NBER WORKING PAPER SERIES

HOW ARE PREFERENCES FOR COMMITMENT REVEALED?

Mariana Carrera
Heather Royer
Mark Stehr
Justin Sydnor
Dmitry Taubinsky

Working Paper 26161
http://www.nber.org/papers/w26161

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2019, Revised November 2019

We are grateful to seminar and conference participants at Harvard, Wharton, UC San Diego, University of Zurich, Dartmouth, Claremont Graduate University, Erasmus University, Economics Science Association conference, American Society of Health Economists conference, Hebrew University, Stanford Institute for Theoretical Economics, and the Stanford-Berkeley mini conference for helpful comments and suggestions, as well as to Doug Bernheim, Stefano DellaVigna, David Molitor, Matthew Rabin, Gautam Rao, Frank Schilbach, Charles Sprenger, Severine Toussaert, and Jonathan Zinman for helpful comments. Paul Fisher, Chang Lee, Priscila de Oliveira, and Afras Sial 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

How are Preferences For Commitment Revealed?
Mariana Carrera, Heather Royer, Mark Stehr, Justin Sydnor, and Dmitry Taubinsky
NBER Working Paper No. 26161
August 2019, Revised November 2019
JEL No. C9,D9,I12

ABSTRACT

A large literature treats take-up of commitment contracts, in the form of choice-set restrictions or penalties, as a smoking gun for time-inconsistency or self-control problems (for short, "present focus"). This paper develops techniques for inspecting this assumption, presents new field-experimental results that challenge it, and provides an alternative approach to measuring present focus. Theoretically, we show that in a standard model of quasi-hyperbolic discounting, demand for commitment is unlikely when there is some uncertainty about the future. In a field experiment with 1292 members of a fitness facility, we test for the presence of imperfect perception of contract value and experimenter demand effects in commitment contract take-up by offering contracts both for going to the gym more and for going to the gym less. We find that there is significant take-up of both types of contracts, and that individuals who take up contracts to go to the gym more are the most likely to take up contracts to go the gym less. These starkly conflicting choices are inconsistent with all standard models of present focus, and suggest that offers of commitment contracts may not be well-targeted policy tools. However, we show that a combination of belief forecasts and elicitations of willingness-to-pay for piece-rate incentives provides more robust identification of present focus and people's awareness of it. We apply the methodology to our experimental results to obtain estimates of the partially sophisticated quasi-hyperbolic discounting model.

Mariana Carrera
Department of Agricultural Economics
and Economics
Montana State University
P.O. Box 172920
Bozeman, MT 59717
mariana.carrera@montana.edu

Heather Royer
Department of Economics
University of California, Santa Barbara
2127 North Hall
Santa Barbara, CA 93106
and NBER
royer@econ.ucsb.edu

Mark Stehr
Drexel University
LeBow College of Business
Ghall 10th Floor
3220 Market Street
Philadelphia, PA 19104
stehr@drexel.edu

Justin Sydnor
Wisconsin School of Business, ASRMI Department
University of Wisconsin at Madison
975 University Avenue, Room 5287
Madison, WI 53726
and NBER
jsydnor@bus.wisc.edu

Dmitry Taubinsky
University of California, Berkeley
Department of Economics
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
dmitry.taubinsky@berkeley.edu

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; Heidhues and Kőszegi, 2009; Sprenger, 2015). 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]

As shown in Table 1, there are thirty-one empirical studies of commitment contract take-up as of the writing of this paper, spanning domains such as savings, health, and work effort, with all but two written in the last ten years. The take-up rates in these studies are remarkable: on average 35% of participants are willing to choose self-imposed penalties and 46% of participants are willing to restrict their own choice-sets.

Demand for commitment is typically interpreted as "smoking gun" evidence for awareness of present focus: if people voluntarily restrict their choice sets or agree to penalties with no financial upside, 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.[2]

This paper presents new field-experimental results that suggest that take-up of commitment contracts may not actually be a reliable indicator of (partial awareness) of present focus, nor a reliable policy solution. Our new measurement techniques and results suggest that alternative price-based mechanisms may be a more effective approach to both detecting and addressing present focus.

Our work is motivated by two separate strands of prior literature. First, in contrast to the popularity of commitment contracts in the literature, Laibson (2015) has highlighted that pure commitment contracts (i.e. those that offer no possibility of financial gain) may be fundamentally unattractive even to those who are aware they are present focused. Laibson uses numerical simulations 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".[3] This insight is consistent with the fact that there are relatively few examples of pure commitment contracts outside of behavioral economics experiments (Laibson, 2015, 2018; Ericson and Laibson, 2019) and suggests that it is something of a puzzle as to why we see *so much* take-up in behavioral economics experiments.

[1] As an example, a penalty-based commitment contract has the following general structure. An individual places X dollars at stake to reach a future goal. If the individual is unsuccessful, the X dollars is lost and if successful, the X dollars is returned to the individual. On net, the expected monetary returns of such contracts are bounded at zero.

[2] A number of theoretical investigations show how commitment contracts can emerge as optimal policy; see, e.g., Bisin et al. (2015), Lizzeri and Yariv (2017), Beshears et al. (2019), Moser and de Souza e Silva (2019).

[3] Bernheim et al. (2015) also show that commitment contracts interfere with internal self-control strategies, which could make them especially undesirable in environments with sufficient uncertainty.

A possible resolution to this puzzle comes from a second literature: the large and growing body of work in psychology, neuroscience, and economics showing that neural constraints generate a form of "stochastic valuation errors" in the perception of incentives (see Woodford 2019 for a recent review). Such imperfect perception implies that when making a choice about taking up a contract, some individuals might erroneously perceive it as valuable even when it is not. Stochastic valuation errors pose a particularly problematic confound in binary-choice experiments, like the those offering commitment contracts, because errors in binary choice will generically not be "mean-zero" (e.g., Aigner, 1973; Hausman, 2001). On top of this, experimenter demand effects could generate similar problems. Experiments that offer contracts that require some mental effort to assess can thus generate substantial take-up even if the contract would not be valuable for many people.

Our field experiment reveals patterns of commitment contract take-up that are inconsistent with standard theories of present focus, but are consistent with imperfect perception and demand effects. At the same time, we develop experimental techniques that are more robust to perceptual constraints, and deploy them to provide new estimates of present focus.

We begin with a model of quasi-hyperbolic discounting that incorporates uncertainty about the future costs of action. We provide formal theoretical results that generalize Laibson's (2015) numerical results and highlight that modest uncertainty about whether an action will be desirable in the future erodes the value of pure commitment contracts for that action. Our field setting, exercising at a fitness facility, is one where uncertainty about the future desirability of action is likely. Although many people may believe that they use the gym less than they should due to present focus, there is enough potential for future shocks (e.g., work schedules, health and injury, weather) that few could say with certainty that committing to a particular goal would allow them to reach that goal with certainty.

When imperfect perception of contract value is introduced to the model, however, even undesirable contracts can be selected by an agent. The model with imperfect perception makes a series of predictions in the context of gym attendance: (1) people will not only take-up commitment contracts to go to the gym more, but also contracts to go to the gym *less*; (2) the demand for both "more" and "less" contracts will be correlated in within-subjects experiments; and (3) increasing people's awareness of their present focus will *decrease* demand for "more" contracts.

These predictions are borne out in a field experiment conducted with 1,292 members of a fitness facility. In a key component of this experiment, 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 unconditional payment of $80 regardless of how often they used the gym and 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).

However, because we were sensitive to the role that imperfect perception and experimenter demand effects could play in experimental choices, we also gave participants a set of novel commitment options to go to the gym *less*. We asked participants to choose between receiving $80 unconditionally or conditional on going to the gym *fewer* than 8, 12, or 16 times over the next month. We find that 28-34% of participants chose these commitment contracts. This take-up is not concentrated only on participants who think these contracts will not be binding for them: those whose expected attendance in the absence of incentives is well above the contract threshold are almost as likely to take-up the contracts as those below the contract threshold.

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 imperfect perception or experimenter demand effects. We rule out other explanations for our results, such as participants simply confusing the "fewer visits" contracts for the "more visits" contracts, or participants simply disengaging and not taking their decisions seriously.

Despite individuals having imperfect perception of contract value, our results are consistent with prior literature in showing that individuals receiving the contracts do change their behavior. This suggests that simply documenting the effect of commitment contracts on behavior change is insufficient for welfare analysis—individuals who take them up and consequently change their behavior may be individuals who most miscalculated the contract value, rather than individuals who are most in need of additional motivation due to their present focus.

A further prediction of our imperfect perception model is that making people more aware of their own present focus can *decrease* their demand for commitment contracts. In the presence of uncertainty about the costs of exercising in the future, penalty-based commitment contracts appear most harmful to individuals who are more aware of their present focus because they know they will fail the contract more often. It will take larger degrees of valuation errors or demand effects to get these individuals to "mistakenly" choose the commitment contract. In order to investigate this hypothesis, we tested information treatments aimed at debiasing overoptimistic beliefs about using the gym. In the first wave of the experiment, 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 then introduced a second, enhanced information treatment that motivated participants to internalize information on past attendance and also informed them about the overestimation by prior participants. This treatment significantly reduced overestimation of future gym attendance.

Consistent with our theoretical predictions, 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, 2019), 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., 2019; John, 2019). Our results suggest that settings with high take-up of commitment contracts could be settings in which individuals are particularly naive and perceive contract value imperfectly, rather than particularly sophisticated.[4]

These empirical results suggest that inferring present focus from commitment take-up can be problematic, but we introduce a new method for estimating present focus models, which we show is more robust to imperfect perception and demand effects. Building on a design feature first introduced by Acland and Levy (2015), as well as theoretical insights introduced in DellaVigna and Malmendier (2004), the method combines people's forecasts of future attendance with their willingness to pay for piece-rate incentives for attendance.

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, 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 between \$0.83 and \$2.10 to induce an additional gym visit.

We show that in the quasi-hyperbolic model of discounting, these 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.

We find that the enhanced information treatment aimed at debiasing overoptimistic beliefs significantly *increased* our measure of WTP for behavior change. This contrasts with how the debiasing intervention *lowered* the demand for commitment contracts and bolsters the interpretation that at least part of the effect of the treatment on beliefs can be attributed to reducing naivete about present focus.

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 one exception, reporting numerical results from a uniform distribution of task completion costs. We provide general mathematical results for arbitrary probability distributions about task completion costs, and for a range of economic environments including static, dynamic, and continuous choice.

Second, we provide evidence that imperfect perception and demand effects could be an important component of the demand for commitment contracts. Remarkably, only sixteen of the thirty-one studies in Table 1 ever even mention potential confounds, and only eight discuss the confounds in depth as potential drivers of demand.[5] We propose what we think is a simple and direct test that

[4] As we discuss in Section 2.2, particular calibrations of future uncertainty could lead to a non-monotonic relationship between perceived present focus $\tilde{\beta}$ and commitment contract demand (as in Heidhues and Kőszegi, 2009; John, 2019), which is not inconsistent with the empirical result. Bernheim et al. (2015) also predict that external commitment devices could harmfully interfere with internal self-control strategies in environments where there is a need for flexibility.

[5] We coded a study as discussing confounds if it used the keywords *experimenter effects, demand effects, alternative considerations, alternative explanations, confusion, noise, desirability bias, or Hawthorne effects.* Eight discuss such

could be incorporated in most future work: additionally offer participants commitment contracts to do the opposite of the goal activity.

Imperfect perception and demand effects may help explain some puzzling patterns in prior studies of commitment contracts. For example, in an experiment on preferences over menus, Toussaert (2019) finds that only one third of the participants in her experiment are fully consistent with any model of time-inconsistency or costly self-control, and that 25% are at least one violation away from all possible models. Imperfect perception and demand effects may also 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., 2019), or why some studies find that over 90% of participants choose commitment contract thresholds 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 valuation errors or demand effects could lead these people to take up commitment contracts.[6]

Third, we provide estimates of both actual and perceived present focus in a field setting where individuals have had opportunities to learn about themselves.[7] Paserman (2008), Laibson et al. (2018), and Martinez et al. (2018) estimate present focus using observational data, after *assuming* either naivete or sophistication. Augenblick and Rabin (2019) estimate both parameters using a "one-shot" laboratory task. Bai et al. (2018) and Skiba and Tobacman (2018) are closest to our work in that they estimate both present focus and perceived present focus. Different from our methods, their methodology relies on decisions to take up commitment contracts and the timing of loan defaults, respectively. Our methods make use of what we think are the most transparent and straightforward moments generated by models of partially sophisticated present focus: how people's expectations line up with reality, and how much they are willing to pay to change their future selves' behavior.

We conclude the paper with a discussion of implications and suggested guidelines for future work, as well as some implications for policy design with present focus. The results of our experiment should not be taken to mean that take-up of commitment contracts only reflects imperfect perception, or that commitment contracts are never a well-targeted intervention. Rather, our results raise the *possibility* of confounds, and our methodology provides some simple experimental

effects but consider them to be relatively minor determinants of commitment demand, and another eight mention that they may in fact play an important role. For example, Exley and Naecker (2017) discuss demand effects, John (2019) discusses intrahousehold conflict, Brune et al. (2016) discuss the desire to shield savings from one's social network, and Bonein and Denant-Boemont (2015) discuss the role of peer pressure in shaping commitment demand. Sadoff et al. (2019) make the point that most of commitment demand in their study remains unexplained. Kaur et al. (2015) and Schilbach (2019)are two exceptions that have noted that demand effects and imperfect perception could affect take-up.

[6]See also Andreoni et al. (forthcoming) for an analogous finding in an experiment on ex-ante versus ex-post fairness. Andreoni et al. (forthcoming) find that participants most likely to take up commitment to stick to their ex-ante choice are the ones least likely to revise their choice; they suggest experimenter demand effects as an explanation for the bulk of commitment take up.

[7]A larger lab experimental literature estimates present focus but not sophistication about present focus. See, e.g., Sprenger (2015) for a review.

techniques for investigating this potential confound that can be easily incorporated into future work on commitment contracts.

The paper proceeds as follows: Section 2 introduces the theoretical model. Section 3 describes the details of our field experiment design. Section 4 presents the results for take-up of commitment contracts and Section 5 presents the results from the willingness to pay exercise. Section 6 presents the results of the debiasing information intervention. Section 7 concludes.

2 Theoretical predictions and measurement techniques

In this section we begin by setting up and characterizing the predictions of a model of quasi-hyperbolic discounters facing a task with stochastic immediate costs and deterministic delayed benefits. We then augment this model to incorporate imperfect perception and experimenter demand effects. We summarize the new implications of imperfect perception and demand effects in Section 2.3, and derive a more robust test of awareness of present focus in Section 2.4.

To keep the exposition as intuitive and concise as possible, we mostly summarize the intuition, and state formal results in the appendix. In Appendix B we provide a step-by-step instructional exposition of the results of a two-period model (one period for decision-making and one period for action), which appeared in an earlier (and longer) version of this section. In Appendix D we then generalize this to the general case of many periods. In Appendix E we also show that the results generalize to continuous choice[8] and to alternative models of limited self-control.

2.1 Quasi-hyperbolic preferences

We consider individuals who in periods $t = 1, \ldots, T$ have the option to take an action $a_t \in \{0, 1\}$. This action generates immediate stochastic costs c realized in period t as well as deterministic delayed benefits b realized in period $T + 1$. We only require that $c > 0$ with positive probability; we do not preclude the possibility that on some "good days" individuals actually find immediate pleasure in activities with delayed benefits such as exercise.

In period 0 individuals choose between contracts (y, P_0, P_1) that consist of a fixed transfer y (which could be negative), a contingent reward $P_0\left(\sum_{t=1}^{T} a_t\right) \geq 0$ that is decreasing in $\sum_{t=1}^{T} a_t$ and a contingent reward $P_1\left(\sum_{t=1}^{T} a_t\right) \geq 0$ that is increasing in $\sum_{t=1}^{T} a_t$. We normalize so that $P_0(T) = P_1(0) = 0$. All financial payments are received period $T + 1$. We label the different components of the contract the way that we do because this formulation is particularly convenient for formalizing the role of imperfect perception in Section 2.3.

To illustrate the possible contracts, if a_t represents going to the gym and $x = \sum_{t=1}^{T} a_t$ is total attendances, then a contract $(-p, 0, P_1)$, with $P_1(x) = p \cdot \mathbf{1}_{x \geq r}$, is a penalty-based commitment contract for attending the gym at least r times: the individual loses p unless she goes to gym at

[8] With continuous choice, some minimal restrictions on behavior are typically desired, as established by Amador et al. (2006) and others. However, this work does not establish how significant optimal restrictions can be in the presence of uncertainty; in particular, all but the mildest restrictions could be suboptimal in the presence of uncertainty. We provide results about how much uncertainty it takes to erode demand for a given penalty contract.

least r times. Conversely, a contract $(-p, P_0, 0)$, with $P_0(x) = p \cdot \mathbf{1}_{x<r}$, is a penalty-based contract for *not* going to the gym r times or more. In Section 2.4 we consider piece-rate incentive contracts such as $(0, 0, P_1)$, with $P_1(x) = px$, but for Sections 2.2 and 2.3 we focus on penalty-based commitment contracts with no financial upside, as well as choice-set restrictions that can be thought of as the limit case of infinite penalties.

We assume that individuals 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 construction, $u_t = -a_t \cdot c$ for $1 \leq t \leq T$ and $u_{T+1} = y + b \sum_{t=1}^{T} a_t + P_0 \left(\sum_{t=1}^{T} a_t \right) + P_1 \left(\sum_{t=1}^{T} a_t \right)$. Following O'Donoghue and Rabin (2001), we allow individuals to mispredict their preferences: in period t, they believe that their period $t+1$ self will have a short-run discount factor $\tilde{\beta} \in [\beta, 1]$. For simplicity, we set $\delta = 1$. We use $V(y, P_0, P_1)$ to denote the individual's subjective expectation (give beliefs $\tilde{\beta}$) about utility under contract (y, P_0, P_1).

