NBER WORKING PAPER SERIES

WHO CHOOSES COMMITMENT? EVIDENCE AND WELFARE IMPLICATIONS

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

A previous version of this paper circulated under "How are Preferences for Commitment Revealed?" We are grateful to seminar and conference participants at Harvard, Wharton, UC San Diego, University of Zurich, Dartmouth, Claremont Graduate University, Erasmus University, the Economics Science Association conference, the 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, Séverine Toussaert, and Jonathan Zinman for helpful comments. Paul Fisher, Max 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.

Who Chooses Commitment? Evidence and Welfare Implications
Mariana Carrera, Heather Royer, Mark Stehr, Justin Sydnor, and Dmitry Taubinsky
NBER Working Paper No. 26161
August 2019, Revised in October 2020
JEL No. C9,D9,I12

ABSTRACT

This paper develops an approach for estimating people's perceived and actual self-control problems in the field, and for investigating whether offers of commitment contracts are well-targeted tools for addressing self-control problems. In an experiment on gym attendance (N=1,248), we estimate a model of partially naive quasi-hyperbolic discounting, and we study (i) how commitment contract take-up relates to structural estimates of people's perceived and actual self-control problems and marginal benefits of behavior change, (ii) if commitment contract take-up is in part due to boundedly-rational understanding of incentives, and (iii) whether commitment contracts increase welfare, and how they compare to linear subsidies or taxes. We find high take-up of commitment contracts to increase gym attendance, which translates to significant increases in exercise. However, contract take-up is not positively related to estimates of actual or perceived self-control problems. A novel test suggests that this may be partly explained by boundedly-rational take-up decisions: many individuals are willing to take up commitment contracts both for higher gym attendance and for lower gym attendance, and there is a significantly positive correlation in people's propensity to take up both types of contracts. Our estimated model of quasi-hyperbolic discounting and gym attendance implies that offering our commitment contracts lowers consumer surplus and is less socially efficient than utilizing linear incentive schemes that achieve the same average change in behavior. We estimate an optimal exercise subsidy of $7.54 per attendance.

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

One of the central insights from economic models of time inconsistency and limited self-control 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). Although this insight has a number of economic implications, the most prominent focus in the field-experimental literature has been on demand for *commitment contracts*—reductions of opportunity sets through choice-set restrictions or self-imposed penalties for certain (mis)behaviors.[1] As shown in Table 1, there are thirty-three 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 high take-up rates (see Table 1) and significant effects on behavior documented in the literature suggest that commitment contracts could be welfare-enhancing, but this is not guaranteed (e.g., Heidhues and Kőszegi, 2009). Nor do existing results shed light on whether other approaches to behavior change, such as taxes or subsidies (e.g., Gruber and Kőszegi, 2001; O'Donoghue and Rabin, 2006), might be more or less efficient.

In this paper, we develop an approach for estimating people's perceived and actual self-control problems in the field, and for using those estimates to evaluate the benefits of commitment contracts versus other mechanisms for counteracting failures of self-control. Our approach allows us to answer three key research questions. First, "who" takes up the commitment contracts? Specifically, how does take-up of commitment contracts relate to people's actual and perceived time inconsistency and marginal benefits of behavior change? What are the *causal* effects of increasing people's awareness of their time inconsistency on their demand for commitment contracts? Second, do other factors—such as stochastic valuation errors in perception of incentives (see, e.g., Woodford, 2019, for a review)—affect take-up of commitment contracts? The existence of these other factors may help reconcile the high take-up rates observed in practice with the low take-up rates predicted by theory (see, e.g., Laibson, 2015). Third, taking into account all of the drivers of commitment contract take-up, do commitment contracts increase consumer surplus and social welfare? Are commitment contracts more or less efficient than the kinds of tax instruments studied by, e.g., Gruber and Kőszegi (2001) and O'Donoghue and Rabin (2006)?

We address these questions through a combination of theory and empirical findings from a field experiment on gym attendance with 1,248 participants. Our approach has three novel features. First, we directly assess how commitment take-up relates to reduced-form and structural estimates of both perceived and actual time inconsistency. In addition to offering commitment contracts, we utilize a separate experimental elicitation to estimate people's perceived and actual time inconsistency, as well as the marginal benefit of behavior change. Second, we utilize a novel approach to detecting stochastic valuation errors or other confounds in take-up of commitment contracts. We offer individuals commitment contracts both for going to the gym more and for going to the gym less, and we study the correlation in people's propensity to take-up both types of contracts. Third,

[1] An example of a penalty-based commitment contract is as follows. An individual places X dollars at stake to reach a future goal. If the individual is unsuccessful, the X dollars are lost and if successful, the X dollars are returned to the individual.

our rich experimental data allows us to estimate a structural model of quasi-hyperbolic discounting and partial naivete (Laibson, 1997; O'Donoghue and Rabin, 1999, 2001), and to validate it with out-of-sample tests—one of the first such estimates using field-experimental data. We then use this model to study the key question of whether it is more socially efficient to use commitment contracts or linear tax instruments to counteract failures of self-control.

Section 2 fleshes out our approach to estimating models of time inconsistency. The empirical content of models of time inconsistency consists of three objects: (i) how people desire to behave in the future, (ii) how people expect to behave in the future, and (iii) how people actually behave in the future. If people are time-consistent, then all three objects are identical. If people are time-inconsistent but fully sophisticated, then (ii) and (iii) will be equal but differ from (i). If people are time-inconsistent and are partially aware of their self-control problems, then all three objects will differ.

Objects (ii) and (iii) can be estimated directly by measuring people's forecasts and actual attendance at different levels of attendance incentives. We show that the wedge between (i) and (ii) can be elicited by extending the insights from DellaVigna and Malmendier (2004) and Acland and Levy (2015). Intuitively, the Envelope Theorem implies that a person who believes herself to be time-inconsistent, and forecasts say 8 attendances over the experimental period at an incentive of $\$p$ per attendance, should value a marginal $\$dp$ per attendance increase in incentives by $\$8dp$. Valuations above $\$8dp$ indicate that the person values the behavior change induced by the incentive increase—i.e., that the person's forecasted behavior does not align with the person's desired behavior. We call the deviation from the time-consistent benchmark the *behavior change premium*.

A key feature of the behavior change premium is that it is a continuous measure of perceived time inconsistency. Thus, although individual-level estimates of the behavior change premium may be noisy, we show that the average of the estimates is an unbiased measure of the mean behavior change premium in the population. In contrast, we show that because commitment contract take-up is a coarse, discontinuous measure of time inconsistency, it is neither a lower nor upper bound on people's time inconsistency when (i) people are not fully aware of their time inconsistency, (ii) there is a nontrivial benefit to flexibility because of uncertainty about the future, or (iii) take-up is partly affected by boundedly-rational stochastic valuation errors or perceived social pressure.

We show that in a model of quasi-hyperbolic preferences, incorporating meaningful levels of uncertainty and stochastic valuation errors generates two predictions about take-up of commitment contracts. First, increasing the level of sophistication about one's time inconsistency should *reduce* the take-up of penalty-based commitment contracts. Second, people will take up commitment contracts that not only target doing *more* of an activity with delayed benefits, but also contracts that target doing *less* of that activity. And if stochastic valuation errors are a strong enough driver of take-up, there will be a positive correlation in demand for the "more" and "less" contracts.

Our experimental design, summarized in Section 3, allows us to test these predictions, to provide reduced-form evidence for perceived and actual time inconsistency, and to provide structural estimates of a model of quasi-hyperbolic discounting. The experiment involved 1,248 members

of a fitness facility in a large city in the midwest of the United States, and consisted of an online elicitation followed by four weeks of observed gym attendance under different attendance incentives.

Following the measurement approach laid out in Section 2, we first elicited people's forecasted attendance over the next four weeks at different levels of piece-rate incentives that ranged from \$0 to \$12 per attendance. We then used an incentive-compatible procedure to elicit participants' willingness to pay (WTP) for different piece-rate incentives. Finally, we randomly assigned different piece-rate incentives to a subset of the subjects and measured the impact on actual gym attendance.

To study commitment contract take-up, we elicited demand for commitment contracts tied to attending the gym *at least* 8, 12, or 16 times over the next four weeks. For each of these thresholds, participants chose between an unconditional payment of \$80 and a conditional payment of \$80 that they received only if their attendance met or exceeded the threshold. We also 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 four weeks.

Finally, to estimate the causal effects of increasing participants' awareness of time inconsistency, we included a randomized information treatment prior to the elicitations, aimed at reducing overestimation of gym attendance.[2] The treatment provided participants with information about their past gym attendance and highlighted (truthfully) that members of this gym tended to overestimate how often they would use the gym.

In Section 5, we report reduced-form results on people's forecasted, desired, and actual attendance. On average, people overestimate their future gym attendance, although there is a highly positive correlation between forecasted and actual attendance. At the same time, we estimate a significantly positive average behavior change premium, which implies partial awareness of time inconsistency, on average. The estimates imply that, on average, participants valued increasing their future selves' gym attendance by \$1.78 to \$3.00 per visit. Reassuringly, the individual-level measures of the behavior change premium correlate positively with simple proxies for sophistication. We also find that our information treatment significantly increased the behavior change premium.

In Section 6, we report results on commitment contract take-up. We find high take-up of commitment contracts to attend the gym more, consistent with the take-up rates observed in other studies with similar designs (64% for 8+ visits, 49% for 12+ visits, and 32% for 16+ visits). We also find that participants who were randomly assigned to receive the conditional \$80 incentive for 12+ visits increased their attendance by 3.51 visits. These kinds of results are often interpreted as evidence for wide-spread awareness of time inconsistency, as well as the welfare benefits of commitment contracts.

However, we present a range of evidence that suggests that such inferences may be inappropriate in the absence of additional evidence. First, we find that there is no relationship between the individual-level measures of the behavior change premium and take-up of the commitment con-

[2] As we describe in Section 3, in our first wave of the experiment we had a simpler information treatment that only provided information about past visits at the gym and found that this did not meaningfully affect beliefs. The second two waves of the experiment used an enhanced information treatment, which we show in Section 5 significantly reduced expectations of gym visits.

tracts. Second—and in contrast to our results about the behavior change premium—our proxies for awareness of time inconsistency are on net unrelated to commitment contract take-up, while the information treatment significantly *decreased* the take-up of commitment contracts for higher gym attendance. Third, we find that 27-34% of participants chose commitment contracts to attend the gym less, and that the take-up of "more" and "less" contracts at each threshold is significantly *positively* correlated.[3] This is inconsistent will *all* take-up of commitment contracts reflecting a self-control strategy, but is consistent with our theoretical predictions about the consequences of stochastic valuation errors. This finding helps explain why commitment contract take-up is unrelated to the behavior change premium or to proxies for sophistication.

In Section 7, we combine our empirical results with a structural model to evaluate the welfare effects of commitment contracts, taking into account that at least some of the take-up reflects mistakes. We first use our data on piece-rate incentives to estimate a structural model of quasi-hyperbolic preferences with partial sophistication. We assume that all future utility is discounted by an additional $\beta \leq 1$, which we refer to as *present focus* in the language of Ericson and Laibson (2019). Following O'Donoghue and Rabin (2001), we parametrize misprediction of time inconsistency by allowing people to believe that their future selves behave as if their present focus is $\tilde{\beta}$. We estimate an actual average present focus parameter of $\hat{\beta} = 0.55$ and an average perceived present focus parameter of $\hat{\tilde{\beta}} = 0.84$. We estimate a (perceived) long-run benefit of exercise of $\hat{b} = \$9.66$ per attendance, which sits comfortably in the range of health benefits estimated in the public health literature. These estimates imply an average internality—the harms people impose on themselves due to present focus—of $(1 - \hat{\beta}) \cdot \hat{b} = \$4.39$.

We also find meaningful heterogeneity. Past attendance is positively correlated with health benefits b and the present focus parameter β, although it is not related to people's awareness of present focus, as measured by Augenblick and Rabin's (2019) awareness statistic, $(1 - \tilde{\beta})/(1 - \beta)$. Our information treatment lowered the perceived present focus parameter from $\hat{\tilde{\beta}} = 0.86$ to $\hat{\tilde{\beta}} = 0.78$ and increased awareness of present focus from $(1 - \hat{\tilde{\beta}})/(1 - \hat{\beta}) = 0.30$ to $(1 - \hat{\tilde{\beta}})/(1 - \hat{\beta}) = 0.49$. Commitment contract take-up, however, is largely unrelated to any of the model parameters.

We show that our estimated model accurately predicts how the threshold incentive schemes from commitment contracts affect people's behavior, which serves as an out-of-sample test because these incentives were not used to estimate our model. Turning to welfare implications, we estimate that people who take up commitment contracts are on average made worse off. On average, consumers who take-up the 8+, 12+, and 16+ commitment contracts incur losses equivalent to −$7.91, −$18.69, and −$10.51 per person, respectively. Although the commitment contracts induce people who don't exercise enough to exercise more, this is counteracted by two other forces. First, a substantial fraction of people fail to meet the contracts' thresholds and incur financial losses. Second,

[3]We present a range of robustness checks for these results. We show that 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. We also 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.

the nonlinear incentives of commitment contracts induce inefficient patterns of gym attendance, where present-focused individuals attend the gym too little at the beginning of the four-week period due to procrastination, but then attend the gym too much toward the end in an effort to avoid the financial penalty.

Of course, the fact that people who take up the contracts on average lose surplus does not imply that the contracts lower *social welfare*, since the revenue collected from people's penalty payments can be "recycled" to benefit people in other ways. We estimate that the contracts lead to modest gains in social welfare. However, these gains pale in comparison to the effects of linear per-attendance incentives that are offered to the entire population and scaled to match the average attendance increases generated by commitment contracts. We show that this is because commitment contracts are not well-targeted.

Our study fleshes out a number of mechanisms for why take-up and behavior change are not sufficient statistics for evaluating the efficacy of commitment contracts, and provides methods for assessing the importance of these mechanisms in other domains. This is illustrated by our specific results about how our commitment contracts are suboptimal tools for both measuring and addressing self-control problems in our exercise setting. Of course, this need not be true for all other domains of behavior or other types of contracts. In Section 8, we summarize a number of caveats to our specific results, and also discuss how extensions of our methods can be usefully utilized for related questions about data-driven incentive design for present-focused individuals.

1 Related literature

First, we contribute to work estimating structural models of time inconsistency, particularly in field settings. While there is a growing set of papers estimating the present focus parameter in the field after *assuming* either naivete or sophistication,[4] there is, to our knowledge, only a handful of papers that provide more complete and direct identification by estimating both people's actual and perceived present focus: Augenblick and Rabin (2019), Bai et al. (Forthcoming), Chaloupka et al. (2019), Skiba and Tobacman (2018), and Allcott et al. (2020). With the exception of Bai et al. (Forthcoming)—who use commitment contract take-up to estimate present focus—our paper is unique in additionally offering people commitment contracts and examining how the structural estimates vary with take-up decisions. Our estimation approach is most similar in spirit to that of Augenblick and Rabin (2019), who provide direct estimates of people's desired, forecasted, and realized effort in a laboratory experiment with college students. But unlike Augenblick and Rabin (2019), our approach does not rely on the assumption that future effort costs are deterministic, and can be tractably applied in many field settings. For example, Allcott et al. (2020) extend our approach to study present focus among payday loan borrowers—a complex decision environment

[4]For field estimates, see, for example, Laibson et al. (2018), Fang and Silverman (2004), Mahajan, Michel, and Tarozzi (2020), Paserman (2008), Martinez et al. (2020), and Shui and Ausubel (2005). There is also a large laboratory literature focused almost exclusively on estimating actual but not perceived time inconsistency; see, e.g., the review in Ericson and Laibson (2019) or the meta-analysis in Imai et al. (2020) on the convex time budget approach of Andreoni and Sprenger (2012) and Augenblick et al. (2015).

with non-separable payoffs and high uncertainty, non-quasilinearity in money, and potentially low financial literacy of experimental subjects. Our approach is also similar to, and builds on the early working-paper version of Acland and Levy (2012), but produces estimates of both β and $\tilde{\beta}$, and develops our behavior change premium statistic to provide a model-free test of perceived time inconsistency that is not tied to specific parametric assumptions.

