which do you prefer? (a) $80 incentive you get only if you go to the gym at least 12 days over the next four weeks or (b) $0 fixed payment – no chance to earn money."[18] Since an incentive of $80 for 12 visits equals $6.67 per visit, we determined $7 to be the most comparable piece-rate incentive. Thus, our assignment probabilities in Wave 3 were 33% for the $80 incentive to make 12 visits, 33% $0 incentive, and 33% $7 piece rate incentive, to allow us to compare their effects.

We discuss the rationale for these changes across waves in more detail in Section 3.3. Although this variation of incentive scheme assignments across waves is not ideal for the analyses in Section 5.3, we do not find significant differences between participant pools in the three waves, as shown in Appendix C.[19]

### 3.1.5 Indicators of inattention

In the previous section, we described two study questions that offered a binary choice in which one of the choices, $0, was clearly dominated by the other. These questions were given high probabilities of selection to steer participants into either the group with a $20 fixed payment and no incentives for attendance or the group with an $80 incentive to make 12 visits. However, they also serve to identify participants who may be inattentive to the study instructions, or simply decided to click through the study at random.

The survey also included an attention check and a comprehension check for the WTP elicitations (see questions in Appendix F). The attention check presented a multiple-choice question to the participants but instructed them to click the "next" button without filling out one of the choices, with the explanation that this would indicate their attention to the question prompts.

Using all of these questions, we categorized a total of 117 participants as potentially inattentive to the question prompts. This group includes 49 participants who failed the attention check, 61

[17] Our initial plan to target only two distinct incentive levels was based on conservative estimates of the number of participants our budget would support and the potential variance of the incentive effects.

[18] Note that this is different from the question we used to elicit demand for commitment contracts, in which participants chose between a fixed payment of $80 and the $80 conditional payment. This enabled us to observe behavior under the incentive among both the participants who would and would not select into commitment contracts on their own. All but five individuals (1.2% of Wave 3 participants) who were asked this question chose the $80 incentive over $0.

[19] We also reiterate that the ex-post incentive assignment is only utilized in the analyses in Section 5.3, and does not affect our other analyses.

who failed the comprehension check about the WTP elicitation twice, 5 who chose $0 over the $80 contingent incentive, and 13 who chose $0 over a $20 fixed payment.

#### 3.1.6 Demographics and other questions

The next section of the survey collected demographics, checked numeracy, and elicited a measure of risk aversion through an incentive-compatible choice between different gambles.

#### 3.1.7 Announcement of incentives

In the final section of the survey, participants learned which incentive, if any, they would receive in the next four weeks. Using the process described in Section 3.1.4, which favored relatively low fixed payments, 98.9 percent of participants targeted for a piece-rate incentive received the incentive, and 99.3 percent of participants targeted for the $80 incentive to make 12 visits received that incentive.

The four week incentive period began the Monday after each participant's completion of the survey. Participants received an email upon completion of the survey that confirmed the incentive they were eligible for and the dates. Afterwards, participants were notified via email of their total number of visits and the total payment they had earned. Final payments were disbursed via mailed checks.

#### 3.1.8 Attendance data

Our measure of attendance is computed from participants' swiping into the gym using their membership ID cards. Gym login records are potentially problematic if participants enter and leave the gym to earn incentives without exercising in their usual manner. We do not believe this possibility is a major concern because this behavior includes many of the costs of attending the gym (e.g. travel) but excludes some benefits (e.g. exercise). We also introduced a new checkout procedure partway through the study (in February 2016). Participants after that time were required to swipe out after attending the gym for at least 10 minutes in order to get credit for a visit toward their incentive. Introducing this procedure did not change visit patterns or the estimated incentive effects in the study and the swipe-out records reveal that the vast majority of gym visits lasted substantially longer than 10 minutes.

### 3.2 Sample

Table 1 summarizes demographics, past attendance, recalled attendance, and desired attendance for all participants in the study, as well as the difference between the information treatment and control groups for Wave 1 and for Waves 2-3. The participant pool is 61% female with a mean age of just under 34 years. 56 percent of the participants are either part or full time students, 57 percent work either part or full time, 27 percent are married, just under half hold an advanced degree, and household income averages fifty five thousand dollars. Participants averaged just over 22 gym visits over the last 100 days in the computerized gym records, but recalled just over 30 visits over this

same period.[20] On average, participants also reported that there were 30.5 days in the last 100 days when they thought they should have gone to the gym but did not.

Column 3 shows the p-values for a test for each variable that the treatment group mean equals that of the control group for Wave 1, and Column 5 shows the analogous p-values for Waves 2 and 3. Overall, the results are consistent with good balance between treatment and control groups with the exception the employment variable in Wave 1, where the treatment group works at a higher rate than the control group. However, given the large number of tests (22), it is not unexpected with a test of 5% size.

Compared to other samples in other field experiments on commitment contract demand—particularly those involving low-income populations—our sample is more educated and numerate, due to being affiliated with a university. For example, 96.4% of our sample correctly answered two numeracy questions from Lusardi and Mitchell (2007), which is significantly higher than the rate in the broader U.S. population.[21] Given this high numeracy, it does not seem likely that our sample is more susceptible to "noisy valuation" than the typical sample in commitment contract field experiments.

### 3.3 Rationale for design decisions

We originally designed the experiment for the purpose of quantifying the WTP for behavior change (section 5.1), producing parameter estimates of the partially sophisticated model of quasi-hyperbolic discounting (Section 5.3), and examining the degree to which sophistication is malleable by information provision (Section 6.4). We included commitment contract take up questions to examine the $\beta$ and $\tilde{\beta}$ of the participants who take up those contracts, and to estimate the welfare effects of offering commitment contracts (i.e., how well-targeted they are). We included the "fewer" questions for the purposes of examining the extent of "noise," if any, which we did not have strong priors about at the initial stage.

This affected our design decisions in a few ways. First, because the estimation of $\beta/\tilde{\beta}$ in Section 5.3 requires exogenous assignment of piece-rate incentives, we initially randomized participants into receiving the piece-rate incentives only (with 99% chance). After observing the surprising patterns in commitment demand in Wave 1, we sought to replicate the patterns in Wave 2 with no changes to the commitment contract component. After the Wave 2 replication, we altered our design in Wave 3 to further investigate the mechanisms of commitment contract demand. We randomized some participants into actually receiving the commitment contracts, to make sure that we could replicate previous findings that the commitment contracts do alter behavior (thereby also confirming that participants were not confused about the terms ex-post). Second, we elicited beliefs about the

[20] The biased recollection is consistent with selective memory (e.g., Benabou and Tirole, 2002, 2004) as a potential mechanism for the overestimation of future visits that we document.

[21] The percentage calculation question asks "If the chance of getting a disease is 10 percent, how many people out of 1,000 would be expected to get the disease?" The lottery division question asks "If 5 people all have the winning number in the lottery and the prize is 2 million dollars, how much will each of them get?" For comparison, in a sample of 1,984 adults aged 51-56 in the 2004 HRS, the percentage answering each question correctly were 83.5% (the percentage calculation) and 56% (the lottery division) (Lusardi and Mitchell, 2007).

likelihood of meeting the thresholds stipulated by the "more" and "fewer" contracts, which we use in Section 4.2 to rule out some alternative hypotheses not consistent with the model we propose in Section 2.3.

# 4 Take up of commitment contracts

Unless otherwise noted, in this section, as well as in Section 5, we focus our analysis on the half of the participant pool who did not receive any information treatment, i.e., the control group. For appropriate analyses, we provide replications using the treated group in the appendices.

## 4.1 Take up of commitment contracts for more versus fewer visits

Before examining patterns in take up of commitment contracts in detail, we examine their effects on behavior. Recall that in wave 3, we randomized some participants into receiving the commitment contracts, and that for most participants this assignment was exogenous to their stated desire to take up the contract. Consistent with existing literature, we find that assignment of a "12 or more" visits contract increased attendance by 3.22 visits ($p$-value $< 0.01$) for those participants who wanted the contract, and by 4.00 visits ($p$-value $< 0.01$) for those who did not.

Table 2 shows the take up rates for "more visits" commitment contracts for the three different visit thresholds (8 days, 12 days, and 16 days). Column (1) shows that substantial shares of participants selected the "more visits" contracts at each threshold. The take up rate is declining in the threshold, from a high of 64% at the 8-visit threshold to a low of 36% at the 16-visit threshold. These results would typically be interpreted as clear evidence of widespread awareness of time-inconsistency, combined with a presumably sensible desire to avoid thresholds that are too demanding.

However, column (2) shows that approximately one third of participants selected the "fewer visits" contracts at each threshold as well. Under the standard interpretation of commitment contracts as indicating a desire to influence one's future behavior, take up of these "fewer visits" contracts would be interpreted as a reasonably large share of the population having either awareness of future bias or perceiving visits to the gym as having immediate benefits and delayed costs. The alternative, and we think more plausible interpretation is that participants selecting these contracts either miscalculated or had some other consideration in mind like just wanting to say "yes"—as in the model in Section 2.3.

The extended model in Section 2.3 not only predicts that some participants will select the "fewer visits" contracts but also makes the stronger prediction that some participants will select both contracts types at the same threshold (Proposition 4). Columns (3) and (4) in the table show the shares of participants selecting each type of contract type conditional on selecting the other contract type for each threshold. Many participants selected both the "more visits" and the "fewer visits" contracts at the same threshold. In particular, among participants who made "more visits" contracts at each threshold, nearly half of them also selected the "fewer visits" contract for the same threshold. Choosing both contracts at the same threshold is inconsistent with decisions driven by

awareness of time inconsistency, and thus a strong indicator that noise or alternative considerations are prevalent in commitment contract take up.

The extended model also predicts that if noise or alternative considerations are strong drivers of contract take up, then we will not only see some participants selecting both types of contracts, but a positive correlation in the take up of the two types of contracts (Proposition 5). That prediction is also borne out in the data. The last two columns show that participants who chose the "fewer" commitment contracts were significantly more likely to choose the "more" commitment contracts, and vice versa.

The results suggest that take up of commitment contracts is a "smoking gun" for awareness of present focus. Of course, that does not imply that all take up of commitment contracts was necessarily influenced by noise or alternative considerations. Just over half of the participants who selected "more visits" commitments at each threshold did not select the fewer visits contracts. Those decisions could be consistent with (partially) sophisticated awareness of present focus.

On the other hand, a sizable share of participants who opted for "fewer visits" contracts did so without also selecting "more visits" contracts. Since those choices would tend to indicate some noise or alternative considerations other than awareness of present focus, it is reasonable to conjecture that some share of the participants selecting "more visits" commitments but not "fewer visits" commitments were subject to similar issues.

This seemingly contradictory behavior of demanding "fewer visits" contracts while simultaneously demanding "more visits" contracts is consistent with the extended model in Section 2.3. However, one could argue that an asymmetric error process could make take up of "fewer visits" contracts noisy while not affecting take up of "more visits" contracts. For example, people could mistake "fewer visits" contracts for "more visits" contracts. But the fact that some people select "fewer visits" contracts without also selecting "more visits" speaks against this possibility as an explanation for all choices. Moreover, the experimental instructions made a clear distinction between the two types of contracts, presenting them together with the differences underlined for emphasis (see Figure 3). For example, at the 12-visit threshold the "more visits" contract underlined "at least 12" and the "less visits" contract underlined "11 or fewer."

**Summary:** Participants take up both "more visits" commitment contracts and "less visits" commitment contracts. There is a positive correlation in the take up of both types of contracts.

### 4.2 Correlates of commitment contract take up

Another strategy for assessing whether asymmetric errors can fully account for the observed patterns in take up of "more visits" and "fewer visits" is to look at the correlates of take up. If participants were simply confusing "fewer" contracts for "more" contracts, then any variable that is positively correlated with perceived success in or take up of a "more" contract should also be positively correlated with perceived success in or take up of a "fewer" contract.

Table 3 shows that participants differentiated between questions about perceived likelihood of

success in a "more" contract versus a "fewer" contract.[22] Participants who expected to attend the gym frequently in the absence of incentives were more likely to believe that they would meet the terms of a "more" contract, and less likely to believe that they would meet the terms of a "fewer" contract. This implies that at least in answering the forecasting questions, participants were not simply confusing the "fewer" contract for the "more" contract.

In Table 4 we then look at how the perceived likelihood of success correlates with actual take up. As columns (1)-(3) of Table 4 show, beliefs about the likelihood of meeting the "12 or more" visits threshold of the "more" contract are positively associated with choice of the "more" contract but negatively associated with choice of the "11 or fewer" visits contract. The converse holds for individuals who think that they are more likely to meet the "11 or fewer" visits threshold. These patterns are consistent with our extended model, which predicts that the more damaging the contracts would appear in the absence of noisy valuation or alternative considerations, the less likely they would be chosen.

In Appendix D.2 we continue to build on this analysis and present correlations of commitment contract take up with (i) expected attendance in the absence of incentives, (ii) past attendance, and (iii) desired "goal" attendance. Each of these three variables is significantly positively correlated with take up of "more" contracts, and significantly negatively correlated with take up of "fewer" contracts.

Finally, we consider how much of commitment contract take up could be driven by participants who were inattentive to the experiment and could have been choosing randomly. Recall that in Section 3.1.5 we identified a total of 117 participants (53 in the no information control group) who were inattentive to at least one part of the study. Excluding these participants, approximately 8% of the sample, lowers overall demand for the "fewer" contracts from 32% to 31%, and has no effect on demand for the "more" contracts.

Recall also that in Wave 3, participants faced the following question: "Which do you prefer? a) \$80 incentive you get only if you go to the gym at least 12 days over the next four weeks or b) \$0 fixed payment – no chance to earn money." All but 5 participants (1.2% of Wave 3) chose option a). This implies that our findings are not induced by random choices of disengaged participants, but rather by the more meaningful forms of noisy valuation or alternative considerations proposed in Section 2.3.

**Summary:** The take up of "fewer" commitment contracts cannot be explained by participants merely confusing "fewer" contracts for "more" contracts. It is also not driven by participants who were simply inattentive and choosing randomly.

[22]For Tables 3 and 4, we restrict our analysis to wave 3–the only wave for which we elicited beliefs about the likelihoods of meeting the commitment contract thresholds.

### 4.3 Replication for participants in the information treatments

The results in this section are not limited to participants in the no information control group. In Appendix D.1 and D.2 we replicate the results for participants in the information treatment group.

## 5 Willingness to pay for behavior change

Our results thus far call into question the assumption that take up of commitment contracts implies awareness of limited self-control. In this section we show how the more robust methodology described in Section 2.4 can be used to provide both reduced-form evidence of awareness of limited self-control, as well as parameter estimates.

### 5.1 Willingness to pay for behavior change

As Proposition 7, and its dynamic generalization in Appendix B.1 shows, average willingness to pay (WTP) data provides an alternative indicator of awareness of time-inconsistency. Figure 4 graphs the average willingness to pay for piece-rate incentives elicited from our participants for each of the six different piece-rate levels. The figure also shows two measures that are derived from elicitation of participants' average expectations about the number of times they would visit the gym at each piece-rate level. The dashed line shows the average subjective expected earnings at that piece rate. The blue line with squares gives the average subjective time-consistent willingness to pay that would be consistent with the participants' beliefs about visits under the different incentive levels. This time-consistent-benchmark line is calculated using the approximation derived in (the dynamic extension of) Proposition 7 (again, see Appendix B.1) with $\tilde{\beta}$ set equal to 1: we use the proposition to compute the difference in WTP for any pair of consecutive incentive levels, and we then take the sum of the differences.[23]

Consistent with theoretical predictions for the case of (partial) awareness of time inconsistency, average WTP is above participants' subjective expected earnings for low incentives and below expected earnings for high incentives. For example, under a \$1 per-visit piece-rate, participants believed that they would attend an average of 13.2 times but had an average willingness to pay for a \$1 piece-rate incentive of \$18.37, \$5 more than their subjective expected earnings. The average willingness to pay is above the time consistent benchmark at all incentive levels.

To create our aggregate measure of WTP for behavior change, we follow (the dynamic extension of) Proposition 7 in Section 2.6. The willingness to pay for behavior change when going from one incentive $p_k$ to the next higher incentive $p_{k+1}$ is defined as the increase in willingness to pay per dollar of incentive rise minus the average visit rate the participant expects at the two different incentive levels. We calculate this value of behavior change for each participant for each of the six piece-rate incentive increases (i.e., \$0 to \$1, \$1 to \$2,..., \$7 to \$12), and normalize it by $(p_{k+1} - p_k)$.

[23] Formally, for incentive levels $p_0 = 0, p_1, \ldots p_K$, we use the proposition to compute the counterfactual $E\left[WTP_i(p_{k+1}) - WTP_i(p_k)\right]$ if $\tilde{\beta}_i = 1\ \forall i$, and we then compute $E[WTP(p_k)] = \sum_{j=1}^{k} E\left[WTP(p_j) - WTP(p_{j-1})\right]$.

Consistent with our conjecture of noisy valuation of contract values, we find substantial variation in these valuation measures at the individual level, with some even negative.[24] Following Proposition 7 and its generalization, our approach is instead to analyze the average valuations for behavior change in the population. Concretely, we average the estimates of WTP for behavior change over all individuals and all incentive levels.

Figure 5 shows the average value across six incentive levels, as well as the average excluding the valuation of increasing the piece-rate from \$0 to \$1, along with 95% confidence intervals computed from estimates of heteroskedasticity-robust standard errors. On average, participants exhibited a valuation for behavior change of \$1.40 per \$1 of incentive increase. However, this valuation is driven in part by an especially large estimated valuation for behavior change going from no incentive to the \$1 incentive. As Proposition 7 shows, if there are fixed effects influencing willingness to pay for contingent incentives, the more robust measure of the valuation of behavior change involves only changes in positive piece-rate amounts. This more conservative average is \$0.55 per dollar of piece-rate increase, and is also statistically significant.

A linear regression of expected attendance on the piece-rate incentives shows that participants expect that on average, a \$1 change in piece-rates will increase attendance by 0.67 visits (participant-cluster-robust s.e. 0.014). This implies that our two measures of WTP of behavior change per dollar of piece-rate incentives translate to WTPs of \$2.10 and \$0.83 per attendance.

**Summary:** Participants' willingness to pay for increases in piece-rate incentives is higher than expected increases in profits generated by those increases. This reveals a willingness to pay for changing their future selves' behavior, and consequently at least partial sophistication.

### 5.2 Correlation with commitment contract take up

Having established our more robust measure of participant's willingness to pay for behavior change, a natural question to ask is how it correlates with take up of the "more" commitment contracts. A positive correlation would be indicative that at least on average, participants who are more likely to take up the contracts have a higher desire to change their behavior.

Table 5 shows that there is no correlation between our measure of willingness to pay for behavior change and the take up of commitment contracts. In these regressions, all commitment contracts for more visits are pooled together and take up is regressed on the z-score of estimated WTP for behavior change. This WTP estimate is based on the WTP values given for all incentive amounts in columns (1)-(2), or for incentive amounts excluding the \$1 incentive in columns (3)-(4). In columns (2) and (4), we control for the elasticity of each individual's visit expectations with incentive size. We find no significant correlations between the WTP values and commitment contract take up. The point estimates are essentially zero when using the measure of average WTP across all incentive

[24] For example, we observe that the estimated value of behavior change is negative for 34 percent of the individual valuation measures. If we took those negative measures at face value, it would imply that participants have a desire to reduce their gym use at some incentive levels 34 percent of the time. However, these negative values more likely represent noise in participants' decisions about willingness to pay and/or their estimates of visit rates.

levels, and slightly negative when using the measure that excludes the $1 incentive level. According to these estimates, one standard deviation increase in WTP for behavior change reduces the take up of commitment contracts by up to 3 percentage points.

One interpretation of these results is that take up of commitment contracts in our experiment is not a strong correlate of actual demand for behavior change. Another possible interpretation is that the correlation between these two measures is low simply because both of these measures are very noisy proxies of demand for behavior change. We can rule out this second possibility by constructing pairwise correlations between (individual level) estimates of WTP for behavior change at each different level of piece-rate incentives (e.g., correlation between WTP for behavior change at a $1 incentive and WTP for behavior change at a $2 incentive, etc.), and the pairwise correlations of demand for the three types of "more" commitment contracts (e.g. correlation between choosing the "8 or more" contract and choosing the "12 or more" contract, etc.). We estimate that the average pairwise correlation of our measures of WTP for behavior change is 0.18 (bootstrapped cluster-robust s.e. 0.055) and the average pairwise correlation of demand for the different "more" contracts is 0.50 (bootstrapped cluster-robust s.e. 0.02). These results show that on average, the WTP for behavior change at one piece rate incentive is not so noisy that it is unassociated with the WTP for behavior change at another piece rate incentive. Likewise, the demand for one "more" contract is not so noisy that it is unassociated with demand for a different "more" contract. As such, if the WTP measures are not correlated with the demand for commitment contracts, it is unlikely that the lack of an association is because the measures are simply too noisy. Rather, it is more plausible that they measure different things.