2.2 With uncertainty about costs, quasi-hyperbolic preferences rarely generate a demand for commitment

In this section we argue that the standard quasi-hyperbolic model rarely predicts a demand for commitment, except in special cases with minimal uncertainty about the costs c.

To begin, note that commitment contracts for choosing $a_t = 1$ more often will be desired when $\tilde{\beta} < 1$ and there is little uncertainty about the action $a_t = 1$ being desirable from the period $t = 0$ perspective. For example, suppose that the costs c are always smaller than the delayed benefits b, but that the individual thinks that because of present focus she may sometimes choose $a = 0$.

More generally, when there is only a small chance that immediate costs will exceed the delayed benefits, individuals with $\tilde{\beta} < 1$ will want penalty-based contracts as long as $\tilde{\beta}$ is not too low. If $\tilde{\beta}$ is too low, then the penalties will lead to financial losses that are too large in magnitude relative to the desired behavior change. This line of logic can be used to to establish that when there is a small chance that costs exceed benefits, there will be demand for commitment by some individuals, and it will be non-monotonic in $\tilde{\beta}$. This result is formally established in Appendix B, and is analogous to the results of Heidhues and Kőszegi (2009) and John (2019).

However, such results about (non-monotonic) demand for commitment depend on strong assumptions about how much uncertainty there is about the costs of doing the action. As a simple illustration for the case of $T = 1$, Figure 1 summarizes 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 $b + (1 - \tilde{\beta})b \leq 1$.[9]

As an example of the numerical results, suppose that the benefit is 0.8, which implies that the costs exceed the benefit only 20% of the time and never by more than 25% in magnitude. Although this is an arguably modest amount of uncertainty, in this case, an individual with $\tilde{\beta} = 0.8$ would

[9] 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 are specific to the uniform distribution.

never have demand for any kind of commitment contract.

Such results are not special to the uniform distribution. Since particularly high draws of c are what make commitment contracts particularly costly, the thin-tailed uniform distribution overstates the amount of uncertainty it would take to erode demand for commitment. In Appendices B and D we provide formal results that generalize the uniform distribution example to a broad class of other distributions, and show that that the standard quasi-hyperbolic model predicts that there should not be demand for commitment when there is at least a moderate chance that costs exceed delayed benefits.

There are two key conditions on the distribution of cost draws under which demand for commitment is eroded. First, the chances of getting a cost draw under which it is suboptimal to take the action ($c > b$) are at least as high as the chances of getting a cost draw under which the time $t = 0$ individual thinks she should choose $a = 1$ but thinks that her time $t = 1$ self will not do so ($c \in [\tilde{\beta}b, b]$). Second the cost draws exceeding b are not all concentrated at a point only slightly higher than b—more formally, this translates to the density function of c not decreasing faster than $1/c^2$ in the region $[b, b + (1 - \tilde{\beta})b]$. The condition that the density does not decrease faster than $1/c^2$ is satisfied by the uniform distribution as well as by most other distributions that put sufficient mass on $[b, b + (1 - \tilde{\beta})b]$.

2.3 Commitment take-up with imperfect perception and demand effects

In light of these results, we reexamine why individuals choose commitment contracts. One possibility is that there is limited uncertainty regarding future costs of the activity. Another possibility is that individuals do not behave according to the standard quasi-hyperbolic model we have presented in Section 2.1.

Because evaluating incentives for future behavior is complicated, individuals may do so imperfectly, based on imperfect perception of the value of commitment contracts. This is in line with a long intellectual history of measuring and modeling stochastic valuation errors in individuals' decisions, starting from Block and Marschak (1960), continuing with Quantal Response Equilibrium (McKelvey and Palfrey, 1995), and recently gaining prominence in a variety of new approaches to bounded rationality (e.g., Woodford, 2012; Wei and Stocker, 2015; Khaw et al., 2017; Natenzon, 2019). We refer to this mechanism as imperfect perception.

Another reason is that some individuals simply like to say "yes" to offers, feel pressure to do so (DellaVigna et al., 2012), or falsely trust that the authority offering the contracts must be offering something valuable We refer to this possibility as "demand effects."

Formally, we suppose that for a given decision j, individual i behaves as if her expected utility under contract (y, P_0, P_1) is

$$\widehat{V}(y, P_0, P_1) = \beta y + \varepsilon_{ij} V(0, P_0, P_1) + \eta_i \mathbf{1}_{(P_0,P_1)\neq 0} \tag{1}$$

where $\mathbf{1}_{(P_0,P_1)\neq 0}$ is an indicator that at least some contingent incentives are involved. The ε_{ij} term captures "stochastic valuation" leading to imperfect perception of contract value, which we assume does not affect the certain incentive y.[10] The η_i term captures experimenter demand effects. We model this term as additive to reflect the common intuition that social motives such as social desirability bias have a smaller percentage effect at larger stakes. We also allow for the possibility $\eta_i < 0$ to capture "reactance effects": doing the opposite of what one perceives the authority wants. For simplicity, we assume that η_i and ε_{ij} are unrelated to β_i and $\tilde{\beta}_i$.

For short, we refer to the choice implied by (1) as the *imperfect perception model.* This reduced-form model, which is most similar to the Quantal Response Equilibrium models, is not consistent with all types of seemingly irrational choices. The individual would never turn down a contract with only financial upsides that incentivizes choosing $a_t = 1$; that is, a commitment contract of the form $(y, 0, P_1)$ with $y \geq 0$. 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."

As we show formally in Appendix B and D, the imperfect perception model generates three intuitive predictions, which we state here in the exercise context for concreteness. The commitment contracts we consider here are threshold penalty contracts—individuals lose an amount p if they do not meet a certain threshold for attendance ("more exercise" contracts) or they lose an amount p if they do exceed a certain threshold for attendance ("less exercise" contracts).

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

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.

3. Increasing individuals' sophistication about their present focus will decrease their demand for commitment contracts to exercise more.[11]

The intuition for the first prediction is that imperfect perception or demand effects can lead individuals to choose undesirable contracts.

The second prediction can result from two different mechanisms. 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 contracts, 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 imperfect perception, then individuals with the highest variance in the stochastic valuation term ε will be the most likely to take-up both types of contracts.

[10]We model error terms ε_{ij} as multiplicative to reflect that, setting aside social motives, individuals are likely to have fairly accurate valuations of contracts in which the contingent incentives are small, and are not likely to perceive contracts with no financial downside as harmful. But, our results hold for any type of mean zero errors around V, including errors that are more substantial at smaller stakes (because of, e.g., mental effort responding to stakes). Formally, we just need $E\left[\widehat{V}(r, p_0, p_1)\right] = V(r, p_0, p_1) + \eta_i \mathbf{1}_{(p_0,p_1)\neq 0}$.

[11]Interestingly, the converse does not hold for commitment contracts for $a = 0$. That is, it does not hold that the likelihood of choosing a commitment contract for $a = 0$ is decreasing in $\tilde{\beta}$. 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.

The intuition for the third prediction is that under moderate to large uncertainty, the perceived harms of a commitment contract are decreasing in $\tilde{\beta}$ in the standard quasi-hyperbolic model without imperfect perception (see Appendix B and D). Intuitively, the less well-behaved an individual thinks he is, the more likely he thinks he is to incur the commitment contract penalty, and with at least moderate uncertainty, that trumps the potential benefits of behavior change induced by the contract. Although in the standard quasi-hyperbolic model these conditions would lead individuals to never choose a commitment contract, in our imperfect perception model individuals still choose the contract, but with a propensity that is decreasing in the expected harms in the standard model.

2.4 More robust tests of demand for behavior change

Although our imperfect perception model implies that commitment contract demand could be a poor indicator of actual awareness of self-control problems, we provide alternative techniques for testing whether the average perceived present focus, $E[\tilde{\beta}_i]$, is less than one.

Consider an incentive contract that pays the agent p every time she chooses $a = 1$, $\left(0, 0, p\sum_{t=1}^{T} a_t\right)$. Define an individual's willingness to pay for the contract, $w_i(p)$, to be the smallest y such that the individual prefers $(y, 0, 0)$ over $\left(0, 0, p\sum_{t=1}^{T} a_t\right)$. Define $\alpha_i(p)$ to be individual i's expectation of $\sum_{t=1}^{T} a_t$ under the contract. In Appendix A we establish the following result:

Proposition 1. *Assume that the costs in each period t are independently distributed according to smooth density functions. Then*

$$\frac{d}{dx}E[w_i(x)]|_{x=p} = E\left[\alpha_i(p) + (b_i + p)(1 - \tilde{\beta}_i)\alpha_i'(p)\right]. \tag{2}$$

for all $p > 0$. If additionally the terms $\left\{(\Delta p)^n \frac{d^m}{dp^m} E[\alpha_i(p)]\right\}_{\{n\geq 2, m\geq 2\}}$ are negligible, and if $E[\eta_i] = 0$ or $p > 0$, then

$$E[w(p+\Delta p) - w(p)] \approx (\Delta p)\frac{E[\alpha_i(p+\Delta p)] + E[\alpha_i(p)]}{2} + E\left[(b_i + p + \Delta p/2)(1 - \tilde{\beta}_i)(\alpha_i(p+\Delta p) - \alpha_i(p))\right] \tag{3}$$

The assumptions about negligible terms that produce (3) are essentially the same as those in the canonical Harberger (1964) formula of the deadweight loss of taxation: the change in incentives is not too large and curvature of the behavior response is negligible in the region of incentive change.

To obtain intuition for the proposition, consider first the case in which $\tilde{\beta}_i = 1$ for all i. Then, Proposition 1 states that on average, time-consistent (or fully naive) individuals should value a marginal \$1 increase in the contingent reward p by approximately the average of their expectations of choosing $a_t = 1$. For example, if an individual perceives that he will choose $a_t = 1$ seven times in expectation, then a marginal \$1 increase should be worth approximately \$7.00 to this individual. This condition is an immediate corollary of the Envelope Theorem when utility is quasilinear in money, and does not depend on any specific assumptions of our model—it holds as long as expected behavior is a differentiable function of incentives.

Valuations that are in excess of the $\tilde{\beta}_i \equiv 1$ benchmark are due to a demand for behavior change that results from $\tilde{\beta}_i < 1$. This demand for behavior change depends on (i) the perceived "internality" from present focus, $(b_i + p + \Delta p/2)(1 - \tilde{\beta}_i)$ and (ii) the degree to which the incentive is perceived to change behavior, $\alpha_i(p + \Delta p) - \alpha_i(p)$.

This formula allows individuals' valuations to be stochastic and subject to demand effects. The model in Section 2.3 implies that the difference in WTP for two different incentive levels, $w(p + \Delta p) - w(p)$, is an unbiased estimate of how much an individual would value the incentive increase in the absence of demand effects or imperfect perception (when $p > 0$). Intuitively, the mean-zero decision error generated by ε_{ij} translates to mean-zero decision error in people's willingness to pay for a Δp increase in piece-rate incentives, $w(p + \Delta p) - w(p)$. The fixed demand effects η_i are differenced out from $w(p + \Delta p) - w(p)$ when $p > 0$. The formula also 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, as long as $\hat{\alpha}_i(p)$ is an unbiased estimate of $\alpha(p)$ for $p > 0$.[12]

Equation (3) motivates our measure of "per-dollar willingness to pay for behavior change" that we utilize in our empirical analysis:

$$\frac{E[w(p + \Delta p) - w(p)]}{\Delta p} - \frac{E[\alpha_i(p + \Delta p)] + E[\alpha_i(p)]}{2} \tag{4}$$

Core to this result is that the range of WTP can range from below to above expected earnings, meaning that the measure of WTP for behavior change can range from negative to positive.[13] Restricting WTP for a *commitment contract,* as in Milkman et al. (2014), would mechanically lead to an upward bias in valuations, since negative draws of errors in valuation would be censored at 0 while positive draws of errors would be uncensored. 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.

Variations of our imperfect perception model in which valuation errors not mean-zero, or in which demand effects rise with stakes, would invalidate the methodology we propose here (along with using commitment demand as a measurement tool). Our claim is that our methodology is *more* robust than using commitment demand as a measurement tool—but it is, of course, not perfect.

Quasilinearity in money is also an important assumption for identification of $\tilde{\beta}$, and is plausible for the relatively modest incentive sizes that are offered in field experiments such as ours.[14] If participants are non-negligibly risk averse over small amounts of money, then this methodology underestimates the WTP for behavior change, and overestimates $\tilde{\beta}$. Allcott et al. (2019b) extend

[12] Systematic over-statement of true beliefs such that $E[\hat{\alpha}_i(p)] > E[\alpha_i(p)]$ would make this a particularly conservative test, as this would bias against us finding a demand for behavior change.

[13] Note that even though our experiment imposed a lower bound of $0 for WTP for a piece-rate incentive, the multiplicative nature of errors in our model imply that the perceived valuations for a piece-rate incentive cannot be below zero. Intuitively, individuals should not perceive the value of a positive piece-rate incentive as negative.

[14] Consistent with this, we do not find any relationship between WTP for the piece-rate incentives and our elicited measure of risk aversion.

our methodology to a richer consumption-savings setting with interdependent payoffs and infinite time horizons, and provide a generalization that allows estimates of $\tilde{\beta}$ in the presence of risk aversion and income effects.

3 Experimental design and sample for analysis

3.1 Design

Our study recruited members of a 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. The university has a separate facility for undergraduate students.

Members of the facility were recruited to participate in a study that consisted of an online component followed by four weeks of observation of gym attendance. The online component elicited beliefs about gym attendance and preferences over contingent incentives for using the gym during the following four weeks. Members were assigned to attendance incentives at the end of the online component. 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]

The study was conducted in three waves. 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.

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

[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 online component 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.

they thought they should have gone, but did not. Next, participants were assigned to receive an information treatment with 50% chance.

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 2(a)). Participants were required to confirm whether they could see the graph in order to proceed to the next page. 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. 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 2(c)).

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 online component), 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) incentive contracts 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 online component. The incentives were \$1/day, \$2/day, \$3/day, \$5/day, \$7/day, and \$12/day. Each incentive was presented on a separate page, and the order of these pages was randomized.

Participants were first 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. Importantly, this WTP could be as low as \$0 and thus substantially below the expected earnings from the incentive. If participants indicated the maximum WTP allowed by the slider (i.e., positioned it all the way to the right), they were taken to a fill-in-the-blank question where they entered their willingness to pay.[16] All payments, both fixed and contingent, were to be paid out after the four-week period.[17]

[16] The minimum value on each slider was zero, and the maximum was the value of the per day incentive multiplied by 30 so as to include (slightly more than) the maximum possible expected earnings. 7.4% of responses were at the slider maximum. Of the subsequent fill in the blank responses, half indicated a willingness to pay that was actually below the maximum, 23% indicated a willingness to pay equal to the maximum, and 27% indicated a willingness to pay that was above the maximum.

[17] 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

The WTP elicitation used the incentive-compatible Becker-DeGroot-Marschak (BDM) mechanism: 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. If their WTP was above the randomly chosen fixed payment, they would receive the piece-rate incentive. If their WTP was below the randomly chosen fixed payment, they would receive the randomly chosen fixed payment.

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 online component. See Appendix I for 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). The "more" and the "fewer" contract choices were presented on separate pages, with the order randomized.

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 online component 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.[18]

3.1.4 Assignment of incentives

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

commitment contract take-up questions to examine the β 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 importance of imperfect perception and demand effects, if any, which we did not have strong priors about at the initial stage.

[18] After observing the surprising patterns in commitment demand in Wave 1 (i.e., many participants chose both "fewer" and "more" contracts), 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 elicited beliefs about the likelihood of meeting the thresholds stipulated by the "more" and "fewer" contracts to rule out some alternative hypotheses not consistent with the model we propose in Section 2.3.

was to avoid the endogenous assignment of incentives to participants with higher WTPs for those incentives.

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). 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 5.3, 6.1 and 6.4, which rely on exogenous assignment of incentives.

We targeted a small number of questions with high probabilities of selection in order to power our comparisons of the incentive effects. In Wave 1, the questions about the \$2 and \$7 piece-rate incentives were each assigned a 0.33 probability of being chosen. To create a group that did not face any incentive to visit the gym, the study also included a choice between a \$0 per day 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.[19]

The targeted incentives were varied in order to document the effects of different incentive sizes.[20] 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. This shift resulted in the following incentive assignment probabilities: 33% for the \$0 incentive; 33% for the \$2 incentive; 16.5% for the \$5 incentive; 16.5% for the \$7 incentive.

In Wave 3, we added a group that would receive \$80 conditional on making 12 or more visits, an attendance 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."[21] 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% for the \$0 incentive, and 33% for the \$7 piece-rate incentive, to allow us to compare their effects.[22]

[19]We informed the participants about this randomization scheme in the instructions by clarifying "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."

[20]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.

[21]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.

[22]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

Although this variation of incentive scheme assignments across waves is not ideal for the analysis in Section 5.3, we find that the participant pools look similar across waves, as shown in Appendix F. We further include wave fixed-effects in relevant analysis below to account for the changing nature of our survey across waves. The ex-post incentive assignment does not affect our other analyses.

3.1.5 Demographics and other questions

The next section of the online component collected demographics and checked numeracy.

3.1.6 Announcement of incentives

In the final section of the online component, participants learned which incentive, if any, they would receive in the next four weeks. Participants received an email upon completion of the online component that confirmed the incentive they were eligible for and explained that the four week incentive period would begin on the Monday after they completed the online component. 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.

3.1.7 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. 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 (no information) 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% of the participants are either part or full time students, 57% work either part or full time, 27% 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

ex-post).

same period.[23] 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. 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 except for 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 expected that one would be significant at the 5% level.

Compared to 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.[24] Given this high numeracy, it does not seem likely that our sample is more susceptible to imperfect perception than the typical sample in commitment contract field experiments.

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. In Appendix G.2 and G.3, we show that the results of this section replicate for participants in the information treatment groups.

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

Table 3 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 present focus, combined with a presumably sensible desire to avoid thresholds that are too demanding.