Second, we contribute to a small set of papers studying the welfare effects of commitment contracts. Heidhues and Kőszegi (2009) show theoretically that partially naive individuals can harm themselves by taking up commitment contracts. John (2019) offers support for this prediction by studying how proxies for naivete relate to commitment contract take-up and the likelihood of failure. Sadoff et al. (2019) investigate how commitment contract take-up relates to intertemporal preferences reversals. Bai et al. (Forthcoming) estimate a parametrized distribution of β and $\tilde{\beta}$ from commitment contract choices under two assumptions that we relax: that individuals are in a fully deterministic environment and that no individuals take up commitment contracts due to stochastic valuation errors or perceived social pressure. Relative to these papers, we are the first to use empirical moments that are separate from contract take-up to directly estimate β, $\tilde{\beta}$, and internalities both for individuals who take-up the contracts and for those who do not. This is crucial for examining whether the contracts are well-targeted. Our welfare evaluation of commitment contracts is also the first to allow both a non-deterministic decision environment and stochastic valuation errors in take-up decisions.

Third, we contribute to a slightly broader set of papers that analyze drivers and correlates of commitment contract take-up. 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; Sadoff et al., 2019), our study is the first to report a *causal* estimate.[5] Our study also provides a uniquely detailed analysis of correlates, relating take-up to both a set of reduced-form proxies as well as to structural estimates of internalities and perceived and actual present focus. Augenblick et al. (2015) relate take-up to an estimate of one of the structural parameters, while the other papers focus on reduced-form proxies.

Relatedly, our paper is the first to directly investigate the possible role of stochastic valuation errors and perceived social pressure in commitment contact take-up.[6] Only seventeen of the thirty-three studies in Table 1 ever even mention potential confounds, and only eight discuss the confounds in depth as potential drivers of take-up.[7]

[5]This prior evidence on correlations is mixed, with some studies finding positive correlations between measured impatience and commitment demand (Ashraf et al., 2006; Augenblick et al., 2015; Kaur et al., 2015) but others finding negative correlations (Sadoff et al., 2019; John, 2019).

[6]Toussaert (2019) reports results consistent with stochastic valuation errors, finding 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 two violations away from all possible models.

[7]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 effects but consider them to be relatively minor determinants of commitment take-up, and another eight mention that they may 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, Bonein and Denant-Boemont (2015) discuss the role of peer pressure, and Kaur et al. (2015) and Schilbach (2019)

2 Theoretical predictions and measurement techniques

In this section, we set up and characterize the predictions of a model of quasi-hyperbolic discounters facing a task with stochastic immediate costs and deterministic delayed benefits. We first lay out our direct approach to assessing awareness of time inconsistency, then discuss implications for commitment contracts, and end by studying the implications of stochastic valuation errors.

2.1 Model setup

We consider individuals who in periods $t = 1, \ldots, T$ have the option to take an action $a_t \in \{0, 1\}$. Choosing $a_t = 1$ generates immediate stochastic costs c_t realized in period t as well as deterministic delayed benefits b realized in period $T + 1$. We assume that $c > 0$ with positive probability, but don't preclude the possibility of draws $c < 0$. For concreteness, we will often refer to $a_t = 1$ as attending the gym and $a_t = 0$ as not attending the gym, with the understanding that our results apply to the general model presented here and not just gym attendance.

For $\bar{a} = \sum_{t=1}^{T} a_t$, we consider incentive contracts that pay out in $T + 1$, denoted as $(y, P(\bar{a}))$, that consist of a fixed transfer y (which could be negative), and a contingent reward $P(\bar{a})$ for certain levels of gym attendance. The contingent component $P(\bar{a})$ is non-negative, with $\min_{\bar{a}\in[0,T]} P(\bar{a}) = 0$. We assume for simplicity that utility is quasilinear in money, given the relatively modest incentives involved in our experiment.

A simple piece-rate incentive contract with per-attendance incentive p has $y = 0$ and $P(\bar{a}) = p\bar{a}$. Penalty-based commitment contracts for attending the gym at least r times are $(-p, P)$, with $P(\bar{a}) = p \cdot \mathbf{1}_{\bar{a}\geq r}$. Conversely, a contract $(-p, P)$, with $P(\bar{a}) = p \cdot \mathbf{1}_{\bar{a}<r}$, is a penalty-based contract for *not* going to the gym r times or more.

We assume that individuals have quasi-hyperbolic preferences given by $U^t(u_t, u_{t+1}, \ldots, u_T, u_{T+1}) = \delta^t u_t + \beta \sum_{\tau=t+1}^{T+1} \delta^\tau u_\tau$, where u_t is the period t utility flow. By construction, $u_t = -a_t \cdot c$ for $1 \leq t \leq T$ and $u_{T+1} = y + b\bar{a} + P(\bar{a})$. 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 $\delta = 1$ given the short time horizons involved in our experiment. We use $V(y, P)$ to denote the individual's subjective expectation (given beliefs $\tilde{\beta}$) about utility under contract (y, P).

2.2 Measuring time inconsistency and the behavior change premium

Figure 1 illustrates the framework motivating our experimental design and analysis of time inconsistency. The x-axis is the agent's attendance, and the y-axis is incentives for that behavior, which for simplicity of illustration we consider to be linear per-attendance incentives. There are three attendance curves: actual, forecasted, and desired. These curves are meant to depict averages over all realizations of c, meaning that, e.g., they correspond to the actual, forecasted, and desired *probabilities* of attending the gym in the one-period model with $T = 1$. We draw the curves as linear

discuss both perceived social pressure and confusion.

for graphical illustration, but our formal results do not require linearity. We use $\tilde{\alpha}(p)$ to denote an agent's forecasted attendance at incentive level p.

In the absence of present focus ($\beta = \tilde{\beta} = 1$), these curves are identical. For a fully sophisticated agent ($\tilde{\beta} = \beta$), the forecasted and actual curves are identical, while for a fully naive agent ($\tilde{\beta} = 1$), the forecasted and desired curves are identical. The three curves intersect at incentive $p = -b$, because at this point the net future benefit of the action ($p + b$) is zero, hence the different discount factors are irrelevant, and behavior is governed solely by the stochastic costs of effort. We assume for our graphical illustration that $c \geq 0$, so that attendance is zero at $p = -b$.

The actual and forecasted attendance curves can be measured directly at the population level by randomizing incentives. The desired attendance curve can be inferred from the agent's willingness to pay (WTP) for a change in the incentive level. In the figure, we consider an increase from $p = p'$ to $p = p' + \Delta$. At incentives p', the person's perceived total surplus is denoted by the area of ABCD in the graph—the difference between marginal benefits p' and marginal costs, integrated between 0 and $\tilde{\alpha}(p')$. As the incentive increases by Δ, the agent's perceived surplus rises to AEFG. The difference between AEFG and ABCD consists of two trapezoids: BEFC and DCFG. The area BEFC corresponds to the increase in total surplus that the agent would receive if she were time-consistent (with actual attendance given by $\tilde{\alpha}(p)$). The area DCFG is what we call the behavior change premium: the additional increase in surplus that results from the fact that the agent would be willing to pay to motivate her future self to attend the gym more because her desired attendance is above her forecasted attendance.

The area of DCFG can be backed out from the forecasted attendance curve and an agent's WTP for an increase from p' to $p' + \Delta$. The area of trapezoid BEFC is simply a function of the agent's forecasts, and is given by $\Delta \cdot (\tilde{\alpha}(p') + \tilde{\alpha}(p' + \Delta))/2$. The WTP for the incentive increase Δ is simply the area of BEFC and DCFG. Thus, the area of DCFG is obtained by differencing out the area of BEFC from the WTP.

Importantly, the quasi-hyperbolic discounting model provides a tight parametrization of the wedges between the curves—which enables our structural estimates in Section 7. Roughly speaking, the wedge between the actual and forecasted curves is proportional to $\tilde{\beta} - \beta$. The wedge between the forecasted and desired curves is proportional to $1 - \tilde{\beta}$, which we show formally below can be backed out from the behavior change premium.

Formally, consider a piece-rate contract that pays the agent p every time she chooses $a_t = 1$. Define an individual's willingness to pay for the contract, $w_i(p)$, to be the smallest y such that she prefers a sure payment of y over this contract. We then have the following:

Proposition 1. *Assume that the costs in each period t are distributed according to smooth density functions, and that terms of order Δ^3 and $\Delta^2\tilde{\alpha}''(p)$ are negligible. If $\tilde{\beta} = 1$, then*

$$\frac{w(p+\Delta) - w(p)}{\Delta} = \frac{\tilde{\alpha}(p+\Delta) + \tilde{\alpha}(p)}{2} \tag{1}$$

If $\tilde{\beta} < 1$ and the costs are distributed independently, then

$$\frac{w(p+\Delta)-w(p)}{\Delta} = \underbrace{\frac{\tilde{\alpha}(p+\Delta)+\tilde{\alpha}(p)}{2}}_{\textit{Surplus if time-consistent}} + \underbrace{(1-\tilde{\beta})(b+p+\Delta/2)\frac{\tilde{\alpha}(p+\Delta)-\tilde{\alpha}(p)}{\Delta}}_{\textit{Behavior change premium}} \quad (2)$$

The assumptions about negligible terms in Proposition 1 are essentially the same as those in the canonical Harberger (1964) formula of the dead-weight loss of taxation: the change in incentives is not too large, particularly relative to the degree of curvature in the region of the incentive change. Plainly, these assumptions always hold in the limit of $\Delta \to 0$, where the left-hand-side of (1) and (2) approaches $w'(p)$ and $(\tilde{\alpha}(p+\Delta)-\tilde{\alpha}(p))/\Delta$ approaches $\alpha'(p)$.

The proposition formally shows that the WTP for an increase in incentives consists of two terms, as in our graphical argument. The first term is the surplus, per dollar of incentive change, that an individual would obtain if she were time-consistent and behaved according to her forecasts. As shown in equation (1) of the proposition, the WTP for individuals with $\tilde{\beta} = 1$ has a simple characterization under general conditions that only require "smooth" behavior—the payoffs in each period t need not be identically or independently distributed. This characterization is a corollary of the Envelope Theorem, and analogs of this expression hold in any stochastic dynamic optimization problem, as shown in extensions by Allcott et al. (2020). Thus, deviations from this expression, which we label

$$BCP(p,\Delta) := \frac{w(p+\Delta)-w(p)}{\Delta} - \frac{\tilde{\alpha}(p+\Delta)+\tilde{\alpha}(p)}{2}, \quad (3)$$

indicate that $\tilde{\beta} \neq 1$. In particular, $BCP > 0$ implies that $\tilde{\beta} < 1$. We call this reduced-form measure the *behavior change premium per dollar of financial incentives.* It is the WTP for a change in behavior that is induced by a \$1 increase in piece-rate incentives. As equation (2) of the proposition shows, if costs are distributed independently (but not necessarily identically), the BCP is a linear function of perceived time inconsistency $(1-\tilde{\beta})$—a feature we use in our structural estimation.[8]

2.3 Commitment contracts are often unattractive when costs are uncertain

The behavior change premium is a monotonic function of $1-\tilde{\beta}$, and is in fact a linear function under additional plausible assumptions. However, here we show that the take-up of commitment contracts is not related to $\tilde{\beta}$ (and plainly β) in any straightforward way, and need not even be a monotonic function of $\tilde{\beta}$. We can illustrate this point by returning to Figure 1. For concreteness, suppose now that there is a single period of action ($T = 1$), so that the attendance curves in Figure 1 give the probability of $a = 1$. Given a CDF F of the costs, the actual, forecasted and desired attendance (probability) curves are given by $F(\beta(b+p))$, $F(\tilde{\beta}(b+p))$,and $F(b+p)$, respectively. We let the vertical line running through points H and I to correspond to the point where the individual attends the gym with probability 1.

[8]Quasilinearity in money is 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. If participants are non-negligibly risk averse over small amounts of money, then the statistic in (3) underestimates the WTP for behavior change, and leads to overestimates of $\tilde{\beta}$ (see Allcott et al., 2020, for further details).

In the one-period model, a commitment contract for attending the gym is then a contract of the form $(-p, pa)$, where the individual loses money p if she does not go to the gym. Binding commitment contracts are special cases with $p = \infty$.

Consider a commitment contract where the individual puts an amount Δ at stake that is returned if and only if the agent chooses $a = 1$. This is equivalent to the individual receiving an increase Δ in attendance incentives, while also having to pay Δ non-contingently. The surplus loss from an non-contingent payment of Δ is the large rectangle BEHI in the figure. Thus, a commitment contract is perceived to be valuable if the behavior change premium DCFG exceeds the loss CFHI. In general, this need not be the case, and as the figure illustrates, this is particularly unlikely to be the case when the probability of attendance is non-negligibly below 100%. The perceived value of commitment contracts also does not increase with perceived present focus $1-\tilde{\beta}$: as perceived present focus increases, the forecasted attendance curve rotates to the left, which increases the financial loss CFHI without necessarily increasing the behavior change premium (since the width CZ shrinks). Moreover, the actual benefits of a commitment contract may be smaller than the perceived benefits, for the same reason that a leftward rotation of the forecasted curve can increase the financial costs more than it increases the behavior change premium.

In Appendix A.2 and A.4 we derive two general results about the demand for commitment contracts when costs are uncertain. First, we show that for a broad class of stochastic cost distributions, the quasi-hyperbolic model predicts that there should not be demand for *any* commitment contract when there is at least a moderate chance that costs exceed delayed benefits. Second, when there is enough uncertainty to make commitment contracts unattractive, the perceived harms of a commitment contract, given by the difference between CFHI and DCFG in the figure, are *increasing* in perceived present focus $1 - \tilde{\beta}$.[9] That is, people who perceive themselves to be more present-focused will find commitment contracts less attractive (i.e., more harmful).

There are two key conditions on the distribution of cost draws under which the value of commitment contracts 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 in a small neighborhood of b.

As a simple illustration for the case of $T = 1$, Figure 2 summarizes commitment contract demand for the case in which c is uniformly distributed on $[0, 1]$. For the uniform case, the attendance curves are linear and the predictions are simple. Individuals prefer binding contracts over penalty-based contracts, and they want them if and only if $b + (1 - \tilde{\beta})b \leq 1$.[10] The figure shows little demand for commitment. For example, no individuals with $\tilde{\beta} \geq 0.8$ desire any kind of commitment contract

[9] Heidhues and Kőszegi (2009) and John (2019) present results implying that the value of commitment contracts will be non-monotonic in $\tilde{\beta}$, with individuals with either low or high $\tilde{\beta}$ choosing not to take up contracts. We replicate this result formally in Appendix A.2 and A.4 in our setting, but show that it only holds for cases where there is little uncertainty about the costs of action.

[10] 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.

when the costs of attendance exceed the benefits at least 20% of the time—an arguably modest degree of uncertainty.

2.4 The consequences of stochastic errors and perceived social pressure

In light of the results above—which replicate and extend points previously made by Laibson (2015) and others—a natural question is why we see *so much* demand for commitment in behavioral economics experiments. One possible reason is that because evaluating incentive schemes may be complicated, individuals may do so imperfectly. 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 perceived social pressure.

To formalize, our reduced-form econometric model of these effects supposes that for a given decision j, individual i behaves as if her forecasted utility under contract (y, P) is

$$\widehat{V}(y, P) = \tilde{\beta}_i y + \varepsilon_{ij} V(0, P) + \eta_i \mathbf{1}_{P \neq 0} \tag{4}$$

where $E[\varepsilon_{ij}] = 1$ and $\mathbf{1}_{P \neq 0}$ 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.[11] The η_i term, which need not be positive, captures perceived social pressure. 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. 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 (4) as the *imperfect perception model.*[12]

2.4.1 Commitment contract take-up is systematically biased by imperfect perception

The take-up of commitment contracts is a particularly problematic measure with imperfect perception because binary take-up decisions are biased by even mean-zero valuation errors (Aigner, 1973; Hausman, 2001). Even if the errors are symmetric—say 10% of the individuals always choose the

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

[12] Note that 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 $a_t = 1$. 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."