**Summary:** Demand for commitment contracts in our experiment is not meaningfully correlated with our more robust measure of demand for behavior change. The lack of correlation in the measures is unlikely to be due merely to noise in the data.

## 5.3 Parameter estimates

Our results so far provide reduced-form evidence of some demand for behavior change, and thus that at least some participants must have $\tilde{\beta} < 1$. Here, we translate these reduced-form results into implications for parameter estimates.

### 5.3.1 Estimates of $\tilde{\beta}$

Our reduced-form measures of willingness to pay for piece-rate incentives provide estimates of the perceived short-run discount factor$\tilde{\beta}$ for a given value of delayed health benefits $b$. We use the generalization of Proposition 7, proven in Appendix B.1, which allows for multiple periods of action. In the dynamic extension, Proposition 7 continues to hold, except with $\alpha$ now denoting total expected attendances. We also make the additional assumption that $b$ and $\tilde{\beta}$ are homogeneous across individuals. Under these assumptions, equation (1) identifies $\tilde{\beta}$ for any two incentive levels given (i) WTP for those incentives, (ii) expected beliefs at those incentives, and (iii) a value of $b$.

Our experiment provides data on all of the requisite statistics other than $b$, which we calibrate using evidence from public health. In Appendix E.4 we provide public health and epidemiological evidence on the value of exercise, which suggests per-attendance health benefits between \$4 and \$20. As such, we provide estimates for $b$ ranging from \$1 to \$20.

Because we have multiple levels of piece-rate incentives, our model is over-identified. Formally, given our six positive piece-rate incentive values $p_1 < \cdots < p_6$, we use (the extension of) equation (1) to generate five moment conditions, one for each adjacent pair of incentives $p_i$, $p_{i+1}$ $(i = 1, \ldots, 5)$. We use the WTP data corresponding to our second, more robust measure of WTP for behavior change, excluding the WTP for increasing piece-rate incentives from \$0 to \$1. We estimate $\tilde{\beta}$ to be the value that minimizes the weighted sum of squared differences between the left-hand-side and the right-hand-side means of the five moments obtained from equation (1). We use a two-step estimator as in Hall (2005) to obtain the efficient weights on the moment conditions, and we cluster standard errors at the participant level. See Appendix E.2 for further details of the moment conditions and the estimator.

Figure 6 presents the point estimates and 95% confidence intervals of $\tilde{\beta}$ for each value of $b$ in the range of \$1 to \$20. Overall, our estimates of $\tilde{\beta}$ range from approximately 0.74 for $b =$ \$1 to approximately 0.93 for $b =$ \$20, with $\tilde{\beta}$ approximately 0.88 for the middling value of $b =$ \$10.

Intuitively, $(1 - \tilde{\beta})$ must decrease in $b$ because the higher is the health value of exercise, the costlier is the perceived misbehavior stemming from $1 - \tilde{\beta} > 0$, and thus the higher must be the WTP for changing one's future behavior through piece-rate incentives. Consequently, for a given set of WTP data, a higher $b$ must imply a lower $1 - \tilde{\beta}$. Despite that, Figure 6 shows that the identified set of $\tilde{\beta}$ is still relatively narrow. Intuitively, this is because the perceived cost of one's future misbehavior stems not only from losing out on the delayed health benefits $b$, but also from losing out on the piece-rate incentives. And, because most of our piece-rate incentives are reasonably large relative to the range of plausible values of $b$, we obtain relatively tight bounds on $\tilde{\beta}$.

### 5.3.2 Estimates of $\beta/\tilde{\beta}$

The variation in piece-rate incentives, combined with forecasts of attendance under those piece-rate incentives, also allows us to set-identify the degree of actual present focus, $\beta$, under additional assumptions. The intuition for identification is as follows: If a piece-rate $p^*$ that increases the delayed benefits by, e.g., 30% leads to average attendance that equals the expected attendance in the absence of piece-rate incentives, then $\beta/\tilde{\beta} = 1/1.3 = 0.77$.

Formally, we show in E.3 that perceived attendance $\alpha(p)$ and actual average attendance $\alpha^*(p)$ can be expressed as $\alpha(p) = A(\tilde{\beta}(b+p))$ and $\alpha^*(p) = A(\beta(b+p))$, for some function $A$. Consequently, if $\alpha(0) = a^*(p^*)$, then $\tilde{\beta}b = \beta(b + p^*)$, and thus

$$\beta/\tilde{\beta} = b/(b + p^*). \tag{2}$$

The key identifying assumption is that all overestimation of future behavior is due to naivete

about present focus; that is, due to $\tilde{\beta} > \beta$. This assumption is probably too strong, as participants may also overestimate future attendance due to planning fallacies that lead to an underestimation of the future hassle costs of gym attendance. Moreover, participants' forecasts may be systematically upward biased of actual beliefs if they see the forecasting prompt as a more aspirational exercise.[25] If some of the overestimation is due to reasons other than naivete about $\beta$, then our procedure generates lower bounds on $\beta/\tilde{\beta}$.

Because the identification strategy here relies on estimates of the impact of incentives on actual behavior, we exclude the 16 participants for whom incentives are not (or would not be) exogenously assigned, as described in Section 3.1.4.[26] This leads to a slightly smaller sample than the one used throughout the paper, though the impacts of these restrictions on any of the other estimates reported in the paper are inconsequential.

To produce an estimate of $p^*$, we begin with Figure 7 , which reports perceived and actual attendance behavior. These statistics are reported for this smaller sample. Figure 7 shows that participants do, indeed, significantly overestimate their future attendance at all levels of piece-rate incentives.

As Figure 7 also shows, the incentive level at which actual attendance equals expected attendance without an incentive is approximately \$5. Formally, we approximate the incentive $p^*$ by first estimating attendance as a quadratic function of piece-rate incentives, and then solving the quadratic equation to find the price $p^*$ at which actual attendance equals the attendance expected at $p = 0$. To compute standard errors that reflect sampling error both in the quadratic fit and in the estimate of perceived attendance in the absence of incentives, we simultaneously estimate two moment conditions: one for the quadratic fit and one for the estimate of perceived attendance in the absence of incentives. We then compute the standard error around $b/(b + p^*)$ using the delta-method, clustering at the participant level. Appendix E.3 provides further details.

Figure 8 shows the resulting estimates of $\beta/\tilde{\beta}$. As equation (2) shows, this statistic is particularly sensitive to calibrations of the health benefits $b$. Consequently, the range in Figure 8 is wider than the range in Figure 6. Overall, however, the figure suggests significant naivete, despite a clear demand for behavior change. At the highest value of $b = \$20$, for example, the estimate of $\beta/\tilde{\beta}$ is approximately 0.81, while at the middling value of $b = \$10$, the estimate of $\beta/\tilde{\beta}$ is approximately 0.67. Combined with our estimates of $\tilde{\beta}$, this implies a $\beta = 0.59$ at the middling value $b = \$10$.

Overall, the results indicate that despite perceiving that they are present-focused ($\tilde{\beta} < 1$), people are more present-focused than they perceive ($\beta < \tilde{\beta}$). Even if some of the overestimation of future visits was due to mechanisms other than overestimation of $\beta$, the general conclusion of non-negligible naivete would hold.[27]

[25] As we discuss in Section 2.4, systematic bias in stated versus actual beliefs does not bias estimates of $\tilde{\beta}$ if the bias is a level shift that does not affect how participants state they respond to *changes* in incentives.

[26] Excluding these participants in our estimation of $\tilde{\beta}$ has no effect on the result.

[27] Roughly speaking, if half of the overestimation was due to other mechanisms, then our estimates of $1-\beta/\tilde{\beta}$ would be approximately half as large.

**Summary:** Our parameter estimates imply a perceived short-run discount factor $\tilde{\beta}$ below 1, indicating some awareness of present focus, but they also imply substantial naivete.

# 6 Debiasing beliefs

## 6.1 Impact of the information treatment on beliefs

As described in Section 3, our experiment included an information treatment aimed at debiasing overoptimistic beliefs about gym attendance. In the first wave of the study (Fall 2015), we tested a basic information treatment which presented participants in the treatment group with a graph of their prior visit patterns over the prior 20 weeks. This treatment was unsuccessful at debiasing overoptimistic beliefs. Subfigure (a)of Figure 9 shows the average expected visit rates participants reported in the survey at each piece-rate incentive level for both the control group (who were given no graph of prior visits) and the treatment group, along with 95% confidence intervals. It is clear from the figure that this treatment had no effect on expectations of future visits.

Having observed this lack of response to the information treatment after Wave 1, we launched an enhanced information treatment for Waves 2 and 3 of the survey, as described in Section 3. For this enhanced information treatment, participants were asked to estimate their average visit rate from the graph of their own past visits and were informed that participants in Wave 1 had on average overestimated their visits at this same fitness facility by about 1 visit per week.[28] As subfigure (b) of Figure 9 shows, this revised information treatment significantly reduced expected visit rates both under no incentive and at each possible incentive level for the treatment group relative to the control group. The reduction in expected visits over the study month for those seeing the information treatment ranged from 1 to 2 visits depending on the incentive.

Figure 10 shows that the net effect of the enhanced information treatment was a partial debiasing that reduced but did not completely eliminate the gap between participants' expectations and the reality of their visit patterns (for this figure, we exclude the 41 participants described in Section 3.1.4 for whom incentives were not exogenously assigned). In this figure we plot both expected and realized visit rates under no incentive and the three incentive levels that were targeted for assignment ($2, $5, and $7). A comparison of realized visits the treatment and control groups reveals that the information treatment had no economically or statistically significant impact on actual visit attendance.[29] As such, the net effect of the information treatment was a partial reduction in participants' level of overconfidence, representing between a one-third to one-fourth reduction in the level of overestimation of visit frequency.[30]

[28] The statement about prior participants was accurate and reflected a comparison of the average expectations (11.4 visits) and the realized average visits for the control group (7 visits) from Wave 1.

[29] In a regression controlling for the incentive level received, we estimate an average affect on visits of receiving the info treatment of -0.18 visits over the 4-week period, with a 95% confidence interval ranging from -1.14 to 0.77.

[30] We are underpowered for analyzing how the information treatment affected the perceived likelihood of meeting the commitment contract thresholds, because we collected beliefs about surpassing the threshold of only one pair of contracts (the 12 ore more visits contract and the 11 or fewer visits contract), and in only one of the waves. Nevertheless we find qualitatively similar results. The information treatment decreased the expected likelihood of

**Summary:** An information treatment that induced participants to engage with their past visit frequency, and informed them that participants in a previous wave of the experiment overestimated their visits, successfully reduced overestimation of gym visits.

### 6.2 Impact of the information treatment on willingness to pay for behavior change

As described in Section 2.6, our model predicts that an increase in an individual's level of sophistication about their present focus (i.e., moving $\tilde{\beta}$ toward $\beta$) should increase the perceived value of behavior change induced by piece-rate incentives. In other words, reducing overoptimism about getting to the gym should (on average, in the presence of noisy valuation) increase willingness to pay for a mechanism that will help motivate more visits.

Consistent with this, we find that participants in the information treatment showed significantly higher valuations for behavior change via their willingness to pay for piece-rate incentives. Table 6 shows the estimated effect of both the basic information treatment, which failed to shift expectations of future visits, and the enhanced information treatment, which reduced overoptimism by approximately one-third, on the average valuation for behavior change. Under the enhanced information treatment, both the average valuation across all incentives and the average excluding the \$1 incentive increased substantially. Across all incentive levels, we estimate that the information treatment increased the average value of behavior change by \$1.15 per dollar of incentive increase [95% CI \$0.29 - \$2.02] and estimate an increase of \$1.33 for the average excluding the \$1 incentive [95% CI \$0.43 - \$2.24].

These results are consistent with the interpretation that the information treatment at least partially increased sophistication about participants' time-inconsistency. If the information treatment affected only other sources of misperceptions, like underestimation of one's future time constraints, it would not be expected to have a pronounced affect on demand for behavior change.

**Summary:** The enhanced information treatment increased participants' willingness to pay for behavior change. This implies that the enhanced information treatment reduced naivete about time-inconsistency.

### 6.3 Impact of the information treatment on commitment contract take up

Table 7 shows the estimated effect of both information treatments on the take up rates of each "more-visits" commitment contract (columns (1)-(3)) and on all "more-visit" contracts pooled together (column (4)). The enhanced information treatment reduced the take up rate by approximately 5 percentage percentage points at both the 8-visit and 12-visit thresholds ($p$-value = .18 and $p$-value = 0.09, respectively) and by 10 percentage points at the 16-visit threshold ($p$-value = 0.02). On average, the information treatment reduced demand for commitment by a statistically significant 7

meeting the 12 more visits contract and increased the likelihood of meeting the 11 or fewer visits contract; the overall difference in these effects 7.7 percentage points ($p = 0.038$).

percentage points ($p$-value = 0.02). This empirical result is consistent with the theoretical prediction in Proposition 6. The finding that only the enhanced information treatment affected commitment contract take up is consistent with the results about beliefs in Section 6.1.

**Summary:** The enhanced information treatment decreased participants' take up of "more" commitment contracts.

### 6.4 Impact of the information treatment on parameter estimates

How do the reduced-form results about the impact of information provision on beliefs translate in to the perceived short-run discount factor $\tilde{\beta}$? To answer this question, we utilize the methodology of Section 5.3 and Appendix E.2 to estimate $\tilde{\beta}$ for both the control group and the enhanced information treatment group. Again, we use the WTP data corresponding to our second, more robust reduced-form measure of WTP for behavior change, excluding the WTP for increasing piece-rate incentives from \$0 to \$1. Figure 11 presents the results, again for a range of health benefits between \$1 and \$20. The implied differences are meaningful. At the middling value of $b$ = \$10, for example, the debiasing intervention decreases $\tilde{\beta}$ from 0.88 to 0.82.

## 7 Conclusion

Why do people take up commitment contracts? The typical revealed preference logic in the literature has been that people are revealing a desire to change their future selves' behavior when they decide to take up a commitment contract. This paper shows that the choices could also revealing calculation errors, or that individuals have alternative considerations arising from demand effects, distrust, or signaling.

Empirically-minded researchers who work with experimental or survey data are well aware that this kind of data is often noisy, and recent work in economics has provided theoretical foundations for some sources of this noise, grounded in neurobiology (Woodford, 2012; Wei and Stocker, 2015; Khaw et al., 2017; Frydman and Jin, 2019). Acknowledging such noise, the usual approach by empirical researchers is to limit analysis to group means, since mean-zero measurement error does not bias estimates of means.

However, mean-zero error is an unrealistic assumption in binary choice data, a fact that has long been known in the econometrics of measurement error literature (Aigner, 1973; Hausman, 2001). Even if the the errors are symmetric—say 10% of the individuals always choose the wrong option—binary choice data will typically introduce bias. For example, if 10% of choices are mistakes, then in a world in which only 5% actually want option A, approximately 14% will end up choosing it.

Mean-zero error assumptions, like those invoked in our derivation of WTP for behavior change, are more realistic when the outcome of interest is continuous and uncensored. Consequently, designs utilizing continuous dependent variables are more likely to be robust to the presence of noise.

This paper provides evidence that noise could be an important factor in commitment contract choices. We provide an extension of the standard quasi-hyperbolic model that explicitly allows for such noise, derive testable predictions of this model, and provide strong support for the predictions in a field experiment on gym attendance.

Better understanding the nature of motives and mistakes in commitment contract demand informs not only positive analysis but also normative analysis. The insights from this study should help to inform thinking about the relative benefits of commitment contracts versus "sin taxes" (O'Donoghue and Rabin, 2006; Allcott et al., forthcoming) as a way of addressing suboptimal decisions arising from present focus. If people are sophisticated, have limited uncertainty about the desirability of target actions, and their decisions are not affected much by noisy mental processing or alternative considerations, commitment contracts can be a well-targeted policy tool. Sin taxes are a more blunt policy tool because they affect everyone, not just those who are present focused. Yet if people are (partially) naive, have a need for flexibility due to uncertainty about about future needs, tastes, and time constraints uncertainty, and sometimes make noisy decisions, the targeting benefits of commitment contracts will be low and sin taxes and subsidies will be more efficient. Our findings suggest that exercise behavior may fall into this latter case.

Of course, our results do not imply that the take up of commitment contracts only reflects noise, or that commitment contracts are never a well-targeted intervention. Instead, our work provides a set of steps that researchers and policy designers can use to better evaluate the effectiveness of commitment contracts in a variety of settings where they might be considered.

First, designers should consider, and ideally try to assess, the extent to which there is uncertainty and a need for flexibility regarding the behavior being considered. Commitment contracts will be most effective when there is little uncertainty about whether the behavior will be desirable. For example, low income individuals who experience a lot of financial instability may not experience large benefits from commitment contracts for future financial plans.

Second, designers should, whenever possible, build in the option for committing to do *less* of the targeted (and presumably) beneficial behavior, as we have introduced in this study. Observing the take up of these alternative commitments, and how it is correlated with take up of "standard" contracts, provides a simple and powerful way to assess the extent to which observed demand for commitment contracts reflects the awareness of self-control problems that they are intended to address.

Other studies have proposed examining the validity of commitment contract take up by analyzing how take up correlates with proxies of present focus, as well as by appealing to prior experience with commitment contracts. However, some studies find a positive correlation between patience and commitment demand (Augenblick et al., 2015; Kaur et al., 2015), while others find a negative correlation (Sadoff et al., forthcoming; John, forthcoming); and both sets of studies argue that their results can be explained by standard present focus models. Moreover, appealing to prior experience with commitment contracts may be problematic because it may in fact amplify alternative considerations by creating a status quo bias, or by amplifying perceived pressure from the experimenter.

Finally, designers can use our approach of measuring the willingness to pay for behavior change to construct an alternative reduced-form measure of participants' desire for behavior change. This can provide a means of cross-validating the extent to which commitment contracts are effectively targeting those with recognized self-control problems. Well-targeted commitment contract programs should result in a strong correlation between commitment contract demand and the measured desire to change behavior from willingness-to-pay data.

At the same time, there are at least a few important questions left open by our study, that future work should address. First, although we theoretically clarify the important role that uncertainty about future costs plays in commitment contract demand, we do not explore it empirically. In part, this is because the initial focus of our design was on obtaining estimates of present focus using WTP for piece-rate incentive data. And in part, this is because eliciting uncertainty about future hassle costs using simple and transparent survey questions is challenging. Yet settings with naturally occurring differences in uncertainty, like Kaur et al. (2015), are clearly in line with our theoretical results. Future work should hone in on this comparative static.

Second, it is natural to expect that in the presence of noisy valuation and alternative considerations, stakes will matter. Although our $80 stakes were not low relative to many other commitment contract experiments, settings like those of Ashraf et al. (2006), Kaur et al. (2015), and Schilbach (2019) feature larger stakes. Although the participants in those studies are likely to be significantly less numerate than the participants in our study, and thus presumably more susceptible to noisy decision processes, it is possible that the larger stakes in those studies lead to less noise than what we observe. Analyzing the impact of stakes, holding the sample constant, is another important question for future research.

Finally, it will be useful to explore analogues of our design when participants can set their own target thresholds. As we have explained in Section 2.4, noise and alternative considerations will still introduce bias in commitment contract take up unless the set of possible targets includes negative values (i.e., options to commit to less of the goal activity). However, some patterns of choice may be quantitatively different. For example, if a lot of commitment contract take up is driven by a simple desire to accept offers, then participants should just select low targets that they would exceed even without the commitment.

Our hope is that our paper serves as a useful foundation for further research into the drivers of commitment contract demand and present focus more broadly. Obtaining a more complete understanding of what commitment contracts do and when they should be deployed is crucial, but can only come about from further research building on tests such as ours.

# References


**Acland, Dan and Matthew R Levy**, "Naiveté, projection bias, and habit formation in gym attendance," *Management Science*, 2015, *61* (1), 146–160.

**_ and Vinci Chow**, "Self-control and demand for commitment in online game playing: evidence from a field experiment," *Journal of the Economic Science Association*, 2018, *4* (1), 46–62.

**Aigner, Dennis J.**, "Regression with a Binary Independent Variable Subject to Errors of Observation," *Journal of Econometrics*, 1973, *1*, 49–60.

**Allcott, Hunt, Benjamin B. Lockwood, and Dmitry Taubinsky**, "Regressive Sin Taxes, with an Application to the Optimal Soda Tax," *Quarterly Journal of Economics*, forthcoming.

**Amador, M., I. Werning, and G.-M. Angeletos**, "Commitment vs. Flexibility," *Econometrica*, 2006, *74*, 365–396.

**Ashraf, Nava, Dean Karlan, and Wesley Yin**, "Tying Odysseus to the Mast: Evidence From a Commitment Savings Product in the Philippines," *The Quarterly Journal of Economics*, May 2006, *121* (2), 635–672.

**Augenblick, Ned and Matthew Rabin**, "An Experiment on Time Preference and Misprediction in Unpleasant Tasks," *The Review of Economic Studies*, 2019.

**_ , Muriel Niederle, and Charles Sprenger**, "Working over Time: Dynamic Inconsistency in Real Effort Tasks," *The Quarterly Journal of Economics*, 2015, *130* (3), 1067–1115.