The take-up rates in our study are comparable to the take-up rates in other studies. As Table 1 shows, while take-up rates are lower for studies that require participants to put their own money at stake, take-up rates are much higher for studies like ours that feature "house money" or other currency like course grade points. Overall, the take-up rates for penalty-based contracts that do

[23] 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.

[24] 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 percentages answering each question correctly were 83.5% (the percentage calculation) and 56% (the lottery division) (Lusardi and Mitchell, 2007).

not require participants to put up their own money range from 36% to 73%, with an average of 44% for house money. The take-up rates for contracts that feature removal of options that do not affect transactions with own money are even higher. Most similar to our contract options, Schilbach (2019) also offers participants a choice between money for sure versus the same amount of money only if participants stay sober, and finds take-up rates ranging from 37% to 55% for the commitment option.

However, column (2) of Table 3 shows that a substantial fraction of this take-up may be due to imperfect perception or demand effects—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.

However, the imperfect perception 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 types of contracts at the same threshold. Our within-subject design allows us to examine this prediction. Columns (3) and (4) in the table show the shares of participants selecting each type of contract 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 selected "more visits" contracts at each threshold, nearly half also selected the "fewer visits" contract at the same threshold. Choosing both contracts at the same threshold is inconsistent with decisions driven by awareness of present focus, and thus a strong indicator that imperfect perception or demand effects are prevalent in commitment contract take-up.

Moreover, an even stronger prediction of the imperfect perception model is that not only will some participants select both types of contracts, but that there will be a positive correlation in the take-up of the two types of contracts. 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.

While these results suggest valuation errors and demand effects, they do not imply that all take-up of commitment contracts is explained by these confounds. Just over half of the participants who selected "more visits" commitments at each threshold did not select the fewer visits contracts and conversely for participants who selected "fewer visits" contracts. These patterns could be consistent with some participants truly wanting to commit to attending the gym more, and some participants wanting to commit to attending the gym less.

Imperfect perception and demand effects do not imply that the commitment contracts do not affect behavior once they are assigned to the individuals. Consistent with the existing literature we find a substantial effect 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. 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. At the same time, and also consistent with prior work, we find that a substantial fraction of participants who took up the contract subsequently failed to reach the target (36%).

Although these effects on behavior change are often interpreted as the commitment contracts "working," our results suggest that the effects on behavior change should not be equated with standard welfare or efficiency metrics. To the extent that imperfect perception or experimenter demand effects are a substantial driver of take-up, the contracts are not taken up by the individuals whose whose attendance is below their individual optima.

4.2 Robustness

4.2.1 Participants don't confuse "fewer visits" for "more visits" contracts

Although the reported patterns of behavior are consistent with the imperfect perception model in Section 2.3, 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 "fewer visits" contract underlined "11 or fewer."

Another strategy for assessing whether asymmetric errors can fully account for the observed patterns in take-up of "more visits" and "fewer visits" contracts 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 4 shows that participants differentiated between questions about perceived likelihood of success in a "more" contract versus a "fewer" contract.[25] 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. Moreover, the positive and negative coefficients are not identified off of different subgroups: when restricting to the subgroup who both chose more and fewer contracts, the coefficients remain essentially unchanged. This implies that at least in answering the forecasting questions, participants were not simply misreading the "fewer" contract to be the "more" contract.

In Table 5 we then look at how the perceived likelihood of success correlates with actual take-up. As columns (1)-(3) of Table 5 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

[25] 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.

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 imperfect perception model, which predicts that the more damaging the contracts would appear in the absence of imperfect perception or demand effects, the less likely they would be chosen.

In Appendix G.3 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.

4.2.2 Results are not a consequence of disengagement from the study

In Section 3.1.4 we described two 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 attend the gym 12 or more times. These questions also serve to identify participants who may have been inattentive to the study instructions, or simply decided to click through the study at random. Only 5 chose \$0 over the \$80 contingent incentive, and 13 chose \$0 over a \$20 fixed payment. This constitutes 1.4% of our participants, a proportion that is far smaller than the fraction of individuals choosing the "fewer" contracts. This is consistent with our modeling of imperfect perception, which predicts frequent choice of "fewer" contracts but no choice of dominated options such as these. While the propensity for random choice is rarely tested in field experiments, including experiments on commitment contracts, the 1.4% number compares very favorably to validity checks in all online and lab experiments that we are aware of.

In addition to this direct test for disengaged random choice, we further included an attention check and a comprehension check for the WTP elicitations (see questions in Appendix I). 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. We find that 49 participants failed the attention check and 61 failed the comprehension check about the WTP elicitation twice. To our knowledge, these comprehension and attention rates compare favorably to all experiments that we are aware of, even though these kinds of checks are rarely included in commitment contract field experiments (indeed, the second one concerns a task that is unrelated to commitment contract choice).

Excluding participants who failed a comprehension check or attention check or chose a dominated option lowers overall demand for the "fewer" contracts from 32% to 31%, and has no effect on demand for the "more" contracts. While these tests cannot be guaranteed to identify all individuals who disengaged or misunderstood some portion of the study, the lack of correlation between these markers and demand for commitment contracts implies that disengagement or misunderstanding is very unlikely to drive any portion of our results.

Taken together, the results strongly suggest that our findings are not induced by random choices of disengaged or confused participants, but rather by the deeper forms of imperfect perception or demand effects proposed in Section 2.3. We infer from this that imperfect perception and demand effects likely influence the take-up not only of the novel fewer contracts but also the more widely used more contracts.

4.2.3 Results are not driven by participants for whom the contracts are not binding

Because our commitment contract offers are only weakly financially dominated, one concern might be that some of our take-up is driven by individuals for whom the contracts are not really binding. For example, individuals who choose the 11 or fewer contract could be individuals who would already attend the gym 11 or fewer times in the absence of any discouragement. Such patterns of choice appear to be prevalent in some studies, such as Augenblick et al. (2015), who find that there is demand for commitments when the contracts have no price but that dramatically fewer people choose them when there is even a small positive cost. However, other studies, such as Schilbach (2019) do not find this phenomenon.

In our data it does not appear that much of the take-up is driven by individuals for whom the contracts would be inconsequential. As shown in Appendix G.1, individuals whose expected attendance exceeds the "fewer" threshold by 2 or 4 times are nearly as likely to select the "fewer visits" contracts as the full sample. The same pattern holds for the "more visits" contracts. Perhaps most importantly, the positive correlation between take-up of "more" and "fewer" contracts remains unchanged when restricting to a subset of participants for whom either the "more" or the "fewer" contract would be at least moderately binding.

At the same time, results such as those of Augenblick et al. (2015) could be an indicator of the kind of imperfect perception that we model in Section 2.3. If commitment contracts offer approximately no value to individuals *and* if the variance of the stochastic valuation error term ε is small, then many individuals would choose commitment contracts at no price, but would abruptly stop choosing them at a positive price. However, if the variance of the stochastic valuation error term ε is non-negligible, then commitment contract demand would decline much less rapidly with the price. Thus, while results such as those of Augenblick et al. (2015) are consistent with imperfect perception of contract value, they by no means constitute a thorough test.

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 utilize the more robust methodology described in Section 2.4 to provide evidence for awareness of limited self-control, as well as estimates of present focus and naivete under additional assumptions.

5.1 There is significant willingness to pay for behavior change

As shown in Proposition 1, a combination of willingness to pay (WTP) for piece-rate incentives as well as beliefs data provides an alternative indicator of awareness of present focus.

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 the average subjective expected earnings at that piece rate (i.e., the piece rate multiplied by the participants' expected number of visits). The WTP is above participants' subjective expected earnings for low incentives, where the subjective earnings roughly correspond to WTP from the $\tilde{\beta} = 1$ benchmark. 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. Consistent with the implication of equation (2), the WTP is below expected earnings for high incentives.[26]

To create our aggregate measure of WTP for behavior change, we follow Proposition 1 to construct our measure of per-dollar WTP for behavior change proposed in equation (4). If participants are time-consistent, then the envelope theorem implies that under any set of assumptions that lead expected earnings to be a smooth function of the incentive level, the value of a marginal increase in the incentive level equals the increase in earnings generated by the increase in the incentive level. To create our measure of WTP for behavior change, we contrast with how aggregate WTP changes relative to that benchmark. In that sense, our measure is a "model-free" statistic that captures deviations from the standard time-consistent model.

Formally, in equation (4) the per-dollar willingness to pay for behavior change when moving from one incentive level p_k to the next higher incentive level 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. In other words, this is the deviation from the value of $WTP(p_{k+1}) - WTP(p_k)$ that would be implied by the $\tilde{\beta} = 1$ benchmark.[27] 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 then take the (unweighted) average across all participants and all incentive pairs.

We take the average, rather than analyze individual differences, because according to Proposition 1, the average statistic is the unbiased measure of mean willingness to pay for behavior change in the presence of imperfect perception. Consistent with our conjecture of imperfect perception of contract values, we find substantial variation in these valuation measures at the individual level.[28]

Figure 5 shows the average value across six incentive levels, as well as the average excluding the

[26]To see this formally, note that the derivative of expected earnings with respect to the incentive level p is given by $E[\alpha_i(p) + \alpha_i'(p)]$. Thus as long as $E[(b_i + p)(1 - \tilde{\beta}_i)] < 1$, which will be the case for moderate levels of perceived present focus, $\frac{d}{dp}E[w_i(p)] < E[\alpha_i(p) + \alpha_i'(p)]$.

[27]For example, the WTP for behavior change when moving from the \$7 to the \$12 incentive is the WTP for the \$12 incentive minus the WTP for the \$7 incentive divided by five, minus the average of the expected days under the \$12 incentive and the expected days under the \$7 incentive.

[28]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 valuation errors in participants' decisions about willingness to pay and/or their estimates of visit rates.

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 when moving from no incentive to the \$1 incentive. As Proposition 1 shows, if there are fixed demand effects (η_i, in the notation of that section) 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 per attendance when we include WTP for the \$1/visit incentive and \$0.83 per attendance using the more conservative estimate that excludes WTP for the \$1/visit incentive.

5.2 Willingness to pay for behavior change does not correlate with commitment contract take-up

Having established a positive 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 on average participants who are more likely to take up the contracts have a stronger desire to change their behavior.

Table 6 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 respect to incentives. We find no significant correlations between the WTP values and commitment contract take-up. The point estimates are close to zero when using the measure of average WTP across all incentive levels, and are slightly negative when using the measure that excludes the \$1 incentive level. According to these estimates, a one standard deviation increase in WTP for behavior change reduces the take-up of commitment contracts by up to 3 percentage points.[29]

We provide some evidence that imprecision of our measures is unlikely to fully explain this lack of correlation by examining the within-person correlation in WTP measures and the within-person correlation in uptake of different commitment contracts. We construct pairwise correlations between

[29] Under our assumption that the valuation error term ε_{ij} is independent across different decisions j for a given individual i, the error term cannot bias the covariance between these two measures upward or downward. The fixed demand effect term η_i is eliminated from WTP measures that exclude the \$1 incentive, and thus also cannot bias the covariance between these two measures provided that the WTP for the \$1 incentive is excluded.

(individual level) estimates of WTP for behavior change at each different piece-rate incentive level (e.g., correlation between WTP for behavior change at a \$1 incentive and WTP for behavior change at a \$2 incentive). We also construct 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). We estimate that the average pairwise correlation of our measures of WTP for behavior change is 0.18 (bootstrapped cluster-robust s.e. 0.06) 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 unpredictable that it is unassociated with demand for a different "more" contract.

5.3 Parameter estimates

Our results so far provide evidence of some demand for behavior change, and thus that at least some participants must have $\tilde{\beta} < 1$. Here, we use these findings as sufficient statistics to derive parameters estimates.

5.3.1 Estimates of $\tilde{\beta}$

We can use our model-free measures of willingness to pay for piece-rate incentives to provide esimtates of $\tilde{\beta}$ for a given value of delayed health benefits b. We use Proposition 1 and make the additional assumption that b and $\tilde{\beta}$ are homogeneous across individuals (but not necessarily the distribution of costs). Under these assumptions, equation (3) 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.

If individuals are heterogeneous in $\tilde{\beta}_i$, then our estimates in this section roughly correspond to the average, $E[\tilde{\beta}_i]$. Our methods obtain an unbiased estimate of $E[\tilde{\beta}_i]$ when

$$E\left[(b_i + p + \Delta p/2)(1-\tilde{\beta}_i)(\alpha_i(p+\Delta p) - \alpha_i(p))\right] = E[b_i+p+\Delta p/2]E[1-\tilde{\beta}_i]E[(\alpha_i(p+\Delta p) - \alpha_i(p)].$$

where Δp is the increase in the piece-rate p. In other words, when there is no correlation between delayed benefits b_i, perceived present focus $\tilde{\beta}_i$, and predicted changes in attendance $\alpha_i(p+\Delta p) - \alpha_i(p)$. Positive correlations lead us to overestimate $E[\tilde{\beta}_i]$ while negative correlations lead us to underestimate $E[\tilde{\beta}_i]$.[30]

Our experiment provides data on all of the requisite statistics other than b, which we calibrate. In Appendix H.4 we provide public health and epidemiological evidence on the value of exercise,

[30] Technically speaking, there is a mechanical relationship between $\tilde{\beta}_i$ and $\alpha_i(p+\Delta p) - \alpha_i(p)$, as $\tilde{\alpha}'_i(p) \propto \tilde{\beta}_i$, which leads us to underestimate $E[\tilde{\beta}_i]$ in the presence of heterogeneity. However, this does not introduce a large bias as this correlation is of order $E[(1-\tilde{\beta}_i)^2]/E[\tilde{\beta}_i]$ which is a small margin of error for $\tilde{\beta}_i$ close to 1. E.g., for present focus parameters 0.8 this introduces a margin of error proportional to 0.04.

which suggests per-attendance health benefits between \$4 and \$20. As such, we provide estimates for b ranging from \$1 to \$20.

Formally, given our six positive piece-rate incentive values $p_1 < \cdots < p_6$, we use equation (3) 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 (3). 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 H.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 $b = \$10$.

To see why our estimates of $\tilde{\beta}$ are increasing in b, recall that the higher is the health value of exercise, the costlier is the perceived misbehavior $1 - \tilde{\beta} > 0$, 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 β, the actual short run discount factor, under additional assumptions. The intuition for identification is as follows: Suppose that expected attendance at no incentives is 20% higher than actual attendance and that a piece-rate incentive p^* increases actual attendance by 20%. Then that must imply that naivete leads people to overestimate how much their future self values delayed benefits b by p^*/b.

Formally, we show in Appendix H.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 a continuous function A. Consequently, if $\alpha(0) = a^*(p^*)$, then $\tilde{\beta}b = \beta(b + p^*)$, and thus

$$\beta/\tilde{\beta} = b/(b + p^*). \tag{5}$$

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 may overstate their actual beliefs if they see the forecasting prompt as a more aspirational exercise.[31] If some of the overestimation is due to reasons other than naivete about β, 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.[32] 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. 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 sets of 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 H.3 provides further details.

Figure 8 shows the resulting estimates of $\beta/\tilde{\beta}$—our measure of sophistication. As equation (5) 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 0.86, while at the middling value of $b = \$10$, the estimate of $\beta/\tilde{\beta}$ is approximately 0.75. Combined with our estimates of $\tilde{\beta}$, this implies a $\beta = 0.66$ 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 realize ($\beta < \tilde{\beta}$). Even if some, but not all, of the overestimation of future visits was due to mechanisms other than overestimation of β, the general conclusion of non-negligible naivete would hold.[33]

[31] As we discuss in Section 2.4, systematic bias in stated versus actual beliefs does not bias estimates of $\tilde{\beta}$ as long as the bias in stated beliefs is a level shift that does not affect the *difference* between how often people think they will attend at two different incentive levels. That is, suppose that a person always gives stated beliefs that are 2 visits higher than their true beliefs. That type of bias will not affect our estimates of $\tilde{\beta}$ because it will not bias the difference in how often the person thinks they will use the gym under say a \$1 incentive and a \$5 incentive.

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

[33] 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.

6 Debiasing beliefs

In Section 2.3, we laid out three predictions derived from the imperfect perception model. The first two predictions concern demand for commitment contracts to exercise less and were upheld in the results of Section 4.1. The third prediction is that increasing an individual's sophistication about their present focus will decrease their demand for commitment contracts to exercise more. In this section, we report the results of a randomized information treatment that allows us to test this hypothesis.

6.1 Impact of the information treatments 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, we tested a basic information treatment which presented participants in the treatment group with a graph of their visit patterns over the prior 20 weeks. This treatment was unsuccessful at debiasing overoptimistic beliefs. Sub-figure (a) of Figure 9 shows the average expected visit rates participants reported in the online component 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 study, 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.[34] As sub-figure (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 by the level of incentive the participant was randomly assigned ($0, $2, $5, and $7). A comparison of realized visits between the treatment and control groups reveals that the information treatment had no economically or statistically significant impact on actual visit attendance.[35] 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

[34] 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.

[35] In a regression controlling for the incentive level received, we estimate an average effect 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.

the level of overestimation of visit frequency.[36]

6.2 Debiasing beliefs increases willingness to pay for behavior change

An increase in an individual's level of sophistication about their present focus (i.e., moving $\tilde{\beta}$ toward β) 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, 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 7 shows the estimated effects of both the basic information treatment and the enhanced information treatment 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 valuation of behavior change by $1.15 per dollar of incentive increase [95% CI $0.29 - $2.02], and we estimate an increase of $1.33 for the average excluding the $1 incentive [95% CI $0.43 - $2.24]. These are relatively large effects that increase baseline WTP for behavior change by a factor of approximately two to three.

These results are consistent with the interpretation that the information treatment at least partially increased sophistication about participants' present focus. 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 effect on WTP for behavior change.

6.3 Debiasing beliefs decreases commitment contract take-up

Table 8 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 visits" 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 $= 0.18$ 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 percentage points (p-value $= 0.02$). This empirical result is consistent with the third prediction summarized in Section 2.3. The finding that only the enhanced information treatment affected commitment contract take-up is consistent with the results about beliefs in Section 6.1.