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.

As we show formally in Appendix A.2 and A.4, the imperfect perception model generates three predictions for penalty-based commitment contracts, where an individual loses an amount p unless she attends the gym at least a certain number of times:

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.[13]

The intuition for the first prediction is that imperfect perception or perceived social pressure 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 valuation errors, then individuals with the highest variance in the stochastic valuation term ε will be the most likely to take up both types of contracts.

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 (see Appendix A.2 and A.4). 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.

We emphasize that these predictions hold even under our plausible assumption that $E[\varepsilon_{ij}] = 1$—i.e., that the stochastic valuation errors are mean-zero—and even when $\eta_i = 0$ for all individuals.

2.4.2 Estimates of the behavior change premium are more robust

Our measure of the behavior change premium is robust at the population level, because it is a continuous variable that preserves the mean-zero nature of people's valuation errors. Specifically, let the subscript i denote each individual i's WTP w, beliefs α, and so forth. Then the results of Proposition 1 can be restated as population averages:

Corollary 1. *Under the assumptions in Proposition 1 and the imperfect perception model, if* $\tilde{\beta}_i = 1$ *for all* i *and* $p > 0$ *then*

[13] 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.

$$E\left[\frac{w_i(p+\Delta)-w_i(p)}{\Delta}\right]=E\left[\frac{\tilde{\alpha}_i(p+\Delta)+\tilde{\alpha}_i(p)}{2}\right] \tag{5}$$

and if $\tilde{\beta}_i<1$ *for some* i *and costs are independent across time then*

$$E\left[\frac{w_i(p+\Delta)-w_i(p)}{\Delta}\right]=E\left[\frac{\tilde{\alpha}_i(p+\Delta)+\tilde{\alpha}_i(p)}{2}+(1-\tilde{\beta}_i)(b_i+p+\Delta/2)\frac{\tilde{\alpha}_i(p+\Delta)-\tilde{\alpha}_i(p)}{\Delta}\right] \tag{6}$$

We condition on $p>0$ in the corollary because that allows the fixed terms η_i to be differenced out. The formula also continues to hold if individuals' elicited beliefs are a noisy function of their true beliefs, as long as the noise is mean-zero.[14] Core to our result is that WTP can range from below to above expected earnings, meaning that the measure of WTP for behavior change can range from negative to positive.[15] Having some, but not full, continuity in a commitment measure is insufficient.[16]

Variations of our imperfect perception model in which valuation errors are not mean-zero, or in which perceived social pressure rises with stakes, would invalidate the methodology we propose here, along with using commitment demand as a measurement tool, and all other approaches to measurement of time inconsistency. Fortunately, the key assumptions behind Corollary 1 are testable: individuals who expect no change in behavior ($\tilde{\alpha}_i(p+\Delta)-\tilde{\alpha}_i(p)=0$), should have an average behavior change premium equal to zero when $p>0$. If instead η_i increased with p, or if $E[\epsilon_{ij}]>1$, then we would estimate a positive behavior change premium even for individuals who expect no behavior change. We implement this test in our reduced-form results in Section 5.4.

3 Experimental 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.[17] The university has a separate facility for undergraduates.

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. Appendix Table A20 shows

[14] Systematic over-statement of true beliefs would make this a particularly conservative test, as this would bias against us finding a demand for behavior change.

[15] 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 implies 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.

[16] For example, 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.

[17] There are three membership types at the gym: regular, graduate student, and members through a wellness program offered by their health insurance company. Graduate students have a subsidized membership fee by semester, included by default with their tuition and fees. Members of a health insurer's wellness program are also able to obtain heavily subsidized memberships. Regular members pay an initiation fee and a monthly membership fee, which varies based on their affiliation with the university or other local employers.

the ordering of all parts of the online component of the study, which we summarize in more detail below. Enrollment was limited to people over the age of 18 who had held memberships over the past eight weeks. 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. Waves 1, 2, and 3 had 350, 528, and 414 participants, respectively.[18]

A key feature of the design is that we elicit preferences for commitment contracts and valuations of linear attendance incentives from *all* participants in an incentive-compatible manner, while at the same time generating random assignment of contracts and attendance incentives for *most* participants.

Information treatment Before answering any of the questions described below, 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 3a). 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 3c).

Participants randomized into the no-information control group proceeded directly to the elicitations described below.

Forecasted attendance and WTP for incentives 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.

Additionally, participants were asked to consider six different incentive contracts for 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.

For each incentive, 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.

[18] 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.

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.[19] Consistent with our theoretical model, all financial rewards were paid out after the four-week period.

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.

We devoted several screens to developing participants' understanding of how to use a slider to indicate WTP and why truth-telling was incentive compatible. We 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 D.1 for details.

We did not incentivize accuracy of people's attendance forecasts because according to standard models of time inconsistency, individuals with $\tilde{\beta} < 1$ could use these forecasts as a means of commitment: stating a forecast higher than one's actual belief would incentivize additional attendance.[20] Truth-telling is strictly dominant in the absence of financial incentives if people have even a small aversion to lying.

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. In all three waves, participants were given the "more visits" commitment choice shown in Figure 4(a) and the "fewer visits" commitment choice shown in Figure 4(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.[21]

[19] 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.

[20] Although Augenblick and Rabin (2019) show that this inflation is theoretically small for small incentives in deterministic environments, this is not generally true in environments featuring some uncertainty, such as ours.

[21] 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

Incentive-compatibility and assignment of attendance incentives One question was randomly chosen to determine each participant's attendance incentive. When the selected question involved a piece-rate incentive, the participant's WTP for that incentive was compared against a randomly drawn fixed payment. Fixed payments were drawn from a mixture distribution with two components: a uniform distribution from \$0-\$7 (mixture weight = 0.99), and a uniform distribution from the full range of slider values (mixture weight = 0.01). The rationale for this distribution was to avoid the endogenous assignment of incentives to participants with higher WTPs for those incentives.

Given this design, incentives were exogenously assigned, with the exception of two rare cases. The first case is when the fixed payment draw exceeded \$7 (n=12). 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 attendance incentive, which would bias our estimation of incentive effects on gym visits due to selection. These 44 observations are excluded from the analyses throughout.

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.[22]

The targeted incentives were varied to document the effects of different incentive sizes.[23] 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."[24] Since an incentive

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 2.4. This also motivated us to randomize some participants into actually receiving the commitment contracts, to make sure that we could replicate previous findings that the commitment contracts do alter behavior (thereby also confirming that participants were not confused about the terms ex-post)—we discuss this randomization below.

[22]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."

[23]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.

[24]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

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.

Although the variation of incentive scheme assignments across waves is not ideal for some of our analyses, we find that the participant pools look similar across waves, as shown in Appendix D. We further include wave fixed effects in relevant analysis below.

Announcement and disbursement 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 their incentive and reminded them 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.

4 Data

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.

Sample Table 2 summarizes characteristics of our sample. The participant pool is 61% female with a mean age of just under 34 years. 57% 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 6.9 visits over the past four weeks.

Column 3 shows the p-values for a test that the information treatment group mean equals that of the information 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.

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

on their own. All but five individuals (1.2% of wave 3 participants) who were asked this question chose the $80 incentive over $0.

affiliated with a university. For example, 95.2% 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.[25] 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.

Attention checks We have a few measures that proxy for engagement and attention to our online elicitations. First, as described in Section 3, we had two questions that offered a binary choice in which one of the choices, \$0, was clearly dominated by the other. Only 1.8% of participants chose a dominated option. Second, we had an attention check question that 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. Only 3.5% of participants failed the attention check. Finally, we had two comprehension checks about the WTP elicitations (see Appendix D.1 for details) and can use failing both as an additional indicator of lack of engagement. We find that only 4.3% of participants failed these comprehension checks twice. Taken together, these statistics suggest that attention and engagement were high, and compare favorably with most other lab-in-the-field studies.

5 Actual, forecasted, and desired attendance

5.1 Actual and forecasted attendance

Figure 5 summarizes the forecasted and actual attendance curves, as introduced in Section 2.2. Both forecasted and actual attendance increase significantly with incentives, and there is a sizeable gap between the two, consistent with naivete ($\tilde{\beta} > \beta$). On average, participants forecasted 11.5 visits in the absence of incentives and 17.7 visits with the \$7 incentive during the four-week study period. In reality, participants' attendance was 7.2 visits in the absence of incentives and 13.3 visits with the \$7 incentive.

Figure 6 presents a binned scatter plot of actual attendance versus expected attendance for the (randomly assigned) incentive people actually received. Although participants are over-optimistic about their attendance, the Figure shows a tight relationship between forecasted and realized attendance.[26]

Figure 7 shows how the information treatments affected expectations and actual visits, splitting the sample into information treatment and control groups. Our simple wave 1 information treatment had no effect on either expectations of visits or realized visit patterns, as shown in panel (a). By

[25]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).

[26]The fact that first point does not lie below the 45-degree line does not imply that some people are under-optimistic. This is consistent with mean-zero noise in stated beliefs generating a form of mean-reversion between actual and forecasted behavior.

contrast, the enhanced information treatment in waves 2 and 3 had a significant effect on beliefs that partially reduced participants' overoptimism, as seen in panel (b). This evidence shows that expectations are malleable, and this "first-stage" allows us to study the causal effects of sophistication on the behavior change premium and commitment contract take-up.

5.2 Willingness to pay for incentives

Figure 8 presents binned scatter plots of how WTP for the incentives varies with people's forecasts about attendance given those incentives. As would be implied by standard models, there is a tight relationship between WTP and both the size of the incentive and people's forecasted attendance with that incentive. Moreover, the size of the incentive changes not only the level of WTP, but also its slope with respect to forecasted attendance.

Figure 9 plots the average WTP 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' forecasted attendance. The WTP is above participants' subjective expected earnings for low incentives. For example, under a \$1 per-visit piece-rate, participants believed that they would attend an average of 12.92 times but had an average willingness to pay for a \$1 piece-rate incentive of \$18.30, \$5.38 more than their subjective expected earnings. The fact that people are willing to pay more for small incentives than they expect to earn is consistent with the theoretical predictions for agents that are aware of present focus (i.e., $\hat{\beta} < 1$). We also observe that the WTP is below the expected earnings on average for high incentives. This is consistent with the implication of equation (2), given moderate perceived present focus ($E[\tilde{\beta}_i]$ reasonably close to 1).[27]

5.3 The behavior change premium

The seven different incentive levels for which we elicited WTP and forecasts allow us to produce a precise estimate of the average behavior change premium. Formally, order the incentive levels $p_0 = 0, p_1, \ldots, p_K$ in ascending order. For each pair of adjacent incentives, p_k and p_{k+1}, we can construct an estimate of the behavior change premium according to equation (3), applied to $p = p_k$ and $\Delta = p_{k+1} - p_k$. We then take the (unweighted average) across all participants and all incentive pairs. We focus primarily on the average, rather than individual differences, because according to Corollary 1, the average statistic is the unbiased measure of the mean behavior change premium 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]

[27] 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, $\frac{d}{dp}E[w_i(p)] < E[\alpha_i(p) + \alpha_i'(p)]$.

[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.

Figure 10 shows the average value across six incentive levels, as well as the average excluding the valuation of increasing the piece-rate from \$0 to \$1, along with 95% confidence intervals. On average, participants exhibited a behavior change premium of \$2.01 per \$1 of incentive increase. However, this valuation is driven in part by an especially large premium assigned to the \$1 incentive. As Corollary 1 shows, if there are fixed social pressure effects (η_i, in the notation of that section) influencing willingness to pay for contingent incentives, the more robust measure of the behavior change premium is calculated only from changes in positive piece-rate levels. This more conservative average is \$1.20 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 the behavior change premium translate to values of \$3.00 per attendance when we include WTP for the \$1/visit incentive and \$1.78 per attendance using the more conservative estimate that excludes WTP for the \$1/visit incentive.

5.4 Correlates and determinants of the behavior change premium

Table 3 examines the relationship between proxies for people's perceived present focus and the behavior change premium. One proxy for people's beliefs about their present focus is the gap between their stated goal for attendance over the study period and their stated beliefs about how often they will go to the gym without incentives during the study period. A larger gap indicates a larger difference between people's desired and forecasted attendance. In column 1 of Table 3 we regress the more conservative measure of the behavior change premium (i.e., excluding the \$1/visit incentive) on a standardized measure of the gap between goal and forecasted attendance. The results show that a one standard deviation increase in the gap between stated goal and expected visits is associated with a \$0.68 increase in the behavior change premium, compared to an overall mean of \$1.17.

Our second measure of people's awareness of present focus is the difference between participants' actual attendance under the incentive they were randomly assigned and their expected attendance under that incentive. This difference is negative on average, reflecting participants' over-optimism, so a larger number would imply a smaller gap between actual and expected behavior. Column 2 of Table 3 shows that this measure of awareness of present focus is also strongly related to the behavior change premium. A one standard deviation decrease in the gap corresponds to a \$0.50 increase in the behavior change premium.

Of course, these patterns could be confounded by a range of issues that emerge when analyzing proxies for constructs like sophistication. Our experimental information treatment allows us to estimate a causal relationship between the behavior change premium and beliefs about present focus. In column 3 of Table 3 we regress the behavior change premium on indicators for the information treatments. Consistent with the strong first stage documented in Section 5.1, the enhanced information treatment significantly increased the average behavior change premium, raising the measure

by \$1.36 from the information control group average of \$0.66. Consistent with the null first stage documented in Section 5.1, the wave 1 information treatment had no effect on the behavior change premium. Column 4 shows that the point estimates are unchanged when including both proxies for sophistication and the information treatment dummies in the regression. In Section 7 we present additional evidence that the information treatment is primarily acting on people's beliefs $\tilde{\beta}$, rather than through other channels that might change their forecasts of future attendance.

Finally, in Appendix B.1, we regress the behavior change premium on people's expected change in behavior. Consistent with the Proposition 1, we find that it is strongly related to the expected change in attendance. Moreover, when excluding the \$1/visit incentive, the constant term in column 1 of Table A1 implies that the behavior change premium is indistinguishable from zero for individuals who expect no change in behavior. By contrast, the behavior change premium is significantly positive for the \$1/ visit incentive for individuals who expect no change in behavior (constant term in column 3 of Table A1). These results are consistent with the presence of fixed effects η_i that are differenced out from the estimated behavior change premium when the \$1/visit incentive is excluded.

In summary, we find that the behavior change premium is small in the control group, is significantly affected by the information treatment, varies strongly with proxies for sophistication, varies strongly with individuals' subjective beliefs about behavior change, and is approximately zero for individuals not expecting behavior change. Taken together, these results provide strong support for the assumptions of Corollary 1.

6 Take-up of commitment contracts

6.1 Take-up of "more" commitment contracts

Participants in our study had high take-up of commitment contracts to visit more than 8, 12, or 16 times. The take-up rates were 64% at the 8 visit threshold, 49% at the 12 visit threshold, and 32% at the 16 visit threshold. These take-up rates fit comfortably in the literature. 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.[29] 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 31% to 55% for the commitment options.

Consistent with the existing literature, we find that commitment contracts had 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 3.51 visits (p-value < 0.01) for those participants who wanted the contract, and by

[29] Overall, the take-up rates for penalty-based contracts that do 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.

4.04 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 (35%).

Our results, like those in prior studies, would typically be interpreted as clear evidence of widespread awareness of present focus. However, we show that such inference may not be warranted without additional stress tests.

6.2 Commitment contract take-up correlates only weakly with the behavior change premium

We find only a weak correlation between take-up of commitment contracts and the behavior change premium. In Table 4 we regress take up of commitment contracts against the individual-level measured behavior change premium. A one standard deviation increase in the behavior change premium is associated with around a 2 percentage point increase in the take-up of commitment contracts. Appendix Table A2 shows that this association is even smaller for the control group who did not receive the information treatments.