**Bai, Liang, Benjamin Handel, and Ted Miguel**, "Self-Control and Demand for Preventive Health: Evidence from Hypertension in India," *working paper*, 2019.

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

**_ and _** , "Willpower and Personal Rules," *Journal of Political Economy*, 2004, *112* (4), 848–887.

**Beshears, John, James J Choi, Christopher Harris, David Laibson, Brigitte C Madrian, and Jung Sakong**, "Self Control and Commitment: Can Decreasing the Liquidity of a Savings Account Increase Deposits?," 2015.

**Blair, Steven N., Harold W. Kohl, Ralph S. Paffenbarger, Debra G. Clark, Kenneth H. Cooper, and Larry W. Gibbons**, "Physical Fitness and All-Cause Mortality A Prospective Study of Healthy Men and Women," *Journal of the American Medical Association*, 1989, *262* (17), 2395–2401.

**Carroll, Gabriel D, James J Choi, David Laibson, Brigitte C Madrian, and Andrew Metrick**, "Optimal defaults and active decisions," *The Quarterly Journal of Economics*, 2009, *124* (4), 1639–1674.

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

**_ , John A List, and Ulrike Malmendier**, "Testing for altruism and social pressure in charitable giving," *Quarterly Journal of Economics*, 2012, *127* (1), 1–56.

**Exley, Christine L. and Jeffrey K. Naecker**, "Observability Increases the Demand for Commitment Devices," *Management Science*, 2016, *63* (10), 3262–3267.

**Frydman, Cary and Lawrence J. Jin**, "Efficient Coding and Risky Choice," *working paper*, 2019.

**Fudenberg, Drew and David K. Levine**, "A Dual-Self Model of Impulse Control," *American Economic Review*, 2006, *96* (5), 1449–1476.

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

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

**Hall, Alistair R.**, "Generalized Method of Moments," 2005.

**Hansen, Lars Peter**, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 1982, *50* (4), 1029–1054.

**Harberger, Arnold**, "Taxation, resource allocation, and welfare," in "The role of direct and indirect taxes in the Federal Reserve System," Princeton University Press, 1964, pp. 25–80.

**Hausman, Jerry**, "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problemsfrom the Left," *Journal of Economic Perspectives*, 2001, *15* (4), 57–67.

**Heidhues, Paul and Botond Kőszegi**, "Futile Attempts at Self-Control," *Journal of the European Economic Association*, 2009, *7* (2), 423–434.

**Houser, Daniel, Daniel Schunk, Joachim Winter, and Erte Xiao**, "Temptation and Commitment in the Laboratory," *Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 488*, 2010.

**John, Anett**, "When Commitment Fails – Evidence from a Field Experiment," *Management Science*, forthcoming.

**Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan**, "Self-Control at Work," *Journal of Political Economy*, 2015, *123* (6), 1227–1277.

**Khaw, Mel Win, Ziang Li, and Michael Woodford**, "Risk Aversion as a Perceptual Bias," 2017.

**Laibson, David**, "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 1997, *112* (2), 443–478.

_ , "Why Don't Present-Baised Agents Make Commitments?," *American Economic Review*, 2015, *105* (5), 267–272.

_ , "Private Paternalism, the Commitment Puzzle, and Model-Free Equilibrum," *AEA Papers and Proceedings*, 2018, *108*, 1–21.

_ **and Keith M. Ericson**, "Intertemporal Choice," in B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, eds., *Handbook of Behavioral Economics*, Vol. 2, Elsevier, 2019.

_ , **Peter Maxted, Andrea Repetto, and Jeremy Tobacman**, "Estimating Discount Functions with Consumption Choices over the Lifecycle," *working paper*, 2018.

**Lusardi, Annamaria and Olivia S. Mitchell**, "Baby Boomer Retirement Security: The Roles of Planning, Financial Literacy, and Housing Wealth," *Journal of Monetary Economics*, 2007, *51* (1), 205–224.

**McKelvey, Richard and Thomas Palfrey**, "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior*, 1995, *10*, 6–38.

**McKelvey, Richard D. and Thomas R. Palfrey**, "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior*, 1995, *10* (1), 6–38.

**Milkman, Katherine L., Julia A. Minson, and Kevin G. M. Volpp**, "Holding the Hunger Games Hostage at the Gym: An Evaluations of Temptation Bundling," *Management Science*, 2014, *60* (2), 283–299.

**Neumann, Peter J., Joushua T. Cohen, and Milton C. Weinstein**, "Updating Cost-Effectiveness: The Curious Resilience of the $50,000 per-QALY-Threshold," *The New England Journal of Medicine*, 2014, *371* (9), 796–797.

**O'Donoghue, Ted and Matthew Rabin**, "Doing It Now or Later," *American Economic Review*, 1999, *89* (1), 103–124.

_ **and** _ , "Choice and Procrastination," *Quarterly Journal of Economics*, 2001, *116* (1), 121–160.

_ **and** _ , "Optimal sin taxes," *Journal of Public Economics*, 2006, *90* (10), 1825–1849.

**Paserman, M Daniele**, "Job Search and Hyperbolic Discounting: Structural Estimation and Policy Evaluation*," *The Economic Journal*, 2008, *118* (531), 1418–1452.

**Royer, Heather, Mark Stehr, and Justin Sydnor**, "Incentives, Commitments, and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company," *American Economic Journal: Applied Economics*, 2015, *7* (3), 51–84.

**Sadoff, Sally, Anya Savikhin Samek, and Charles Sprenger**, "Dynamic Inconsistency in Food Choice: Experimental Evidence from a Food Desert," *Review of Economic Studies*, forthcoming.

**Schilbach, Frank**, "Alcohol and Self-Control: A Field Experiment in India," *American Economic Review*, 2019, *109* (4), 1290–1322.

**Schwartz, Janet, Daniel Mochon, Lauren Wyper, Josiase Maroba, Deepak Patel, and Dan Ariely**, "Healthier by Precommitment," *Psychological Science*, 2014, *25* (2), 538–546.

**Strotz, R. H.**, "Myopia and Inconsistency in Dynamic Utility Maximization," *The Review of Economic Studies*, 1955, *23* (3), 165–180.

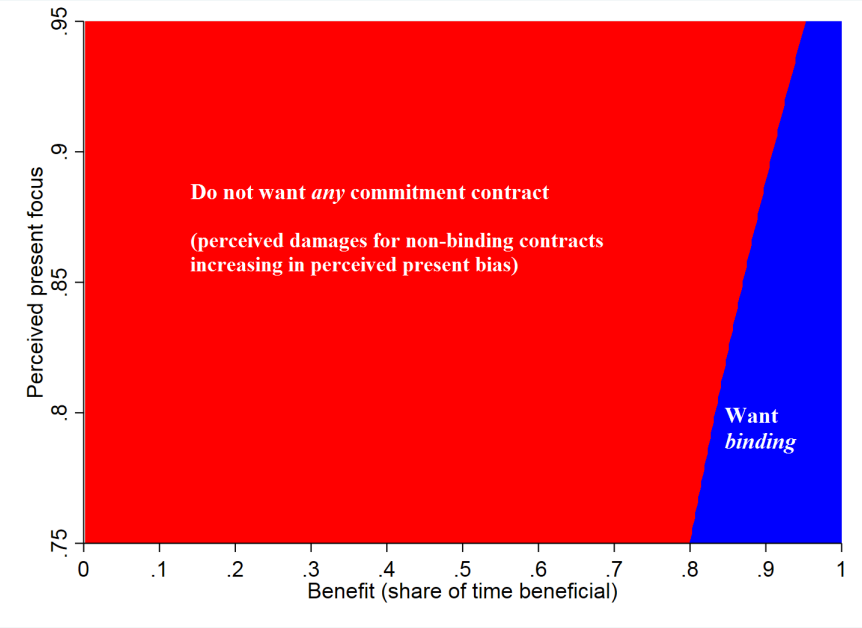
**Sun, Kai, Jing Song, Larry M. Manheim, Rowland W. Chang, Kent C. Kwoh, Pamela A. Semanik, Charles B. Eaton, and Dorothy D. Dunlop**, "Relationship of Meeting Physical Activity Guidelines with Quality Adjusted Life Years," *Seminars in Arthritis and Rheumatism*, 2014, *44* (3), 264–270.

**Toussaert, Séverine**, "Eliciting Temptation and Self-Control Through Menu Choices: A Lab Experiment," *Econometrica*, 2018, *86* (3), 859–889.

**Wei, Xue-Xin and Alan A. Stocker**, "A Bayesian Observer Model Constrained by Efficient Coding Can Explain Anti-Bayesian Percepts," *Nature Neuroscience*, 2015, *18*, 1509–1517.

**Woodford, Michael**, "Inattentive Valuation and Reference-Dependent Choice," *Working Paper*, 2012.

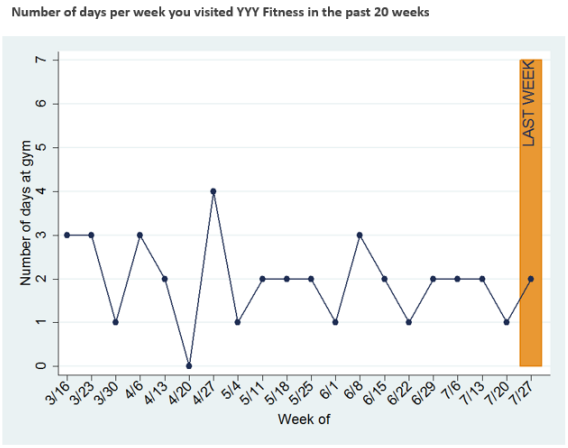
Figure 1: Commitment contract demand for uniform distribution of costs




Notes: This figure illustrates the commitment contract demand for the case in which costs are distributed uniformly on the unit interval ($c \sim U[0,1]$). Commitment contract demand is a function of delayed benefits $b$ and perceived self-control $\tilde{\beta}$. As can be seen, for $\tilde{\beta} \geq 0.75$ and $b \leq 0.8$, individuals do not want any commitment contract. In that case, the perceived damages from a commitment contract are increasing in the degree of perceived present focus, $1 - \tilde{\beta}$. When individuals do want a commitment contract, they prefer that it is binding, a sharp result that holds for uniform distributions but is not generally true.

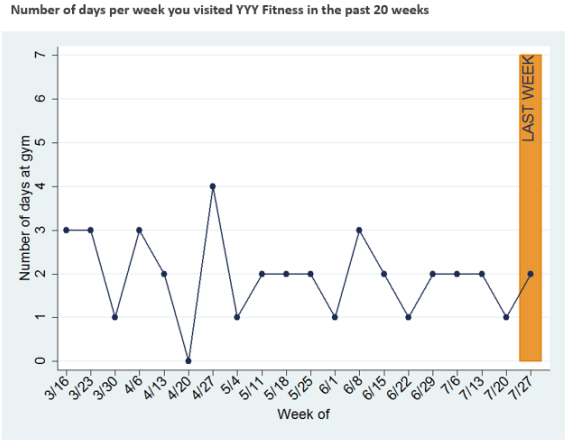
Figure 2: Information treatment

(a) Basic information treatment



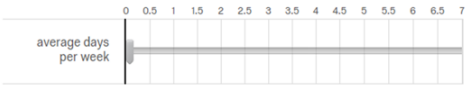

Notes: This graph is calculated based on the check-in records from the front desk. If you joined YYY Fitness within the last 20 weeks, the graph will show zero visits for weeks prior to when you joined.

(b) Enhanced info treatment - first screen




Looking at the graph, what do you estimate is the average number of days per week that you attended YYY Fitness over the past 20 weeks?

If you joined within the past 20 weeks, please just choose what you think the average was for you from the time you started at YYY.




(c) Enhanced info treatment - second screen

Next, we will ask you to estimate how many days you will visit YYY Fitness in the next 4 weeks. In forming your best estimate, here is some information from the 350 participants who took this survey last fall:

Participants estimated that they would visit YYY Fitness **4 more days** over 4 weeks than they actually did. On average, that means they overestimated their attendance by **1 visit per week.**

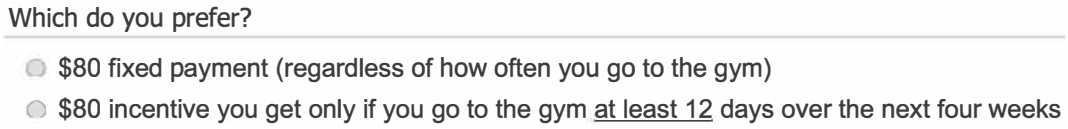
How useful do you think this information about previous participants will be as you think about how often you will attend?

| | Not at all useful | Not very useful | Somewhat useful | Useful | Very useful |
|---|---|---|---|---|---|
| Usefulness of information provided | ○ | ○ | ○ | ○ | ○ |

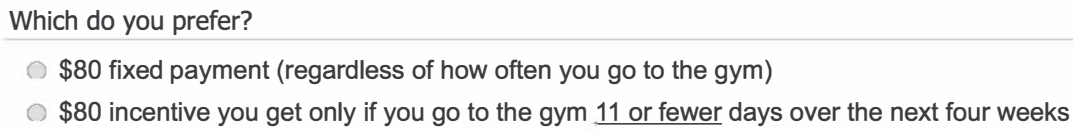
Notes: Sub-Figure a of this figure shows the basic information treatment of the history of past attendance shown to participants. Sub-figures b and c show the enhanced information treatment information which includes the history of past attendance and an engagement activity; Sub-Figure b displays the initial screen and Sub-Figure c shows the engagement activity.

Figure 3: Screen Shots of "More Visits" and "Fewer Visits" Commitment Choices

(a) "More visits" commitment contract

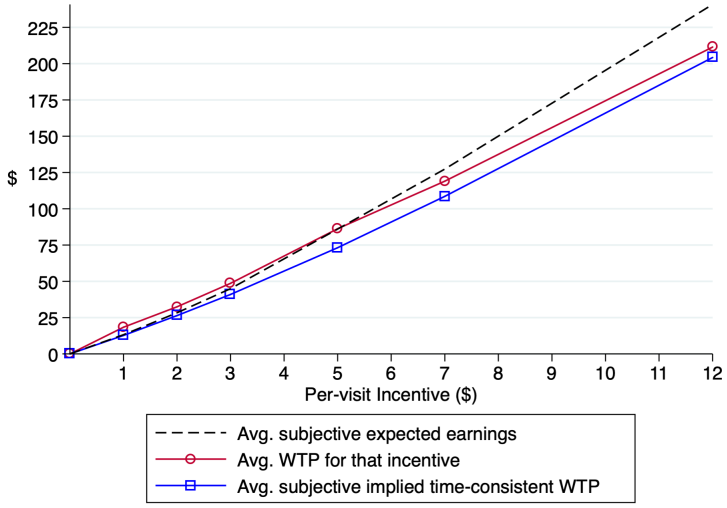



(b) "Fewer visits" commitment contract



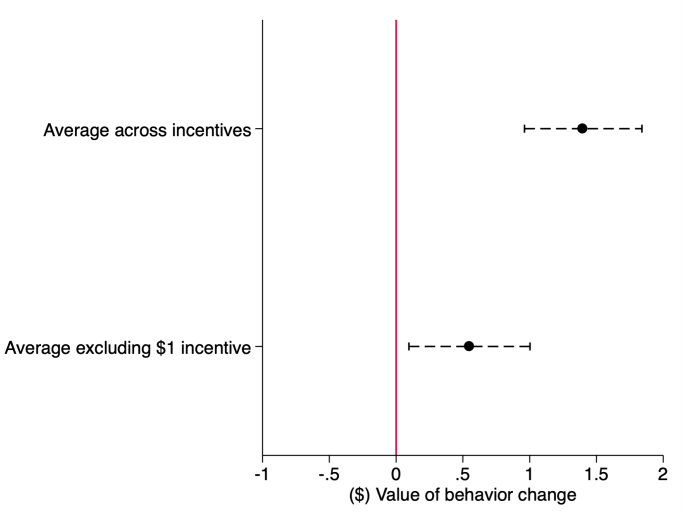

Notes: This figure provides a screenshot of the commitment contracts offered to participants. Sub-Figure a provides an example of a commitment contract to attend the gym more (i.e., the "more visits" contract). Sub-Figure b provides an example of a commitment contract to attend the gym less (i.e., the "fewer visits" contract).

Figure 4: Expected earnings and willingness to pay for piece-rate incentives



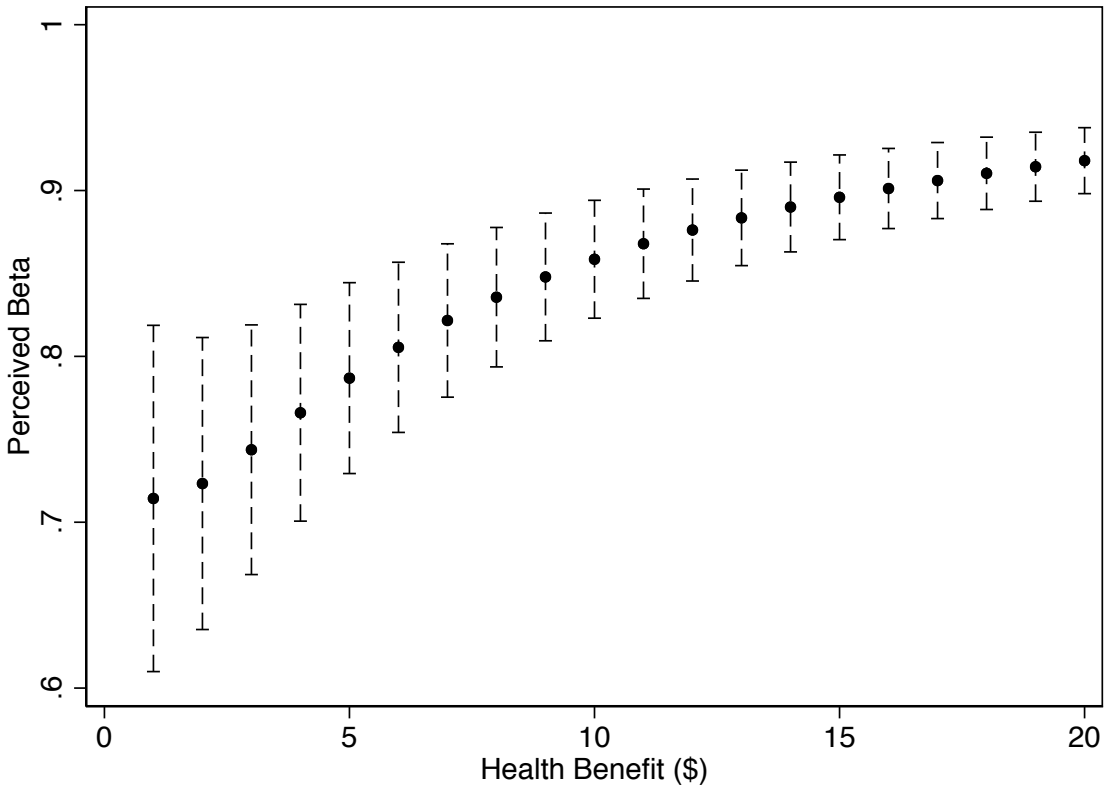

Notes: For each incentive, subjective expected earnings are the product of the piece rate (i.e., per day incentive) and subjects' beliefs about the number of days they would visit under that incentive. WTP is the average willingness to pay for each incentive, elicited with sliders as described in Section 3.1. The subjective implied time-consistent WTP is derived from the participants' beliefs about their visits under the different incentive levels using the approximation derived in Proposition 7 in Subsection 2.4.

Figure 5: Willingness to pay for behavior change



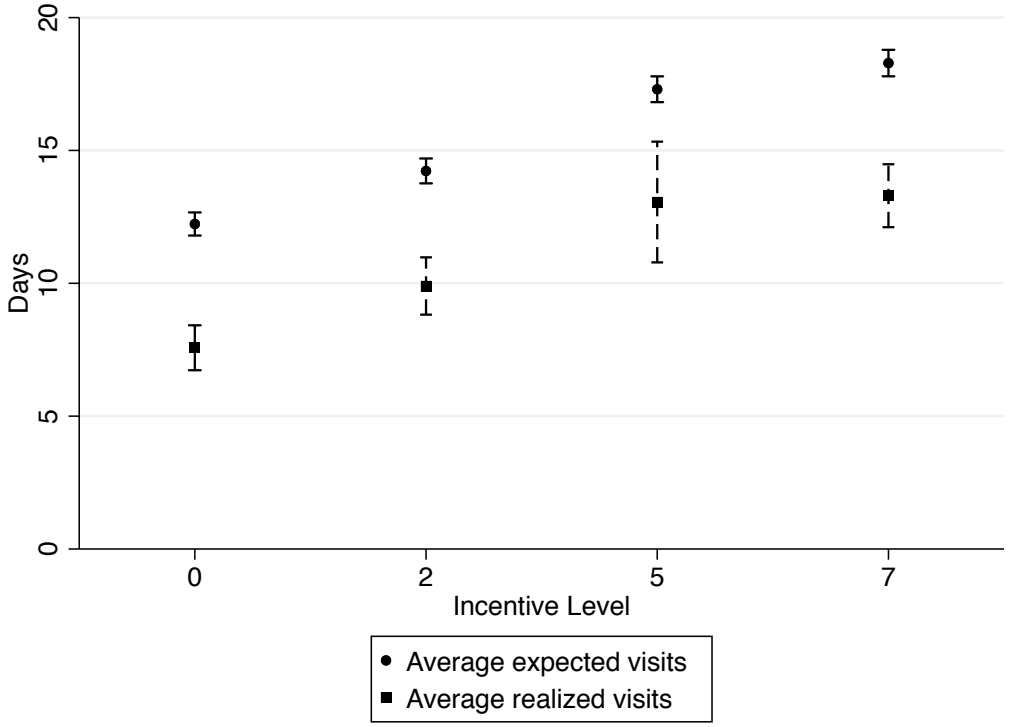

Notes: This figure shows the participants' average perceived value of behavior change as described in Section 5.1 (across all incentive levels in the top of the figure and across all incentive levels excluding $1 incentive in the bottom of the figure), with 95% confidence intervals obtained from heteroskedasticity-robust standard errors.