[36] 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 or more visits contract and the 11 or fewer visits contract) for only one of the waves. Nevertheless we find qualitatively similar results. The information treatment decreased the expected likelihood of meeting the 12 or more visits contract and increased the likelihood of meeting the 11 or fewer visits contract; the overall difference in these effects is 7.7 percentage points (p-value $= 0.038$).

6.4 Impact of the information treatments on parameter estimates

How do the reduced-form results about the impact of information provision on beliefs translate into the perceived short-run discount factor $\tilde{\beta}$? To answer this question, we utilize the methodology of Section 5.3 (see also Appendix H.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, though noisy. At the middling value of $b = \$10$, for example, the debiasing intervention decreases $\tilde{\beta}$ from 0.88 to 0.82.

7 Concluding remarks and implications for future work

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 pure commitment contract. This paper shows that such choices could also reveal imperfect perception or experimenter demand effects.

Empirically-minded researchers who work with experimental or survey data are well aware that this kind of data often features stochasticity in individuals' decisions, 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 valuation errors, 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, 14% will end up choosing it.

Mean-zero error assumptions are more realistic when the outcome of interest is continuous and uncensored—like our WTP for behavior change measure. Consequently, designs utilizing continuous dependent variables are more likely to be robust to imperfect perception. The Convex Time Budget technique (Andreoni and Sprenger, 2012; Augenblick et al., 2015) is another such example.

Better understanding the motives and mistakes in commitment contract demand informs not only positive analysis but also normative analysis. The insights from this study should inform thinking about the relative benefits of commitment contracts versus "sin taxes" (e.g., O'Donoghue and Rabin, 2006; Allcott et al., 2019a) 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 stochastic valuation errors or demand effects, 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 future needs, tastes, and time constraints, and sometimes imperfectly value contracts, then 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.

Practitioners and researchers hoping to test the effect of incentives on overcoming self control problems may, however, be limited in their ability to impose population-level taxes or subsidies. Our results suggest that in these settings it may be useful to consider contract structures that have both financial upside and downside, rather than the purely downside structure of classic commitment contracts. Allowing for an upside can generate contract structures that are attractive for those who recognize a self-control problem even with non-negligible uncertainty about the costs of the activity. Of course, contracts with upside will likely also attract those with no self-control problems who simply want to make money. This type of "adverse selection" problem makes this type of program more challenging to design and implement. On the other hand, contracts with upside will also seem attractive to those who are very naive about their behavior, which can offset some of these selection concerns. Ultimately, our work suggests that better understanding the design of incentive contract programs of this type would be a valuable direction for future work.

Our results do not imply that the take-up of commitment contracts only reflects imperfect perception, 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.

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 ostensibly beneficial behavior, as we have introduced in this study. Observing the take-up of these alternative commitments, especially when it is correlated with the take-up of "standard" commitment contracts, can provide a cautionary signal that factors other than awareness of self-control problems are influencing demand.

Finally, designers can use our approach of measuring the willingness to pay for piece-rate incentives to construct an alternative 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 willingness to pay for behavior change.

These steps are not the only tools for examining the robustness of commitment contract demand, but we think they are an improvement on prior approaches. Other studies have proposed examining the validity of commitment contract take-up by analyzing how take-up correlates with proxies of present focus. 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 (John, 2019; Sadoff et al., 2019); and both sets of studies argue that their results can be explained by standard present focus models.

Prior work has also argued that providing people experience with a specific commitment contract can be helpful under the assumption that this prior experience helps to ensure that any subsequent commitment contract take-up reflects an understanding of their benefits as a self-control strategy. However, relying on prior experience with commitment contracts may be problematic because it may create a status quo bias, or amplify perceived pressure from the experimenter. Our results suggest that it may be particularly informative to study how experience interacts with take-up of contracts to do both more and less of the activity. If experience only reduces take-up of the latter type of contract, and reduces the propensity for people to want both, then that would constitute powerful evidence that experience helps eliminate the choice of undesirable contracts and that "more" contracts are truly desirable.

Finally, while not studying commitment contract take-up, recent work by de Quidt et al. (2018) has proposed a methodology for bounding the importance of demand effects. Although this method cannot be extended to study the role of imperfect perception, it may be a useful complement to our methods in future work.[37]

There are a number of 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. In addition, eliciting uncertainty about future hassle costs using simple and transparent survey questions is challenging. Yet results from 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 imperfect perception and demand effects, 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 valuation errors, 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 demand effects 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

[37]We were not aware of this recent work when we ran our experiment in 2015.

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 this paper serves as a useful foundation for further research into the drivers of commitment contract demand, and present focus more broadly. As the empirical literature deploying commitment contracts continues to grow at a rapid pace, obtaining a more complete understanding of what commitment contracts do, and when they should be deployed, will be crucial.

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.

Afzal, Uzma, Giovanna D'Adda, Marcel Fafchamps, Simon R Quinn, and Farah Said, "Implicit and Explicit Commitment in Credit and Saving Contracts: A Field Experiment," NBER Working Paper 25802, 2019.

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

Alan, Sule and Seda Ertac, "Patience, self-control and the demand for commitment: Evidence from a large-scale field experiment," *Journal of Economic Behavior and Organization*, 2015, *115*, 111–122.

Allcott, Hunt, Benjamin B. Lockwood, and Dmitry Taubinsky, "Regressive Sin Taxes, with an Application to the Optimal Soda Tax," *Quarterly Journal of Economics*, 2019, *134* (3), 1557–1626.

_ **, Joshua Kim, Dmitry Taubinsky, and Jonathan Zinman**, "Payday Lending, Self-Control, and Consumer Protection," 2019. Unpublished.

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

Andreoni, James and Charles Sprenger, "Estimating Time Preferences from Convex Budgets," *American Economic Review*, 2012, *102* (7), 3333–3356.

_ **, Deniz Aydin, Blake Barton, B. Douglas Bernheim, and Jeffrey Naecker**, "When fair isn't fair: Understanding choice reversals involving social preferences," *Journal of Political Economy*, forthcoming.

Ariely, Dan and Klaus Wertenbroch, "Procrastination, Deadlines, and Performance: Self-Control by Precommitment," *Psychological Science*, 2002, *13* (3), 219–224.

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, Ted Miguel, and Gautam Rao, "Self-Control and Demand for Preventive Health: Evidence from Hypertension in India," 2018. Unpublished.

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.

Bernheim, B. Douglas, Debraj Ray, and Sevin Yeltekin, “Poverty and Self-Control,” *Econometrica*, 2015, *83* (5), 1877–1911.

Beshears, John, James Choi, Chirstopher Clayton, Christopher Harris, David Laibson, and Brigitte Madrian, “Optimal Illuquidity,” *working paper*, 2019.

— **, 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?,” NBER Working Paper 21474, 2015.

Bhattacharya, Jay, Alan M Garber, and Jeremy D Goldhaber-Fiebert, “Nudges in Exercise Commitment Contracts: A Randomized Trial,” NBER Working Paper 21406, 2015.

Bisin, Alberto, Alessandro Lizzeri, and Leeat Yariv, “Government Policy with Time Inconsistent Voters,” *American Economic Review*, 2015, *105* (6), 1711–1737.

— **and Kyle Hyndman**, “Present-Bias, Procrastination and Deadlines in a Field Experiment,” 2018. Unpublished.

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.

Block, H.D. and Jacob Marschak, “Random Orderings and Stochastic Theories of Response,” in Ingram Olkin, ed., *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling*, Stanford University Press, 1960.

Bonein, Aurélie and Laurent Denant-Boèmont, “Self-Control, Commitment and Peer Pressure: A Laboratory Experiment,” *Experimental Economics*, 2015, *18* (4), 543–568.

Brune, Lasse, Eric Chyn, and Jason T Kerwin, “Pay Me Later: A Simple Employer-Based Saving Scheme,” 2018. Northwestern University Global Poverty Research Lab Working Paper.

— **, Xavier Giné, Jessica Goldberg, and Dean Yang**, “Facilitating Savings for Agriculture: Field Experimental Evidence from Malawi,” *Economic Development and Cultural Change*, 2016, *64* (2), 187–220.

Casaburi, Lorenzo and Rocco Macchiavello, “Demand and Supply of Infrequent Payments as a Commitment Device: Evidence from Kenya,” *American Economic Review*, 2019, *109* (2), 523–55.

Chow, Vinci YC, “Demand for a Commitment Device in Online Gaming,” 2011. Unpublished.

de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth, “Measuring and Bounding Experimenter Demand,” *American Economic Review*, 2018, *108* (11), 3266–3302.

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.

Dupas, Pascaline and Jonathan Robinson, "Why Don't the Poor Save More? Evidence from Health Savings Experiments," *American Economic Review*, 2013, *103* (4), 1138–71.

Ek, Claes and Margaret Samahita, "Pessimism and Overcommitment," 2019. UCD Centre for Economic Research Working Paper.

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

Exley, Christine L. and Jeffrey K. Naecker, "Observability Increases the Demand for Commitment Devices," *Management Science*, 2017, *63* (10), 3262–3267.

Frydman, Cary and Lawrence J. Jin, "Efficient Coding and Risky Choice," 2019. Unpublished.

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 Problems from 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," *Games and Economic Behavior*, 2018, *107*, 329–344.

John, Anett, "When Commitment Fails: Evidence from a Field Experiment," *Management Science*, 2019.

Karlan, Dean and Leigh L Linden, "Loose Knots: Strong Versus Weak Commitments to Save for Education in Uganda," 2017. Unpublished.

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," NBER Working Paper 23294, 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.

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

Lizzeri, Allesandro and Leeat Yariv, "Collective Self-Control," *American Economic Journal: Microeconomics*, 2017, *9* (3).

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.

Martinez, Seung-Keun, Stephan Meier, and Charles Sprenger, "Procrastination in the Field: Evidence from Tax Filing," 2018. Unpublished.

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.

Moser, Christian and Pedro Olea de Souza e Silva, "Optimal Paternalistic Savings Policies," *working paper*, 2019.

Natenzon, Pauolo, "Random Choice and Learning," *Journal of Political Economy*, 2019, *127* (1), 419–457.

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 and Anya Samek, "Can Interventions Affect Commitment Demand? A Field Experiment on Food Choice," *Journal of Economic Behavior and Organization*, 2019, *158*, 90–109.

_ , Anya Savikhin Samek, and Charles Sprenger, "Dynamic Inconsistency in Food Choice: Experimental Evidence from a Food Desert," *Review of Economic Studies*, 2019, pp. 1–35.

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.

Skiba, Paige Mart and Jeremy Tobacman, "Payday Loans, Uncertainty, and Discounting: Explaing Patterns of Borrowing, Repayment, and Default," 2018. Unpublished.

Sprenger, Charles, "Judging Experimental Evidence on Dynamic Inconsistency," *American Economic Revew: Papers and Proceedings*, 2015, *105* (5), 280–285.

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.

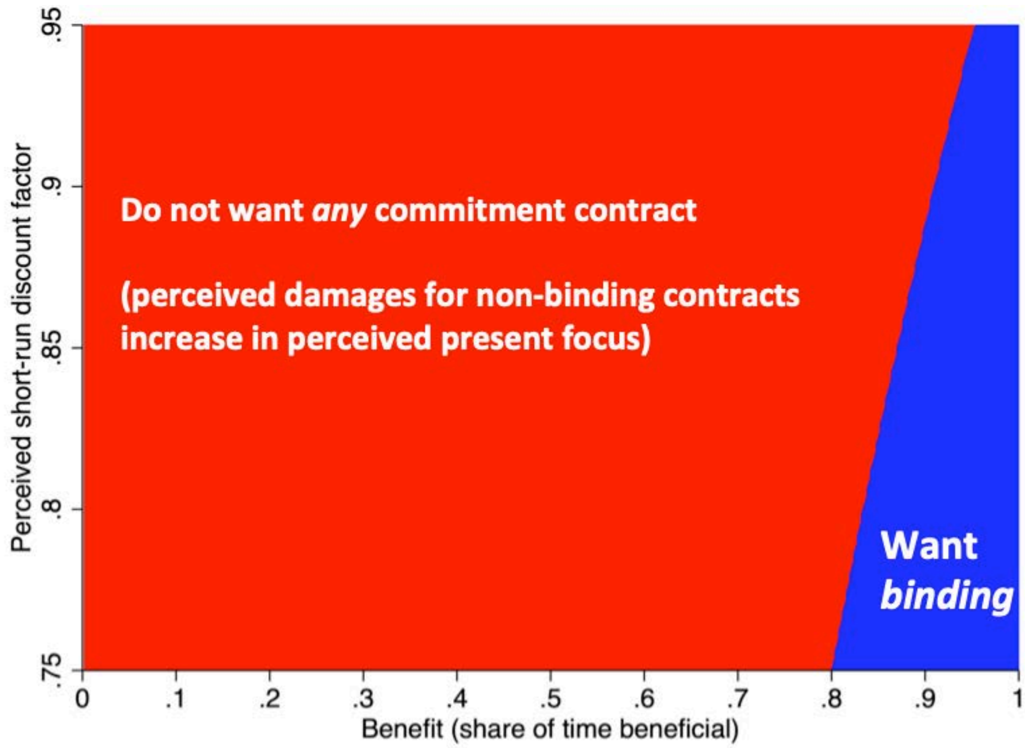
Toussaert, Severine, "Revealing temptation through menu choice: field evidence," 2019. Unpublished.

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," 2012. Unpublished.

_ , "Modeling Imprecision in Perception, Valuation and Choice," 2019. Unpublished.

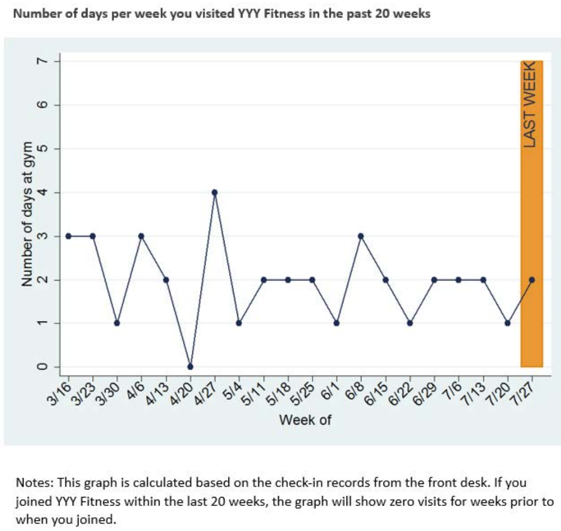
Figure 1: Commitment contract demand for uniform distribution of costs



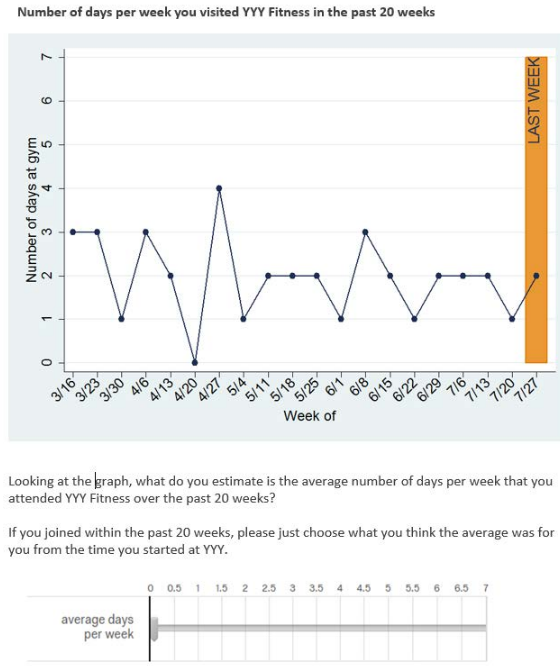
Notes: This figure illustrates the commitment contract demand for the case in which costs are distributed uniformly on the unit interval ($c \sim U[0, 1]$). Commitment contract demand is a function of delayed benefits b and perceived short-run discount factor $\tilde{\beta}$. As can be seen, for $\tilde{\beta} \geq 0.75$ 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

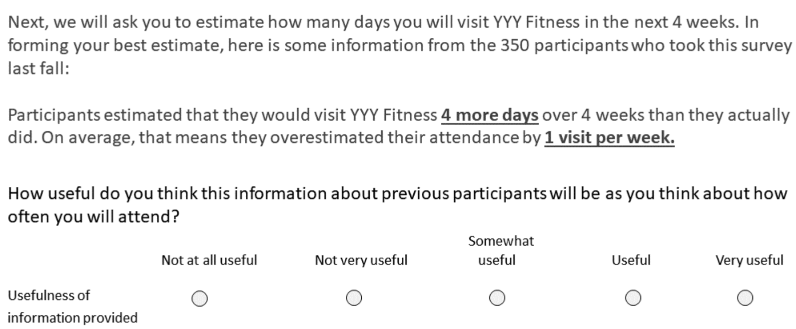
(a) Basic information treatment



(b) Enhanced info treatment - first screen



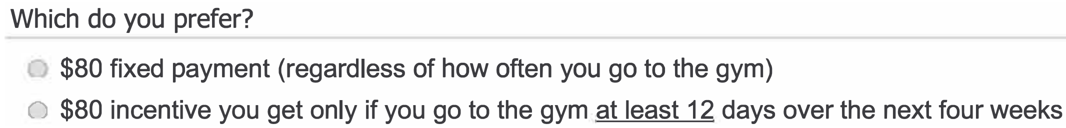
(c) Enhanced info treatment - second screen



Notes: Sub-figure (a) shows the basic information treatment of the history of past attendance shown to participants. Sub-figures (b) and (c) show the enhanced information treatment. Sub-figure (b) displays the first screen of the enhanced information treatment, which was similar to the basic information treatment but also included a question asking participants what they thought their average past weekly attendance was in the last 20 weeks. Sub-figure (c) shows the second screen, which informed that participants in the first wave of the experiment overestimated their attendance.