One potential reason for the lack of association between the behavior change premium and commitment contract take-up could be that both measures are noisy and there is attenuation bias in the relationship. However, the analysis in Table 3 showed very strong associations between the behavior change premium and our proxy for sophistication, suggesting the measure is not so noisy as to attenuate all relationships. Moreover, the average pairwise correlation of the individual-level behavior change premium at different incentive levels is 0.17 (bootstrapped cluster-robust s.e. 0.06) and the average pairwise correlation of demand for the different "more" contracts is 0.49 (bootstrapped cluster-robust s.e. 0.02).

Next, we examine how take-up of "more" commitment contracts correlates with our proxies for sophistication introduced in Section 5.4. Table 5 shows that the the gap between goal and expected attendance is possitively associated with take-up of commitment contracts. However, in contrast to the relationship with the behavior change premium, the association with commitment contract take-up is relatively small in magnitude: a one standard deviation increase in the gap between goal and expected attendance is associated with a 3.9 percentage point increase in the take up of commitment contracts, from an average take-up rate of 0.49. Moreover, and in starker contrast to our results on the behavior change premium, column 2 shows that participants who are more over-optimistic about their gym attendance are actually *more* likely to take up commitment contracts for higher gym attendance.

Finally, Section 2 introduced the prediction that if there is enough uncertainty about the costs and benefits of future gym attendance, then the perceived harms of commitment contracts will be increasing in the degree of perceived present focus ($1 - \tilde{\beta}$). Thus, if contract demand is influenced by imperfect perception, an (exogenous) increase in sophistication should decrease the take-up of contracts to attend the gym more. Consistent with this prediction, column 3 of Table 5 shows that the enhanced information treatment *decreased* the take-up of "more" commitment contracts by

8.0 percentage points. This is in stark contrast to the strong positive effect that the information treatment had on the measured behavior change premium. Consistent with a null first stage, the wave 1 information treatment had no effect on commitment contract take-up.

Collectively, these results are consistent with the hypotheses introduced in Sections 2.3 and 2.4—that commitment contracts might be most unattractive to those with stronger perceived present focus, or that their take-up may be influenced by noisy valuation and perceived social pressure. The next section provides a more direct test of whether stochastic valuation errors are affecting the take-up of commitment contracts.

6.3 Commitment contract take-up appears to reflect imperfect perception

To test for potential noisy valuation and perceived social pressure, we use our novel design feature of offering people "fewer visits" commitment contracts. Table 6 summarizes take-up of both "more" commitments and "fewer" commitments at each of the visit thresholds. Column 2 shows that approximately one-third of participants selected the "fewer visits" contracts. 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 2.4 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 stochastic valuation errors or perceived social pressure are prevalent in commitment contract take-up.

An even stronger prediction of our 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. Consistent with this, the last two columns of Table 6 show that participants who chose the "fewer" commitment contracts were significantly more likely to choose the "more" commitment contracts, and vice versa. Appendix Table A3 shows that these patterns are consistent in both our information control group and the group receiving the enhanced information treatment.

While these results suggest the presence of stochastic valuation errors or social pressure effects, they do not imply that all take-up of commitment contracts is explained by these confounds. For example, 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. However, in Appendix Table A4 we investigate the association between the measured behavior change premium and taking up "more" commitment contracts but not also "fewer" contracts and do not find any positive association. This suggests that it may not be possible to reliably identify the behavior change premium by simply restricting to individuals who take-up "more" contracts but not "fewer" contracts.

6.4 Robustness of commitment contract results

6.4.1 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 2.4, 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 4). For example, at the 12-visit threshold, the "more visits" contract description underlined "at least 12" and the "fewer visits" contract description 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 7 shows that participants differentiated between questions about perceived likelihood of success in a "more" contract versus a "fewer" contract.[30] 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, as shown in column 4. This implies that at least in answering the forecasting questions, participants were not simply misreading the "fewer" contract to be the "more" contract. In Appendix 6.4.1, we continue to build on this analysis and present correlations of commitment contract take-up with (i) perceived likelihood of success under "more" commitment contracts (Appendix Table A5), (ii) subjective expected attendance in the absence of incentives (Appendix Table A6), (iii) past attendance (Appendix Table A6), and (iv) desired goal

[30] For Table 7, we restrict our analysis to wave 3, the only wave for which we elicited beliefs about the likelihoods of meeting the commitment contract thresholds.

attendance (Appendix Table A6). Each of these variables is significantly positively correlated with take-up of "more" contracts, and significantly negatively correlated with take-up of "fewer" contracts.

6.4.2 Results are not a consequence of disengagement from the study

In Section 4 we summarized results from attention and comprehension checks, which suggest strong engagement and attention. When we exclude the small percentage of participants who failed a comprehension check or attention check or chose a dominated option, overall demand for the "fewer" contracts falls from 31% to 30%, and this exclusion has no effect on demand for the "more" contracts. While these proxies cannot be guaranteed to identify all individuals who disengaged or misunderstood some portion of the study, the lack of correlation between the proxies and demand for commitment contracts implies that disengagement or misunderstanding is unlikely to drive our results.

6.4.3 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 visits 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 demand for choice-set restrictions drops decreases substantially when a small price is introduced. 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 6.4.3, individuals whose expected attendance exceeds the "fewer" threshold by 2 or 4 visits 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 (Appendix Table A8).[31]

6.5 Summary of reduced-form results

Sections 5 and 6 provide a core set of reduced-form results. First, participants in our study perceive themselves to be time-inconsistent. Second, participants appear to be only partially aware of their time inconsistency, as they overestimate their future gym attendance. Third, take-up of commitment

[31] 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 2.4. If commitment contracts offer approximately no perceived 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.

contracts is not strongly related to perceived present focus and appears to be influenced by stochastic valuation errors or perceived social pressure. This suggests that commitment contracts are unlikely to be a well-targeted tool for addressing time inconsistency in this setting. In the next section we estimate the parameters of the quasi-hyperbolic discounting model, and we use the model to evaluate the welfare effects of commitment contracts.

7 Structural estimates and welfare implications

In this section, we first estimate the model of present focus introduced in Section 2.1, examining heterogeneity by commitment contract choice and several other covariates. We then use the estimated model to evaluate the welfare effects of commitment contracts, and compare them to the welfare effects of linear financial incentives.

7.1 Parameter estimates

7.1.1 Summary of methodology

We estimate the model of present focus introduced in Section 2.1 using our data on forecasted and actual attendance, and the WTP for the piece-rate incentives. We estimate the model both by pooling over the full population, as well as for various subsamples to incorporate heterogeneity. For simplicity, we assume that once people have financial incentives in place, their daily gym attendance decisions are not biased by stochastic valuation errors, although our welfare results do incorporate people's possible errors in contract *take-up* decisions. We discuss this assumption in detail in Section 7.3.

We assume that each day corresponds to a period, and we thus set $T = 28$ to correspond to the four-week study period. We assume the attendance costs in each period are distributed independently and identically according to the exponential distribution with rate parameter λ.[32] This assumption implies that the costs of attending the gym are always non-negative, which, together with the fact that all membership contracts did not involve per-attendance fees, implies that individuals never experience immediate pleasure from attending the gym. We discuss robustness to variations of this assumption in Section 7.3, but also show that this assumption is most consistent with our data.

The free parameters in our model are the perceived and actual present focus parameters $\tilde{\beta}$ and β, the (perceived) delayed health benefits b, and the rate parameter λ.[33] The parametric assumptions imply that actual and forecasted average attendance at incentive p are given by $28 \cdot \left[1 - e^{-\lambda\beta(b+p)}\right]$ and $28 \cdot \left[1 - e^{-\lambda\tilde{\beta}(b+p)}\right]$, respectively.

Because we have rich information about the perceived and actual attendance curves and the behavior change premium, and because these objects are functions of only four parameters $(\beta, \tilde{\beta}, b, \lambda)$,

[32] The CDF of an exponential distribution with rate parameter λ is $F(x) = 1 - e^{-\lambda x}$.

[33] Formally, people's behavior is determined by their perceptions of the per-attendance health benefits, not the actual health benefits. If the two are different, our methodology identifies the *perceived* health benefits.

identification of our parametric model follows straightforwardly from the logic introduced in Section 2.2. More detailed algebraic intuition is as follows. First, note that the intercepts at $p = 0$ of the forecasted and actual attendance curves are determined by $\lambda\beta b$ and $\lambda\tilde{\beta}b$, respectively, and the slopes of the curves are determined by $\lambda\beta$ and $\lambda\tilde{\beta}$, respectively. Thus, if $\lambda\beta$ is identified from the slope of the actual attendance curve and $\lambda\beta b$ is identified from the intercept, then the health benefits b are identified as well. More generally, as depicted in Figure 1, the health benefits b correspond to the projected level of attendance disincentives at which the forecasted and actual attendance curves intersect. Second, note that Proposition 1 shows that the behavior change premium is a linear function of $(1 - \tilde{\beta})b$. Thus, since we can identify b, we can also identify $\tilde{\beta}$. Third, note that the vertical wedge between the forecasted and actual attendance curves is $(\tilde{\beta} - \beta)(b + p)$. Thus, since we can identify b and $\tilde{\beta}$, the wedge between the two curves identifies β. In sum, we have four parameters, and we have five sets of moments identifying them: the average behavior change premium, the intercepts of the forecasted and actual attendance curves, and the slopes of the forecasted and actual attendance curves.

Formally, we estimate the parameters using the generalized method of moments (GMM), with the moment equations and the estimation procedure detailed in Appendix C.1. Since the forecasted attendance curve and the behavior change premium utilizes multiple observations per person, we cluster all standard errors at the subject level. In Appendix C.2 we show that, to a first order, our parameter estimates can be regarded as estimates of population averages, under the assumption that the health benefits b and the cost parameter λ are independent of each other, and independent of actual and perceived present focus parameters β and $\tilde{\beta}$. We provide empirical evidence for this in the results we summarize below.

7.1.2 Results

Table 8 presents our parameter estimates. Column 1 presents our estimate of the (average) present focus parameter β, column 2 presents our estimate of the (average) perceived present focus parameter $\tilde{\beta}$, and column 3 presents our estimate of the (average) perceived health benefits b. Columns 4-6 present three key functions of these parameters. Column 4 presents our estimate of the average internality $(1 - \beta)b$, which is the wedge between forecasted and desired attendance, in units of marginal utility. Column 5 presents our estimate of the perceived internality, $(1 - \tilde{\beta})b$. Column 6 presents a measure—introduced by Augenblick and Rabin (2019)—of the degree to which people are aware of their present focus: $(1 - \tilde{\beta})/(1 - \beta)$. A value of 1 corresponds to complete awareness, and a value of 0 corresponds to complete naivete.

Row 1 presents our estimates for all participants in the study. We estimate actual and perceived present focus parameters $\hat{\beta} = 0.55$ and $\hat{\tilde{\beta}} = 0.84$, respectively, and health benefits $\hat{b} = 9.66$. Our estimates of $(\beta, \tilde{\beta})$ are approximately in the middle of the range of estimates from studies estimating both parameters: (0.31, 0.73) in Mahajan et al. (2020), (0.37, 0.8) in Bai et al. (Forthcoming), (0.67, 0.85) in Chaloupka et al. (2019), (0.74, 0.77) in Allcott et al. (2020), and (0.85, 1) in Augenblick and Rabin (2019). As reviewed in Appendix C.8, our estimate $\hat{b}$ of health benefits is close to the

middle of the range of public health estimates.

Rows 2 and 3 present parameter estimates for participants in the information control group and participants who received the enhanced information treatment. Consistent with our interpretation that the information treatment affects awareness of present focus, the two rows show a significant difference in the estimated $\hat{\tilde{\beta}}$, but essentially identical estimates $\hat{\beta}$ and $\hat{b}$. The remarkable similarity of the $\hat{\beta}$ and $\hat{b}$ estimates across the two rows would be a highly unlikely coincidence if our model were misspecified—e.g., if overestimation of future attendance was due to underestimation of future cost shocks, but we incorrectly modeled that overestimation as coming from naivete about present focus. If this were the case, the information treatment would not change the behavior change premium. Thus, the reduced gap between forecasted and actual attendance would be interpreted as the information treatment increasing β and/or decreasing b, which would lead the estimates $\hat{\beta}, \hat{b}$ to be significantly impacted by the information treatment.

Rows 4 and 5 explore heterogeneity by gym attendance over the past four weeks. Past attendance is highly predictive of future attendance, suggesting that there are stable "attendance types": the regression coefficient from a regression of realized attendance on past attendance is 0.685 (robust s.e. 0.028). Here, there are significant differences in $\hat{\beta}$ and $\hat{b}$. Consistent with economic intuition, lower β and lower b are associated with lower past attendance. On the other hand, there is no theoretical reason to expect that past attendance should be related to awareness of present focus, and correspondingly, we find that $(1-\hat{\tilde{\beta}})/(1-\hat{\beta})$ is remarkably stable across the two past attendance groups.

In rows 6-8, we estimate the model for the subsamples of participants who indicated that they wanted the 8+, 12+, and 16+ contracts, respectively. Consistent with our reduced-form results, we find slightly lower estimates of β and $\tilde{\beta}$ for these individuals, but the differences are economically small. We find no evidence that commitment contracts are chosen by those with particularly high perceived or actual self-control problems, or those with particularly high internalities $(1-\beta)b$.

Row 9 explores the potential bias that might result from ignoring heterogeneity. We assume that there are eight types of individuals corresponding to eight subgroups: below- or above-median past attendance, crossed with receiving either the enhanced information treatment or no information treatment, crossed with willing to take-up the 12+ commitment contract.[34] We exclude individuals who received the ineffective information treatment in wave 1, although treating these individuals as being in the information control group leads to essentially identical results. We estimate the parameters separately for these eight groups, and then report the average, with each group weighted in proportion to its size. As rows 2-5 show, there is significant heterogeneity along these dimensions. However, the estimates in row 9 show that averaging over these eight subgroups produces essentially the same estimates as in row 1. Of course, there is likely additional heterogeneity not captured by the subsample splits in row 9, but the exercise illustrates the econometric result from Appendix C.2 that our estimates can be regarded as sample averages.

Figure 11 shows the in-sample fit of our model to the actual and forecasted attendance curves.

[34] We focus on the 12+ commitment contract since the other contracts were offered only in the first two waves.

Panel (a) uses the representative agent specification from row 1 of Table 8, while panel (b) allows for eight different types as in row 9 of Table 8. As can be seen, our estimates produce a tight in-sample fit under either specification. The fact that the in-sample fit is nearly identical in both panels is consistent with the Appendix C.2 result that our parameter estimates can be regarded as sample averages.

7.2 Implications for commitment contracts

Appendix C.3 presents the derivations for how present-focused individuals behave in the presence of commitment contracts, and how commitment contracts affect their period 0 surplus. To summarize, the threshold incentives of the commitment contracts generate payoffs that are non-separable over time, and we solve for individuals' equilibrium strategies by backwards induction—formalized as the Perception Perfect Equilibrium by O'Donoghue and Rabin (2001). Given an incentive scheme, a person's perceived and actual expected utility of starting out in period t with h_t prior attendances can be computed recursively. These value functions allow us to conduct welfare analysis, and to obtain analytic solutions for a person's strategy in each period t. Our welfare analyses take the long-run preferences of present-focused individuals as the normative criterion, which is a common but not uncontroversial assumption (Bernheim and Rangel, 2009; Bernheim, 2016; Bernheim and Taubinsky, 2018).

7.2.1 Out-of-sample validation tests

Before turning to welfare implications, we show that our model delivers accurate out-of-sample predictions about the effects of commitment contracts on gym attendance. Recall that in wave 3, we elicited preferences for commitment from all participants, but only a subset of participants were randomized to actually receive the 12+ contract. This allows us to empirically estimate how commitment contracts affect those who want them.

Row 1 of Table 9 reports our empirical estimates of how the 12+ commitment contract affects the behavior of those who want it. Column 1 reports the change in average attendance, column 2 reports the likelihood of attending 12 or more times with the contract, and column 3 reports the likelihood of attending 12 or more times without the contract. Column 4 reports the difference between columns 3 and 2: the impact of the commitment contract on the likelihood of attending 12 or more times.