Figure 6: Estimates of $\tilde{\beta}$ for different values of delayed health benefits



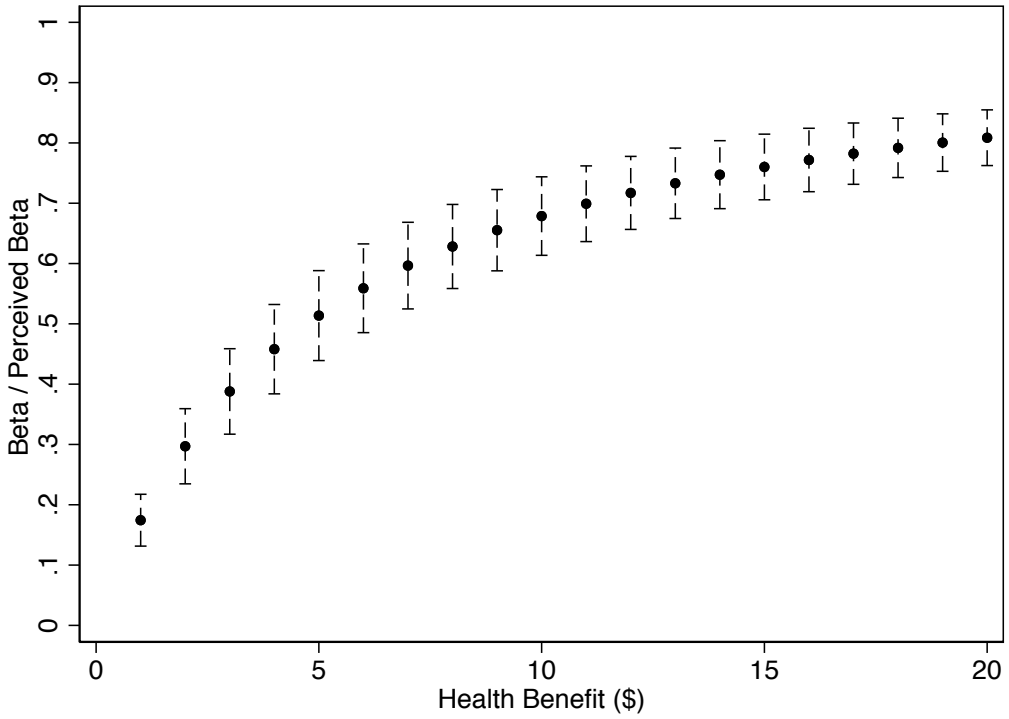

Notes: This figure shows estimated perceived short-run discount factor $\tilde{\beta}$ for a given value of delayed health benefits per attendance ranging from \$0 to \$20. Alongside the estimates, the corresponding 95% confidence intervals are displayed. Standard errors are clustered at the participant level.

Figure 7: Estimated vs. actual attendance




Notes: This figure shows the means and 95% confidence intervals for participants' expected number of days visiting the gym ("Best guess of days I would attend over the next four weeks") under no incentive ($0) and with piece-rate (i.e., per day) incentives of $2, $5, or $7. Expectations under no incentive were elicited prior to the description of how piece-rate incentives would be implemented. Statistics in the figure are based on data from control group participants excluding those with either low willingness-to-pay (12 participants) or those randomly assigned a high fixed payment (4 participants). Average realized visits are based on the subsets of participants who received each incentive level. Incentives were offered over the same four-week period for which expectations were elicited. Section 3.1.4 describes how different incentive levels were probabilistically targeted in each of the three survey waves. Because the incentive levels shown here were not all targeted in every wave, the sample sizes underlying the average realized visits statistics differ (N=211 ($0); N=147 ($2); N=34 ($5); N=169 ($7)).

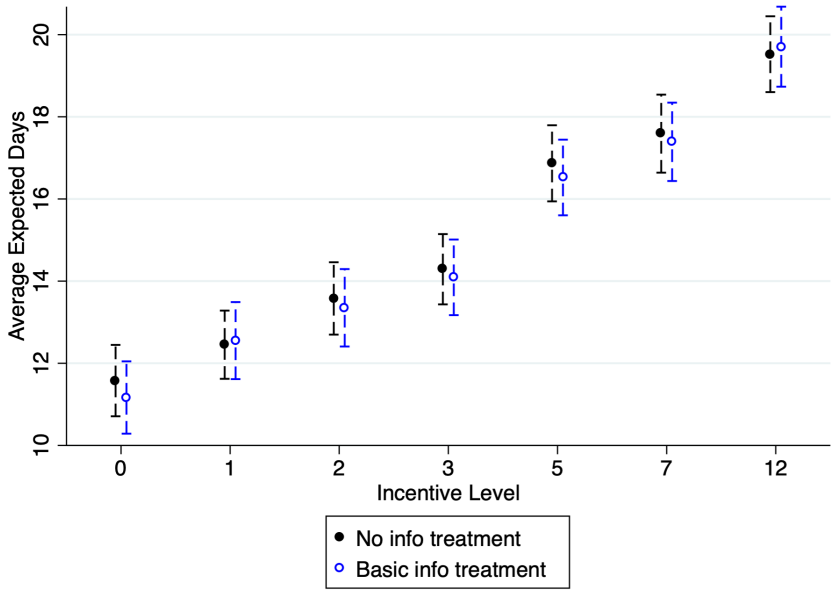
Figure 8: Estimates of $\beta/\tilde{\beta}$ for different values of delayed health benefits



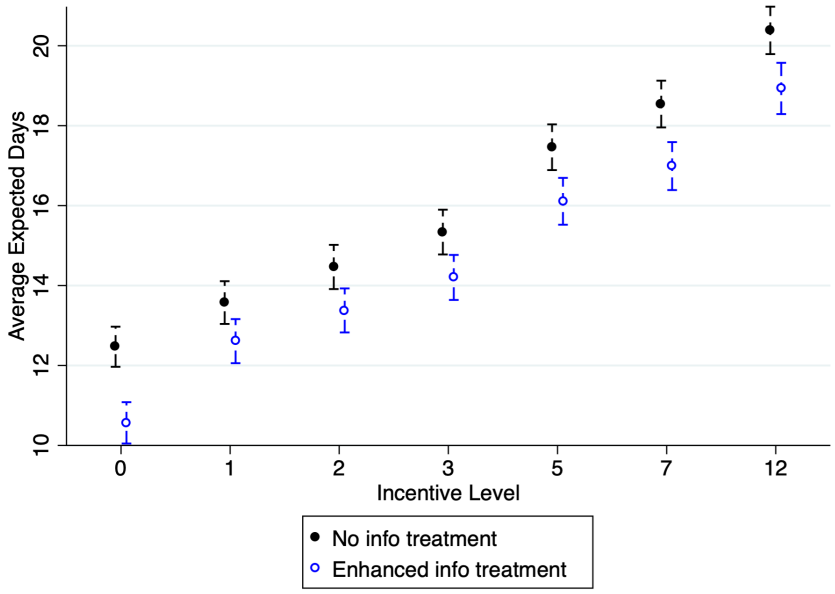

Notes: This figure shows the estimated ratio of actual short-run discount factor to perceived short-run discount factor, $\beta/\tilde{\beta}$, for a given value of delayed health benefits per attendance ranging from \$0 to \$20. Alongside the estimates, the corresponding 95% confidence intervals are displayed. Standard errors are clustered at the participant level.

Figure 9: Effect of information treatments on expected visits

(a) Impact of basic information treatment

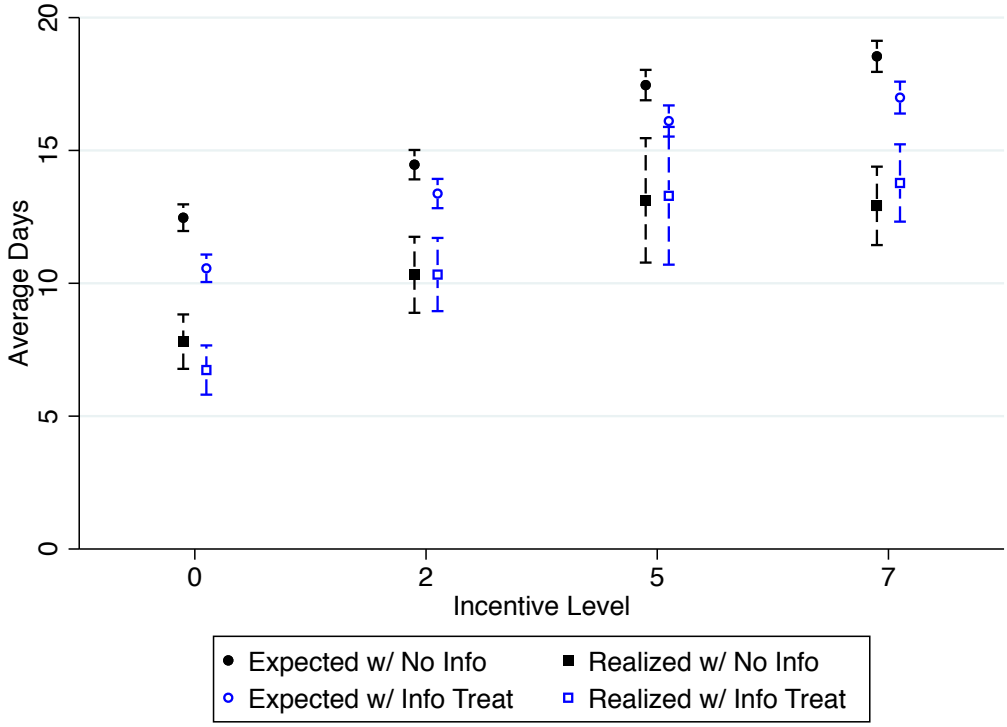



(b) Impact of enhanced information treatment



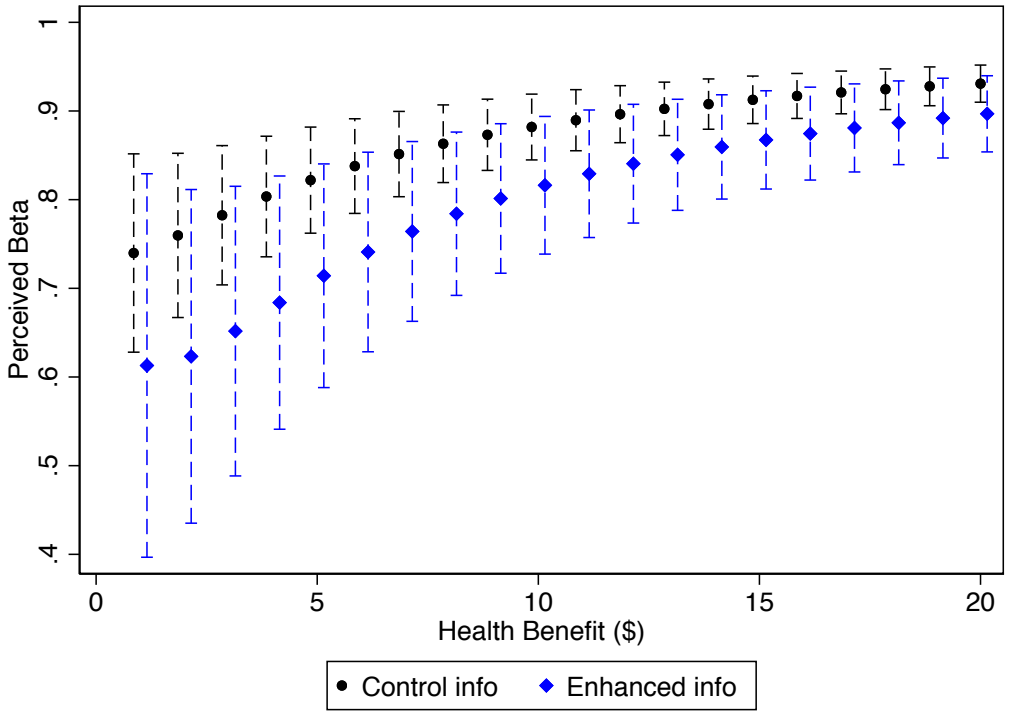

Notes: Sub-Figure a of this figure shows the mean and 95% confidence intervals for expected visits in the next four weeks for participants randomly selected to receive the basic information treatment (N=174) and participants randomly assigned to no information control group (N=174) from wave 1. Sub-Figure b of this figure shows the mean and 95% confidence intervals for expected visits in the next four weeks among participants randomly selected to receive the enhanced information treatment (N=453) and participants in the no information control group (N=458) from waves 2 and 3. Excluded from all calculations are those participants with low willingness to pay for incentives or those participants randomly assigned a high fixed payment (N=41). See Figure 10 for further details on these exclusion restrictions.

Figure 10: Estimated vs. actual attendance by enhanced information treatment




Notes: This figure shows average expected visits for the 4-week incentive period and average realized visits for participants in Waves 2 and 3 along the corresponding 95% confidence intervals. "No Info" refers to the control group whereas "Info Treat" refers to those who received the enhanced information treatment.

Figure 11: Estimates of $\tilde{\beta}$ by info treatment, for different values of delayed health benefits




Notes: This figure shows estimates of the perceived short-run discount factor $\tilde{\beta}$ for a given value of delayed health benefits per attendance for two groups: the control group and those receiving the enhanced information treatment. Alongside the estimates the corresponding 95% confidence intervals are displayed. As the enhanced information treatment was only part of Waves 2 and 3, the statistics in the figure are based on data from Wave 2 and 3 participants.

Table 1: Demographics and balance

| | Overall Mean | Difference in Means: Treatment - Control | | | |
|---|---|---|---|---|---|
| | Waves 1-3 | Wave 1 | P-value | Waves 2-3 | P-value |
| Female | 0.61 | –0.04 | 0.44 | –0.04 | 0.22 |
| Age[a] | 33.63 | –0.37 | 0.79 | –1.07 | 0.29 |
| Student, full-time | 0.56 | –0.09 | 0.07 | 0.01 | 0.87 |
| Working, full or part-time | 0.57 | 0.14 | 0.01 | 0.00 | 0.95 |
| Married | 0.27 | 0.08 | 0.09 | –0.01 | 0.83 |
| Advanced degree[b] | 0.46 | 0.06 | 0.28 | –0.01 | 0.79 |
| Household Income[a] | 55,434 | 2,842 | 0.56 | –4,798 | 0.17 |
| Visits in the past 4 weeks | | | | | |
| Days visited, recorded | 6.92 | 0.25 | 0.70 | –0.21 | 0.58 |
| Visits in the past 100 days | | | | | |
| Days visited, recorded | 22.13 | –0.21 | 0.91 | –0.23 | 0.84 |
| Days visited, self-recollection | 30.51 | –1.80 | 0.46 | –1.33 | 0.33 |
| Days that *I should have gone, but didn't* | 30.52 | 0.00 | 1.00 | –0.92 | 0.56 |
| Indicator for inattention during survey | 0.09 | –0.02 | 0.39 | 0.03 | 0.14 |
| N | 1,292 | 169 Control<br>181 Treated | | 471 Control<br>471 Treated | |

a. Imputed from categorical ranges.

b. A graduate degree beyond a B.A. or B.S.

Notes: This table shows the means of demographic variables reported in the survey, as well as differences in treatment and control group means. In wave 1 of the experiment, the treatment group received the "basic" information treatment. In waves 2 and 3, treated participants received the "enhanced" information treatment. The table also summarizes data on past visit frequencies to the gym. "Recorded" visits are obtained from the fitness center's log-in records, while "self-recollection" refers to participants' survey-reported estimates of their own past visits.

Table 2: Take up of commitment contracts

| Threshold | Chose "More" Contract (1) | Chose "Fewer" Contract (2) | Chose "More" Given Chose "Fewer" (3) | Chose "Fewer" Given Chose "More" (4) | Diff (3)-(1) | Diff (4)-(2) |
|---|---|---|---|---|---|---|
| 8 visits | 0.64 | 0.35 | 0.88 | 0.49 | 0.24*** | 0.14*** |
| 12 visits | 0.52 | 0.33 | 0.72 | 0.46 | 0.20*** | 0.13*** |
| 16 visits | 0.36 | 0.31 | 0.56 | 0.48 | 0.20*** | 0.17*** |

Notes: Column (1) reports take up rates of commitment contracts to visit the gym at least 8, 12, or 16 days over the next four weeks (i.e., take up of the "more" contract). Column (2) shows take up rates of commitment contracts to visit the gym less than 8, 12, or 16 days over the same period (i.e., take up of the "fewer" contract). Columns (3) and (4) shows the take up rates of each type of commitment contract conditional on having chosen the other type of commitment contract. Columns (5) and (6) display the difference in the takeup rates of column (3) versus column (1) in column (5) and the difference in the take up rates of column (4) versus column (2) in column (6). All take up rates are computed for control group participants exclusively. Over three survey waves, all participants faced the choice of both commitment contracts at the 12 visit threshold (N=640) while the 8 visit and 16 visit commitment contracts were only shown in the first two waves (N=441). *** denotes those differences that are statistically significantly different from 0 at the 1% level.

Table 3: Correlation between perceived success in contracts and expected attendance

| | Subj. expected attendance w/ out incentives | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Subj. prob succeed in "more" contract | 0.11*** | | 0.12*** |
| | (0.02) | | (0.01) |
| Subj. prob succeed in "fewer" contract | | –0.03*** | –0.05*** |
| | | (0.01) | (0.01) |
| N | 199 | 199 | 199 |
| "More" − "Fewer" | | | 0.16*** |
| | | | (0.02) |

Notes: This table displays the association between subjective beliefs about commitment contract success and expected attendance with no incentives. Each column presents coefficient estimates and heteroskedasticity-robust standard errors in parentheses from a separate OLS regression. Prob succeed in "more" contract is the ex-ante self-assessed probability of attending the gym 12 or more days during the 4-week incentive period. Prob success in "fewer" contract is ex-ante self-assessed probability of attending the gym fewer than 12 days during the 4-week incentive period. The sample consists exclusively of control group participants in wave 3, the only wave in which we elicited the probabilities of contract success. The "More" - "Fewer"' row denotes a test of the difference between the coefficient on the probability of success under the "more"' contract versus the coefficient on the probability of success under the "fewer" contract. *** denotes statistics that are statistically significantly different from 0 at the 1% level.

Table 4: Correlation between perceived success in contracts and take up of contracts

| | Prob succeed in "more" attendance contract | | | Prob succeed in "less" attendance contract | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Commit to "more" | 12.09*** | | 13.41*** | –10.84** | | –16.04*** |
| | (3.03) | | (2.89) | (4.87) | | (5.10) |
| Commit to "fewer" | | –2.47 | –6.01* | | 19.38*** | 23.61*** |
| | | (3.42) | (3.21) | | (4.48) | (5.07) |
| N | 199 | 199 | 199 | 199 | 199 | 199 |
| "More" - "Fewer" | | | 19.42*** | | | -39.65*** |
| | | | (4.17) | | | (8.82) |

Notes: This table displays the association between commitment contract take up and subjective beliefs about success in the commitment contract. Each column presents coefficient estimates and heteroskedasticity-robust standard errors in parentheses from separate OLS regression. Columns (1)-(3) display associations with participants' expectations of following through on the commitment contract requiring attendance at the gym 12 or more days during the 4-week incentive period. Columns (4)-(6) display associations with participants' expectations of following through on the commitment contract requiring attendance at the gym fewer than 12 days during the 4-week incentive period. The sample consists exclusively of control group participants in wave 3, the only wave we elicited the probabilities of contract success. *,**,*** denote statistics that are statistically significantly different from 0 at the 10%, 5%, and 1% level respectively.