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

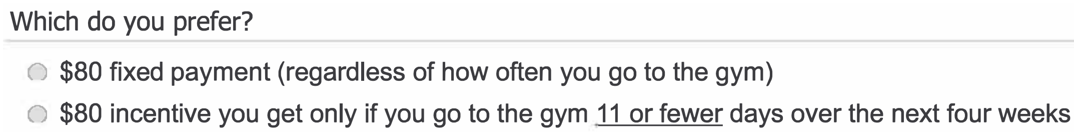
(a) "More visits" commitment contract


Which do you prefer?

$80 fixed payment (regardless of how often you go to the gym)

$80 incentive you get only if you go to the gym at least 12 days over the next four weeks

(b) "Fewer visits" commitment contract

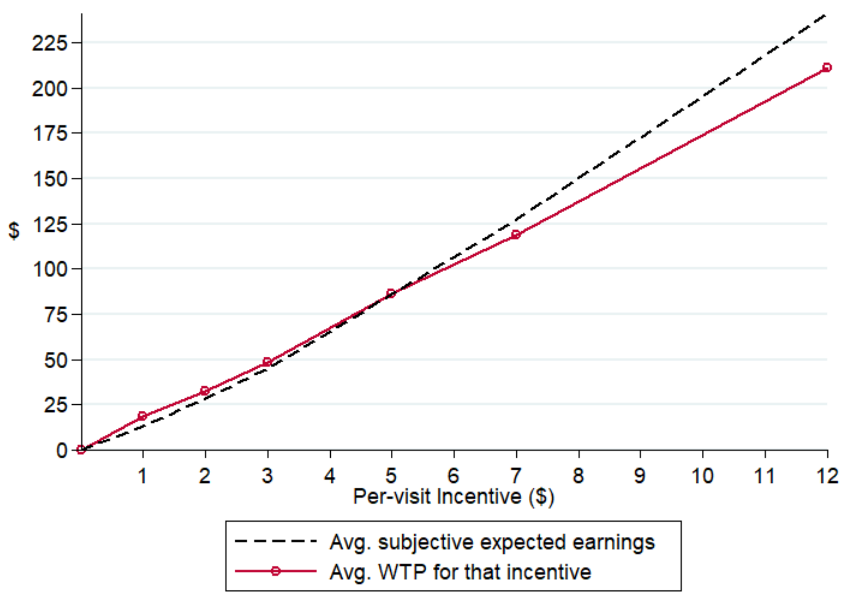

Which do you prefer?

$80 fixed payment (regardless of how often you go to the gym)

$80 incentive you get only if you go to the gym 11 or fewer days over the next four weeks

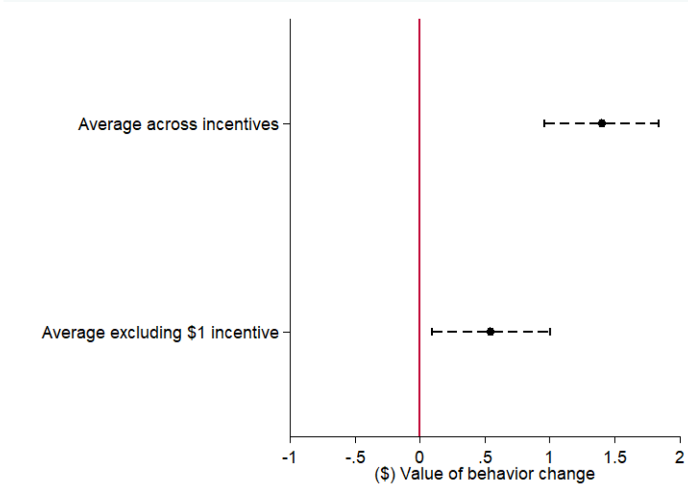
Notes: This figure provides a screenshot of the commitment contracts offered to participants. Sub-figure (a) provides an example of a commitment contract to attend the gym more (i.e., the "more visits" contract). Sub-figure (b) provides an example of a commitment contract to attend the gym less (i.e.,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

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



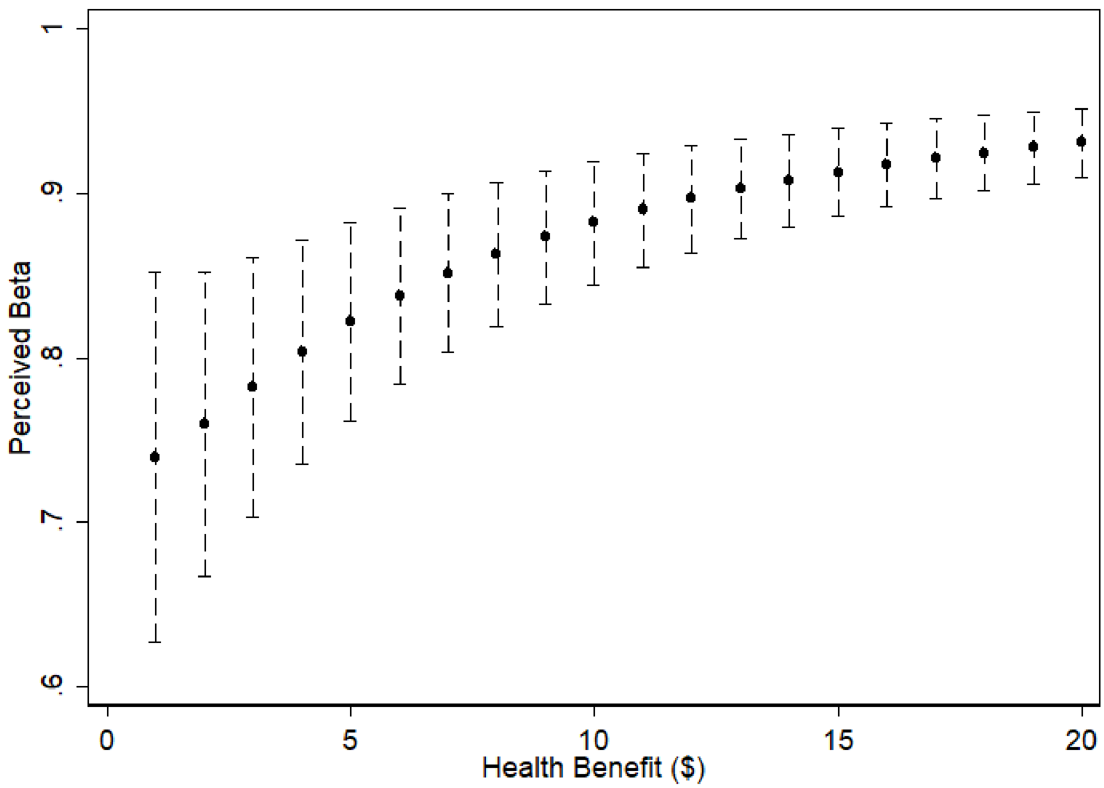
Notes: For each incentive, subjective expected earnings are the product of the piece rate (i.e., per day incentive) and participants' beliefs about the number of days they would visit under that incentive. WTP is the average willingness to pay for each incentive, elicited with sliders as described in 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 1 in Subsection 2.4. The sample consists exclusively of control group participants.

Figure 5: Willingness to pay for behavior change



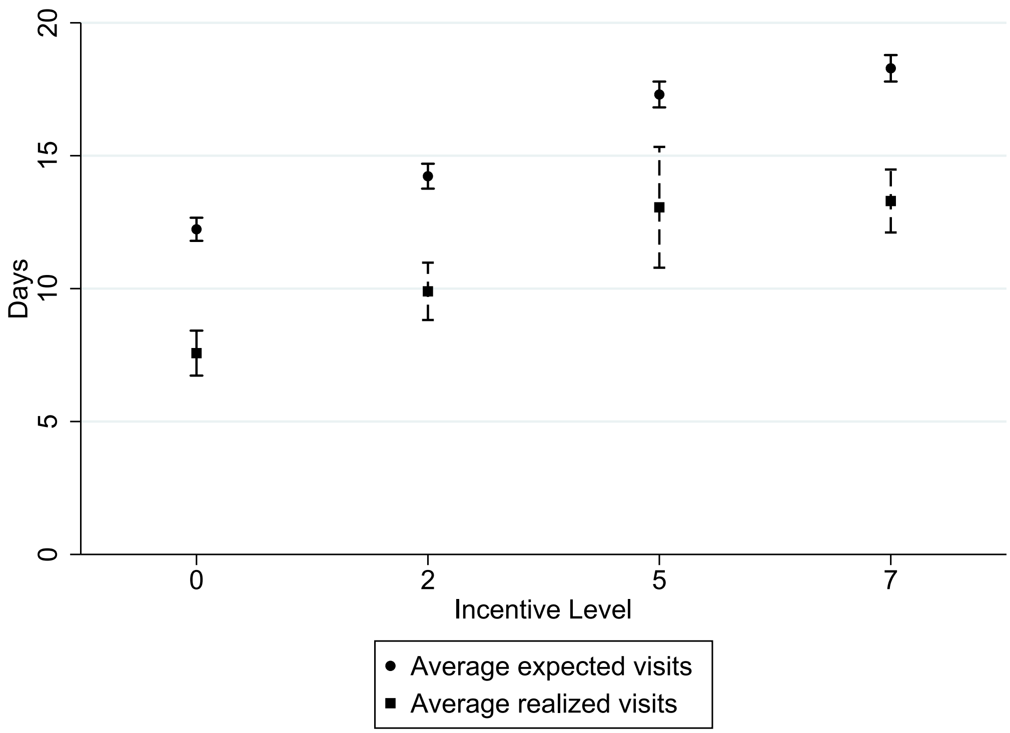
Notes: This figure shows the participants' average perceived value of behavior change as described in Section 5.1 (across all incentive levels in the top of the figure and across all incentive levels excluding the $1 incentive in the bottom of the figure), with 95% confidence intervals obtained from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

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



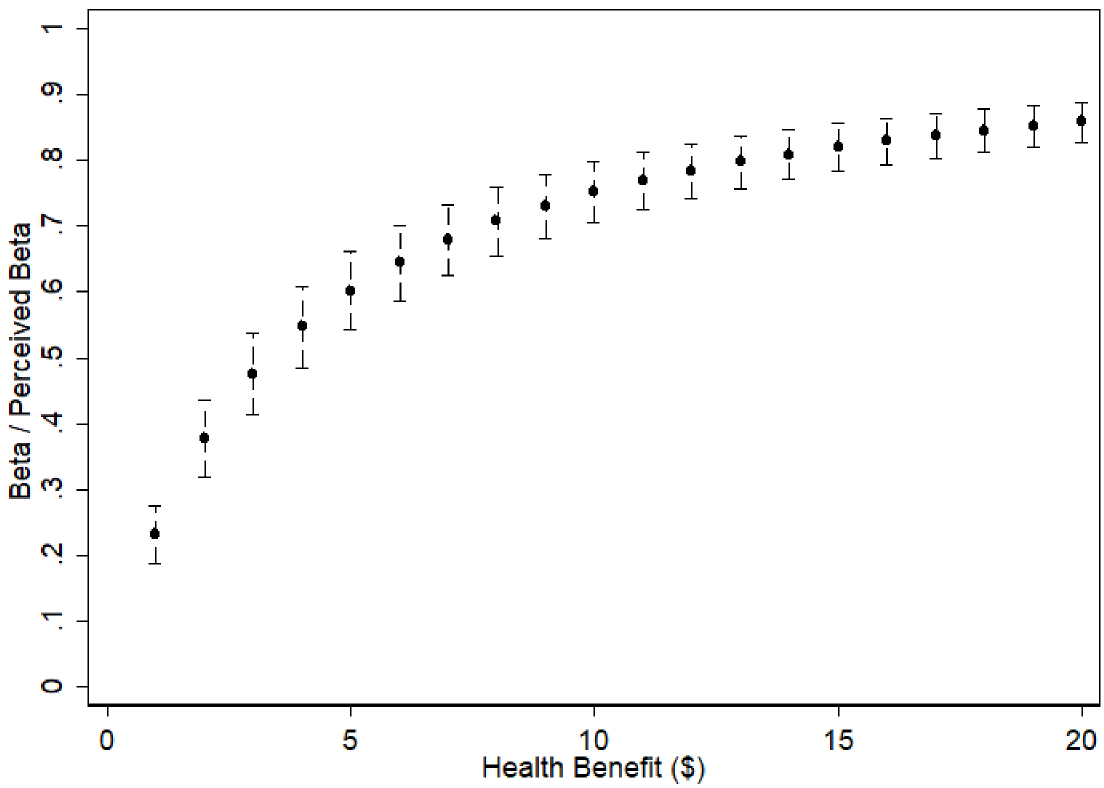
Notes: This figure shows the estimated perceived short-run discount factor $\tilde{\beta}$ for a given value of delayed health benefits per attendance ranging from $0 to $20. Alongside the estimates, the corresponding 95% confidence intervals are displayed. Standard errors are clustered at the participant level. See Appendix H.2 for the GMM methodology used to obtain the point estimates and the standard errors. The sample consists exclusively of control group participants.

Figure 7: Actual attendance versus subjects' expected 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 for whom the piece-rates were assigned exogenously (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 study 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)).

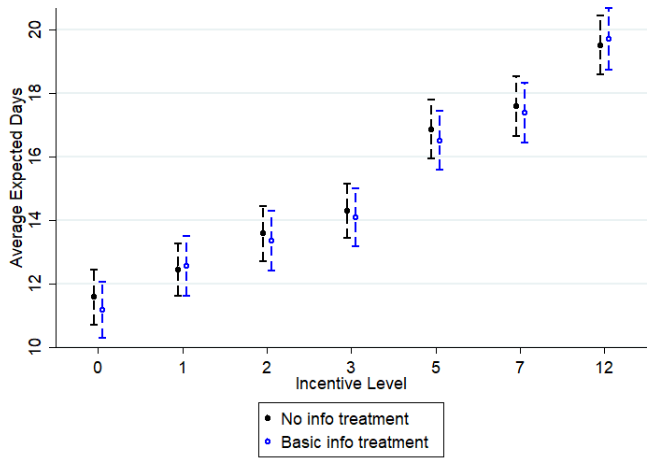
Figure 8: Estimates of $\beta/\tilde{\beta}$ for different values of delayed health benefits



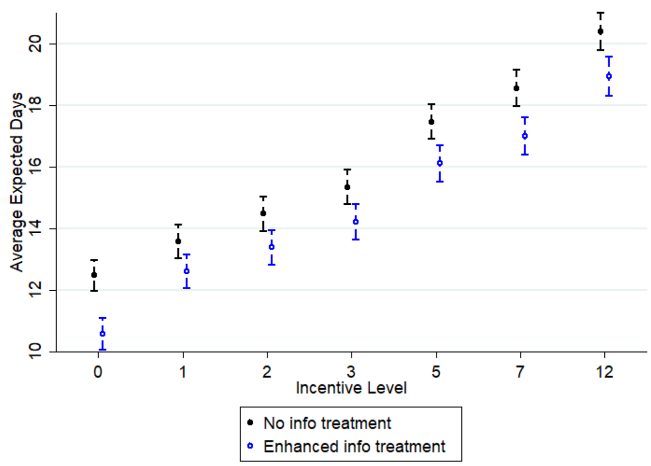
Notes: This figure shows the estimated ratio of the actual short-run discount factor to the perceived short-run discount factor, $\beta/\tilde{\beta}$, for a given value of delayed health benefits per attendance ranging from \$0 to \$20. Alongside the estimates, the corresponding 95% confidence intervals are displayed. Standard errors are clustered at the participant level. See Appendix H.3 for the GMM methodology used to obtain the point estimates and the standard errors. Statistics in the figure are based on data from control group participants for whom the piece-rates were assigned exogenously (excluding those with either low willingness-to-pay (12 participants) or those randomly assigned a high fixed payment (4 participants)).