Rows 2-5 report our model's predictions under different assumptions about heterogeneity, still restricting to those individuals who chose to take up the contract offer. These are out-of-sample predictions in the sense that information about commitment contracts was not used in the structural estimation of model parameters. Row 2 assumes homogeneity conditional on taking up the 12+ contract, which is analogous to the specification in row 7 of Table 8. Row 3 allows for more heterogenous parameters, allowing them to vary by the attendance and information subgroups considered in Row 9 of Table 8.[35] Rows 4-6 consider robustness to alternative heterogeneity assumptions—in

[35]Note that because Table 9 studies individuals who take up the 12+ contract, there are four rather than eight

particular, heterogeneity by median past attendance only, by quartile of past attendance only, or by quartile of past attendance crossed with receipt of the information treatment.

Table 9 shows that while all specifications accurately predict the impact on average attendance, more realistic heterogeneity assumptions are required to match the impact of the 12+ commitment contract on the likelihood of attending the gym 12 or more times. When individuals are assumed to be homogeneous, the model predicts that individuals who take up the contract almost always meet its 12-visit threshold—in contrast to the 35% failure rate we observe in the data. These homogeneous individuals are also unlikely to exceed the 12-visit threshold in the absence of any incentives, given their baseline attendance of 7.2 visits. Allowing for heterogeneity substantially changes the predictions, because individuals with high β and b are likely to attend the gym 12 or more times both with and without the commitment contract, while individuals with low β and b are unlikely to attend the gym 12 or more times both with and without the commitment contract. Thus, allowing heterogeneity in attendance decreases the predicted impact of the commitment contract, in line with our empirical estimates. As illustrated by the similar predictions of rows 4-6, the exact modeling of heterogeneity is largely inconsequential, as long the model allows for both "low"- and "high"-attendance types.

7.2.2 Welfare implications

Table 10 presents our welfare estimates for different types of incentive schemes. We conduct these calculations under the assumption of eight heterogeneous types, as in row 9 of Table 8. The welfare results are similar for other assumptions about heterogeneity, and are reported in Appendix C.5. The results for the 8+ and 16+ contracts, which were offered only in waves 1 and 2, are also very similar, and reported in Appendix C.4.

Note that our procedure does not require us to use our model to predict who will take up the contracts—take-up decisions are directly obtained from our data, and are directly incorporated into our construction of eight heterogeneous types. We only use our model to predict how the contracts affect the welfare of those who choose them.

Column 1 of Table 10 reports the predicted impact on average gym attendance. Column 2 reports the average impact on individuals' long-run utility. Column 3 reports the average impact on health benefits. Specifically, if Δ_k is the average impact on attendance of type k individuals who have delayed health benefits b_k, then the average impact on health benefits is $\sum_k \Delta_k b_k / \left(\sum_k \Delta_k\right)$. Column (4) reports the average increase in attendance costs that results from an increase in attendance. Any incentive scheme that increases the likelihood of attendance each day must mechanically increase the incurred attendance costs. Column 5 reports the difference between columns 3 and 4. The number reported in column 5 is the social surplus from an incentive scheme, and corresponds to a standard utilitarian welfare criterion, such as the one used in Gruber and Kőszegi (2001) or O'Donoghue and Rabin (2006).[36] All of the results reported in the table are averaged over all participants in the

heterogeneous types considered in row 3.

[36] Here, we make the implicit assumption that the marginal cost to the gym of an additional attendance is negligible.

study and not just, e.g., those who take-up the contracts.

The difference between individual surplus (column 2) and social surplus (column 5) is due to how the individuals' financial outcomes are treated. Financial losses or gains affect individual surplus, and thus are included in the statistics reported in column 2. However, the statistic in column 2 is an inappropriate measure of social surplus because, e.g., the penalty payments incurred by individuals who take up commitment contracts are not "burned." These penalty payments are a financial transfer from individuals taking up the contract to the provider of the contracts (e.g., a public organization). Analogously, subsidies mechanically increase individual surplus, but they are costly. Our social surplus measure incorporates these considerations, focusing on the efficiency of behavior change: the impact on health benefits minus the impact on attendance costs.

Our social surplus measure corresponds to several commonly used metrics. First, it corresponds to a utilitarian objective function, where provider and consumer incomes are weighted equally. Second, it corresponds to a consumer surplus metric when providers fund the subsidies through lump-sum taxes or fees and return commitment contract penalties through lump-sum rebates. For example, employers might provide gym attendance subsidies at the ultimate expense of less generous bonuses or other benefits, such that on net, the subsidies only change behavior and do not create a financial transfer between employees and employers.[37]

Row 1 presents the estimated surplus of offering a commitment contract for 12 or more gym attendances. Offering this commitment contract lowers individuals' private surplus, as shown in column 2. Indivuduals who take-up this contract incur a surplus loss of −$18.69 per person. Averaging over all participants (not just those who take-up the contract), this implies that offering this contract lowers overall consumer surplus by an average of −$9.23 per person.

Although individuals are made worse off by taking up the contract, the increased gym attendance generated by this contract—2.47 for those who take it up, 1.22 averaged over all participants—increases social efficiency. However, the 12+ contract is not the most efficient means of generating the average 1.22 visits increase. As reported in row 2, a gym attendance subsidy of $1.90 per attendance generates the same change in average attendance, but in a more socially efficient manner. This subsidy generates both a higher increase in health benefits and a smaller increase in attendance costs, leading to a net social surplus gain of $4.39 per person.[38] The fact that this subsidy generates

If the gym incurs non-negligible costs from additional attendances, the social efficiency criterion in column 5 would need to be modified to include those costs as well.

[37]For example, our social surplus metric is identical to the utilitarian criterion in O'Donoghue and Rabin (2006), where the provider is the government, which balances its budget with lump-sum transfers. That is, exercise subsidies would be funded by a lump-sum income tax, while the revenues from commitment contract penalties would be redistributed through a lump-sum tax refund. Thus, the assumption of lump-sum revenue recycling mechanically guarantees that pure monetary transfers between individuals and the provider have no effect on social welfare.

In principle, there may be cases where provider revenue is weighted more heavily than consumer incomes. Such cases push against subsidies and toward commitment contracts. However, such cases also push most strongly toward Pigovian *taxes*. E.g., "sin taxes" would compare particularly favorably to commitment contracts in, e.g., the case of reducing sugary drinks consumption. Thus, a high marginal value of provider funds does not mechanically favor using commitment contracts as a policy tool.

[38]Additionally, column (3) of Table 10 reveals that a linear attendance subsidy not only minimizes costs, but is also more targeted to people with the highest estimates of health benefits b_i. This is because the subsidy targets individuals with higher b_i. This is not a general property of subsidies, and is not true for the 16+ contract, as shown

higher surplus to individuals is mechanical and not economically interesting.

The results are similar for the 8+ and 16+ contracts, as reported in Appendix C.4. Both contracts lower individuals' private surplus, and both generate positive but small increases in social efficiency. In both cases, linear attendance subsidies that generate the same average increase in attendance are far more socially efficient.

Row 3 considers the per-attendance subsidy that maximizes social surplus, which approximately equals the average value of $(1-\beta_i)b_i/\beta_i$. We calculate this subsidy to be \$7.54 per attendance, and we find that the subsidy increases social surplus by \$9.36 per person. We do not compare to the "optimal" commitment contract because theory does not provide clear guidance about what this would be, particularly in light of our findings about stochastic valuation errors. By contrast, the optimal subsidy is straightforward to calculate and implement, and is estimated to yield large social gains. This illustrates the potential benefits of using structural estimates to inform the design of simple incentive schemes.

Linear incentives are estimated to be more socially efficient than commitment contracts for two basic reasons. First, although commitment contracts are not more likely to be taken up by those with the largest internalities $(1-\beta_i)b_i$, they nevertheless change behavior unevenly across people. Mechanically, only those who take up the contracts increase their attendance. However, the efficiency gains from behavior change are concave: it is more efficient to increase everyone's attendance by 1.5 visits than to increase half of the population's attendance by 3.0 visits, if that half of the population does not differ from the broader population. The intuition is simply that if c_i^* is the marginal cost draw at which a person is indifferent between attending the gym or not, then a marginal change in this person's motivation to attend the gym generates social benefits of $b_i - c_i^*$. Thus, the more motivated a person is to attend in the first place, the higher is c_i^*, and thus the lower are the social benefits of providing this person with additional motivation to exercise.

Second, commitment contracts change behavior unevenly across time. By definition, a linear attendance subsidy increases a person's motivation to attend the gym by the same degree each day. Commitment contracts, however, introduce time-varying incentives because financial rewards are discontinuous at the threshold.[39] The incentives to attend the gym are relatively small at the beginning, when there are many remaining opportunities for meeting the threshold. Moreover, present-focused individuals will "procrastinate" on fulfilling the threshold requirement. As shown in Figure A1 in Appendix C.6, our structural model predicts that on average, commitment contracts will have a limited effect on behavior at the beginning of the four-week period and a large effect on behavior at the end of the four-week period. Appendix Figure A2 shows that this prediction is borne out in the data: the 12+ commitment contract has a larger effect on people's behavior at the end of the four-week period. For reasons summarized above, this unequal distribution of treatment effects over time is less efficient than the constant effects of linear attendance subsidies.

Both of these principles apply to non-stationary cost distributions, including situations where

in Appendix Table A9.

[39] A similar argument would apply to financial rewards that are kinked at the threshold, as in, e.g., Kaur et al. (2015).

costs might decreases or increase over time. More generally, it is most efficient for *incentives* for behavior change to be distributed evenly.

7.3 Further robustness considerations

Alternative assumptions about the cost distribution We have assumed that the smallest value of a cost draw c is zero. That is, that the "good" days are those on which attending the gym is not immediately unpleasant. We consider robustness to this assumption in Appendix C.7, where we consider a model in which $c \sim -\$5 + X$ or $c \sim \$10 + X$, where X is exponentially distributed with rate λ. The first assumption corresponds to the gym being immediately pleasurable on some "good" days, while the second assumption corresponds to the minimal hassle cost being equivalent to \$10.

As Appendix C.7 shows, our conclusions about individual and social surplus are largely the same under these alternative assumptions—commitment contracts on net harm those who take them up, and linear incentives are a more efficient means of changing behavior. The parameter estimates naturally change—but in a manner that worsens both the in-sample and out-of-sample fit of the model. Higher mean costs lead to a higher estimate of perceived health benefits b; this, in turn, leads to lower estimates of $(1-\tilde{\beta})$ and $(1-\beta)$ because the wedges between the actual, forecasted, and desired attendance are functions of $(1-\beta)b$ and $(1-\tilde{\beta})b$. The in-sample fit to the actual and forecasted attendance curves does not suffer when we assume the higher cost-draw distribution, but it worsens significantly when assume the lower cost-draw distribution, as shown in Appendix Figure A4. The out-of-sample fit to the effects of the 12+ commitment contracts worsens dramatically for both assumptions. The higher distribution of cost draws leads the model to predict that commitment contracts have too high of an effect on the probability of attending the gym 12 or more times, while the lower distribution of cost draws leads the model to predict that commitment contracts have too small of an effect on both average attendance and the probability of attending the gym 12 or more times.

Imperfect perception of incentives on the "intensive" margin Although we have allowed for stochastic valuation errors in people's choice of incentives, we have assumed that stochastic valuation errors are not present in people's daily gym attendance decisions once the chosen incentives are instituted. This assumption is plausible for at least the linear piece-rate incentives, where a person's daily attendance decision involves comparing the costs c to the benefits $b+p$ for a single day, and does not involve any kind of complex aggregation over a longer horizon. This assumption is also consistent with our model's tight fit to various moments of the data. For example, the stability of our estimates of b and β in rows 2 and 3 of Table 8, or the out-of-sample validation in Table 9, would be less likely in a misspecified model.

At the same time, this assumption may be less realistic for the dynamic incentives generated by the threshold incentives of commitment contracts, since reacting to these incentives requires people to solve the dynamic programming problem detailed in Appendix C.3. If this complexity injects

noise in people's decisions about gym attendance, it would strengthen our qualitative results about commitment contracts' negative effects on consumer surplus, and the greater social efficiency of simple linear subsidies.

8 Concluding remarks and implications for future work

Better understanding how present-focused individuals make choices between various incentives, including commitment contracts, informs both positive and normative analysis. In addition to producing new estimates of present focus and new evidence about who takes up commitment contracts, the insights from this study can help inform policy design aimed at counteracting limited self-control. For example, while economists have long-studied "sin taxes" (e.g., O'Donoghue and Rabin, 2006; Allcott et al., 2019), there is little work on when the optimal policy mix should involve such taxes instead of commitment contracts, or when the two tools are complementary. Our theoretical results suggest that commitment contracts can be a well-targeted policy tool 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 perceived social pressure. Sin taxes are a more blunt policy tool because they affect everyone, not just those who are present-focused. Yet our results suggest that in settings with uncertainty and stochastic valuation errors, commitment contracts are not well-targeted. Our results illustrate how sin taxes can be more socially efficient in these settings, even when there is high take-up of commitment contracts that have significant effects on behavior.

Of course, our results come with many caveats and leave open many questions. First, our estimates are local to the participants of our fitness center. Even within the exercise domain, it will be valuable to apply our methodology to other populations. More broadly, it will be valuable to extend our methods to other domains of behavior, such as food choice, education, and saving and borrowing decisions. For example, Allcott et al. (2020) extend our method for estimating present focus parameters to consumer lending markets, though they do not examine offers of commitment contracts.

Second, our analyses focus on a particular set of commitment contracts, and it will be important for future work to apply our methodology to evaluate other types of commitment contracts. Although our results illustrate that high take-up and high treatment effects on behavior do not imply that commitment contracts are welfare-enhancing, our results do not preclude the possibility that commitment contracts different from ours may be more beneficial.

Third, 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.

Fourth, it is natural to expect that in the presence of noisy valuation and other frictions such as perceived social pressure, 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.

Fifth, our analyses assume the long-run criterion is the normative standard, which has been challenged by Bernheim and Rangel (2009) and others. Exploring welfare implications under alternative criteria—as in, e.g., Sadoff et al. (2019)—could be fruitful.

Sixth, we evaluate the welfare effects of only several types of incentive schemes. Our structural estimates can be used to explore the welfare effects of other types of incentive schemes.

These open questions and caveats illustrate the need for further testing, refinement, and critiquing of our approach. Our results illustrate the potential value of theoretically-grounded quantitative methods such as ours in helping improve incentive design for people with limited self-control.

References

Acland, Dan, and Vinci Chow. 2018. "Self-Control and Demand for Commitment in Online Game Playing: Evidence from a Field Experiment." *Journal of the Economic Science Association* 4 (1): 46–62, https://ideas.repec.org/a/spr/jesaex/v4y2018i1d10.1007_s40881-018-0048-3.html.

Acland, Dan, and Matthew R. Levy. 2012. "Naiveté, Projection Bias, and Habit Formation in Gym Attendance." Working Paper: GSPP13-002.

Acland, Dan, and Matthew R Levy. 2015. "Naiveté, Projection Bias, and Habit Formation in Gym Attendance." *Management Science* 61 (1): 146–160.

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

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

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

Allcott, Hunt, Joshua Kim, Dmitry Taubinsky, and Jonathan Zinman. 2020. "Are High-Interest Loans Predatory? Theory and Evidence from Payday Lending." Working Paper.

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

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

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

Ashraf, Nava, Dean Karlan, and Wesley Yin. 2006. "Tying Odysseus to the Mast: Evidence From a Commitment Savings Product in the Philippines." *The Quarterly Journal of Economics* 121 (2): 635–672. 10.1162/qjec.2006.121.2.635.

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. 2015. "Working Over Time: Dynamic Inconsistency in Real Effort Tasks." *The Quarterly Journal of Economics* 130 (3): 1067–1115, https://econpapers.repec.org/article/oupqjecon/v_3a130_3ay_3a2015_3ai_3a3_3ap_3a1067-1115..htm.