Table 5: Correlation between WTP for behavior change and take up of "more" contracts

| | Take-up of more-visits contract (mean = .51) | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| WTP for behavior change (z-score) | 0.006 | –0.002 | | |
| | (0.023) | (0.023) | | |
| Expected-visits elasticity (z-score) | | 0.056*** | | |
| | | (0.020) | | |
| WTP for behavior change excl. $1 incentive (z-score) | | | –0.024 | –0.031 |
| | | | (0.022) | (0.022) |
| Expected-visits elasticity excl. $1 incentive (z-score) | | | | 0.027 |
| | | | | (0.017) |
| N | 1,522 | 1,522 | 1,522 | 1,522 |

Notes: This table displays the association between estimated WTP for behavior change (expressed as a z-score) and take up of "more" commitment contracts. Each column presents coefficient estimates and standard errors clustered at the participant level in parentheses from separate OLS regressions. The sample consists exclusively of control group subjects (N=640). In columns (1) and (2), WTP is calculated based on all incentive levels whereas in columns (3) and (4), WTP is calculated excluding the $1 incentive. In columns (2) and (4), the average elasticity of each individual's visit expectations with incentive size (expressed as a z-score) is also included. All regressions include wave fixed effects and commitment contract threshold fixed effects (i.e., 8, 12, 16 visit thresholds). *** denotes a statistic that is statistically significantly different from 0 at the 1% level.

Table 6: Effect of information provision on willingness to pay for behavior change

| | All incentives | Excluding $1 incentive |
|---|---|---|
| | (1) | (2) |
| Basic information treatment | 0.24 | 0.19 |
| | (0.52) | (0.55) |
| Enhanced information treatment | 1.15** | 1.33*** |
| | (0.48) | (0.51) |
| N | 1,292 | 1,292 |

Notes: This table displays the association between the information treatments and estimated WTP for behavior change (expressed as a z-score). Each column presents coefficient estimates and heteroskedasticity-robust standard errors in parentheses from separate OLS regressions. In column (1), WTP is calculated based on all incentive levels whereas in column (2), WTP is calculated excluding the $1 incentive. All regressions include wave fixed effects. **, *** denote statistics that are statistically significantly different from 0 at the 5% and 1% level respectively.

Table 7: Effect of information provision on take up of "more" contracts

| | 8+ visits (1) | 12+ visits (2) | 16+ visits (3) | Pooled (4) |
|---|---|---|---|---|
| Basic info treatment | 0.048 | –0.067 | –0.024 | –0.014 |
| | (0.052) | (0.053) | (0.047) | (0.040) |
| Enhanced info treatment | –0.056 | –0.054* | –0.096** | –0.066** |
| | (0.042) | (0.033) | (0.042) | (0.030) |
| N | 878 | 1,292 | 878 | 3,048 |
| Take-up Mean | 0.64 | 0.52 | 0.36 | 0.48 |

Notes: This table displays the association between the information treatments and take up of the "more" commitment contracts. Each column presents coefficient estimates and heteroskedasticity-robust standard errors in parentheses from separate OLS regressions. Columns (1)-(3) consider "more" commitment contracts for 8 or more days of attendance, 12 or more days of attendance, and 16 or more days of attendance. Column (4) pools the 3 contracts together. All participants are included in the column (2) regression because the commitment contract for making 12+ visits was offered in all three survey waves. In columns (1) and (3), the sample size is smaller because these contracts were only offered in waves 1 and 2. The sample size in column (4), in which all take up decisions are pooled, equals the sum of the previous three columns. In the column (4) specification, fixed effects for each contract are included and standard errors are clustered by participant. Across all columns, wave fixed effects are included.

# Appendices (not for publication)

## A Proofs of Propositions in the Body of the Paper

### Proof of Proposition 1

We provide a proof of the proposition in the paper, as well as the generalization to $\max_{p\in[0,\bar{p}]} \Delta V$.

*Proof.* The perceived gains from a commitment contract are

$$\begin{aligned}\Delta V/\beta &= -p + \int_{c\leq\tilde{\beta}(p+b)} (p+b-c)dF - \int_{c\leq\tilde{\beta}b} (b-c)dF \\ &= -p(1-F(\tilde{\beta}(b+p))) + \int_{c=\tilde{\beta}b}^{c=\tilde{\beta}(b+p)} (b-c)dF \end{aligned} \tag{3}$$

Now $-p(1-F(\tilde{\beta}(p+b))) \to -p$ as $\tilde{\beta}\to 0$ since $F(\tilde{\beta}(p+b))\to 0$ as $\tilde{\beta}\to 0$. For this same reason, $\int_{c=\tilde{\beta}b}^{c=\tilde{\beta}(b+p)}(b-c)\to 0$ as $\tilde{\beta}\to 0$. Thus, $\Delta V/\beta \to -p$ as $\tilde{\beta}\to 0$, which establishes that there exists $\underline{\beta}$ such that $\Delta V<0$ for each $p$. Because $\underline{\beta}$ is continuous in $p$, there must also exist a $\underline{\beta}>0$ such that $\max_{p\in[0,\bar{p}]}\Delta V<0$ if $\tilde{\beta}<\underline{\beta}$.

Because $\Delta V$ is continuous in $\tilde{\beta}$, and because $\Delta V<0$ for $\tilde{\beta}=1$, we also have that there exists a $\bar{\beta}$ such that $\Delta V<0$ if $\tilde{\beta}>\bar{\beta}$. Again, the result generalizes immediately to $\max_{p\in[0,\bar{p}]}\Delta V$ as well.

Next, suppose that $c\in\{\underline{c},\bar{c}\}$, where $\bar{c}>b$ and $\underline{c}<b$. Let $\mu$ denote the probability of $c=\bar{c}$. If $\tilde{\beta}(b+p)<\underline{c}$ then clearly the commitment contract is perceived not worthwhile, since it only increases penalties incurred. If $\tilde{\beta}b>\underline{c}$ then the commitment contract is also perceived not worthwhile, since the agent already believes that he will choose $a=1$ when $c=\underline{c}$.

The commitment contract has a chance of being worthwhile when $\tilde{\beta}b<\underline{c}<\tilde{\beta}(b+p)$. In this case, if $\tilde{\beta}(b+p)<\bar{c}$ then the agent incurs the cost $p$ with probability $\mu$. If $\tilde{\beta}(b+p)>\bar{c}$ then the agent incurs a utility loss of $\bar{c}-b$ with probability $\mu$. Either way, $\Delta V>0$ for small enough $\mu$and $\Delta V<0$ for large enough $\mu$.

Since there exist bounds $\underline{\beta}(p)$ and $\bar{\beta}(p)$ for each $p\in[0,\bar{p}]$, the union of the intervals $I(p)=(\underline{\beta}(p),\bar{\beta}(p))$ over $p\in[0,\bar{p}]$ produces an interval $(\underline{\beta},\bar{\beta})$ such that $\max_p\Delta V>0$ iff $\tilde{\beta}\in(\underline{\beta},\bar{\beta})$. □

### Proof of Proposition 2

*Proof.* From 3, we have

$$\begin{aligned}\frac{d}{d\tilde{\beta}}\Delta V/\beta &= p(b+p)f(\tilde{\beta}(p+b)) + (b+p)(b-\tilde{\beta}(b+p))f(\tilde{\beta}(b+p)) - b(b-\tilde{\beta}b)f(\tilde{\beta}b) \\ &= (1-\tilde{\beta})(b+p)^2 f(\tilde{\beta}(b+p)) - (1-\tilde{\beta})b^2 f(\tilde{\beta}b)\end{aligned} \tag{4}$$

The expression (4) is positive if $\frac{f(\tilde{\beta}(p+b))}{f(\tilde{\beta}b)} \geq \frac{b^2}{(p+b)^2}$.

Since the condition implies $Pr(c > b)$ when $\tilde{\beta} = 1$, Proposition 1 implies that $\tilde{\beta} = 1$ agents have $\Delta V < 0$. The first part of the lemma then implies that $\Delta V < 0$ for all $\tilde{\beta}$. □

## Proof of Proposition 3

We begin with a Lemma:

**Lemma 1.** *Under the assumptions of the proposition, no individuals will want commitment contracts that force $a = 1$.*

*Proof.* To shorten equations, set $\gamma = (1 - \tilde{\beta})b$. The perceived expected gains from a binding commitment contract are given by

$$\Delta V/\beta = \int_{c \geq \tilde{\beta}b} (b - c) f(c) dc.$$

The goal is thus to show that $\int_{c \geq \tilde{\beta}b} (b - c) f(c) dc < 0$ under the assumptions of the proposition

**Case 1:** Suppose that $f$ is increasing on $[b, b + \gamma]$. Then by the single-peak assumption, $f$ is increasing on $[b - \gamma, b + \gamma]$ . Then the value of the fully binding contract is

$$\begin{aligned}
\int_{c=\beta b}^{\infty} (b - c) f(c) dc &\leq \int_{c=\beta b}^{c=b+(1-\beta)b} (b - c) f(c) dc \\
&= \int_{c=\beta b}^{b} (b - c) f(c) dc + \int_{c=b}^{b+(1-\beta)b} (b - c) f(c) dc \\
&\leq \int_{c=\beta b}^{b} (b - c) f(c) dc + \int_{c=b}^{b+(1-\beta)b} (b - c) f(2b - c) dc \\
&= \int_{c=\beta b}^{b} (b - c) f(c) dc - \int_{c=\beta b}^{b} (b - c) f(c) dc \\
&= 0
\end{aligned}$$

where to get to the second-to-last line we perform a change-of-variable on the second integral via the function $\varphi(x) = 2b - x$.

Case 2: Suppose now that $f$ is decreasing on $[b - \gamma, b + \gamma]$. Define $\mu := F(b) - F(b - \gamma)$, and recall that the second assumption requires that $1 - F(b) \geq \mu$. On the other hand, $\mu = \int_{x=b-\gamma}^{b} f(x) \geq \int_{x=b-\gamma}^{b} f(b) = \gamma f(b)$.

Now

$$\begin{aligned}
\int_{c=\beta b}^{b}(b-c)f(c)dc &= \int_{c=\beta b}^{b}(b-c)f(b)dc + \int_{c=\beta b}^{b}(b-c)(f(c)-f(b))dc \\
&= \frac{\gamma^2}{2}f(b) + \int_{c=\beta b}^{b}(b-c)(f(c)-f(b))dc \\
&\leq \frac{\gamma^2}{2}f(b) + \int_{c=\beta b}^{b}\gamma(f(c)-f(b))dc \\
&= \frac{\gamma^2}{2}f(b) + (\mu-\gamma f(b))\gamma \\
&= \gamma\mu - \frac{\gamma^2}{2}f(b) \qquad (5)
\end{aligned}$$

Intuitively, all of the mass that is in excess of a uniform distribution on $[b-\gamma, b]$ with density $f(c) = f(b)$ is concentrated on the point adding the most to the mean: $c = \beta b$.

Next,

$$\begin{aligned}
\int_{c\geq b}(b-c)f(c)dc &= \int_{c=b}^{b+\gamma}(b-c)f(c)dc + \int_{c\geq b+\gamma}(b-c)f(c)dc \\
&\leq \int_{c=b}^{b+\gamma}(b-c)f(c)dc - \int_{c\geq b+\gamma}\gamma f(c)dc \\
&= \int_{c=b}^{b+\gamma}(b-c)f(c)dc - \gamma(1-F(b+\gamma)) \\
&= \int_{c=b}^{b+\gamma}(b-c)f(c)dc - \gamma\left[(1-F(b)-(F(b+\gamma))-F(b))\right] \\
&\leq \int_{c=b}^{b+\gamma}(b-c)f(c)dc - \gamma\left(\mu - \int_{c=b}^{b+\gamma}f(c)dc\right) \\
&= \int_{c=b}^{b+\gamma}(b+\gamma-c)f(c)dc - \gamma\mu \\
&\leq \int_{c=b}^{b+\gamma}(b+\gamma-c)f(b)dc - \gamma\mu \\
&= \frac{\gamma^2}{2}f(b) - \gamma\mu \qquad (6)
\end{aligned}$$

Intuitively, the quantity $-\int_{c=b}^{b+\gamma}(b-c)f(c)dc$ is minimized when $1-F(b) = \mu$ and as much of the mass $\mu$ as possible belongs to $[b, b+\gamma]$. So to minimize $-\int_{c=b}^{b+\gamma}(b-c)f(c)dc$, we need to maximize the mass of $F$ on $[b, b+\gamma]$, and the way to do that is to let it be uniform on $[b, b+\gamma]$,with density $f(c) := f(b)$. In this case, the rest lies on points $c \geq b+\gamma$ and has to integrate to at least $(\mu-\gamma f(b))\gamma$.

Putting (5) and (6) together shows that $\int_{c\geq\tilde{\beta}b}(b-c)f(c)dc \leq 0$.

Case 3: Suppose that the mode of $f$ lies in $[b-\gamma, b]$ and that $\mu \geq \gamma f(b)$. Equation (6) holds

because as in case 2, $f$ is decreasing on $[b, b+\gamma]$.

Next, we consider the maximum of the function $A$ given by $A(f) := \int_{c=\tilde{\beta}b}^{b}(b-c)f(c)dc$, over all $f$ that have a mode on $[b-\gamma, b]$. Suppose for a given $f$ that the mode is at $c^* > \tilde{\beta}b$, and that $\int_{c=\tilde{\beta}b}^{b}(f(c^*) - f(c))dc > 0$. Then consider $\tilde{f}$ given by $\tilde{f}(c) = f(c)$ for $c \geq c^*$, and $\tilde{f}(c) = \frac{\int(f(c^*)-f(\tilde{\beta}b))dc}{c^*-\tilde{\beta}b}$ for $c < c^*$. Since $f$ is increasing on $[\tilde{\beta}b, c^*]$, $f$ stochastically dominates $\tilde{f}$. Consequently, since $b-c$ is positive and decreasing in $c$, $A(\tilde{f}) > A(f)$. This establishes that the $f$ that maximizes $A$ must be decreasing almost everywhere on $[\tilde{\beta}b, b]$ (except for a set of zero Lebesgue measure). We can then proceed as in Case 2 to establish that $\int_{c=\tilde{\beta}b}^{b}(b-c)f(c)dc \leq \gamma\mu - \frac{\gamma^2}{2}f(b)$.

Case 4: Suppose that the mode lies in $[b-\gamma, b]$ and that $\mu < \gamma f(b)$. As in case 3, we have shown that $A$ is maximized when $f$ is decreasing almost everywhere. But sine $\mu < \gamma f(b)$, this means that $f$ must be uniform almost everywhere, with density $f(c) = \mu/\gamma$. Thus in this case

$$\int_{c=\tilde{\beta}b}^{b}(b-c)f(c)dc \leq \gamma\mu/2. \tag{7}$$

Now the highest value of $\int_{c\geq b}(b-c)f(c)dc$ is obtained by a density function $f$ that puts as much mass toward $b$ as possible, and minimizes the value of $f(b)$. That is, $f(c) = (b/c)^2 f(b)$ for $c \geq b$, with $\bar{c} = b+\gamma$, and $f(b)$ large enough to satisfy the constraint $\int_{c\geq b} f(c) = \mu/\tilde{\beta}$. The constraint on $f(b)$ is

$$\begin{aligned}
\mu/\tilde{\beta} &\leq \int_{x=b}^{x=b+\gamma} \frac{b^2}{x^2} f(b)dx \\
&= -\frac{b^2}{x} f(b)\Big|_b^{b+\gamma} \\
&= b - \frac{b^2}{b+\gamma} \\
&= bf(b)\left[\frac{\gamma}{b+\gamma}\right]
\end{aligned}$$

Now for $k = 1 - \tilde{\beta}$,

$$
\begin{aligned}
-\int_{x=b}^{x=b+\gamma} (b-x) f(c) dc &= \int_{x=b}^{x=b+\gamma} (x-b) \frac{b^2}{x^2} f(b) dx \\
&= b^2 f(b) \int \left( \frac{1}{x} - \frac{b}{x^2} \right) dx \\
&= b^2 f(b) \left[ \ln(x) + \frac{b}{x} \right]_{x=b}^{b+\gamma} \\
&= b^2 f(b) \left[ \ln(b+\gamma) + \frac{b}{b+\gamma} - \ln(b) - 1 \right] \\
&= b^2 f(b) \left[ \ln(1+k) - \frac{k}{1+k} \right] \\
&\geq b^2 f(b) \left[ k - \frac{k^2}{2} - \frac{k}{1+k} \right] \qquad (8) \\
&= b^2 f(b) \left[ \frac{k + k^2 - k}{1+k} - \frac{k^2}{2} \right] \\
&= b^2 f(b) \left[ \frac{k^2}{1+k} - \frac{k^2}{2} \right] \\
&= f(b) \left[ \frac{\gamma^2}{1+k} - \frac{\gamma^2}{2} \right] \\
&= f(b) \left[ \frac{\gamma^2 (1-k)}{2(1+k)} \right] \\
&= \frac{\tilde{\beta} \gamma^2}{2(1+k)} f(b) \\
&= \frac{1}{2} \tilde{\beta} \gamma \frac{\gamma}{b+\gamma} b f(b) \\
&\geq \frac{\tilde{\beta} \gamma}{2} \frac{\mu}{\tilde{\beta}} \\
&= \gamma \mu / 2 \qquad (9)
\end{aligned}
$$

To obtain (8), we need to show that $\log(1+x) \geq x - x^2/2$ for $x \geq 0$. To that end, note that equality holds when $x = 0$. The derivatives with respect to $x$ are $\frac{1}{1+x}$ and $1 - x$, respectively, so it is enough to show that $\frac{1}{1+x} \geq 1 - x$. This holds iff $1 \geq 1 - x^2$, which follows because $x^2 \geq 0$.

The combination of (7) and (9) implies that $\int_{c \geq \tilde{\beta} b} (b-c) f(c) dc \leq 0$.

Case 5. Suppose that the mode is in $[b, b+\gamma]$. Since this implies that $f$ is increasing on $[b-\gamma, b]$, the highest possible value of $\int_{c=\tilde{\beta}b}^{b} (b-c) f(c) dc$, given a that $F(b) - F(\tilde{\beta} b) = \mu$, is obtained when $f$ is almost everywhere uniform, with density $f(c) = \mu/\gamma$. As in Case 4, this implies that $\int_{c=\tilde{\beta}b}^{b} (b-c) f(c) dc \leq \gamma\mu/2$. And as in Case 4, the highest value of $\int_{c \geq b} (b-c) f(c) dc$ is obtained by a density function $f$ that puts as much mass toward $b$ as possible, and minimizes the value of $f(b)$. That is, $f(c) = (b/c)^2 f(b)$ for $c \geq b$, with $\bar{c} = b + \gamma$, and $f(b)$ large enough to satisfy the constraint $\int_{c \geq b} f(c) = \mu/\tilde{\beta}$. Proceeding as in that case establishes the result. □

With the Lemma in hand, we are ready to prove the Proposition 3.

*Proof.* Case 1: Suppose that $\bar{c} = \infty$. Then Proposition 2 implies that for any value of $p$, the value of the commitment contract is increasing in $\tilde{\beta}$. But since $\Delta V < 0$ for $\tilde{\beta} = 1$ agents, it must be that $\Delta V < 0$ for all $\tilde{\beta}$.

Case 2: Suppose that $\bar{c} < \infty$. Set $\beta^{\dagger} = \min(1, \bar{c}/(b+p))$. If $\beta^{\dagger} < \tilde{\beta}$ then this commitment contract generates the same utility as a fully binding commitment contract. The previous lemma implies that it is undesirable. If $\beta^{\dagger} > \tilde{\beta}$ then Proposition 2 implies that an individual with perceived present focus $\beta^{\dagger}$ expects higher gains from this contract than an individual with perceived present focus $\tilde{\beta}$. However, to an individual with perceived present focus $\beta^{\dagger}$, this is equivalent to a fully binding commitment contract. It is thus enough to show that a fully binding commitment contract is undesirable to an agent with perceived present focus $\beta^{\dagger}$.

But a binding commitment contract is less attractive to this individual than to an individual with perceived present focus $\tilde{\beta}$. But Lemma 1 implies that a fully binding commitment contract is undesirable to an agent with perceived present focus $\tilde{\beta}$. Consequently, it is undesirable to an agent with perceived present focus $\beta^{\dagger}$.