Figure 9: Effect of information treatments on expected visits

(a) Impact of basic information treatment

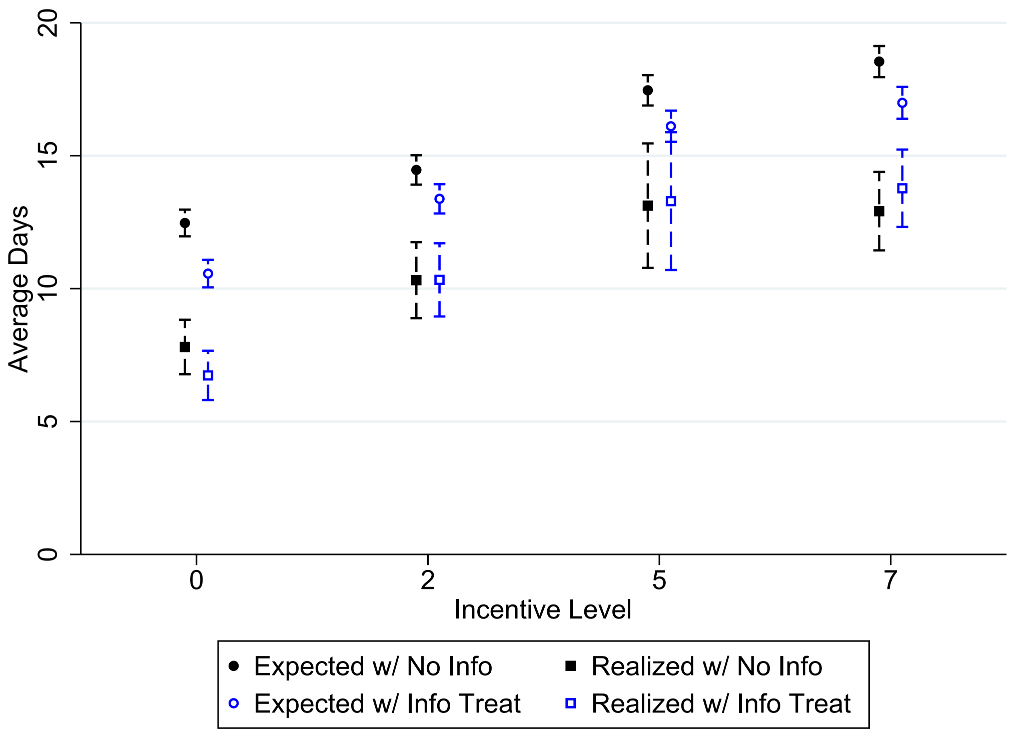


(b) Impact of enhanced information treatment



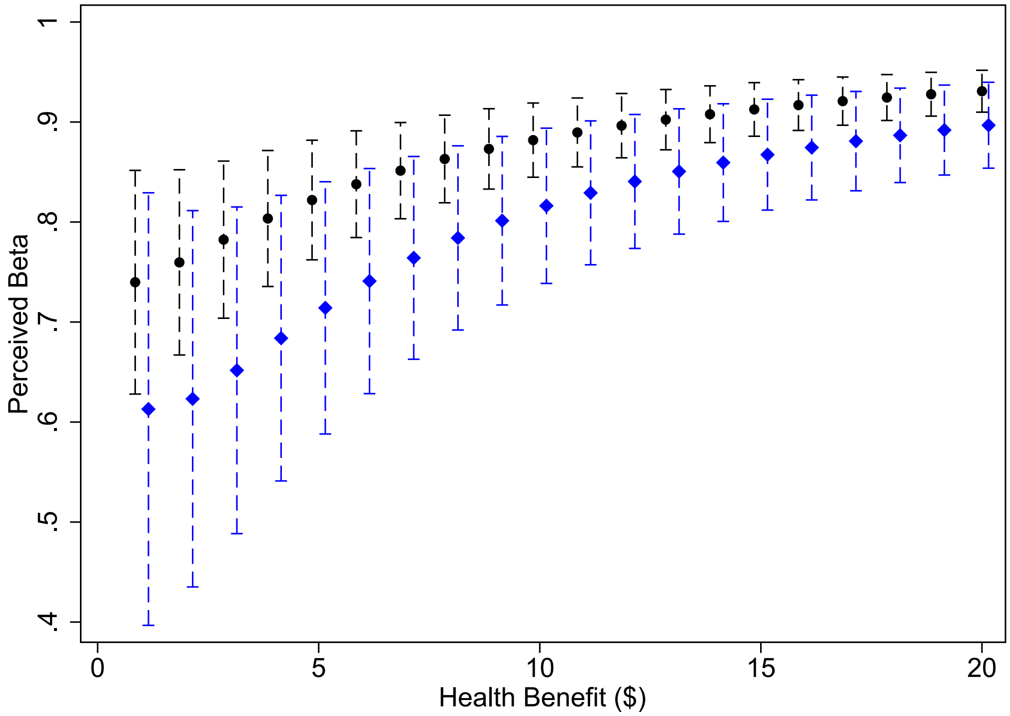
Notes: Subfigure (a) of this figure shows the mean and 95% confidence intervals for expected visits in the next four weeks for participants randomly selected to receive the basic information treatment (N=174) and participants randomly assigned to the no information control group (N=174) from Wave 1. 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.

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 with 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: Summary of commitment contact studies

Type of contract Authors (year)	Take-up rate	At stake
A. Penalty based:		
Gine, Karlan and Zinman (2010)	11%	own money
Royer, Stehr and Sydnor (2015)	12%	earned money
Bai et al. (2018)	14%	own money
Bhattacharya, Garber and Goldhaber-Fiebert (2015)	23%	own money
John (2019)	27%	own money
Kaur, Kremer and Mullainathan (2015)	36%	own money
Schwartz et al. (2014)	36%	house money
Bonein and Denant-Boèmont (2015)	42%	other[1]
Besheard et al. (2015)	39-46%[2]	house money
Toussaert (2019)	21-65%	house money
Schilbach (2019)	49%	house money
Exley and Naecker (2017)	41-65%	house money
Ariely and Wertenbroch (2002)	73%	other[3]
Average take-up rates (Penalty based contracts)		
Own money at stake	22%	
House money at stake	44%	
Other stakes	42%	
Overall	35%	
B. Removing options:		Restricted access to
Brune et al. (2016)	6%	own money
Afzal et al. (2019)	4-9%	own money
Sadoff and Samek (2019)	20-50%	other
Ek and Samahita (2019)	27%[4]	other
Ashraf, Karlan and Yin (2006)	28%	own money
Sadoff, Samek and Sprenger (2019)	33%	other
Acland and Chow (2018)	35%	other
John (2019)	42%	own money
Karlan and Linden (2017)	44%	own money
Toussaert (2018)	45%	other
Bisin and Hyndman (2018)	31-62%	other
Houser et al. (2018)	48%	other
Brune, Chyn and Kerwin (2018)	50%	own money
Augenblick, Niederle and Sprenger (2015)	59%	other
Milkman, Minson and Volpp (2014)	61%[4]	other
Dupas and Robinson (2013)	65%	own money
Besheard et al. (2015)	56%[5]	house money
Alan and Ertac (2015)	69%	house chocolates
Chow (2011)	79%	other
Casaburi and Macchiavello (2019)	91%	own money
Average take-up rates (Option removal contracts)		
Own money at stake	42%	
House money/object at stake	63%	
Other stakes	45%	
Overall	46%	

[1] Points in a two-part experiment
[2] Fraction of endowment put into account with early withdrawal penalty
[3] Grade points
[4] Percent of participants with WTP>0
[5] Fraction of endowment put into account with early withdrawal prohibited

Notes: This table reports the take-up rates for weakly dominated commitment contracts offered at no cost. We include studies appearing in Table 1 of Schilbach (2019) or Table 1 of John (2019) as well as four more recent studies. Panel A represents contracts that imposed a penalty when the commitment threshold was not reached, i.e. non-binding contracts, while Panel B represents fully binding commitments. For studies that reported take-up rates from different waves or treatment groups, the range of relevant take-up rates is shown. At the bottom of each panel, we report unweighted averages across the studies of each type.

Table 2: 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 181 Treated		471 Control 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 online component, 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' reported estimates of their own past visits.

Table 3: Take-up of "more" and "fewer" 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***	0.13***
12 visits	0.52	0.33	0.72	0.46	0.21***	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 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). All take-up rates are computed for control group participants exclusively. Over three study 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 4: 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 the ex-ante self-assessed probability of attending the gym fewer than 12 days during the 4-week incentive period. The dependent variable is the expected attendance in the absence of any incentives. 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 shows 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 5: Correlation between perceived success in contracts and take-up of contracts

	Subj. prob succeed in "more" contract			Subj. prob succeed in "fewer" 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 regressions. 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 from which 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 6: Correlation between WTP for behavior change and take-up of "more" contracts

	Take-up of more-visit contract			
	(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
Take-up mean:	0.51	0.51	0.51	0.51

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 participants (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 7: Effect of information provision on willingness to pay for behavior change

	All	Ex. $1
	(1)	(2)
Basic info treatment	0.24	0.19
	(0.52)	(0.55)
Enhanced info treatment	1.15 **	1.33 ***
	(0.48)	(0.51)
N	1,292	1,292
Control mean	1.40	0.55

Notes: This table displays the impact of the information treatments on estimated WTP for behavior change. 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 8: Effect of information provision on take-up of "more" contracts

	8+ visits (1)	12+ visits (2)	16+ visits (3)	Pooled (4)
Basic information treatment	0.048	–0.067	–0.024	–0.014
	(0.052)	(0.053)	(0.047)	(0.040)
Enhanced information treatment	–0.056	–0.054*	–0.096**	–0.066**
	(0.042)	(0.033)	(0.042)	(0.030)
N	878	1,292	878	3,048
Control mean:	0.64	0.52	0.36	0.51

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 study 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 Proof of Proposition 1

Proof. Let F_t and f_t denote the CDF and PDF, respectively, of the cost draws in period t. We have

$$\begin{aligned}
\frac{d}{dp}V(0,0,p\sum a_t) &= \frac{d}{dp}\sum_t \int_{c\leq\tilde{\beta}(p+b)} (p+b-c)f_t(c)dc \\
&= \sum_t F_t(\tilde{\beta}(b+p)) + (1-\tilde{\beta})(p+b)\tilde{\beta}\sum_t f_t(\tilde{\beta}(p+b)) \\
&= \alpha(p) + (1-\tilde{\beta})(p+b)\alpha'(p)
\end{aligned}$$

$$\begin{aligned}
\frac{d^2}{dp^2}V(0,0,p\sum a_t) &= \alpha'(p) + (1-\tilde{\beta})(p+b)\alpha''(p) + (1-\tilde{\beta})\alpha'(p) \\
\frac{d^3}{dp^3}V(0,0,p\sum a_t) &= O(\alpha''(p))
\end{aligned}$$

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

$$\begin{aligned}
V(0,0,(p+\Delta p)\sum a_t) - V(0,0,p\sum a_t) &\approx (\Delta p)\frac{d}{dp}V(0,0,p\sum a_t) + \frac{(\Delta p)^2}{2}\frac{d^2}{dp^2}V(0,0,p\sum a_t) + O\left((\Delta p)^2\frac{d^2}{dp^2}\alpha(\right. \\
&= (\Delta p)\alpha(p) + (\Delta p)(1-\tilde{\beta})(p+b)\alpha'(p) + \frac{(\Delta p)^2}{2}(2-\tilde{\beta})\alpha'(p) + O\left((\Delta p)^2\right. \\
&= (\Delta p)\left(\alpha(p) + \frac{\Delta p}{2}\alpha'(p)\right) + (\Delta p)(1-\tilde{\beta})(p+\Delta p/2+b)\alpha'(p) + O\left((\Delta p\right. \\
&= (\Delta p)\frac{\alpha(p+\Delta p)+\alpha(p)}{2} + (b+p+\Delta p/2)(1-\tilde{\beta})(\alpha(p+\Delta p)-\alpha(p))O
\end{aligned}$$

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

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

If $p=0$,

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

$\square$

B Formal results for $T = 1$

B.1 Behavior in the perfect perception model

In period 0 individuals 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 $p_1 \geq 0$ for taking the actions $a = 0$ and $a = 1$, respectively. The action a is chosen in period 1. All financial payments are received in period 2. For example, if $a = 1$ represents going to the gym and $a = 0$ represents not going to the gym, then a contract $(0, 0, p)$ is a contract that pays the agent p every time she goes to the gym. A contract $(-p, 0, p)$ is a penalty-based commitment contract for attending the gym: the individual loses p unless she goes to the gym ($a = 1$). Conversely, a contract $(-p, p, 0)$ is a penalty-based contract for *not* going to the gym ($a = 0$).

In period 1, individuals choose $a = 1$ if $\beta(p_1 + b) - c \geq \beta p_0$, or equivalently if $c \leq \beta(p_1 + b - p_0)$. 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, an individual's perceived expected utility given contract (r, p_0, p_1) is

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

The first term, r, is just the fixed payment. The second term is the payoff from choosing $a = 0$, which the individual expects will happen when $c > \tilde{\beta}(p_1 + b - p_0)$. The third term is the payoff from choosing $a = 1$, which the individual expects will happen when $c \leq \tilde{\beta}(p_1 + b - p_0)$

B.2 With uncertainty about costs, quasi-hyperbolic preferences rarely generate a demand for commitment

Commitment contracts for $a = 1$ will be desired when $\tilde{\beta} < 1$ and there is little uncertainty about the action $a = 1$ being desirable from the period $t = 0$ perspective. For example, suppose that the costs c are always smaller than the delayed benefits b, but that the individual thinks that because of present focus she may sometimes choose $a = 0$. In this case, the individual will always want a commitment contract with a high enough penalty p that guarantees that he will always choose $a = 1$. In our notation, this is a contract $(-p, 0, p)$ with $p \geq \frac{(1-\tilde{\beta})b}{\tilde{\beta}}$.

More generally, when there is only a small chance that immediate costs will exceed the delayed benefits, individuals with $\tilde{\beta} < 1$ will want penalty-based contracts as long as $\tilde{\beta}$ is not too low. If $\tilde{\beta}$ is too low, then the penalties will lead to financial losses that are too large in magnitude relative to the desired behavior change. This line of logic can be used to to establish that when there is a small chance that costs exceed benefits, there will be demand for commitment by some individuals, and it will be non-monotonic in $\tilde{\beta}$.

We define $\Delta V = V(-p, 0, p) - V(0, 0, 0)$.

Proposition 2. *Suppose that $Pr(c > b)$ is positive for all realizations of c.*

1. For a given $\bar{p} > 0$: there exist $\underline{\beta} > 0$ and $\bar{\beta} < 1$ such that $\max_{p\in[0,\bar{p}]} \Delta V(-p, 0, p) \leq 0$ *(i.e., commitment contract with penalty p for action a=1 is undesirable) if $\tilde{\beta} < \underline{\beta}$ or if $\tilde{\beta} > \bar{\beta}$.*

2. For a given $p > 0$: When the distribution of c is Bernoulli, there exist thresholds $\underline{\beta} \leq \bar{\beta}$ such that $\Delta V(-p, 0, p) > 0$ if and only if $\beta \in (\underline{\beta}, \bar{\beta})$, with $\bar{\beta} > \underline{\beta}$ if $Pr(c > b)$ is sufficiently small.

3. For a given $\bar{p} > 0$: When the distribution of c is Bernoulli, there exist thresholds $\underline{\beta} \leq \bar{\beta}$ such that $\max_{p\in[0,\bar{p}]} \Delta V(-p, 0, p) > 0$ *if and only if $\beta \in (\underline{\beta}, \bar{\beta})$, with $\bar{\beta} > \underline{\beta}$ if $Pr(c > b)$ is sufficiently small.*

Proposition 2 captures the intuition of non-monotonic demand for commitment, analogous to the results of Heidhues and Kőszegi (2009) and **?**. 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 support of c is Bernoulli, $c \in \{\underline{c}, \bar{c}\}$, is analogous to **?**, who derives this non-monotonicity in the context of consumption and savings. In line with Heidhues and Kőszegi (2009) and **?**, the results also generalize to the question of whether there exists any commitment contract of size $p \in [0, \bar{p}]$ that is worthwhile.

However, such results about (non-monotonic) demand for commitment depend on strong assumptions about how much uncertainty there is about the costs of doing the action. We now show that that the standard quasi-hyperbolic model predicts that there should not be demand for commitment when there is at least a moderate chance that costs exceed delayed benefits.

We consider first whether for a fixed penalty p there exists any $\tilde{\beta}$ such that individuals will want the contract. Second, we consider whether for a given $\tilde{\beta}$ there exists any commitment contract (including fully binding ones) that will be desirable. Throughout, we will assume that the distribution of costs can be characterized by a continuous density function f with support on $[\underline{c}, \bar{c}]$. For shorthand, we will define $\Delta V(-p, 0, p) := V(-p, 0, p) - V(0, 0, 0)$ to be the individual's perceived expected utility from taking up the contract.

Proposition 3. *Fix p and assume that $f(c_2)/f(c_1) \geq (c_1/c_2)^2$ for all $c_2 > c_1$ in some 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 $\Delta V(-p, 0, p)$ is strictly increasing in $\tilde{\beta}$ for all $\tilde{\beta}$, and thus no individual will want the contract.*

The economic content of the assumption in Proposition 3 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, the chances of a cost draw that exceeds the benefits 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 instance, suppose that $c \sim U[0, 1.5b]$, so that time-consistent individuals do not want to take the action 33% of the time. In this case, there does not exist any $\tilde{\beta}$ for which a commitment contract with penalty $p < b/2$ is desirable.

In fact, the uniform distribution example overstates how big the probability of costs exceeding benefits must be to erode demand for commitment. Proposition 3 shows that even if the density of cost draws between b and $1.5b$ is decreasing at rate $1/x^2$, individuals will still not want commitment.

We complement our first result with a proposition that fixes $\tilde{\beta}$ and gives sufficient conditions for there to exist no desirable commitment contract at any value of p. This includes commitment contracts that simply restrict choice to $a = 1$ with infinite penalties $p = \infty$ for choosing $a = 0$.

Proposition 4. *Fix $\tilde{\beta}$ and assume that (i) f is unimodal;*[38] *(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.*

The economic content of the assumptions of Proposition 4 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) provide bounds on uncertainty. The key assumption is assumption (iv), which says that the chances of getting a cost draw under which it is suboptimal to take the action ($c > b$) are at least as high as the chances of getting a cost draw under which the time $t = 0$ individual thinks she should choose $a = 1$, but thinks that her time $t = 1$ self will not do so ($c \in [\tilde{\beta}b, b]$). Assumptions (ii) and (iii) strengthen the content of assumption (iv) by ensuring that the cost draws exceeding b are not all concentrated at a point only slightly higher than b.

All four of the assumptions of Proposition 4 are satisfied by a uniform distribution with support $[0, \bar{c}]$, where $\bar{c} \geq b + (1-\tilde{\beta})b$. For example, with $\tilde{\beta} = 0.8$, the assumptions are satisfied by a uniform distribution with support $[0, 1.2b]$. For this distribution, a time-consistent individual would not want to take the action only 17% of the time, and in those 17% of cases, the cost draws do not exceed the delayed benefits by more than 20%. This is an arguably modest amount of uncertainty. Yet this modest amount of uncertainty erodes demand for all possible commitment contracts.