Augenblick, Ned, and Matthew Rabin. 2019. "An Experiment on Time Preference and Misprediction in Unpleasant Tasks." *The Review of Economic Studies* 86 (3): 941–975. 10.1093/restud/rdy019.

Avery, Mallory, Osea Giuntella, and Peiran Jiao. 2019. "Why Don't We Sleep Enough? A Field Experiment among College Students." IZA Discussion Paper, No. 12772.

Bai, Liang, Benjamin Handel, Ted Miguel, and Gautam Rao. Forthcoming. "Self-Control and Demand for Preventive Health: Evidence from Hypertension in India." *Review of Economics and Statistics*.

Bernheim, B. Douglas. 2016. "The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics." *Journal of Benefit-Cost Analysis* 7 (1): 12–68.

Bernheim, B. Douglas, and Antonio Rangel. 2009. "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics." *Quarterly Journal of Economics* 124 (1): 51–104.

Bernheim, B. Douglas, and Dmitry Taubinsky. 2018. "Behavioral Public Economics." In *The Handbook of Behavioral Economics*, edited by Bernheim, B. Douglas, Stefano DellaVigna, and David Laibson Volume 1. New York: Elsevier.

Beshears, John, James J Choi, Christopher Harris, David Laibson, Brigitte C Madrian, and Jung Sakong. 2020. "Which Early Withdrawal Penalty Attracts the Most Deposits to a Commitment Savings Account?" *Journal of Public Economics* 183 Article 104144.

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

Bisin, Alberto, and Kyle Hyndman. 2020. "Present-Bias, Procrastination and Deadlines in a Field Experiment." *Games and Economic Behavior* 119 339–357.

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

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

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

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

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

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

Chaloupka, Frank J., Matthew R. Levy, and Justin S. White. 2019. "Estimating Biases in Smoking Cessation: Evidence from a Field Experiment." NBER Working Paper 26522.

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

DellaVigna, Stefano, John A List, and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics* 127 (1): 1–56.

DellaVigna, Stefano, and Ulrike Malmendier. 2004. "Contract Design and Self-Control: Theory and Evidence." *The Quarterly Journal of Economics* 119 (2): 353–402. 10.1162/0033553041382111.

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

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

Ericson, Keith M., and David Laibson. 2019. "Intertemporal Choice." In *Handbook of Behavioral Economics*, edited by Bernheim, B. Douglas, Stefano DellaVigna, and David Laibson Volume 2. Elsevier.

Exley, Christine L., and Jeffrey K. Naecker. 2017. "Observability Increases the Demand for Commitment Devices." *Management Science* 63 (10): 3262–3267. 10.1287/mnsc.2016.2501.

Fang, Hanming, and Dan Silverman. 2004. "Time Inconsistency and Welfare Program Participation: Evidence from the NLSY." July, Cowles Foundation Discussion Paper No. 1465.

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

Gine, Xavier, Dean Karlan, and Jonathan Zinman. 2010. "Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation." *American Economic Journal: Applied Economics* 2 (4): 213–235. 10.1257/app.2.4.213.

Gruber, Jonathan, and Botond Kőszegi. 2001. "Is Addiction Rational? Theory and Evidence?" *Quarterly Journal of Economics* 116 (4): 1261–1305.

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

Hall, Alistair R. 2005. *Generalized Method of Moments*. Oxford University Press.

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

Harberger, Arnold. 1964. "Taxation, Resource Allocation, and Welfare." In *The role of direct and indirect taxes in the Federal Reserve System*, 25–80, Princeton University Press.

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

Heidhues, Paul, and Botond Kőszegi. 2009. "Futile Attempts at Self-Control." *Journal of the European Economic Association* 7 (2): 423–434, `https://academic.oup.com/jeea/article-lookup/doi/10.1162/JEEA.2009.7.2-3.423`.

Houser, Daniel, Daniel Schunk, Joachim Winter, and Erte Xiao. 2018. "Temptation and Commitment in the Laboratory." *Games and Economic Behavior* 107 329–344.

Imai, Taisuke, Tom Rutter, and Colin Camerer. 2020. "Meta-Analysis of Present-Bias Estimation Using Convex Time Budget." April, Working Paper.

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

Karlan, Dean, and Leigh L Linden. 2017. "Loose Knots: Strong Versus Weak Commitments to Save for Education in Uganda." NBER Working Paper 19863.

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan. 2015. "Self-Control at Work." *Journal of Political Economy* 123 (6): 1227–1277. 10.1086/683822.

Khaw, Mel Win, Ziang Li, and Michael Woodford. 2017. "Risk Aversion as a Perceptual Bias." 10.3386/w23294, NBER Working Paper 23294.

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

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

Laibson, David, Peter Maxted, Andrea Repetto, and Jeremy Tobacman. 2018. "Estimating Discount Functions with Consumption Choices over the Lifecycle." Working Paper.

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

Mahajan, Aprajit, Christian Michel, and Alessandro Tarozzi. 2020. "Identification of Time-Inconsistent Models: The Case of Insecticide Treated Nets." NBER Working Paper 27198.

Martinez, Seung-Keun, Stephan Meier, and Charles Sprenger. 2020. "Procrastination in the Field: Evidence from Tax Filing." Working Paper.

McKelvey, Richard D., and Thomas R. Palfrey. 1995. "Quantal Response Equilibria for Normal Form Games." *Games and Economic Behavior* 10 (1): 6–38. 10.1006/game.1995.1023.

Milgrom, Paul, and Ilya Segal. 2002. "Envelope Theorems for Arbitrary Choice sets." *Econometrica* 70 (2): 583–601.

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

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

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

O'Donoghue, Ted, and Matthew Rabin. 1999. "Doing It Now or Later." *American Economic Review* 89 (1): 103–124. 10.1257/aer.89.1.103.

O'Donoghue, Ted, and Matthew Rabin. 2001. "Choice and Procrastination." *Quarterly Journal of Economics* 116 (1): 121–160.

O'Donoghue, Ted, and Matthew Rabin. 2006. "Optimal Sin Taxes." *Journal of Public Economics* 90 (10): 1825–1849, `https://econpapers.repec.org/article/eeepubeco/v_3a90_3ay_3a2006_3ai_3a10-11_3ap_3a1825-1849.htm`.

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

Royer, Heather, Mark Stehr, and Justin Sydnor. 2015. "Incentives, Commitments, and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company." *American Economic Journal: Applied Economics* 7 (3): 51–84. 10.1257/app.20130327.

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

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

Schilbach, Frank. 2019. "Alcohol and Self-Control: A Field Experiment in India." *American Economic Review* 109 (4): 1290–1322. 10.1257/aer.20170458.

Schwartz, Janet, Daniel Mochon, Lauren Wyper, Josiase Maroba, Deepak Patel, and Dan Ariely. 2014. "Healthier by Precommitment." *Psychological Science* 25 (2): 538–546. 10.1177/0956797613510950.

Shui, Haiyan, and Lawrence M. Ausubel. 2005. "Time Inconsistency in the Credit Card Market." Working Paper.

Skiba, Paige Marta, and Jeremy Tobacman. 2018. "Payday Loans, Uncertainty, and Discounting: Explaining Patterns of Borrowing, Repayment, and Default." Working Paper.

Strotz, R. H. 1955. "Myopia and Inconsistency in Dynamic Utility Maximization." *The Review of Economic Studies* 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. 2014. "Relationship of Meeting Physical Activity Guidelines with Quality Adjusted Life Years." *Seminars in Arthritis and Rheumatism* 44 (3): 264–270.

Toussaert, Séverine. 2018. "Eliciting Temptation and Self-Control Through Menu Choices: A Lab Experiment." *Econometrica* 86 (3): 859–889. 10.3982/ECTA14172.

Toussaert, Séverine. 2019. "Revealing Temptation Through Menu Choice: Field Evidence." Unpublished.

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

Woodford, Michael. 2012. "Inattentive Valuation and Reference-Dependent Choice." Unpublished.

Woodford, Michael. 2019. "Modeling Imprecision in Perception, Valuation and Choice." *Annual Review of Economics* 12 579–601.

Zhang, Qing ⓡ Ben Greiner. 2020. "Time Inconsistency, Sophistication, and Commitment–An Experimental Study." Vienna University of Economics and Business, Department of Strategy and Innovation Working Paper No. 12.

Table 1: Summary of commitment contract 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. (Forthcoming)	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]
Beshears et al. (2020)	39-46%[2]	house money
Toussaert (2019)	21-65%	house money
Schilbach (2019)	31-55%	house money
Exley and Naecker (2017)	41-65%	house money
Avery, Giuntella, and Jiao (2019)	63%	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	47%	
Other stakes	42%	
Overall	37%	
B. Removing options:		Restricted access to
Brune et al. (2016)	6%	own money
Afzal et al. (2019)	4-9%	own money
Zhang ⓡ Greiner (2020)	16-31%	other
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 (2020)	31-62%	other
Houser et al. (2018)	48%	other
Brune, Chyn, and Kerwin (2018)	50%	own money
Beshears et al. (2020)	56%[5]	house money
Augenblick, Niederle, and Sprenger (2015)	59%	other
Milkman, Minson, and Volpp (2014)	61%[4]	other
Dupas and Robinson (2013)	65%	own money
Alan and Ertac (2015)	69%	house chocolates
Chow (2011)	79%	other
Casaburi and Macchiavello (2019)	93%	own money
Average take-up rates (Option removal contracts)		
Own money at stake	42%	
House money/object at stake	63%	
Other stakes	43%	
Overall	45%	

[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 six 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 (1)	Wave 1 (2)	P-value (3)	Waves 2-3 (4)	P-value (5)
Female	0.613	–0.043	0.41	–0.042	0.20
Age[a]	33.51	–0.47	0.73	–0.83	0.42
Student, full-time	0.569	–0.089	0.09	0.004	0.91
Working, full- or part-time	0.571	0.141	0.01	–0.004	0.91
Married	0.272	0.082	0.08	–0.004	0.89
Advanced degree[b]	0.457	0.045	0.40	–0.002	0.94
Household income[a]	55,139	1,637	0.74	–4,399	0.21
Visits in the past 4 weeks, recorded	6.91	0.21	0.74	–0.10	0.79
N	1,248	166 Control 174 Treated		456 Control 452 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 elicited in our online survey, as well as differences in treatment and control group means. In wave 1 of the experiment, the treatment group received the basic information treatment (see Figure 3a). In waves 2 and 3, treated participants received the enhanced information treatment (see Figures 3b-c). The table also summarizes data on past visit frequencies to the gym. Recorded visits are obtained from the fitness center's log-in records.

Table 3: Association between the behavior change premium and proxies for sophistication

	Behavior change premium			
	(1)	(2)	(3)	(4)
Goal − exp. attend. (z-score)	0.68**			0.73**
	(0.28)			(0.29)
Actual − exp. attend. (z-score)		0.50**		0.49**
		(0.21)		(0.22)
Basic info. treatment			0.30	0.43
			(0.56)	(0.56)
Enhanced info. treatment			1.36**	1.30**
			(0.57)	(0.59)
Dep. var. mean:	1.17	1.17	1.17	1.17
	(0.22)	(0.22)	(0.22)	(0.22)
Dep. var. mean, info. control group:	0.66	0.66	0.66	0.66
	(0.24)	(0.24)	(0.24)	(0.24)
Wave FEs	Yes	Yes	Yes	Yes
N	1,126	1,126	1,126	1,126

Notes: This table reports the association between the estimated behavior change premium (calculated excluding the $1 incentive) and proxies for sophistication. *Goal − exp. attend.* is the standardized (z-score) difference between participants' goal attendance and their subjective expectations of attendance in the absence of incentives (unstandardized mean: 3.34, SD: 3.64). *Actual − exp. attend.* is the standardized (z-score) difference between participants' actual attendance and their subjective expectations of attendance for the incentive assigned to them (unstandardized mean: −4.17, SD: 6.61). *Basic info. treatment* and *Enhanced info. treatment* are dummies for whether participants received the basic information treatment (see Figure 3a) or the enhanced information treatment (see Figures 3b-c), respectively. Each column presents coefficient estimates from OLS regressions with heteroskedasticity-robust standard errors in parentheses. Dependent variable means, with standard errors in parentheses, are reported for the full sample and information control group. The sample excludes participants in wave 3 assigned a commitment contract (122 participants) rather than a piece-rate incentive, since the *Actual − exp. attend.* proxy cannot be computed for those participants. ** denotes a statistic that is statistically significantly different from 0 at the 5% level.

Table 4: Association between take-up of "more" contracts and the behavior change premium

	Take-up of "more" visits contract			
	8+ visits	12+ visits	16+ visits	Pooled
	(1)	(2)	(3)	(4)
Behavior change premium	0.019*	0.020*	0.026*	0.022**
(z-score)	(0.011)	(0.012)	(0.013)	(0.010)
Dep. var. mean:	0.64	0.49	0.32	0.49
	(0.02)	(0.01)	(0.02)	(0.01)
Wave FEs	Yes	Yes	Yes	Yes
Contract FEs	No	No	No	Yes
N	849	1,248	849	2,946
Clusters	849	1,248	849	1,248

Notes: This table reports the association between the take-up of "more" commitment contracts and the estimated average behavior change premium (calculated excluding the $1 incentive and expressed as a z-score; unstandardized mean: 1.20, SD: 7.08). Each column reports coefficient estimates from OLS regressions. Dependent variable means with standard errors in parentheses are also reported. In columns 1, 2, and 3, the dependent variables are the take-up of the "more" visit contract with a threshold of 8, 12, and 16 visits, respectively. In column 4, the dependent variable is the take-up of a "more" visit contract, with observations pooled across the three contracts, controlling for commitment contract threshold fixed effects (i.e., 8, 12, 16 visit thresholds). Standard errors are heteroskedasticity-robust in columns 1-3, and are clustered at the subject level in column (4). *,** denote statistics that are statistically significantly different from 0 at the 10% and 5% levels respectively.

Table 5: Association between take-up of "more" commitment contracts and proxies for sophistication

	Take-up of "more" visits contracts			
	(1)	(2)	(3)	(4)
Goal − exp. attend. (z-score)	0.039***			0.036***
	(0.013)			(0.013)
Actual − exp. attend. (z-score)		–0.045***		–0.041***
		(0.013)		(0.014)
Basic info. treatment			–0.022	–0.012
			(0.041)	(0.041)
Enhanced info. treatment			–0.080**	–0.071**
			(0.031)	(0.031)
Dep. var. mean:	0.49	0.49	0.49	0.49
	(0.01)	(0.01)	(0.01)	(0.01)
Dep. var. mean, info. control group:	0.52	0.52	0.52	0.52
	(0.01)	(0.01)	(0.01)	(0.01)
Wave FEs	Yes	Yes	Yes	Yes
Contract FEs	Yes	Yes	Yes	Yes
N	2,824	2,824	2,824	2,824
Clusters	1,126	1,126	1,126	1,126

Notes: This table reports the association between take-up of a "more" visits commitment contract and proxies for sophistication. We pool the data by participant and include commitment contract threshold fixed effects (i.e., 8-, 12-, 16-visit thresholds). The independent variables in this table are defined exactly as in Table 3. Each column presents coefficient estimates from OLS regressions with standard errors, clustered by subject, in parentheses. Dependent variable means, with standard errors in parentheses, are reported for the full sample and information control group. The sample excludes participants in wave 3 assigned a commitment contract (122 participants) rather than a piece-rate incentive, since the *Actual − exp. attend.* proxy cannot be computed for those participants. **,*** denote statistics that are statistically significantly different from 0 at the 5% and 1% level respectively.

Table 6: 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	0.34	0.89	0.47	0.25***	0.13***
12 visits	0.49	0.31	0.67	0.43	0.18***	0.12***
16 visits	0.32	0.27	0.50	0.43	0.18***	0.15***

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 reports 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, for each threshold. Columns 5 and 6 display the difference in the take-up rates of column 3 versus column 1 and the difference in the take-up rates of column 4 versus column 2, respectively. Over three study waves, all participants faced the choice of a commitment contract at the 12-visit threshold (N=1,248) while the 8-visit and 16-visit commitment contracts were only presented in the first two waves (N=849). *** denotes differences that are statistically significantly different from 0 at the 1% level.