Moreover, if the choice of commitment contracts for $a = 1$ is primarily driven by noise rather than a real demand for commitment, then there will be a positive correlation between demand for $CC(p,1)$ and $CC(p,0)$. □

## Proof of Proposition 4

*Proof.* An agent will choose $CC(p,1)$ if

$$\left[ pF(\hat{\beta}_i b) + \int_{\hat{\beta}_i b}^{\hat{\beta}_i(p+b)} (p+b-c)dF \right] \varepsilon_{ij} \geq p - \eta_i/\beta_i \tag{10}$$

and will choose $CC(p,0)$ if

$$\left[ p[1 - F(\hat{\beta}_i b)] - \int_{\hat{\beta}_i(b-p)}^{\hat{\beta}_i b} (b-p-c)dF \right] \varepsilon_{ij} \geq p - \eta_i/\beta \tag{11}$$

Since $pF(\hat{\beta}_i b) + \int_{\hat{\beta}_i b}^{\hat{\beta}_i(p+b)}(p+b-c)dF > 0$, condition (10) will be satisfied if either $\eta_i > \beta_i p$, or if $\mu > 0$ and the draw $\varepsilon_{ij}$ is sufficiently high. Similarly, (11) will hold if either $\eta_i > \beta_i p$ or if $p[1 - F(\hat{\beta}_i b)] - \int_{\hat{\beta}_i(b-p)}^{\hat{\beta}_i b}(b-p-c)dF > 0$ and the draw of $\varepsilon_{ij}$ is sufficiently high. If $\eta_i > \beta_i p$ then the individual will choose both $CC(p,1)$ and $CC(p,0)$ with positive probability (with the former probability being 1). If $p[1 - F(\hat{\beta}_i b)] - \int_{\hat{\beta}_i(b-p)}^{\hat{\beta}_i b}(b-p-c)dF > 0$ then there is again a positive probability that the $\varepsilon$ draws for both the $CC(p,1)$ and $CC(p,0)$ are high enough such that the agent would want to choose both. □

## Proof of Proposition 5

*Proof.* Let $\nu_i$ be an indicator for whether individual $i$ is a noisy thinker, so that $\nu_i = 0$ iff $\varepsilon_{ij} \equiv 1$. When $\tilde{\beta}_i = 1$, the probability of choosing $CC(p,1)$ and $CC(p,0)$ is increasing in both $\nu_i$ and $\eta_i$. Consequently, the result must hold when $E[\tilde{\beta}_i] = 1$. By continuity, it holds for $E[\tilde{\beta}_i]$ sufficiently close to 1. □

## Proof of Proposition 6

*Proof.* Since the probability of choosing a commitment contract is increasing in $\Delta V$, the result follows if we show that $\Delta V$ is increasing in $\tilde{\beta}_i$ and in $b$. By Proposition (2), $\Delta V$ is increasing in $\tilde{\beta}_i$ . □

## Proof of Proposition 7

*Proof.* We begin by characterizing $\frac{d}{dp}V(0,0,p)$. This is

$$\begin{aligned}\frac{d}{dp}\int_{c\leq\tilde{\beta}(p+b)}(p+b-c)f(c)dc &= F(\tilde{\beta}(b+p)) + (1-\tilde{\beta})(p+b)\tilde{\beta}f(\tilde{\beta}(p+b)) \\ &= \alpha(p) + (1-\tilde{\beta})(p+b)\alpha'(p)\end{aligned}$$

Consequently, if the terms $\left\{(\Delta p)^n \frac{d^m}{dp^m}\alpha(0,0,p)_{p=p_1}\right\}_{\{n\geq 1, m\geq 2\}}$ are negligible,

$$\begin{aligned}V(0,0,p+\Delta p) - V(0,0,p) &\approx (\Delta p)w'(p) + \frac{(\Delta p)^2}{2}w''(p) \\ &= (\Delta p)\alpha(p) + (\Delta p)(1-\tilde{\beta})(p+b)\alpha'(p) + \frac{(\Delta p)^2}{2}\alpha'(p) \\ &= (\Delta p)\left(\alpha(p) + \frac{\Delta p}{2}\alpha'(p)\right) + (\Delta p)(1-\tilde{\beta})(p+b)\alpha'(p) \\ &\approx (\Delta p)\frac{\alpha_i(p+\Delta p) + \alpha_i(p)}{2} + (b_i+p)(1-\tilde{\beta}_i)(\alpha_i(p+\Delta p) - \alpha_i(p))\end{aligned}$$

Now if $p > 0$, then $w_i(p+\Delta p) - w_i(p) = V_i(0,0,p+\Delta p)\varepsilon_{ij} - V_i(0,0,p)\varepsilon_{ij'}$ and thus

$$E[w_i(p+\Delta p) - w_i(p)] = E[V_i(0,0,p+\Delta p) - V_i(0,0,p)].$$

If $p = 0$,

$$E[w_i(p+\Delta p) - w_i(p)] = E[V_i(0,0,p+\Delta p) - V_i(0,0,0)] + E[\eta_i].$$

□

# B Further theoretical results

## B.1 Generalizations to the Dynamic Case

We now consider a dynamic environment in which the agent can chooses $a_t \in \{0, 1\}$ in each period $t = 1, \ldots, T$, and chooses commitment contracts in period $t = 0$. The delayed benefit from choosing $a_t = 1$ is $b$, which is realized in period $T + 1$. The costs $c_t$ for choosing $a_t = 1$ are drawn from a distribution $F(c|h_t)$, where $h_t$ is the history of actions up to period $t$. Commitment contracts for more attendance involve a penalty $p$ that is paid if $\sum a_t < X$, while commitment contracts for less attendance involve a penalty that is paid if $\sum a_t \geq X$.

### B.1.1 Generalization of Proposition 2

We generalize Proposition 2 by considering commitment contracts like those in our experiment, which involve a penalty $p$ if the agent does not choose $a_t = 1$ at least $X \leq T$ times.

**Proposition 8.** *Fix $p$ and suppose that $F(\cdot|h_t)$ has a density function $f(\cdot|h_t)$ for each $h_t$, which satisfies $f(c_2|h_t)/f(c_1|h_t) \geq (c_1/c_2)^2$ for all $c_1 < c_2 < b + p$. Then the perceived utility loss of a commitment contract that involves a penalty $p$ for $\sum a_t < X$ is decreasing in $\tilde{\beta}$. Consequently, no agents should desire commitment contracts.*

Throughout, we use the following straightforward but useful extension of Proposition 2:

**Lemma 2.** *Consider a density function $f(\cdot)$ of $c$ such that $f(c_2)/f(c_1) \geq (c_1/c_2)^2$ for all $c_1 < c_2 < B$. Let the payoffs for choosing $a = 0$ and $a = 1$ be $b_0$ and $b_1$, respectively, with $B = b_1 - b_0$. Define $W = b_0 + \int_0^{\tilde{\beta}(b_1 - b_0)} (b_1 - b_0 - c) f(c) dc$ . Then $\frac{\partial^2 W}{\partial \tilde{\beta} \partial b_0} < 0$, and consequently $\frac{\partial W}{\partial b_0} > 0$.*

*Proof.* The first part, $\frac{\partial^2 W}{\partial \tilde{\beta} \partial b_0} < 0$, is an immediate consequence of Proposition 2, since decreasing $b_0$ is equivalent to instituting a penalty for choosing $a = 0$. The second part follows because $\frac{\partial W}{\partial b_0} > 0$ clearly holds for $\tilde{\beta} = 1$, and thus by the first statement must hold for any $\tilde{\beta} < 1$. □

We now prove the proposition:

*Proof.* Let $V_t(h_t)$ denote the period 0 expectation of period $t$ self's utility, following $h_t = \sum_{\tau=1}^{t-1} a_\tau$ choices of $a_\tau = 1$. Note that $V_t(h_t)$ is also the period $t - 1$ expectation of self-$t$ utility, since both period 0 and period $t - 1$ selves have the same beliefs about period $t$self's behavior.

Step 1. We first show that $V_t(h + 1) \geq V_t(h)$ for all $h$. We do this by induction. Consider $t = T$. If $h \geq X$ or if $h \leq X - 2$ then $V_t(h + 1) = V_t(h)$, since in the former case the agent meets the threshold regardless and in the latter case the agent fails to meet the threshold regardless. If $h_t = X - 1$ then Proposition 2 implies that $V_t(h + 1) > V_t(h)$, since in the former case there is no penalty for choosing $a_t = 1$ while in the latter case there is. Now suppose that $V_{t+1}(h)$ is increasing in $h$. In period $t$, this means that the delayed payoffs from choosing $a_t = 1$ and $a_t = 0$, respectively, are $V_{t+1}(h_t + 1)$ and $V_{t+1}(h_t)$. Clearly, period $t$ utility is increasing in $V_{t+1}(h_t + 1)$. Lemma 2 establishes that period $t$ utility must also be increasing in $V_{t+1}(h_t)$, the payoff from choosing $a_t = 0$.

And since $V_{t+1}$ is increasing in $h_t$ by the induction hypothesis, this establishes that $V_t$ must also be increasing in $h_t$.

STEP 2. We now show that $V_t(h_t)$ is increasing in $\tilde{\beta}$ for all $h_t$. We again do this by induction. Consider first $t = T$. If $h_T \geq X$ or if $h_T \leq X-2$ then the penalty does not matter. If $h_T = X-1$ then Proposition 2 implies that $\frac{\partial^2}{\partial p} V_T(h_T) < 0$ and $\frac{\partial^2}{\partial \tilde{\beta} \partial p} V_T(h_T) > 0$ . Now suppose that $\frac{\partial^2}{\partial p} V_{t+1}(h_{t+1}) < 0$ and $\frac{\partial^2}{\partial \tilde{\beta} \partial p} V_{t+1}(h_{t+1}) > 0$. In period $t$, the delayed payoffs from choosing $a_t = 1$ and $a_t = 0$, respectively, are $V_{t+1}(h_t + 1)$ and $V_{t+1}(h_t)$. The induction hypothesis implies that these delayed payoffs decrease with $p$, which by Lemma 2 implies that $V_t$ is decreasing in $p$. Moreover, the induction hypothesis implies that these payoffs decrease the most for those with the lowest $\tilde{\beta}$. Lemma 2 therefore also implies that $V_t$ decreases the most in $p$ for those with the lowest $\tilde{\beta}$. □

### B.1.2 Generalizations of Propositions 4, 5, 6

The generalizations of these propositions follow also verbatim. To establish the generalization of Proposition 6 we only need the stronger assumptions that lead to 8.

### B.1.3 Generalization of Proposition 7

We consider piece-rate incentives $p$ that payout $p \sum_t a_t$, and we let $w_i(p)$ denote agent $i$'s willingness to pay for these piece-rate incentives. Let $A_i(p)$ denote the expected attendance with piece-rate incentive $p$. If $\tilde{\beta}_i = 1$ for all agents, then the Envelope Theorem implies that for $p > 0$ $\frac{d}{dp} E[w_i(p)] = \frac{d}{dp} E[A_i(p)]$. Consequently, by the reasoning of Proposition 7, it follows that $E[w_i(p+\delta) - w_i(p)] \approx \delta E\left[\frac{A_i(p+\delta)] + A_i(p)}{2}\right]$. Thus, if average WTP is above expected value, the null of $\tilde{\beta}_i = 1 \; \forall i$ is rejected.

In the case in which the period-$t$ distribution of $c$ does not depend on the history of past actions, let $\alpha_{it}$ denote the agent $i's$ expectation of choosing $a_t = 1$. Then by Proposition 7, for $p > 0$

$$\begin{aligned} \frac{d}{dp} E[w_i(p)] &= \sum_t E\left[\alpha_{it}(p) + (1 - \tilde{\beta}_i)(p + b_i)\alpha_i'(p)\right] \\ &= E\left[A_i(p) + (1 - \tilde{\beta}_i)(p + b_i)A_i'(p)\right]. \end{aligned}$$

Identical reasoning then establishes that for $p > 0$:

$$E[w_i(p+\delta) - w_i(p)] = E\left[\delta \frac{A_i(p+\delta)] + A_i(p)}{2} + (b_i + p)(1 - \tilde{\beta}_i)(A_i(p+\delta) - A_i(p))\right]$$

and for $p = 0$:

$$E[w_i(p+\delta) - w_i(p)] = E\left[\delta \frac{A_i(p+\delta)] + A_i(p)}{2} + (b_i + p)(1 - \tilde{\beta}_i)(A_i(p+\delta) - A_i(p))\right] + E[\eta_i].$$

## B.2 Generalization to continuous choice

We now generalize our results about the (lack of) desirability of commitment contracts to continuous choices. We consider two models.

**Model I: Costly effort.** We consider a costly effort model as in Kaur et al. (2015), generalized to allow for uncertainty in effort costs. Workers earn future salary $y = wx$ at some cost of effort $C(x)$. In period 0, workers believe that in period 1 they will choose $e$ to maximize $\tilde{\beta}wx - \theta C(x)$, where $\theta \sim F$ is an effort cost shock. However, in period 0 their preferred choice of effort is to maximize $wx - \theta C(x)$. For simplicity, we follow Kaur et al. (2015) in assuming an isoelastic cost of effort function, which produces a constant elasticity of earnings with respect to the wage, denoted by $\varepsilon$.

**Model II: Saving for the future.** In the savings choice model, the agent chooses an amount $x$ to save for the future, given initial endowment $Y$. In period 0, agents believe that in period 1 they will choose $x$ to maximize $\theta(Y - rx) + \tilde{\beta}u(x)$, where $\theta \sim F$ is the uncertainty in the need for funds in period 1, and $r$ is the price of period 1 consumption. However, their preferred level of savings maximizes $\theta(Y - rx) + u(x)$. As before, we simplify by assuming a CRRA functional form, which produces a constant elasticity of saving with respect to $r$, denoted $\varepsilon$.

### Continuous penalties

We begin with contracts that specify specify a penalty $p(X - x)$ for choices $x$ below a target $X$ ($x \leq X$).

**Proposition 9.** *Consider model 1.*

*1. If for a given commitment contract $(p, X)$ there is a positive measure of $\theta$ for which the period 0 self would choose $x^* < X$, then the commitment contract cannot be desired by anyone, and its expected damages are decreasing in $\tilde{\beta}$.*

*2. Let $E[x(p)|x(p) \leq X]$ denote the average effort conditional on it being less than $X$, given penalty $p$ for working less than $X$. If $E[x(p)|x(p) < X] < \frac{X}{(1-\tilde{\beta})\varepsilon+1}$ for all $p \in [0, \bar{p}]$, then expected utility under the commitment contract is decreasing in $p \in [0, \bar{p}]$. Consequently, no commitment contracts of the form $(p, X), p \in (0, \bar{p}]$ are desirable.*

An important implication of the part 2 of the proposition is that what effects the possible desirability of a commitment contract is not the likelihood that the agent will fail to meet it, but rather the expected costs of failing to meet it. Intuitively, this is because a marginal change in the penalty $p$ has no effects on an agent's utility in states of the world in which he does not fail to meet the contract. Both the benefits—which derive from behavior change—and the costs—which derive from the paying the penalty—of the marginal change lie only in the region in which the agent fails to meet it. Consequently, if conditional on failing to meet the contract the agent fails to meet it by a lot, a marginal change in $p$ decreases expected period 0 utility. If this is true for all marginal

changes between 0 and $\bar{p}$, then integration of the marginal changes implies that no penalties in $[0, \bar{p}]$ can be welfare enhancing.

*Proof.* For a realization $\theta$, suppose that the period-0 expected choice under the contract is $x^*(\theta, p) < X$. Now for this taste-shock,

$$\begin{aligned}
\frac{d}{dp}((w+p)x^* - \theta C(x^*) - Xp) &= \frac{dx^*}{dp}(w + p - \theta C'(x^*)) - (X - x^*) \\
&= \frac{dx^*}{dp}(w + p - \tilde{\beta}(w+p)) - (X - x^*) \\
&= (1-\tilde{\beta})(w+p)\frac{dx^*}{dp} - (X - x^*) \\
&= (1-\tilde{\beta})x^*\varepsilon - (X - x^*) \\
&= ((1-\tilde{\beta})\varepsilon + 1)x^* - X \qquad (12)
\end{aligned}$$

where $\varepsilon = \frac{dx}{dw} \cdot \frac{p+w}{x}$ is the elasticity of effort with respect to the wage. Clearly, increasing $p$ has no effect for states of the world in which $x^* \geq X$. Integrating over $\theta$, the net impact of increasing $p$ is thus

$$Pr(x^* < X)\left(((1-\tilde{\beta})\varepsilon + 1)E\left[x^* | x^* < X\right] - X\right)$$

Next, taking the derivative of (12) with respect to $\tilde{\beta}$ gives

$$\begin{aligned}
-\varepsilon x^* + ((1-\tilde{\beta})\varepsilon + 1)\frac{dx^*}{d\tilde{\beta}} &= -\varepsilon x^* + ((1-\tilde{\beta})\varepsilon + 1)\frac{x^*\varepsilon}{\tilde{\beta}} \\
&= \varepsilon x^* \left[\frac{1 + (1-\tilde{\beta})\varepsilon}{\tilde{\beta}} - 1\right] \\
&> 0
\end{aligned}$$

Taking expectations, this implies that the expected utility $V(p, X)$ from the contract satisfies $\frac{d}{d\tilde{\beta}dp}V > 0$ as long as there is a positive measure of states for which $x^* < X$. This implies that if at some value $p = q$ there is a positive measure of states for which a $\tilde{\beta} = 1$ agent would expect to choose $x^* < X$, $\frac{d}{d\tilde{\beta}dp}V > 0$ for all $\tilde{\beta} \in [0, 1]$ and $p \leq q$. But since $\frac{d}{dp}V < 0$ for $\tilde{\beta} = 1$, this implies that $\frac{d}{dp}V < 0$ for all $\tilde{\beta} \in [0, 1]$. □

**Proposition 10.** *Consider model 2.*

*1. If for a given commitment contract $(p, X)$ there is a positive measure of $\theta$ for which the period 0 self would choose $x^* < X$, then the commitment contract cannot be desired by anyone, and its expected damages are decreasing in $\tilde{\beta}$.*

*2. Let $E[x(p)|x(p) \leq X]$ denote the average effort conditional on it being less than $X$, given penalty $p$ for working less than $X$. If $E[x(p)|x(p) < X] < \frac{X}{(1-\tilde{\beta})\varepsilon+1}$ for all $p \in [0, \bar{p}]$, then expected utility under the commitment contract is decreasing in $p \in [0, \bar{p}]$. Consequently, no commitment contracts of the form $(p, X), p \in (0, \bar{p}]$ are desirable.*

*Proof.* For a realization $\theta$, suppose that the period-0 expected choice under the contract is $x^*(\theta, p) < X$. Now for this taste-shock,

$$\begin{aligned}
\frac{d}{dp}(u(x^* + \theta(Y - (r+p)x^* - pX)) &= \frac{dx^*}{dp}(u'(x^*) - \theta(r+p)) - \theta(X - x^*) \\
&= \frac{dx^*}{dp}\left(\frac{1}{\tilde{\beta}}\theta(r+p) - \theta(r+p)\right) - \theta(X - x^*) \\
&= (1/\tilde{\beta} - 1)\theta(r+p)\frac{dx^*}{dp} - \theta(X - x^*) \\
&= (1/\tilde{\beta} - 1)\theta x^*\varepsilon - \theta(X - x^*) \\
&= \theta(1/\tilde{\beta} - 1)\varepsilon + 1)x^* - \theta X \qquad (13)
\end{aligned}$$

where $\varepsilon = \frac{dx}{dr} \cdot \frac{r+p}{x}$ is the elasticity. Clearly, increasing $p$ has no effect for states of the world in which $x^* \geq X$. Integrating over $\theta$, the net impact of increasing $p$ is thus

$$\theta Pr(x^* < X)\left(1/\tilde{\beta} - 1)E\left[x^* | x^* < X\right] - X\right)$$

Next, taking the derivative of (13) with respect to $1/\tilde{\beta}$ gives

$$\begin{aligned}
-\varepsilon\theta x^* + \theta((1/\tilde{\beta} - 1)\varepsilon + 1)\frac{dx^*}{d(1/\tilde{\beta})} &= -\varepsilon\theta x^* + ((1/\tilde{\beta} - 1)\varepsilon + 1)\frac{x^*\varepsilon}{\tilde{\beta}} \\
&= \varepsilon x^*\left[\frac{(1/\tilde{\beta} - 1)\varepsilon + 1}{\tilde{\beta}} - 1\right] \\
&> 0
\end{aligned}$$

Taking expectations, this implies that the expected utility $V(p, X)$ from the contract satisfies $\frac{d}{d\tilde{\beta}dp}V > 0$ as long as there is a positive measure of states for which $x^* < X$. This implies that if at some value $p = q$ there is a positive measure of states for which a $\tilde{\beta} = 1$ agent would expect to choose $x^* < X$, $\frac{d}{d\tilde{\beta}dp}V > 0$ for all $\tilde{\beta} \in [0, 1]$ and $p \leq q$. But since $\frac{d}{dp}V < 0$ for $\tilde{\beta} = 1$, this implies that $\frac{d}{dp}V < 0$ for all $\tilde{\beta} \in [0, 1]$. □

**Discontinuous penalties**

**Proposition 11.** *Consider model 1 and fix a contract $(p, X)$. Let $\theta^\dagger(\tilde{\beta})$ be the taste-shock for which an agent with perceived present focus $\tilde{\beta}$ is indifferent between choosing $X$ versus some amount $x < X$. If $f'(\theta)/f(\theta) \geq -1/\theta$ for $\theta \in [\theta^\dagger(\underline{\tilde{\beta}}), \theta^\dagger(1)]$ then then the commitment contract cannot be desired by anyone with $\tilde{\beta} > \underline{\tilde{\beta}}$, and its expected damages are decreasing in $\tilde{\beta}$. An analogous result holds for model 2.*

*Proof.* Consider now contracts that specify a fixed penalty $p$ as long as $x < X$. This means that in

model 1, for each $p$ and $\tilde{\beta}$, there is a "marginal" taste-shock $\theta^\dagger(p,\tilde{\beta})$ satisfying

$$\tilde{\beta}(wx(\theta^\dagger)-p)-\theta^\dagger C(x(\theta^\dagger))=\tilde{\beta}wX-\theta^\dagger C(X) \tag{14}$$

where $x$ satisfies $\theta^\dagger C'(x)=\tilde{\beta}w$. Differentiating the first condition with respect to $\tilde{\beta}$ gives and using the envelope theorem

$$wx-p-\frac{d\theta^\dagger}{d\tilde{\beta}}C(x)=wX-\frac{d\theta^\dagger}{d\tilde{\beta}}C(X)$$

or

$$\begin{aligned}\frac{d\theta^\dagger}{d\tilde{\beta}}&=\frac{wX+p-wx}{C(X)-C(x(\theta^\dagger))}\\&=\theta^\dagger/\tilde{\beta}\end{aligned}$$

This implies that $\theta^\dagger$ is a linear function of $\tilde{\beta}$, and that $\frac{d\theta^\dagger}{d\tilde{\beta}}$ is a constant; we define it to be $\gamma$. Now the perceived gains from having $\tilde{\beta}$ increased are

$$(1-\tilde{\beta})(wX+p-wx(\theta^\dagger))f(\theta^\dagger)\gamma$$

These gains are increasing in $p$ if $(wX+p-wx(\theta^\dagger))f(\theta^\dagger)$ is increasing in $p$. Now (14) is equivalent to

$$\tilde{\beta}(wX+p-wx(\theta^\dagger))=\theta^\dagger C(X)-\theta^\dagger C(x(\theta^\dagger))$$

The derivative of the right-hand side with respect to $p$ is

$$\frac{d\theta^\dagger}{dp}\left(C(X)-C(x)-\theta^\dagger C'(x)\frac{dx}{d\theta^\dagger}\right)$$

But since $x$ is decreasing in $\theta^\dagger$, this means that $C(X)-C(x)-\theta^\dagger C'(x)\frac{dx}{d\theta^\dagger}$ is positive. In particular, differentiating the FOC yields $C'(x)+\theta^\dagger C''(x)\frac{dx}{d\theta^\dagger}=0$, or $\frac{dx}{d\theta^\dagger}=\frac{-C'}{\theta^\dagger C''}=-\frac{\tilde{\beta}w}{\theta^2 C''}$. Since $\frac{dx}{dw}=\frac{\tilde{\beta}}{\theta C''}$, it follows that $\frac{dx}{d\theta^\dagger}=-\frac{w}{\theta^\dagger}\frac{dx}{dw}=-\frac{x}{\theta^\dagger}\varepsilon$.