B.3 Commitment take-up with imperfect perception and demand effects"

Formally, we suppose that for a given decision j, individual i behaves as if her expected utility under contract (r, p_0, p_1) is

$$\widehat{V}(r, p_0, p_1) = \beta \left[r + \varepsilon_{ij} \int_{c > \tilde{\beta}(p_1 + b - p_0)} p_0 dF + \varepsilon_{ij} \int_{c \leq \tilde{\beta}(p_1 + b - p_0)} (p_1 + b - c) dF \right] + \eta_i \mathbf{1}_{(p_0, p_1) \neq 0}$$

where $\mathbf{1}_{(p_0,p_1)\neq 0}$ is an indicator that at least some contingent incentives are involved. The ε_{ij} term captures stochastic valuation error, which we assume does not affect the certain incentive r. y. To allow for some heterogeneity in the propensity for stochastic valuation, we assume that for a fraction μ of individuals $\varepsilon_{ij} \sim G$ is i.i.d. with G supported on $[0, \infty)$ and $E_G[\varepsilon] = 1$, while for a fraction $1 - \mu$ of individuals $\varepsilon_{ij} \equiv 1$.[39]

To characterize the new implications of the model, we begin with the observation that in the standard quasi-hyperbolic model in Section B.1, no individuals would ever choose commitment

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

[39] More generally, our results hold as long as there is between-person heterogeneity in the variance of the error term ε_{ij}, and we make this assumption only for notational simplicity.

contracts for $a = 0$. This is simply because individuals would not choose to commit to take actions that in effect have immediate benefits and delayed costs. However, choice of commitment contracts for $a = 0$ can be consistent with our imperfect perception model in this section. As can be choice of commitment contracts for $a = 1$ and $a = 0$ by the same person, even when the conditions of Proposition 4 are met.

Proposition 5. *1. Assume that $\mu > 0$ or $Pr(\eta_i > \tilde{\beta}_i p) > 0$. Then a positive mass of individuals will choose a commitment contract for $a = 1$ with penalty p (i.e., contracts $V(-p, 0, p)$), even when $\tilde{\beta}_i = 1$ for all i.*

2. Assume (i) that $\mu > 0$ and (ii) that either there are some $\tilde{\beta}_i$ close enough to 1 or that $Pr(\eta_i > \beta_i p) > 0$. Then a positive mass of individuals will choose a commitment contract for $a = 0$ with penalty p (i.e., contracts $V(-p, p, 0)$). In this case, a positive mass of individuals would choose both commitment contracts for $a = 1$ and for $a = 0$.

3. Assume that (i) $\mu > 0$, (ii) either $\mu < 1$ or that the η_i are heterogeneous across i, and (iii) $E[\tilde{\beta}_i]$ is sufficiently close to 1. Then there will be a positive correlation between demand for commitment contracts for $a = 1$ and commitment contracts for $a = 0$.

Parts 1 and 2 of Proposition 5 establish that imperfect perception and demand effects can lead individuals to choose commitments contracts both for $a = 0$ and for $a = 1$, even when there is significant uncertainty about the cost of doing the activity.

Part 3 shows that in experiments in which individuals are faced with a number of decisions, with only one decision randomly selected to be implemented, there can be a positive correlation between demand for commitment contracts to do more of an activity and to do less of an activity. Intuitively, there are two types of mechanisms that lead to the correlation. 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 contracts, 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 imperfect perception, then only the fraction $\mu \in (0, 1)$ of individuals with imperfect perception 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 the choices of contracts.[40]

As we show below, the imperfect perception model also implies that with at least moderate uncertainty about future costs, the likelihood of choosing a penalty-based commitment contract for $a = 1$ will be monotonically increasing in $\tilde{\beta}$. This is in contrast to the more standard results such as those of Heidhues and Kőszegi (2009) and John (2019), and our analogous derivation in Proposition 2. The typical finding in the standard quasi-hyperbolic model is that if there is demand for commitment, it is non-monotonic in $\tilde{\beta}$, and is decreasing in $\tilde{\beta}$ when $\tilde{\beta}$ is sufficiently high.

[40] It is also helpful to note that even if individuals are observed to be more likely to choose commitments for $a = 1$ than for $a = 0$, 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 a commitment contract for $a = 1$ involves a higher financial loss than choosing a commitment contract for $a = 0$.

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 the contract $(-p, 0, p)$ is increasing in $\tilde{\beta}$.*

This result is a corollary of Proposition 3, which shows that under moderate to large uncertainty, the perceived harms of a commitment contract are decreasing in $\tilde{\beta}$ in the standard quasi-hyperbolic model. Although in the standard quasi-hyperbolic model these conditions would lead individuals to never choose a commitment contract, in our imperfect perception model individuals still choose the contract, but with a propensity that is decreasing in the expected harms in the standard model.[41] Intuitively, the less harmful the contracts would seem in the absence of noise and demand effects, the less noise and demand effects it takes to generate take-up.

C Proofs of Propositions for $T = 1$

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 individual 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 $CC(p, 0)$.

C.1 Proof of Proposition 2

Proof. The perceived gains from a commitment contract are

$$\begin{aligned} \Delta V/\beta &= -p + \int_{c \leq \tilde{\beta}(p+b)} (p + b - c) dF - \int_{c \leq \tilde{\beta} b} (b - c) dF \\ &= -p(1 - F(\tilde{\beta}(b + p))) + \int_{c=\tilde{\beta} b}^{c=\tilde{\beta}(b+p)} (b - c) dF \end{aligned} \tag{6}$$

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, $\int_{c=\tilde{\beta} b}^{c=\tilde{\beta}(b+p)} (b - c) dF \to 0$ 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 ΔV 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 μ 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 individual already believes that he will choose $a = 1$ when $c = \underline{c}$.

[41] Interestingly, the converse of Proposition 6 does not hold for commitment contracts for $a = 0$. That is, it does not hold that the likelihood of choosing a commitment contract for $a = 0$ is decreasing in $\tilde{\beta}$. 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.

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 individual incurs the cost p with probability μ. If $\tilde{\beta}(b+p) > \bar{c}$ then the individual incurs a utility loss of $\bar{c} - b$ with probability μ. Either way, $\Delta V > 0$ for small enough μ and $\Delta V < 0$ for large enough μ.

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 $\tilde{\beta} \in (\underline{\beta}, \bar{\beta})$. $\square$

C.2 Proof of Proposition 3

Proof. We have

$$\begin{aligned} \frac{d}{d\tilde{\beta}} \Delta V/\beta &= p(b+p)f(\tilde{\beta}(p+b)) + (b+p)(b - \tilde{\beta}(b+p))f(\tilde{\beta}(b+p)) - b(b - \tilde{\beta}b)f(\tilde{\beta}b) \\ &= (1-\tilde{\beta})(b+p)^2 f(\tilde{\beta}(b+p)) - (1-\tilde{\beta})b^2 f(\tilde{\beta}b) \end{aligned} \tag{7}$$

The expression (7) is positive if $\frac{f(\tilde{\beta}(p+b))}{f(\tilde{\beta}b)} \geq \frac{b^2}{(p+b)^2}$.

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

C.3 Proof of Proposition 4

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, $\mu = \int_{x=b-\gamma}^{b} f(x)dx \geq \int_{x=b-\gamma}^{b} f(b)dx = \gamma f(b)$.

Now

$$
\begin{aligned}
\int_{c=\beta b}^{b}(b-c)f(c)dc &= \int_{c=\beta b}^{b}(b-c)f(b)dc + \int_{c=\beta b}^{b}(b-c)(f(c)-f(b))dc \\
&= \frac{\gamma^2}{2}f(b) + \int_{c=\beta b}^{b}(b-c)(f(c)-f(b))dc \\
&\leq \frac{\gamma^2}{2}f(b) + \int_{c=\beta b}^{b}\gamma(f(c)-f(b))dc \\
&= \frac{\gamma^2}{2}f(b) + (\mu - \gamma f(b))\gamma \\
&= \gamma\mu - \frac{\gamma^2}{2}f(b) \qquad (8)
\end{aligned}
$$

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,

$$
\begin{aligned}
\int_{c\geq b}(b-c)f(c)dc &= \int_{c=b}^{b+\gamma}(b-c)f(c)dc + \int_{c\geq b+\gamma}(b-c)f(c)dc \\
&\leq \int_{c=b}^{b+\gamma}(b-c)f(c)dc - \int_{c\geq b+\gamma}\gamma f(c)dc \\
&= \int_{c=b}^{b+\gamma}(b-c)f(c)dc - \gamma(1-F(b+\gamma)) \\
&= \int_{c=b}^{b+\gamma}(b-c)f(c)dc - \gamma\left[(1-F(b)-(F(b+\gamma))-F(b))\right] \\
&\leq \int_{c=b}^{b+\gamma}(b-c)f(c)dc - \gamma\left(\mu - \int_{c=b}^{b+\gamma}f(c)dc\right) \\
&= \int_{c=b}^{b+\gamma}(b+\gamma-c)f(c)dc - \gamma\mu \\
&\leq \int_{c=b}^{b+\gamma}(b+\gamma-c)f(b)dc - \gamma\mu \\
&= \frac{\gamma^2}{2}f(b) - \gamma\mu \qquad (9)
\end{aligned}
$$

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 μ 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 (8) and (9) 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 (9) 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 since $\mu < \gamma f(b)$, this means that f must be uniform almost everywhere, with density $f(c) = \mu/\gamma$. Thus in this case

$$
\int_{c=\tilde{\beta}b}^{b}(b-c)f(c)dc \leq \gamma\mu/2. \qquad (10)
$$

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

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

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

$$\begin{aligned}
-\int_{x=b}^{x=b+\gamma} (b-x)f(c)dc &= \int_{x=b}^{x=b+\gamma} (x-b)\frac{b^2}{x^2}f(b)dx \\
&= b^2 f(b) \int \left(\frac{1}{x} - \frac{b}{x^2}\right) dx \\
&= b^2 f(b) \left[\ln(x) + \frac{b}{x}\right]_{x=b}^{b+\gamma} \\
&= b^2 f(b) \left[\ln(b+\gamma) + \frac{b}{b+\gamma} - \ln(b) - 1\right] \\
&= b^2 f(b) \left[\ln(1+k) - \frac{k}{1+k}\right] \\
&\geq b^2 f(b) \left[k - \frac{k^2}{2} - \frac{k}{1+k}\right] \qquad (11) \\
&= b^2 f(b) \left[\frac{k+k^2-k}{1+k} - \frac{k^2}{2}\right] \\
&= b^2 f(b) \left[\frac{k^2}{1+k} - \frac{k^2}{2}\right] \\
&= f(b) \left[\frac{\gamma^2}{1+k} - \frac{\gamma^2}{2}\right] \\
&= f(b) \left[\frac{\gamma^2(1-k)}{2(1+k)}\right] \\
&= \frac{\tilde{\beta}\gamma^2}{2(1+k)} f(b) \\
&= \frac{1}{2}\tilde{\beta}\gamma \frac{\gamma}{b+\gamma} b f(b) \\
&\geq \frac{\tilde{\beta}\gamma}{2}\frac{\mu}{\tilde{\beta}} \\
&= \gamma\mu/2 \qquad (12)
\end{aligned}$$

To obtain (11), 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 of the left and right side side of the inequality 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 (10) and (12) 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 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. $\square$

With the Lemma in hand, we are ready to prove Proposition 4.

Proof of the proposition

Proof. CASE 1: Suppose that $\bar{c} = \infty$. Then Proposition 3 implies that for any value of p, the value of the commitment contract is increasing in $\tilde{\beta}$. But since $\Delta V < 0$ for $\tilde{\beta} = 1$ individuals, 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 3 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 individual 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 individual with perceived present focus $\tilde{\beta}$. Consequently, it is undesirable to an individual 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 $CC(p, 0)$. $\square$

C.4 Proof of Proposition 5

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

$$\left[pF(\hat{\beta}_i b) + \int_{\hat{\beta}_i b}^{\hat{\beta}_i (p+b)} (p + b - c) dF \right] \varepsilon_{ij} \geq p - \eta_i/\beta_i \tag{13}$$

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

$$\left[p[1 - F(\hat{\beta}_i b)] - \int_{\hat{\beta}_i (b-p)}^{\hat{\beta}_i b} (b - p - c) dF \right] \varepsilon_{ij} \geq p - \eta_i/\beta_i \tag{14}$$

Since $pF(\hat{\beta}_i b) + \int_{\hat{\beta}_i b}^{\hat{\beta}_i (p+b)} (p + b - c) dF > 0$, condition (13) will be satisfied if either $\eta_i > \beta_i p$, or if $\mu > 0$ and the draw ε_{ij} is sufficiently high. Similarly, (14) 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 ε_{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 ε_{ij} draws for both the $CC(p,1)$ and $CC(p,0)$ are high enough such that the individual would want to choose both.

Finally, to prove part 3, let ν_i be an indicator for whether individual i makes valuation errors, 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 ν_i and η_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. □

C.5 Proof of Proposition 6

Proof. Since the probability of choosing a commitment contract is increasing in ΔV, the result follows if we show that ΔV is increasing in $\tilde{\beta}_i$ and in b. By Proposition 3, ΔV is increasing in $\tilde{\beta}_i$. □

D Generalizations to the dynamic case

We now consider a dynamic environment in which the individual 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 $\sum a_t \geq X$.

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 3. Under this condition, backwards induction using repeated application of Proposition 3 establishes a result analogous to Proposition 3. 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. However, 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; this non-separability is generated by commitment contracts to meet a certain threshold.

D.1 Generalization of Proposition 3

We generalize Proposition 3 by considering commitment contracts like those in our experiment, which involve a penalty p if the individual does not choose $a_t = 1$ at least $X \leq T$ times.

Proposition 7. *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 individuals should desire commitment contracts.*

Throughout, we use the following straightforward but useful extension of Proposition 3:

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 3, 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 individual meets the threshold regardless and in the latter case the individual fails to meet the threshold regardless. If $h_t = X - 1$ then Proposition 3 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 3 implies that $\frac{\partial}{\partial p}V_T(h_T) < 0$ and $\frac{\partial^2}{\partial\tilde{\beta}\partial p}V_T(h_T) > 0$. Now suppose that $\frac{\partial}{\partial p}V_{t+1}(h_{t+1}) < 0$ 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 $\tilde{\beta}$. □

D.2 Generalizations of Propositions 5 and 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 Proposition 7.

E Generalizations to other environments

E.1 Summary of generalizations

Continuous choice We continue to explore the robustness of our Section 2.2 results about the undesirability of commitment contracts in Appendix E.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 saving for the future. In Appendix E.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 withdrawal 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 Bernoulli model: If the density of cost shocks does not decrease "too quickly" in a region of cost shocks at which individuals with $\tilde{\beta} \in [\underline{\tilde{\beta}}, 1]$ are on the margin for choosing $x = X$, then the commitment contract is unappealing to all individuals with $\tilde{\beta}$ in that region, and its perceived damages are decreasing in $\tilde{\beta}$.[42]

Other models Finally, in Appendix E.3, we consider the robustness of our results about the 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,[43] though uncertainty about future costs erodes the benefits of those contracts as well.

E.2 Generalization to continuous choice

We now generalize our results about the (lack of) desirability of commitment contracts to continuous choices. 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 withdrawal from a savings account), we derive the following striking result both for models of effort provision and savings for

[42] We recognize that with continuous choice, the space of possible commitment contracts is very large. A general penalty-based commitment contract is a function π, $\pi(x) \geq 0$, 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.

[43] 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.

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 individuals with $\tilde{\beta} \in [\underline{\tilde{\beta}}, 1]$ are on the margin for choosing $x = X$, then the commitment contract is unappealing to all individuals with $\tilde{\beta}$ in that region, and its perceived damages are decreasing in $\tilde{\beta}$.[44]

Formally, 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 x 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 ε.

Model II: Saving for the future. In the savings choice model, the individual chooses an amount x to save for the future, given initial endowment Y. In period 0, individuals 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 ε.

Continuous penalties

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

Proposition 8. *Consider model I.*

1. If for a given commitment contract (p, X) there is a positive measure of θ 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.

[44] We recognize that with continuous choice, the space of possible commitment contracts is very large. A general penalty-based commitment contract is a function π, $\pi(x) \geq 0$, 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.

An important implication of part 2 of the proposition is that what affects the possible desirability of a commitment contract is not the likelihood that the individual 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 effect on an individual'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 individual fails to meet it. Consequently, if conditional on failing to meet the contract the individual 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 θ, suppose that the period 0 expected choice under the contract is $x^*(\theta, p) < X$. Now for this taste-shock,

$$\begin{aligned}
\frac{d}{dp}((w+p)x^* - \theta C(x^*) - Xp) &= \frac{dx^*}{dp}(w + p - \theta C'(x^*)) - (X - x^*) \\
&= \frac{dx^*}{dp}(w + p - \tilde{\beta}(w+p)) - (X - x^*) \\
&= (1 - \tilde{\beta})(w+p)\frac{dx^*}{dp} - (X - x^*) \\
&= (1 - \tilde{\beta})x^* \varepsilon - (X - x^*) \\
&= ((1 - \tilde{\beta})\varepsilon + 1)x^* - X
\end{aligned} \tag{15}$$

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 θ, the net impact of increasing p is thus

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

Next, taking the derivative of (15) 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 $\frac{d^2}{d\tilde{\beta}dp}V > 0$ 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 $\tilde{\beta} = 1$ individual would expect to choose $x^* < X$, $\frac{d^2}{d\tilde{\beta}dp}V > 0$ 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 $\tilde{\beta} \in [0, 1]$. $\square$

Proposition 9. *Consider model II.*

1. If for a given commitment contract (p, X) there is a positive measure of θ 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 θ, suppose that the period 0 expected choice under the contract is $x^*(\theta, p) < X$. Now for this taste-shock,

$$\begin{aligned}
\frac{d}{dp}(u(x^* + \theta(Y - (r+p)x^* - pX))) &= \frac{dx^*}{dp}(u'(x^*) - \theta(r+p)) - \theta(X - x^*) \\
&= \frac{dx^*}{dp}\left(\frac{1}{\tilde{\beta}}\theta(r+p) - \theta(r+p)\right) - \theta(X - x^*) \\
&= (1/\tilde{\beta} - 1)\theta(r+p)\frac{dx^*}{dp} - \theta(X - x^*) \\
&= (1/\tilde{\beta} - 1)\theta x^* \varepsilon - \theta(X - x^*) \\
&= \theta((1/\tilde{\beta} - 1)\varepsilon + 1)x^* - \theta X
\end{aligned} \tag{16}$$

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 θ, 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 (16) 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 $\tilde{\beta} = 1$ individual would expect to choose $x^* < X$, $\frac{d^2}{d\tilde{\beta}dp}V > 0$ 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 $\tilde{\beta} \in [0, 1]$. $\square$

Discontinuous penalties

Proposition 10. *Consider model I and fix a contract (p, X). Let $\theta^\dagger(\tilde{\beta})$ be the taste-shock for which an individual 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 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 II.*

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

$$\tilde{\beta}(wx(\theta^\dagger) - p) - \theta^\dagger C(x(\theta^\dagger)) = \tilde{\beta}wX - \theta^\dagger C(X) \tag{17}$$

where x satisfies $\theta^\dagger C'(x) = \tilde{\beta}w$. Differentiating (17) with respect to $\tilde{\beta}$ using the condition $\theta^\dagger C'(x) - \tilde{\beta}w$ gives

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

or equivalently

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

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 γ. 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 (17) 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 (17) 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 $\frac{f'(\theta)}{f(\theta)} \geq -1/\theta$. □

E.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 is linear (as in Gul and Pesendorfer (2001), or Assumption 5' of Fudenberg and Levine (2006)). This means that in period 1, the individual's utility in a contract with penalty p for choosing $a = 0$ is given by $-p + a \cdot [b + p - (1+\gamma)c]$,where γ is the marginal cost of self control. The individual's utility in period 1 when the choice-set is restricted to $A = \{1\}$ is given by $(b-c)$. In period 0, the individual 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 $\int_{c \le b/(1+\gamma)} (b - (1+\gamma)c) dF(c)$. The impact of the restriction is thus

$$\int_{c \le b/(1+\gamma)} \gamma c dF(c) + \int_{c \ge b/(1+\gamma)} (b - c) dF(c) \le \gamma \int_{c \le b} c dF(c) + \int_{c \ge b} (b - c) dF(c)$$

The inequality follows because $\gamma c \ge b - c$ iff $c \ge b/(1+\gamma)$. 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 $\gamma(1/2)(1/\bar{c}) \le \frac{\bar{c}-1}{\bar{c}} \frac{\bar{c}-1}{2}$, or if $\gamma \le (\bar{c} - 1)^2$. 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 individuals with costly self-control, for a given "weight" on delayed benefits versus immediate costs.