Table 7: Correlation between perceived success in contracts and expected attendance

	Subjective expected attendance without incentives			
	(1)	(2)	(3)	(4)
Subj. prob. succeed in	8.46***		9.17***	9.68**
"more" contract	(1.31)		(1.17)	(3.79)
Subj. prob. succeed in		–3.96***	–4.64***	–9.97***
"fewer" contract		(0.91)	(0.85)	(3.10)
N	399	399	399	76
"More" − "Fewer"			13.81***	19.64***
			(1.37)	(6.02)

Notes: This table reports the association between subjective beliefs about commitment contract success and expected attendance with no incentives. Each column presents coefficient estimates from OLS regressions with heteroskedasticity-robust standard errors in parentheses. *Subj. prob. succeed in "more" contract* is participants' subjective expectations of attending the gym 12 or more days during the 4-week incentive period, coded as a probability between 0 and 1. *Subj. prob. succeed in "fewer" contract* is participants' subjective expectations of attending the gym fewer than 12 times during the 4-week incentive period, coded as a probability between 0 and 1. The dependent variable is participants' subjective expectations of attendance in the absence of any incentives. The *"More" − "Fewer"* row shows the estimated 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. The sample in columns 1-3 consists of all participants in wave 3, the only wave in which we elicited the probabilities of contract success. The sample in column 4 is restricted to participants in wave 3 indicated that they wanted both the "more" and "fewer" contract with a threshold of 12 visits. **,*** denote statistics that are statistically significantly different from 0 at the 5% and 1% level respectively.

Table 8: Parameter estimates

		(1) $\hat{\beta}$	(2) $\hat{\tilde{\beta}}$	(3) $\hat{b}$	(4) $(1-\hat{\beta})\cdot\hat{b}$	(5) $(1-\hat{\tilde{\beta}})\cdot\hat{b}$	(6) $\frac{(1-\hat{\tilde{\beta}})}{(1-\hat{\beta})}$
1	All	0.55	0.84	9.66	4.39	1.58	0.36
	(N=1,126)	(0.51, 0.58)	(0.80, 0.88)	(9.05, 10.28)	(4.02, 4.77)	(1.20, 1.96)	(0.29, 0.43)
2	Information control	0.54	0.86	10.03	4.63	1.38	0.30
	(N=560)	(0.50, 0.58)	(0.82, 0.90)	(9.13, 10.93)	(4.15, 5.11)	(1.00, 1.75)	(0.22, 0.37)
3	Enhanced information	0.54	0.78	9.83	4.49	2.19	0.49
	treatment (N=392)	(0.47, 0.62)	(0.69, 0.87)	(8.77, 10.89)	(3.73, 5.26)	(1.32, 3.06)	(0.35, 0.63)
4	Below median past	0.38	0.78	7.07	4.39	1.57	0.36
	attendance (N=550)	(0.33, 0.43)	(0.70, 0.86)	(6.45, 7.68)	(3.92, 4.86)	(1.04, 2.09)	(0.25, 0.46)
5	Above median past	0.68	0.88	12.57	4.08	1.46	0.36
	attendance (N=576)	(0.63, 0.72)	(0.84, 0.92)	(11.45, 13.69)	(3.54, 4.63)	(0.98, 1.94)	(0.26, 0.45)
6	Chose 8+ visit	0.54	0.84	9.16	4.23	1.50	0.36
	contract (N=546)	(0.49, 0.59)	(0.77, 0.90)	(8.34, 9.98)	(3.70, 4.76)	(0.90, 2.11)	(0.24, 0.47)
7	Chose 12+ visit	0.50	0.81	9.62	4.84	1.79	0.37
	contract (N=556)	(0.45, 0.54)	(0.75, 0.88)	(8.78, 10.47)	(4.31, 5.38)	(1.17, 2.40)	(0.26, 0.47)
8	Chose 16+ visit	0.47	0.75	10.30	5.46	2.63	0.48
	contract (N=275)	(0.39, 0.55)	(0.63, 0.86)	(8.94, 11.67)	(4.57, 6.34)	(1.50, 3.75)	(0.33, 0.64)
9	Averaging heterogeneity	0.55	0.85	10.24	4.21	1.39	0.35
	(N=952)	(0.52, 0.58)	(0.81, 0.89)	(9.50, 10.98)	(3.83, 4.59)	(1.05, 1.74)	(0.27, 0.42)

Notes: This table reports parameter estimates and respective 95% confidence intervals for various subsamples. The subsamples are determined by the participants' days of attendance over the 4 weeks prior, selection into the enhanced information treatment group, and their take-up of the various commitment contracts for more visits. Section 7.1 describes how the parameter estimation was performed. The present focus parameter is denoted by β, the perceived present focus parameter is denoted by $\tilde{\beta}$, and people's (perceived) health benefits of a gym attendance are denoted by b. Row 9 averages estimates across eight subsamples corresponding to (i) assignment to either the enhanced information treatment or the information control group, crossed with (ii) whether days of attendance over the 4 weeks prior to the experiment is below or above the median, crossed with (iii) take-up of the more-visit contract with a threshold of 12 visits. Over the three study waves, only participants in waves 2 and 3 (N=908) were eligible for random assignment to the enhanced information treatment group, and thus row 9 excludes participants assigned to the "basic" information treatment in wave 1. Inference for the statistics in columns 4-6, and for the averages reported in row 9, is conducted using the Delta method. All participants faced a take-up decision about a commitment contract with a 12-visit threshold (N=1,248), while the 8-visit and 16-visit commitment contracts were only presented in the first two waves (N=849). The samples exclude participants in wave 3 assigned a commitment contract (122 participants), rather than a piece-rate incentive, as our structural estimates only make use of data about how participants behave under piece-rate incentives.

Table 9: Estimated impact of 12+ contract on attendance

		(1) Δ in att.	(2) Pr(att. ≥ 12) with contract	(3) Pr(att. ≥ 12) without contract	(4) Δ in Pr(att. ≥ 12)
1	Empirical	3.51 (1.38, 5.65)	0.65 (0.52, 0.78)	0.22 (0.10, 0.35)	0.42 (0.26, 0.58)
2	Homogeneous	3.05	0.91	0.15	0.76
3	Heterogeneous by median past att., info. treatment	2.47	0.74	0.34	0.40
4	Heterogeneous by median past att.	2.61	0.74	0.33	0.41
5	Heterogeneous by quartile past att.	2.74	0.73	0.31	0.41
6	Heterogeneous by quartile past att., info. treatment	2.65	0.73	0.32	0.41

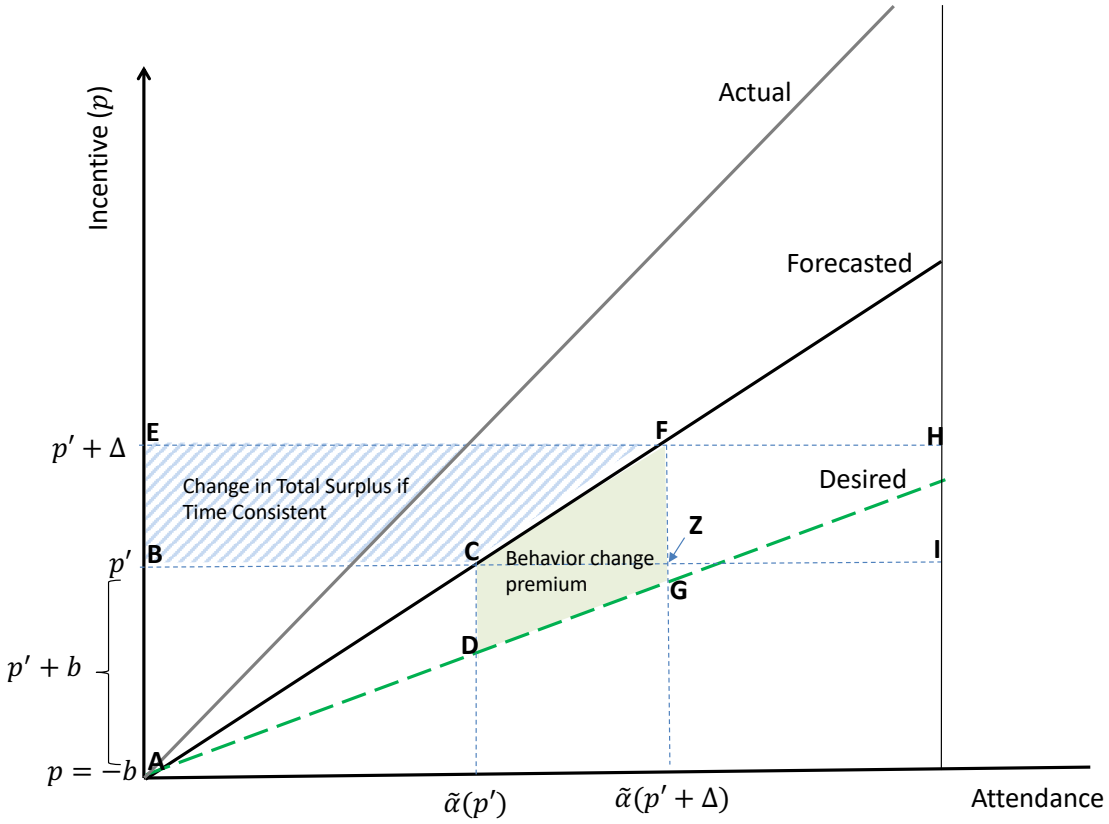
Notes: This table assesses our estimated models' predictions about how the "12 visits or more" contract affects the behavior of participants who indicated that they would take it up. All calculations are for the four-week period in our experiment. Row 1 reports empirical estimates from OLS regressions with wave fixed effects, with 95% confidence intervals in parentheses. In row 2, we assume that participants are homogeneous conditional on taking up the 12+ contract. Thus, row 2 assumes that there are only two types of individuals: those who take up the 12+ contract and those who don't. In row 3, we estimate a heterogeneous model, as in row 9 of Table 8. In rows 4-6, we consider alternative heterogeneity assumptions. Row 4 divides individuals only according to their median past attendance. Row 5 divides individuals by quartile of past attendance. Row 6 divides individuals by quartile of past attendance crossed with receiving the enhanced information treatment.

Table 10: Estimated welfare effects of piece-rates and commitment contracts

		(1) Avg. Δ in attendance	(2) Δ Agent surplus	(3) Δ Health benefits	(4) Δ Attendance costs	(5) Δ Social Surplus
1	12+ visits contract	1.22	−$9.23	$10.88	$9.68	$1.21
2	Linear incentive, $p = \$1.90$	1.22	$22.95	$12.45	$8.06	$4.39
3	Optimal linear incentive, $p = \$7.54$	4.38	$106.71	$44.46	$35.10	$9.36

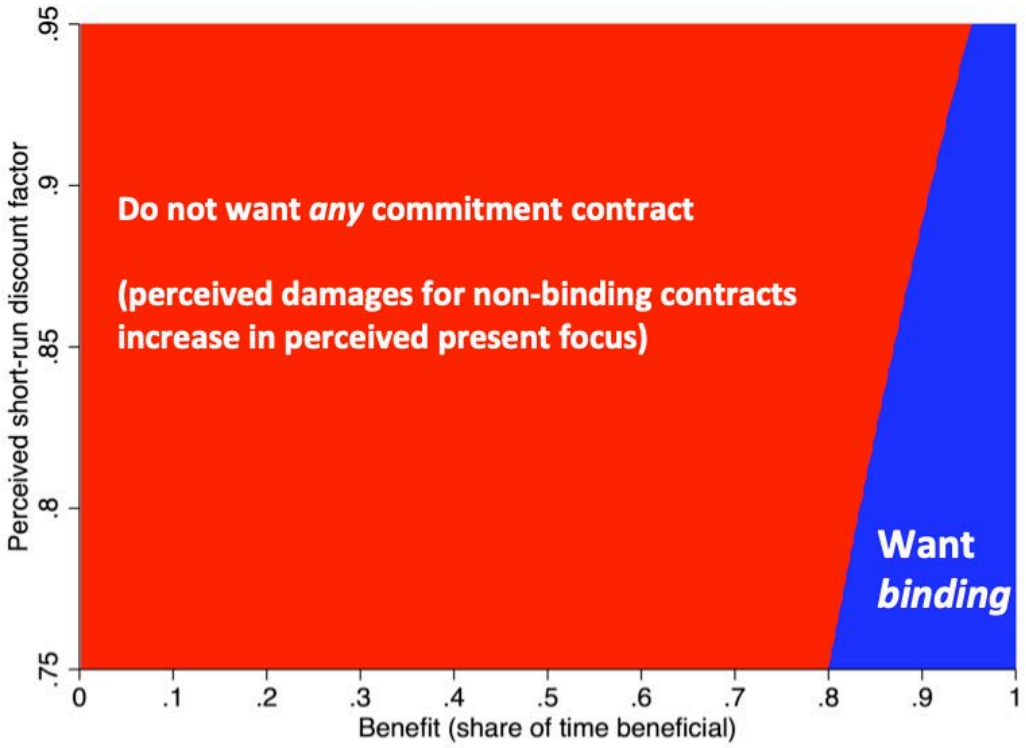
Notes: This table reports the estimated effects of three different incentive schemes, averaged over the full population, using the heterogeneity assumptions from row 9 of Table 8. Row 1 reports the estimated effect of offering individuals the 12+ commitment contract. All calculations are for a four-week period, as in our experiment. The numbers reported in row 1 are averages over those who take-up the contract (and thus affected by it) and those who do not. Row 2 reports the estimated effects of a linear per-attendance subsidy of $p = \$1.90$, which has the same impact on average population attendance as does the 12+ contract. Row 3 reports the effects of the optimal per-attendance subsidy. The formula for this subsidy is derived in Appendix C.3.3.

Figure 1: Illustration of the behavior change premium for a present-focused agent



Notes: This figure gives a representation of actual, forecasted, and desired attendance curves as a function of incentives. See Section 2.2 for a detailed description of this figure.

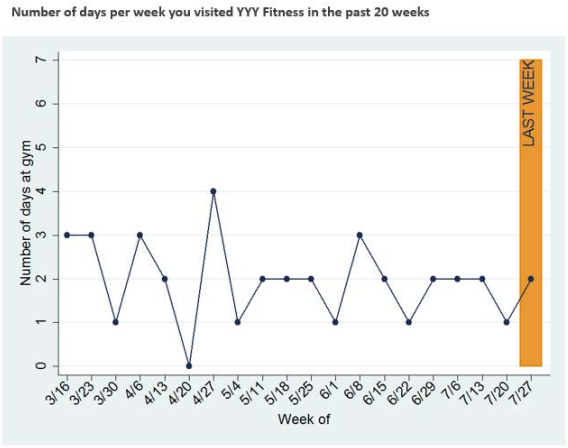
Figure 2: Commitment contract demand for uniform distribution of costs



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.

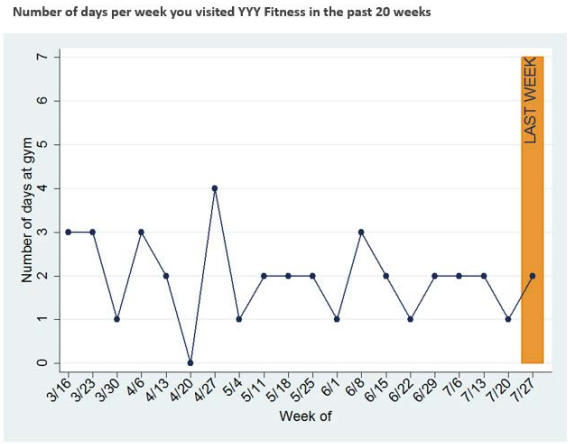
Figure 3: Information treatment

(a) Basic information treatment



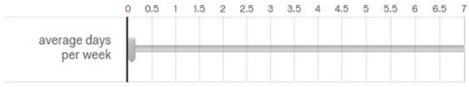
Notes: This graph is calculated based on the check-in records from the front desk. If you joined YYY Fitness within the last 20 weeks, the graph will show zero visits for weeks prior to when you joined.