Consequently $\frac{d}{dp}(X+p-wx(\theta^\dagger))$ has the same sign as $\frac{d\theta^\dagger}{dp}$. Now by the envelope theorem, the derivative of (14) with respect to $p$ is

$$-\tilde{\beta}-C(x)\frac{d\theta^\dagger}{dp}=-C(X)\frac{d\theta^\dagger}{dp}$$

which shows that

$$\frac{d\theta^\dagger}{dp}=\frac{\tilde{\beta}}{C(X)-C(x(\theta^\dagger)}>0$$

Consequently,

$$\frac{d\theta^\dagger}{dp}\left(C(X)-C(x)-\theta^\dagger C'(x)\frac{dx}{d\theta^\dagger}\right)=\tilde{\beta}\frac{C(X)-C(x)+x\varepsilon C'(x)}{C(X)-C(x)}$$

and thus

$$\frac{d}{dp}(wX + p - wx(\theta^\dagger)) = \frac{C(X) - C(x) + x\varepsilon C'(x)}{C(X) - C(x)} \geq 1$$

By the chain rule, the condition for $(wX + p - wx(\theta^\dagger))f(\theta^\dagger)$ to be non-decreasing in $p$ is that

$$\begin{aligned}\frac{f'(\theta^\dagger)}{f(\theta^\dagger)} &\geq -\frac{C(X) - C(x) + x\varepsilon C'(x)}{C(X) - C(x)} \cdot \frac{1}{w(X-x) + p} \frac{1}{\frac{d\theta^\dagger}{dp}} \\ &= -\frac{1}{\tilde{\beta}} \frac{C(X) - C(x) + x\varepsilon C'(x)}{w(X-x) + p} \\ &= -\frac{1}{\theta^\dagger} \frac{w(X-x) + p + x\varepsilon w}{w(X-x) + p}\end{aligned}$$

A sufficient condition is thus that $\frac{f'(\theta)}{f(\theta)} \geq -1/\theta$. □

### B.3 Costly self control

Finally, we consider whether our predictions about the impact of uncertainty on commitment demand carry over to alternative models of self-control problems; in particular, models of costly self-control, as in Fudenberg and Levine (2006) and Gul and Pesendorfer (2001). We assume that the tempting option is to choose $a = 0$, which incurs no immediate costs, and we assume that the self control cost as linear (as in Gul and Pesendorfer, 2001, or Assumption 5' of Fudenberg and Levine, 2006). This means that in period 1, the agent's utility in a contract with penalty $p$ for choosing $a = 0$ is given by $-p + a \cdot [b + p - (1 + \gamma)c]$ ,where $\gamma$ is the marginal cost of self control. The agent's utility in period 1 when the choice set is restricted to $A = \{1\}$ is given by $(b - c)$. In period 0, the agent chooses the contract if it increases expected period 1 utility. The expected utility from a $p$-penalty-contract is

$$F(c^\dagger)(b + p - (1 + \gamma)E[c|c \leq c^\dagger) - p$$

where $c^\dagger = \frac{b+p}{1+\gamma}$. By the envelope theorem, the derivative of that with respect to $p$ is $-(1 - F(c^\dagger))$. Thus, utility is strictly decreasing in $p$ when $F\left(\frac{b+p}{1+\gamma}\right) < 1$. This means that as long as there is some chance that $c < b/(1 + \gamma)$, a penalty-based contract can only decrease utility. Moreover, since the loss $(1 - F(c^\dagger))$ is decreasing in $c^\dagger$, this means that penalties are least attractive to those with the highest (perceived) costs of self-control.

Consider now choice-set restrictions. The utility with a choice-set restriction is $b - E[c]$, while the utility without it is $\int_{c \leq b/(1+\gamma)} (b - (1 + \gamma)c)dF(c)$. The impact of the restriction is thus

$$\int_{c \leq b/(1+\gamma)} \gamma c dF(c) + \int_{c \geq b/(1+\gamma)} (b - c)dF(c) \leq \gamma \int_{c \leq b} c dF(c) + \int_{c \geq b} (b - c)dF(c)$$

The inequality follows because $\gamma c \geq b - c$ iff $c \geq b/(1+\gamma)$. 'To get a quantitative sense of this, suppose

that $c$ is uniform on $[0, \bar{c}]$, and normalize $b = 1$. Then $E[c|c > 1] - 1 = \frac{\bar{c}-1}{2}$ and $E[c|c < b] = b/2$. Then the gains are negative if $\gamma(1/2)(1/\bar{c}) \leq \frac{\bar{c}-1}{\bar{c}}\frac{\bar{c}-1}{2}$, or if $\gamma \leq (\bar{c}-1)^2$. For example, suppose that $\gamma = 0.3$, which is equivalent to weighting delayed benefits relative to costs by a factor of $\beta = 0.77$. In this case, the gains from full commitment are negative if $\bar{c} > 1.55$. Compared to the uniform costs case in the present focus model, this implies that binding commitment contracts are more desirable for agents with costly self-control, for a given "weight" on delayed benefits versus immediate costs.

# C Further study details

Table A.1: Survey Details by Wave

| Wave (Survey dates) | N | Information Treatment | Commitment Contracts Presented | Elicited Perceived Probabilities | Check-out scanner | Targeted Incentives |
|---|---|---|---|---|---|---|
| **Wave 1 (Oct.-Nov. 2015)** | 350 | Basic (Graph of past visits only) | More/Less than 8 days<br>More/Less than 12 days<br>More/Less than 16 days | N/A | N/A | \$0 (33%); \$2 (33%); \$7 (33%) |
| **Wave 2 (Jan.-Feb. 2016)** | 528 | Enhanced (Graph, forced engagement, information on aggregate overconfidence) | | | Participants asked to swipe out upon leaving the gym. | \$0 (33%); \$2 (33%); \$5 (16.5%); \$7 (16.5%) |
| **Wave 3 (Mar.-Apr. 2016)** | 414 | | More/Less than 12 days | More/Less than 12 | | \$0 (33%); \$7 (33%); \$80 if 12+ visits (33%) |

Notes: This table describes the variations in the survey across the three waves of implementation.

Table A.2: Survey Demographics by Wave

| | Wave 1 | Wave 2 | Wave 3 | Total |
|---|---|---|---|---|
| Female | 0.64 | 0.61 | 0.56 | 0.60 |
| | (0.48) | (0.49) | (0.50) | (0.49) |
| Age[a] | 29.94 | 34.20 | 33.49 | 32.84 |
| | (11.59) | (14.83) | (14.89) | (14.16) |
| Student, full-time | 0.65 | 0.52 | 0.54 | 0.56 |
| | (0.48) | (0.50) | (0.50) | (0.50) |
| Working, full or part-time | 0.51 | 0.64 | 0.64 | 0.61 |
| | (0.50) | (0.48) | (0.48) | (0.49) |
| Married | 0.23 | 0.28 | 0.27 | 0.26 |
| | (0.42) | (0.45) | (0.44) | (0.44) |
| Advanced degree[b] | 0.40 | 0.48 | 0.49 | 0.46 |
| | (0.49) | (0.50) | (0.50) | (0.50) |
| Income[a] | 46457 | 59227 | 58122 | 55483 |
| | (41190) | (48413) | (49491) | (47238) |
| Days visited in past 4 weeks, recorded | 7.00 | 7.87 | 6.06 | 7.06 |
| | (5.77) | (6.21) | (5.49) | (5.92) |
| *Visits in past 100 days* | | | | |
| Days visited, recorded | 24.55 | 21.53 | 20.86 | 22.13 |
| | (18.30) | (17.55) | (17.24) | (17.71) |
| Days visited, self-recollection | 33.83 | 30.16 | 28.16 | 30.51 |
| | (22.68) | (21.13) | (20.39) | (21.42) |
| Days that *I should have gone, but didn't* | 29.62 | 31.20 | 30.39 | 30.52 |
| | (22.83) | (24.09) | (23.92) | (23.69) |

Notes:
a. Imputed from categorical ranges.
b. A graduate degree beyond a B.A. or B.S.
This table shows the means of demographic variables reported in the survey across the three waves of implementation. The table also summarizes data on past visit frequencies to the gym. "Recorded" visits are obtained from the fitness center's log-in records, while "self-recollection" refers to participants' survey-reported estimates of their own past visits.

# D Further results on take up of commitment contracts

## D.1 Commitment contract take up in treatment group and its correlates

Table A.3: Commitment contract take up by treated individuals

| Threshold | Chose "More" Contract (1) | Chose "Fewer" Contract (2) | Chose "More" Given Chose "Fewer" (3) | Chose "Fewer" Given Chose "More" (4) | Diff (3)-(1) | Diff (4)-(2) |
|---|---|---|---|---|---|---|
| 8 visits | 0.63 | 0.33 | 0.89 | 0.47 | 0.26*** | 0.14*** |
| 12 visits | 0.46 | 0.31 | 0.62 | 0.41 | 0.16*** | 0.10*** |
| 16 visits | 0.29 | 0.24 | 0.45 | 0.38 | 0.16*** | 0.14*** |

Notes: Column (1) reports take up rates of commitment contracts to visit the gym at least 8, 12, or 16 days over the next four weeks (i.e., take up of the "more" contract). Column (2) shows take up rates of commitment contracts to visit the gym less than 8, 12, or 16 days over the same period (i.e., take up of the "fewer" contract). Columns (3) and (4) shows the take up rates of each type of commitment contract conditional on having chosen the other type of commitment contract. Columns (5) and (6) display the difference in the take up rates of column (3) versus column (1) in column (5) and the difference in the take up rates of column (4) versus column (2) in column (6). *** denotes those differences that are statistically significantly different from 0 at the 1% level. All take up rates are computed for the treatment group subjects exclusively. Over three survey waves, all participants faced the choice of both commitment contracts at the 12 visit threshold (N=652) while the 8 visit and 16 visit commitment contracts were only shown in the first two waves (N=437).

Table A.4: Correlation between perceived success in contracts and expected attendance; treated group

| | Expected Attendance w/ out incentives | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Prob succeed in "more" contract | 0.07*** | | 0.07*** |
| | (0.02) | | (0.02) |
| Prob succeed in "less" contract | | –0.04*** | –0.05*** |
| | | (0.01) | (0.01) |
| N | 215 | 215 | 215 |

Notes: This table displays the association between subjective beliefs about commitment contract success and expected attendance with no incentives. Each column presents coefficient estimates and heteroskedastic-consistent standard errors in parentheses from a separate OLS regression. Prob succeed in "more" contract is the ex-ante self-assessed probability of attending the gym 12 or more days during the 4-week incentive period. Prob success is "fewer" contract is ex-ante self-assessed probability of attending the gym fewer than 12 days during the 4-week incentive period. The sample consists exclusively of treatment group subjects in wave 3, the only wave we elicited the probabilities of contract success. The "More" - "Fewer"' row denotes a test of the difference between the coefficient on the probability of success under the "more" contract versus the probability of success under the "fewer" contract. *** denotes statistics that are statistically significantly different from 0 at the 1% level.

Table A.5: Correlation between perceived success in contracts and take up of contracts; treated group

| | Prob succeed in *more attendance* contract | | | Prob succeed in *less attendance* contract | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Commit to "more" | 12.35*** | | 14.68*** | –6.39 | | –9.60** |
| | (3.03) | | (3.01) | (4.18) | | (4.12) |
| Commit to "less" | | –8.21** | –11.67*** | | 13.79*** | 16.05*** |
| | | (3.80) | (3.57) | | (3.77) | (3.72) |
| N | 215 | 215 | 215 | 215 | 215 | 215 |

Notes: This table displays the association between commitment contract take up and subjective beliefs about success in the commitment contract. Each column presents coefficient estimates and heteroskedastic-consistent standard errors in parentheses from separate OLS regression. Columns (1)-(3) display associations with subjects' expectations of following through on the commitment contract requiring attendance at the gym 12 or more days during the 4-week incentive period. Columns (4)-(6) display associations with subjects' expectations of following through on the commitment contract requiring attendance at the gym fewer than 12 days during the 4-week incentive period. The sample consists exclusively of treatment group subjects in wave 3, the only wave we elicited the probabilities of contract success. **,*** denote statistics that are statistically significantly different from 0 at the 5% and 1% level respectively.

## D.2 Other correlates of commitment contract take up

Table A.6: Other correlates of commitment contract take up; control group

| | Expected attendance | Past attendance | Goal attendance |
|---|---|---|---|
| | (1) | (2) | (3) |
| Chose "more" contract | 2.58*** | 1.63*** | 3.05*** |
| | (0.29) | (0.32) | (0.29) |
| Chose "less" contract | –0.60* | –1.67*** | –1.10*** |
| | (0.32) | (0.33) | (0.32) |
| N | 1,522 | 1,522 | 1,522 |

Notes: This table displays the association between commitment contract take up and attendance patterns. Each column presents coefficient estimates and heteroskedastic-consistent standard errors in parentheses from separate OLS regression. Column (1) displays the correlations between commitment contract choice and expected days of attendance under no incentive over the 4-week incentive period. Column (2) displays the correlations between commitment contract choice and days of attendance over the past 4 weeks. Column (3) presents the correlations between commitment contract choice and goal days of attendance over the next 4 weeks. The sample consists exclusively of control group subjects. Since subjects were asked about multiple commitment contracts, each subject contributes more than 1 observation to these regressions (i.e., the data are pooled across the different commitment contract questions). *,*** denote statistics that are statistically significantly different from 0 at the 10% and 1% level respectively.

Table A.7: Other correlates of commitment contract take up; treated group

| | Expected attendance | Past attendance | Goal attendance |
|---|---|---|---|
| | (1) | (2) | (3) |
| Chose "more" contract | 1.41*** | 1.16*** | 2.18*** |
| | (0.30) | (0.30) | (0.32) |
| Chose "less" contract | –1.32*** | –2.28*** | –1.22*** |
| | (0.32) | (0.32) | (0.36) |
| N | 1,526 | 1,526 | 1,526 |

Notes: This table displays the association between commitment contract take up and attendance patterns. Each column presents coefficient estimates and heteroskedastic-consistent standard errors in parentheses from separate OLS regression. Column (1) displays the correlations between commitment contract choice and expected days of attendance under no incentive over the 4-week incentive period. Column (2) displays the correlations between commitment contract choice and days of attendance over the past 4 weeks. Column (3) presents the correlations between commitment contract choice and goal days of attendance over the next 4 weeks. The sample consists exclusively of treatment group subjects. Since subjects were asked about multiple commitment contracts, each subject contributes more than 1 observation to these regressions (i.e., the data are pooled across the different commitment contract questions). *** denotes statistics that are statistically significantly different from 0 at the 1% level.

# E Further results on willingness to pay for behavior change

## E.1 Correlations with fewer commitment contracts

Table A.8: Correlation between WTP for behavior change and take up of "fewer" contracts

| | Take-up of less-visits contract (mean = .33) | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| WTP for behavior change (z-score) | 0.005 | 0.009 | | |
| | (0.025) | (0.025) | | |
| Expected-visits elasticity (z-score) | | –0.031* | | |
| | | (0.019) | | |
| WTP for behavior change excl. $1 incentive (z-score) | | | 0.000 | 0.007 |
| | | | (0.025) | (0.025) |
| Expected-visits elasticity excl. $1 incentive (z-score) | | | | –0.027 |
| | | | | (0.020) |
| N | 1,522 | 1,522 | 1,522 | 1,522 |

Notes: This table displays the association between estimated WTP for behavior change (expressed as a z-score) and take up of "fewer" commitment contracts. Each column presents coefficient estimates and cluster-robust standard errors in parentheses from separate OLS regressions. The sample consists exclusively of control group subjects (N=640). In columns (1) and (2), WTP is calculated based on all incentive levels whereas in columns (3) and (4), WTP is calculated excluding the $1 incentive. In columns (2) and (4), the average elasticity of each individual's visit expectations with incentive size (expressed as a z-score) is also included. All regressions include wave fixed effects and commitment contract threshold fixed effects (i.e., 8, 12, 16 visit thresholds). * denotes a statistic that is statistically significantly different from 0 at the 10% level.

## E.2 Estimates of $\tilde{\beta}$

Formally, for each set of incentives $p_k$ and $p_{k+1}$, WTP estimates $w_i(p_k)$ and $w_i(p_{k+1})$, and expected attendances $\alpha_i(p_k)$ and $\alpha_i(p_{k+1})$, the moment condition is

$$E\left[w_i(p_k) - w_i(p_{k+1}) - (p_{k+1} - p_k)\frac{E[\alpha_i(p_k)] + E[\alpha_i(p_{k+1})]}{2} + (b_i + p_1)(1 - \tilde{\beta}_i)(\alpha_i(p_k) - \alpha_i(p_{k+1}))\right] = 0$$

Since there are five pairs of adjacent incentives, we have five such moment conditions. Letting $\hat{\tilde{\beta}}$ denote the parameter, the GMM estimator chooses the parameter $\hat{\tilde{\beta}}$ that minimizes $\left(m(\tilde{\beta}) - m(\hat{\tilde{\beta}})\right)' W \left(m(\tilde{\beta}) - m(\hat{\tilde{\beta}})\right)$ where $m(\xi)$ are the theoretical moments, $m(\hat{\tilde{\beta}})$ are the empirical moments, and $W$ is the optimal weighting matrix given by the inverse of the variance-covariance matrix of the moment conditions. We approximate $W$ using a two-step estimator outlined in Hall (2005). In the first step, we set $W$

equal to the identity matrix,[31] and use this to solve the moment conditions for $\hat{\tilde{\beta}}$, which we denote $\hat{\tilde{\beta}}_1$. Since $\hat{\tilde{\beta}}_1$ is consistent, by Slutsky's theorem the sample residuals $\hat{u}$ will also be consistent. We then use these residuals to estimate the variance-covariance matrix of the moment conditions, $S$, given by $Cov(\mathbf{z}u)$, where $\mathbf{z}$ are the instruments for the moment conditions. We then minimize $\left(m(\tilde{\beta}) - m(\hat{\tilde{\beta}})\right)' \hat{W} \left(m(\tilde{\beta}) - m(\hat{\tilde{\beta}})\right)$ using $\hat{W} = \hat{S}^{-1}$, which gives the optimal $\hat{\tilde{\beta}}$ (Hansen, 1982).