F Further study details

Table A.1: Study 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 More/Less than 12 days 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 study across the three waves of implementation.

Table A.2: Study 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)
N	350	528	414	1,292

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 study 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' reported estimates of their own past visits.

G Further results on take-up of commitment contracts

G.1 Commitment contract take-up rates by expected attendance

Table A.3: Take-up rate by expected attendance (α); Non-treated

Threshold (T)	Chose "More" Contract (1)	Chose "More" Given α $\leq$ T-2 (2)	Chose "More" Given α $\leq$ T-4 (3)	Chose "Fewer" Contract (4)	Chose "Fewer" Given α $\geq$ T+1 (5)	Chose "Fewer" Given α $\geq$ T+3 (6)
8 visits	0.64	0.56	0.55	0.35	0.35	0.34
12 visits	0.52	0.38	0.32	0.33	0.34	0.33
16 visits	0.36	0.26	0.25	0.31	0.36	0.39

Notes: Sample includes all participants not given the information treatment. Expected attendance is in the absence of incentives. Each column reports the take-up rate of a commitment contract for either more or fewer visits than the threshold, with the sample limited in columns (2), (3), (5), and (6) by participants' expectations of gym visits. All take-up rates are computed for control group participants exclusively. *** denotes those differences that are statistically significantly different from 0 at the 1% level.

Table A.4: Take-up rate by expected attendance (α); All

Threshold (T)	Chose "More" Contract (1)	Chose "More" Given α $\leq$ T-2 (2)	Chose "More" Given α $\leq$ T-4 (3)	Chose "Fewer" Contract (4)	Chose "Fewer" Given α $\geq$ T+1 (5)	Chose "Fewer" Given α $\geq$ T+3 (6)
8 visits	0.64	0.60	0.60	0.35	0.31	0.30
12 visits	0.49	0.39	0.34	0.32	0.31	0.30
16 visits	0.32	0.24	0.24	0.28	0.32	0.33

Notes: Sample includes all participants, including those given the information treatment. Expected attendance is in the absence of incentives. Each column reports the take-up rate of a commitment contract for either more or fewer visits than the threshold, with the sample limited in columns (2), (3), (5), and (6) by participants' expectations of gym visits.

Table A.5: Correlation of take-up by expected attendance (α); Non-treated

Threshold (T)	All (1)	$\alpha \leq$ T-2 (2)	$\alpha \leq$ T-4 (3)	$\alpha \geq$ T+1 (4)	$\alpha \geq$ T+3 (5)	$\alpha \leq 6$ (6)	$\alpha \geq 17$ (7)
8 visits	0.37***	0.40***	0.54***	0.36***	0.39***	0.40***	0.40***
12 visits	0.29***	0.28***	0.32***	0.29***	0.23***	0.37***	0.26***
16 visits	0.28***	0.29***	0.30***	0.31***	0.24***	0.33***	0.31***

Notes: Sample includes all participants not given the information treatment. Expected attendance is in the absence of incentives. Each column reports the correlation between the take-up of commitment contracts for more and fewer visits than the threshold, with the sample limited in columns (2)-(7) by participants' expectations of gym visits. All take-up rates are computed for control group participants exclusively. *** denotes those differences that are statistically significantly different from 0 at the 1% level.

Table A.6: Correlation of take-up by expected attendance (α); All

Threshold (T)	All (1)	$\alpha \leq$ T-2 (2)	$\alpha \leq$ T-4 (3)	$\alpha \geq$ T+1 (4)	$\alpha \geq$ T+3 (5)	$\alpha \leq 6$ (6)	$\alpha \geq 17$ (7)
8 visits	0.38***	0.40***	0.46***	0.38***	0.39***	0.39***	0.41***
12 visits	0.25***	0.24***	0.29***	0.31***	0.26***	0.31***	0.32***
16 visits	0.25***	0.24***	0.24***	0.34***	0.34***	0.27***	0.34***

Notes: Sample includes all participants, including those given the information treatment. Expected attendance is in the absence of incentives. Each column reports the correlation between the take-up of commitment contracts for more and fewer visits than the threshold, with the sample limited in columns (2)-(7) by participants' expectations of gym visits. *** denotes those differences that are statistically significantly different from 0 at the 1% level.

G.2 Commitment contract take-up in treatment group and its correlates

Table A.7: 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***	0.11***
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 participants exclusively. Over three study 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.8: Correlation between perceived success in contracts and expected attendance; treated group

	Subj. expected attendance w/ out incentives		
	(1)	(2)	(3)
Subj. prob succeed in "more" contract	0.07***		0.07***
	(0.02)		(0.02)
Subj. prob succeed in "fewer" contract		–0.04***	–0.05***
		(0.01)	(0.01)
N	215	215	215
"More" − "Fewer"			0.12***
			(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 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 in "fewer" contract is the 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 participants in Wave 3, the only wave we elicited the probabilities of contract success. *** denotes statistics that are statistically significantly different from 0 at the 1% level.

Table A.9: Correlation between perceived success in contracts and take-up of contracts; treated group

	Subj. prob succeed in "more" contract			Subj. prob succeed in "fewer" 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 "fewer"		–8.21**	–11.67***		13.79***	16.05***
		(3.80)	(3.57)		(3.77)	(3.72)
N	215	215	215	215	215	215
"More" - "Fewer"			26.35***			-25.65***
			(5.05)			(5.97)

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 regressions. 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 treatment 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 5% and 1% level respectively.

G.3 Other correlates of commitment contract take-up

Table A.10: Other correlates of commitment contract take-up; control group

	Expected attendance (1)	Past attendance (2)	Goal attendance (3)
Chose “more" contract	2.58***	1.63***	3.05***
	(0.29)	(0.32)	(0.29)
Chose “fewer" contract	–0.60*	–1.67***	–1.10***
	(0.32)	(0.33)	(0.32)
N	1,522	1,522	1,522
“More” - “Fewer”	3.18***	3.31***	4.16***
	(0.48)	(0.52)	(0.49)

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 regressions. 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 participants. Since participants were asked about multiple commitment contracts, each participant 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.11: Other correlates of commitment contract take-up; treated group

	Expected attendance (1)	Past attendance (2)	Goal attendance (3)
Chose “more" contract	1.41***	1.16***	2.18***
	(0.30)	(0.30)	(0.32)
Chose “fewer" contract	–1.32***	–2.28***	–1.22***
	(0.32)	(0.32)	(0.36)
N	1,526	1,526	1,526
“More” - “Fewer”	2.73***	3.44***	3.40***
	(0.47)	(0.47)	(0.52)

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 regressions. 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 participants. Since participants were asked about multiple commitment contracts, each participant 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.

H Further results on willingness to pay for behavior change

H.1 Correlations with fewer commitment contracts

Table A.12: Correlation between WTP for behavior change and take-up of "fewer" contracts

	Take-up of fewer-visit contract			
	(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
Take-up mean:	0.32	0.32	0.32	0.32

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 participants (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.

H.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,[45] 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).

H.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, the 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 (H.2).

H.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,

[45] 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.

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:

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

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[46]

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

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:

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

[46]NCHS, National Vital Statistics System, Mortality. "United States Life Tables, 2014". National Vital Statistics Reports Vol. 66 No. 4. August 14, 2017.

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.

I Elicitation of WTP for Piece-Rate Incentives - Instructions

Our online component contained a section designed to elicit willingness to pay for incentive programs. This section began by explaining to participants 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 online component 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.

Incentive programs:

As part of the study you may receive an incentive program that will pay you money based on the number of days you exercise at YYY Fitness over the next 4 weeks (starting Monday, ${e://Field/mondaydate}).

For example, you could get selected for a program that pays you $5 per day you visit YYY Fitness in the next 4 weeks.

We want to know how valuable you find these types of incentives and how often you think you will go if you get each incentive program.

We will first do a few practice questions and then will explain more.

Figure A.1: Introduction to Willingness to Pay Section of the Study

Next, the study explained to participants 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.

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

Since participants 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.

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

Decision 1	$1 per-day incentive	$0 Fixed payment
Decision 2	$1 per-day incentive	$2 Fixed payment
Decision 3	$1 per-day incentive	$4 Fixed payment
Decision 4	$1 per-day incentive	$6 Fixed payment
Decision 5	$1 per-day incentive	$8 Fixed payment
Decision 6	$1 per-day incentive	$10 Fixed payment
Decision 7	$1 per-day incentive	$12 Fixed payment
Decision 8	$1 per-day incentive	$14 Fixed payment
Decision 9	$1 per-day incentive	$16 Fixed payment
Decision 10	$1 per-day incentive	$18 Fixed payment
Decision 11	$1 per-day incentive	$20 Fixed payment
Decision 12	$1 per-day incentive	$22 Fixed payment
Decision 13	$1 per-day incentive	$24 Fixed payment
Decision 14	$1 per-day incentive	$26 Fixed payment
Decision 15	$1 per-day incentive	$28 Fixed payment
Decision 16	$1 per-day incentive	$30 Fixed payment

Figure A.4: Decision Table

The study then asked participants 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 participant who switched from the one dollar incentive to the fixed payment at Decision 6 indicating a willingness to pay between eight and ten dollars.

Ok, the way you filled out the table says that you would prefer the incentive program if the available fixed payment is $8 or less. But if the fixed payment were at least $10 you would trade the incentive program for the fixed payment.

Does that sound right about what you prefer?

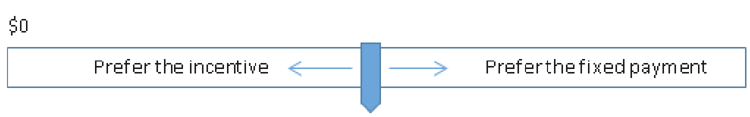
- Yes, that's right (go on to the next question)
- No, that's not right (fill out the table again)

Figure A.5: Comprehension Check for Table

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

Use a slider to answer these questions more quickly.

A faster way to figure out what you prefer between fixed payments and the incentive program is to use a slider.



The line below represents a range of fixed payments that correspond to the table of decisions on the previous page. Instead of checking off your preference in each decision row, you can indicate the same preferences by positioning the slider at the ***smallest fixed payment*** *that you prefer to the incentive program. Go ahead and position the slider:*

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

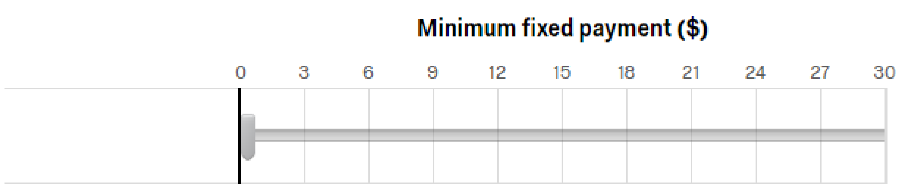


Figure A.6: Slider for WTP for $1 per Day Incentive Program

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

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

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

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

If the participant answered correctly on her second try, she was advanced directly to the next set of instructions. If the participant 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 online component, one of the incentive programs would be randomly selected and the participant would either be given that program or a fixed payment with the choice to be determined by the preferences she had indicated on the online component. See Figure A.9. After being presented with some answers to frequently asked questions (see Figure A.10), participants were instructed to use sliders to 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 participants.

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

If participants 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 one dollar per day incentive where the highest possible value on the slider was thirty dollars.

At the end of the online component, an incentive program and fixed payment were randomly drawn for each participant and the online component explained to the participant whether, in accordance with their preferences, they would receive the fixed payment or the incentive program. The example in Figure A.14 is for a participant whose choices revealed that she would prefer the fixed payment that was drawn to the incentive program that was drawn.

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.

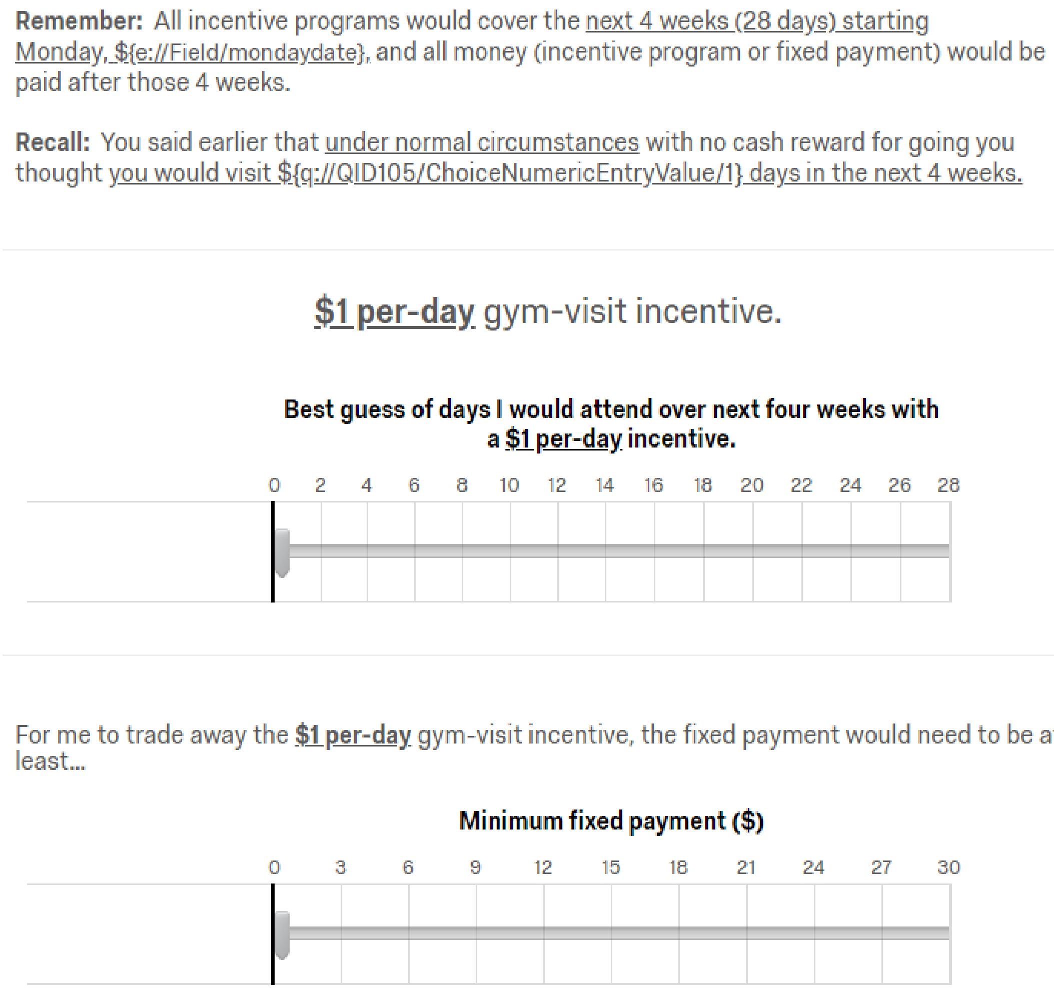
Your answers to these questions will not affect which incentive you get. So please simply give us your best realistic estimate of how many days you would attend in the next 4 weeks with that incentive.

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.

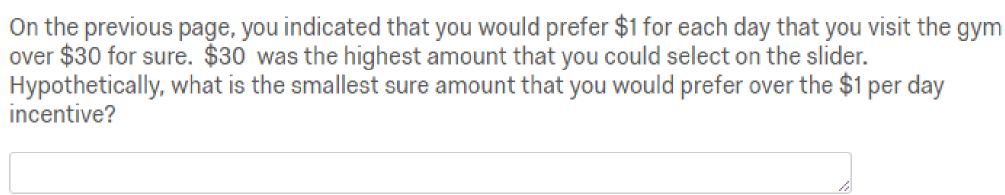
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28

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

Minimum fixed payment ($)

0 3 6 9 12 15 18 21 24 27 30

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



On the previous page, you indicated that you would prefer $1 for each day that you visit the gym over $30 for sure. $30 was the highest amount that you could select on the slider. Hypothetically, what is the smallest sure amount that you would prefer over the $1 per day 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 Online Component Announcement of Fixed Payment or Incentive Program