### E.3 Estimates of $\beta/\tilde{\beta}$

Under the maintained assumption that $c$ is i.i.d. across time, an individual's expected number of attendances is given by $TF(\tilde{\beta}(b+p))$, where $T$ is the number of periods. In contrast, the rational expectation is $TF(\beta(b+p))$. Consequently, we perceived attendance $\alpha(p)$ and actual average attendance $\alpha^*(p)$ can be expressed as $\alpha(p) = A(\tilde{\beta}(b+p))$ and $\alpha^*(p) = A(\beta(b+p))$, for $A = TF$. So if $\alpha(0) = a^*(p^*)$, then $\tilde{\beta}b = \beta(b+p^*)$, and thus

$$\beta/\tilde{\beta} = b/(b+p^*).$$

To implement the estimator, we estimate four moment conditions. First, we model actual average attendance as quadratic in $p$: $\alpha^*(p) = a_0 + a_1 p + a_2 p^2$, which leads to the three moment conditions $E[\alpha_i^*(p) - (a_0 + a_1 p + a_2 p^2)]p^k = 0$ for $k = 0, 1, 2$. The fourth moment condition for average expected attendance is simply $E[\alpha_i(0) - \bar{\alpha}_0] = 0$. Our estimate of naivete is then given by $\hat{n} = b/(b+\hat{p}^*)$, where $\hat{p}^*$ is the solution to $\hat{a}_0 + \hat{a}_1 p + \hat{a}_2 p^2 = \bar{\alpha}_0$.We compute the standard error for $\hat{n}$ using the delta method.

We compute the standard errors of the parameter vector $(\hat{a}_0, \hat{a}_1, \hat{a}_2, \hat{\bar{\alpha}}_0)$ using the two-step estimator described in the preceding appendix section (E.2).

### E.4 Dollar value of exercise

We provide two "back of the envelope calculations" of the dollar benefit of an hour of exercise. Our goal is not to provide a comprehensive review of the literature on the value of exercise, but to demonstrate that the literature provides a range of possible values. We then use that range when calculating values for$\tilde{\beta}$.

Sun et al. (2014)find a median difference of 0.112 Quality Adjusted Life Years (QALYs) between a group that was inactive over a two-year period and a group that exercised on average at least 2.5 hours per week over the two-year period controlling for sociodemographic characteristics (age, race/ethnicity, living arrangement, income, and education) and health status (e.g. smoking and BMI). If we adopt 50,000 dollars as the value for a QALY (Neumann et al., 2014), the benefit from an hour of exercise is:

[31] One other common approach is to use $(\mathbf{z}\mathbf{z}')^{-1}$as the weighting matrix in the first-stage, where $\mathbf{z}$ is a vector of the instruments in the moment equations. We confirmed our standard errors and point estimates are the same under both choices.

$$0.112 \times (\$50000)/(2.5 \times 104) = \$21.5$$

Despite the inclusion of control variables, this study likely overstates the causal effect of exercise because it does not control for other factors that may affect the difference in QALYs between the two groups such as diet before and during the period of study and exercise before the period of study.

Blair et al. (1989) examine the association between mortality risk and exercise over a fifteen-year period among a population of healthy non-geriatric adults. They find that a male who moved from the least fit quintile to the average of the other four quintiles would reduce his chances of dying by 36.7%, and a female who made a similar move would reduce her chances of dying by 48.4%. The authors also find that a brisk walk of 30 to 60 minutes each day would be sufficient to move an individual to a plateau where further exercise would not further lower the risk of death. If we assume that 45 minutes per day of exercise would at least move a person out of the lowest quintile of exercise and into the upper four quintiles (a smaller change than reaching the plateau), then it would lead to the reported reductions in mortality (36.7% for men and 48.4% for women). The paper reports an age adjusted all-cause mortality rate of 64 per 10,000 person years among men in the lowest quintile of exercise and 39.5 per 10,000 person years among women in the lowest quintile. The sample in our study is 61.2% female and 38.8% male with an average age of 34 years. Assuming men age 34 years have a death rate of 161 per 100,000 and women age 34 have a death rate of 85 per 100,000, the weighted average reduction in the death rate from this level of exercise for an individual of age 34 in our sample is,

$$\textit{reduction in deathrate} = 0.388 * 0.367 * 161/100,000 + 0.612 * 0.484 * 85/100,000 = 48.1/100,000$$

The value of the exercise then depends on the value of remaining life for a 34-year-old. If we adopt the SVL (statistical value of life) used by the US Environmental Protection Agency of 9.0 million dollars , we obtain

$$48.1/100,000 \times 9,000,000 = \$4329$$

Since the exercise required to achieve this gain was 45 minutes per day, the value of an hour of exercise is:

$$\$4329/(0.75 \times 365) = \$15.81$$

Alternatively, we could assume that a QALY is worth \$50,000, use life tables to calculate the probability of survival to each age beyond 34, and calculate the present discounted value (PDV) of life remaining. Using a discount rate of 2%, we calculate \$1,431,000 for men and \$1,519,000 for women. Performing similar calculations to the ones above for men and women and then taking the

weighted average based on the fraction of each gender in the sample, we obtain $2.60 per hour of exercise. Since part of the reason for discounting is to take account of the decreasing probability of survival at higher ages, it may be appropriate to apply an even lower discount rate. If we assume a discount rate of 0% so that the decrease in the contribution of QALYs at higher ages is entirely attributable to a decreased probability of survival, the value of life remaining past age 34 increases to 2,189,000 for men and 2,390,000 for women, and the value of an hour of exercise increases to $4.01.

# F Elicitation of WTP for Piece-Rate Incentives - Instructions

Our survey contained a section designed to elicit willingness to pay for incentive programs. This section began by explaining to subjects that as part of the study, they might receive an incentive program that would pay them based on the number of days they exercise at their gym (the fitness gym we partnered with). The survey then explained that we wanted to know the value they placed on different incentive programs and how often they thought they might go to the gym under these programs. See Figure A.1.

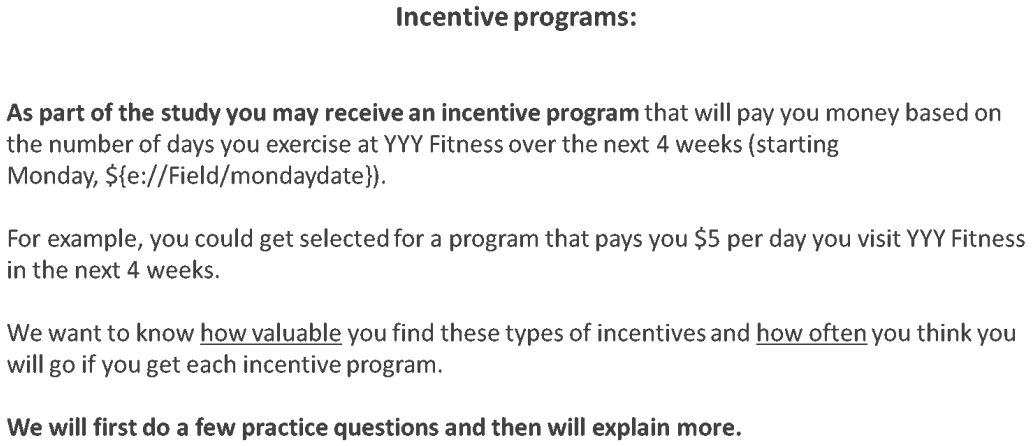



Figure A.1: Introduction to Willingness to Pay Section of Survey

Next, the survey explained to subjects the concept of willingness to pay, drawing on the example of a one dollar per day incentive that ran over the next four weeks. See Figure A.2.

Since subjects may not have been familiar with the idea of willingness to pay, we presented them with a row of decisions arranged in a table, where each decision asked them whether they preferred the one dollar per day incentive or a fixed payment. See Figures A.3 and A.4.

**Example**

**The possible incentive:**
Let's start with the incentive program that would pay you **$1 per day** that you visit YYY Fitness over the next 4 weeks (starting Monday, ${e://Field/mondaydate}). You could earn anywhere between $0 (with no visits) to $28 (if you went every day) with this program. Any earnings would be paid to you via a check along with your $10 survey payment after the 4 weeks are done.

**What is this $1 per-day incentive program worth to you?**

Suppose you knew you could have this incentive program, but you also had the possibility to trade the incentive for a fixed payment that does not depend on how often you visit YYY Fitness. How high does that fixed payment have to be for you to want to trade away the incentive?

For some people the answer might simply be the amount of money they thought they would earn with the incentive. However, for other people it could be more or less than that. For example, some people might like having the incentive program as extra motivation to come to the gym and would need a higher amount of money to give up the incentive. Other people might not like having their payment based on visits to the gym and would be willing to give it up for lower amounts.

There is no right answer here. We simply want to know what you think for yourself.

Figure A.2: Explanation of Willingness to Pay for $1 Incentive Program

**How big would a fixed payment need to be for you to want to trade away the incentive?**

For each decision below, please choose whether you would prefer to have the $1 per-day incentive over the next 4 weeks or instead the fixed payment in that row. As you click, the software will automatically fill in some options where it makes sense.

Figure A.3: Instructions for Decision Table

| | $1 per-day incentive | Fixed payment |
|---|---|---|
| Decision 1 | ○ | $0 ○ |
| Decision 2 | ○ | $2 ○ |
| Decision 3 | ○ | $4 ○ |
| Decision 4 | ○ | $6 ○ |
| Decision 5 | ○ | $8 ○ |
| Decision 6 | ○ | $10 ○ |
| Decision 7 | ○ | $12 ○ |
| Decision 8 | ○ | $14 ○ |
| Decision 9 | ○ | $16 ○ |
| Decision 10 | ○ | $18 ○ |

The survey then asked subjects whether their answers matched their preferences and gave them the chance to fill out the table again if they did not. The example in Figure A.5 is for a subject who switched from the one dollar incentive to the fixed payment at Decision 6 indicating a willingness to pay between eight and ten dollars.

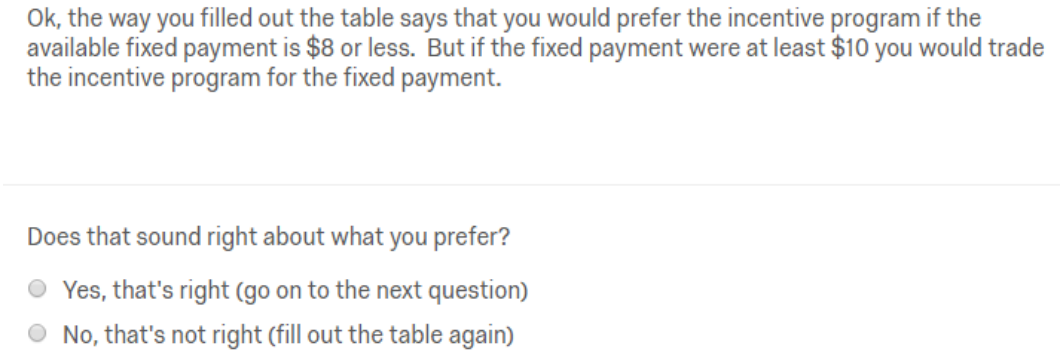



Figure A.5: Comprehension Check for Table

From this point, the survey explained that a slider is a faster way to answer these types of questions, instructed subjects on its use, and asked them to position a slider to indicate their willingness to pay for a $1 dollar per day incentive program that would last 4 weeks. See Figure A.6.

Once the subjects positioned the slider, the survey asked them the two questions shown below to determine whether their answers were consistent with their preferences. See Figure A.7.

If the subject answered correctly, she was taken to instructions for filling out the rest of the willingness to pay section of the survey. If the subject answered incorrectly, she was shown the following explanation (see Figure A.8) and given the change to try again.

If the subject answered correctly on her second try, she was advanced directly to the next set of instructions. If the subject answered incorrectly on her second try, she was given another explanation of the correct answer and then advanced to the instructions. The instructions explained that at the end of the survey, one of the incentive programs would be randomly selected and the subject would either be given that program or a fixed payment with the choice to be determined by the preferences she had indicated on the survey. See Figure A.9. After being presented with some answers to frequently asked questions (see Figure A.10), subjects were instructed to use sliders to

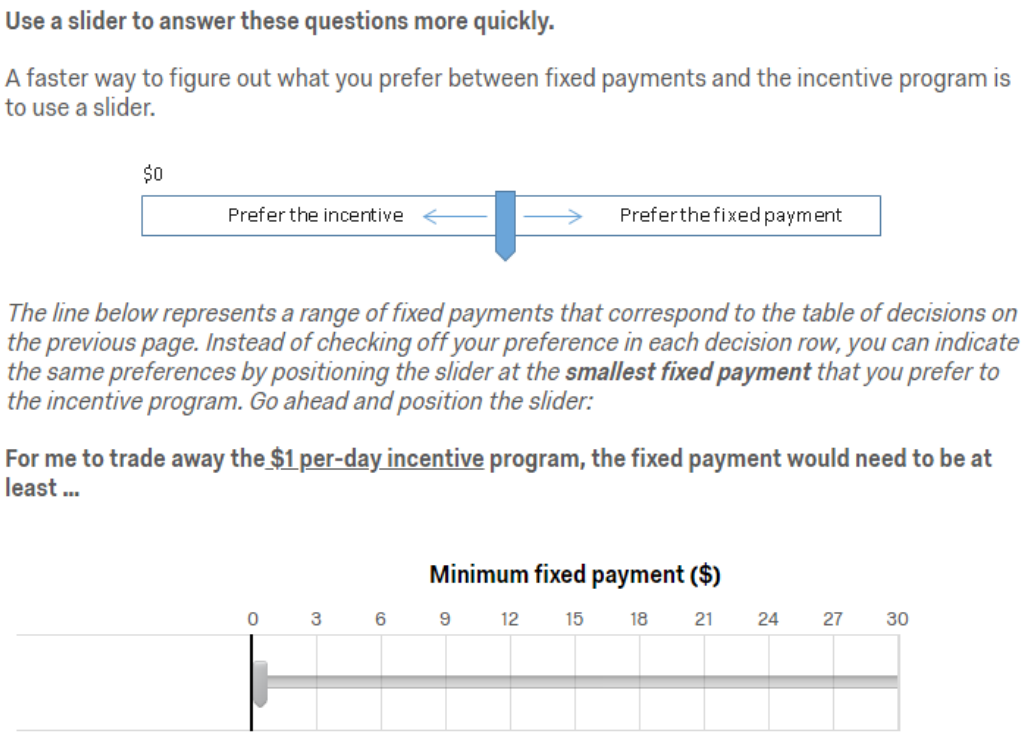



Figure A.6: Slider for WTP for $1 per Day Incentive Program

indicate their attendance projections and willingness to pay for programs paying 1, 2, 3, 5, 7, or 12 dollars per day. See Figures A.11 and A.12. The order of presentation was randomized across subjects.

If subjects positioned the slider on its highest possible value, they were taken to a separate fill in the blank question where they were asked to indicate the smallest fixed amount they would prefer over the incentive program. The example in Figure A.13 is taken from the question that would have been the follow up to the question above for the $1 per-day incentive where the highest possible value on the slider was thirty dollars.

At the end of the survey, an incentive program and fixed payment were randomly drawn for each subject and the survey explained to the subject whether, in accordance with their preferences,

Let's make sure you understand how the slider is working. Suppose we used your answer to the slider above to decide on giving you either the $1 per-day gym-visit incentive or a fixed-payment option.

If the fixed payment option were $5, based on your slider you would prefer to receive:

- The fixed payment of $5
- The $1 per-day gym-visit incentive

Figure A.7: Comprehension Check for Slider

I'm sorry, that's not correct. You put the slider at $${q://QID311/ChoiceNumericEntryValue/1}. So that means you would not be willing to trade the $1 per-day incentive for a fixed payment of less than $${q://QID311/ChoiceNumericEntryValue/1}. But you would trade and take the fixed payment if it were any amount $${q://QID311/ChoiceNumericEntryValue/1} and above.
Let's try one more time.

Figure A.8: Explanation of Incorrect Answer

they would receive the fixed payment or the incentive program. The example in Figure A.14 is for a subject whose choices revealed that she would prefer the fixed payment that was drawn to the incentive program that was drawn.

**How you answer the questions will help determine what you get:**
This study is designed so that it is in your best interest to think carefully about each question and simply tell us what you think and prefer. Each question has the chance to determine what you get from the study.

At the end of the survey you will see a randomly selected incentive program from the set of programs we ask you about. The survey will also randomly select a possible fixed payment that the incentive could be traded for. You will then either keep the incentive program or trade it for the fixed payment depending on which you said you preferred.

For example, suppose a $4 per-day incentive were randomly chosen as your possible incentive and a $10 fixed payment were randomly chosen as your possible fixed payment. The computer would look at your slider for the $4 per-day incentive. If you set the slider at or below $10, you would get the $10 fixed payment. If instead you put the slider higher than $10, you would get the $4 per-day incentive.

Figure A.9: Explanation of Incentive Program Selection

**Frequently asked questions:**

**1) Can I get a better incentive program if I answer questions a certain way?** No. The possible incentive is randomly selected. It is in your best interest to simply answer all questions truthfully based on what you think and prefer.

**2) When will I find out which incentive or fixed payment I get?** This will be shown to you on the last page of the survey.

**3) When will I get the money?** All money from the study will be paid out after the 4-week incentive period is over. You will get a check with your $10 survey payment and either an additional fixed payment or earnings from the incentive program. However, it can take up to another 2 months after that for the check to go through the accounting process for our grant and arrive to you.

**4) Do I have to do something special for the incentive program?** We ask only that you exercise for at least 10 minutes on any day you visit YYY Fitness over the next 4 weeks. To verify that, we have installed a new "check out" scanner by the front door of YYY Fitness. All you need to do to get credit for a visit day with the incentive program is to check in at the YYY front desk, as you normally would, and then swipe your card under the checkout scanner after you are done with at least a 10-minute workout.

**5) Are all possible incentives equally likely?** No. To keep within our grant budget, incentives and fixed payments with lower amounts are more likely to be randomly selected, but every incentive and fixed amount we ask you about has some chance of being selected.

Figure A.10: Frequently Asked Questions

**How often will you go?**

For each incentive we also want to know how often you think you will go to YYY Fitness over the next 4 weeks if you get that incentive program.

Your answers to these questions will not affect which incentive you get. So please simply give us your best realistic estimate of how many days you would attend in the next 4 weeks with that incentive.

**The following pages will ask you about 6 different per-day incentive programs. The possible per-day incentive amounts are $1, $2, $3, $5, $7 and $12. You will see them in a random order.**

There are no right answers -- simply tell us what you think and prefer.

Each of the next 6 pages will be the same except that the incentive amount will vary.

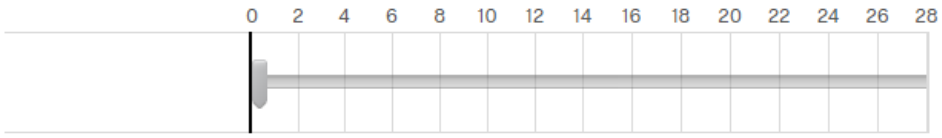
Figure A.11: Instructions for Incentive Program Questions

**Remember:** All incentive programs would cover the next 4 weeks (28 days) starting Monday, ${e://Field/mondaydate}, and all money (incentive program or fixed payment) would be paid after those 4 weeks.

**Recall:** You said earlier that under normal circumstances with no cash reward for going you thought you would visit ${q://QID105/ChoiceNumericEntryValue/1} days in the next 4 weeks.

**$1 per-day** gym-visit incentive.

**Best guess of days I would attend over next four weeks with a $1 per-day incentive.**




For me to trade away the **$1 per-day** gym-visit incentive, the fixed payment would need to be at least...

**Minimum fixed payment ($)**


Figure A.12: Page for $1 Per-Day Gym Visit Incentive

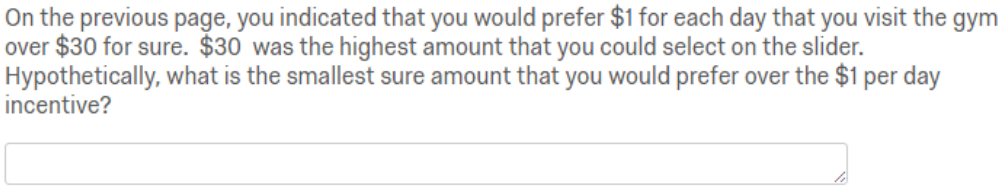



Figure A.13: Fill In the Blank for Off Slider WTP

**End of Survey – Let's see what you get.**

Thank you for taking the survey.

**Possible incentive**: The computer randomly selected $${e://Field/incentive} for each day you visit YYY Fitness over the next 4 weeks as your possible incentive program.

**Possible fixed payment**: The computer randomly selected $${e://Field/fixedpayment} as your possible fixed payment.

**What you get**: According to how you answered the questions, you prefer $${e://Field/fixedpayment} to $${e://Field/incentive} for each day you visit YYY Fitness over the next 4 weeks**. Therefore, in addition to the $10 survey participation payment, you are eligible for $${e://Field/fixedpayment}**.

**When you get it**: Your total payment is $$e{ 10 + e://Field/fixedpayment}. Due to processing, it may take up to 3 months for your check to arrive. You will receive an email confirming these details.

**Click to the next page** to give us the address where we can send your payment.

Figure A.14: End of Survey Announcement of Fixed Payment or Incentive Program