(b) Enhanced info treatment - first screen



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

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



(c) Enhanced info treatment - second screen

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

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

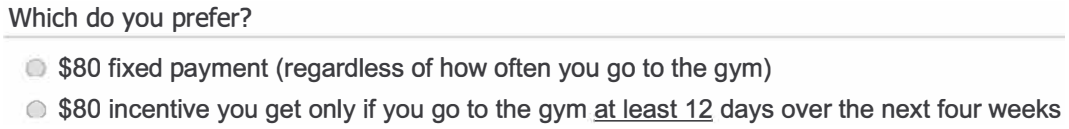
How useful do you think this information about previous participants will be as you think about how often you will attend?

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

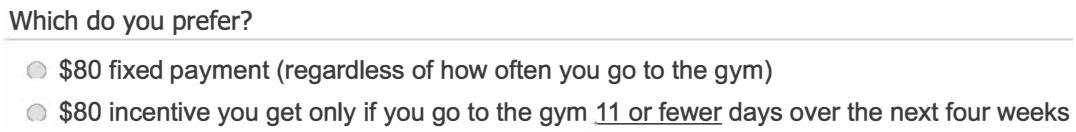
Notes: Panel (a) shows the basic information treatment of the history of past attendance shown to participants. Panels (b) and (c) show the enhanced information treatment. Panel (b) displays the first screen of the enhanced information treatment, which was similar to the basic information treatment but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

Figure 4: Screenshots of "more visits" and "fewer visits" commitment choices

(a) "More visits" commitment contract

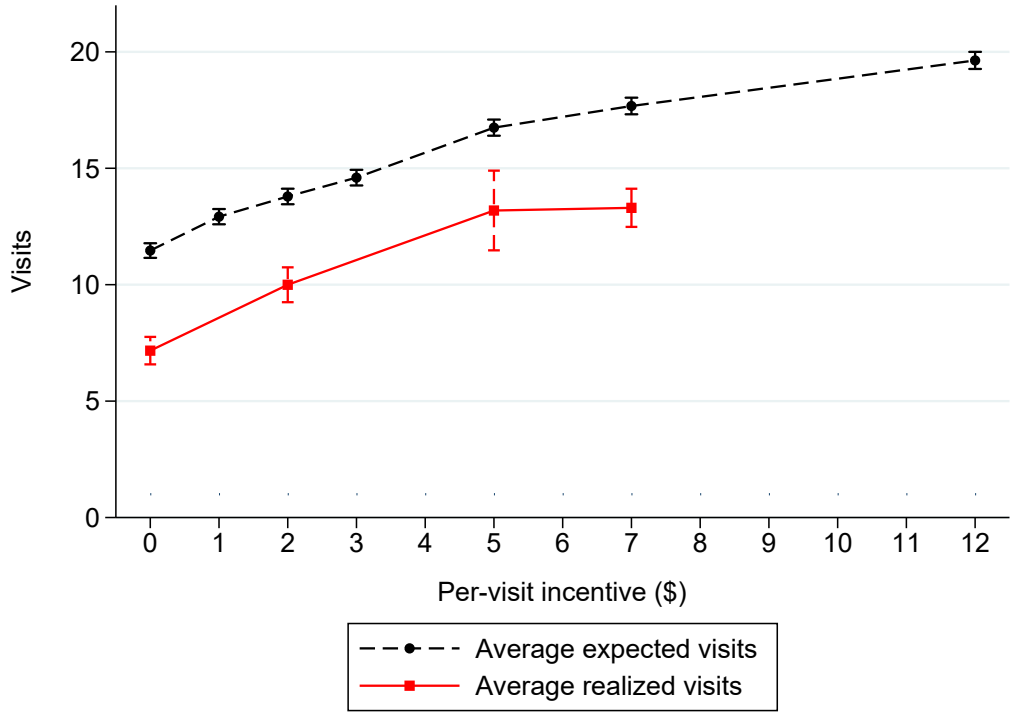


(b) "Fewer visits" commitment contract



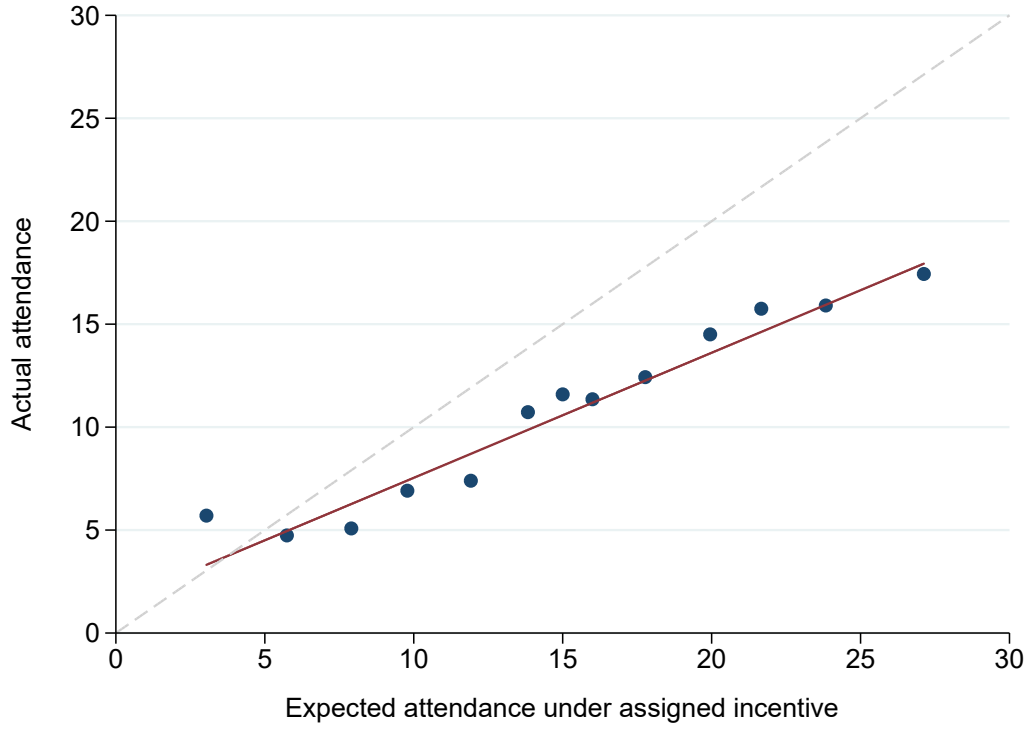
Notes: This figure provides screenshots of the commitment contracts offered to participants. Panel (a) provides an example of a commitment contract to attend the gym more (i.e., the "more visits" contract). Panel (b) provides an example of a commitment contract to attend the gym less (i.e., the "fewer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

Figure 5: Actual attendance and subjects' expected attendance by incentive



Notes: This figure reports the means and 95% confidence intervals for participants' subjective expectations of gym attendance ("Best guess of days I would attend over the next four weeks") and realized attendance, for different levels of piece-rate incentives. Subjective expectations are averaged over all participants in the analysis sample, while average realized visits are based on the subsets of participants who were randomized to receive each incentive. Section 3 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=413 ($0); N=293 ($2); N=75 ($5); N=342 ($7)).

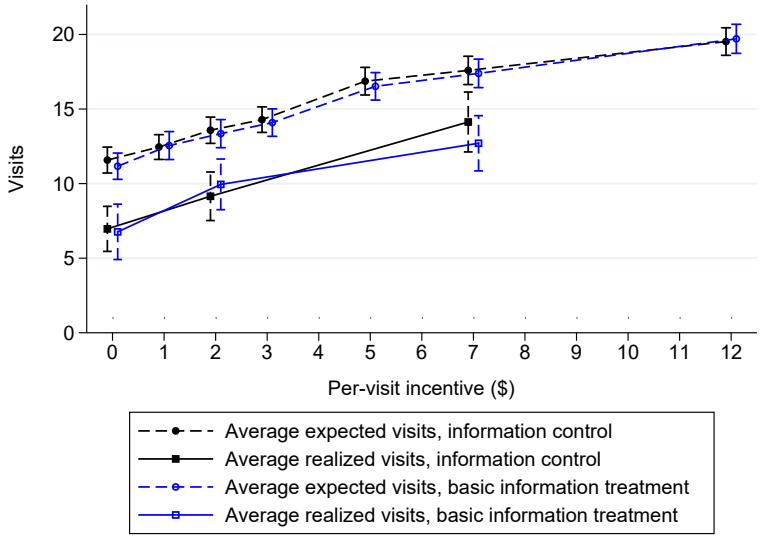
Figure 6: Actual attendance versus participants' subjective expectations of attendance



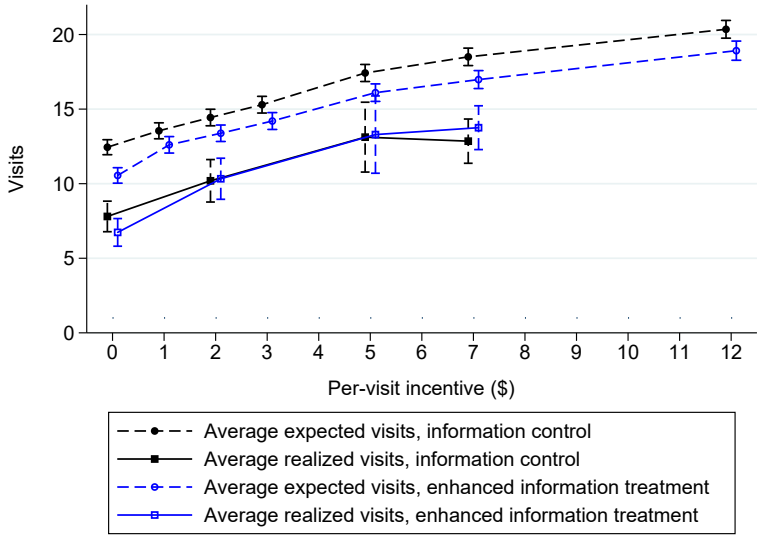
Notes: This figure shows a binned scatterplot comparing participants' actual attendance to their subjective expectations of gym attendance under the incentives they received. A dashed 45-degree line is included for reference. The sample excludes participants in wave 3 assigned a commitment contract (122 participants) rather than a piece-rate incentive.

Figure 7: Effect of information treatments on actual attendance and subjective expectations of attendance

(a) Impact of basic information treatment

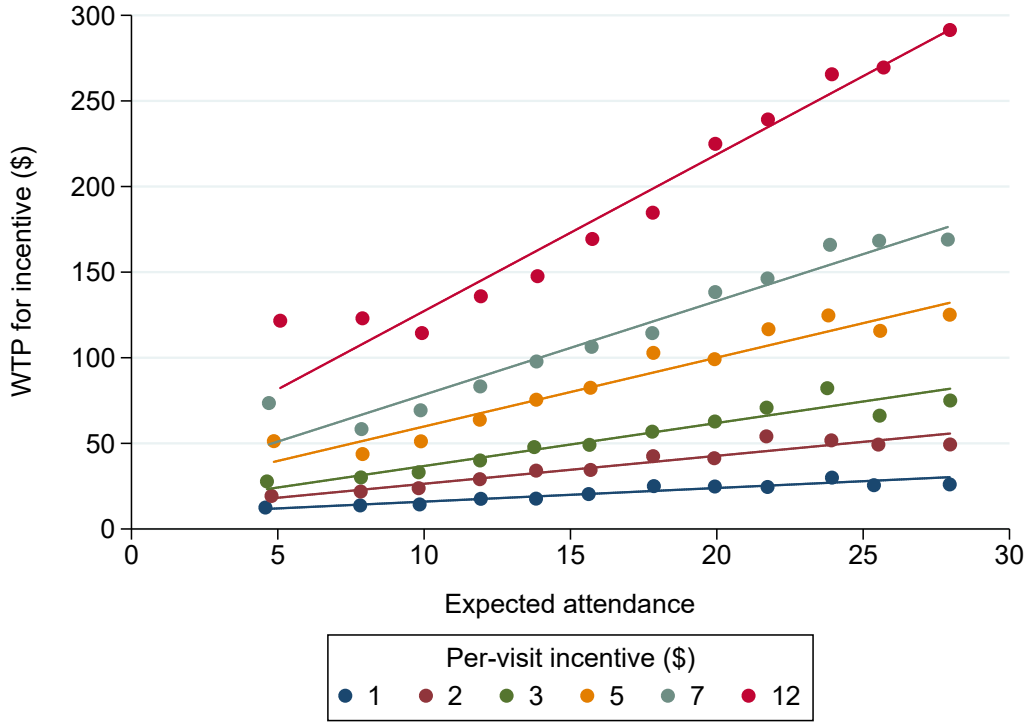


(b) Impact of enhanced information treatment



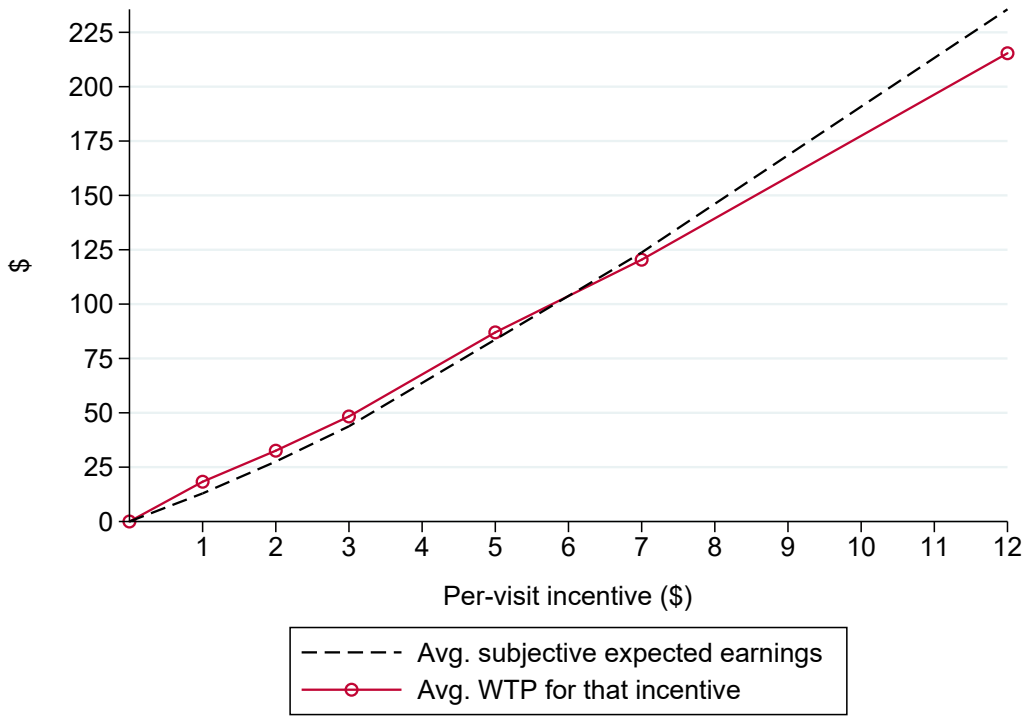
Notes: This figure presents the effects of the basic and enhanced information treatments on participants' subjective expectations of attendance, as well as their actual attendance. Panel (a) presents results from wave 1, where the basic information treatment was randomized. Panel (b) presents results from waves 2 and 3, where the enhanced information treatment was randomized. Subjective expectations are averaged over all participants in the analysis sample, while average realized visits are based on the subsets of participants who were randomized to receive each incentive. Section 3 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 (Panel (a): N=105 ($0), N=112 ($2), N=121 ($7); Panel (b): N=308 ($0); N=181 ($2); N=74 ($5); N=221 ($7)).

Figure 8: Willingness to pay versus participants' subjective expectations of attendance



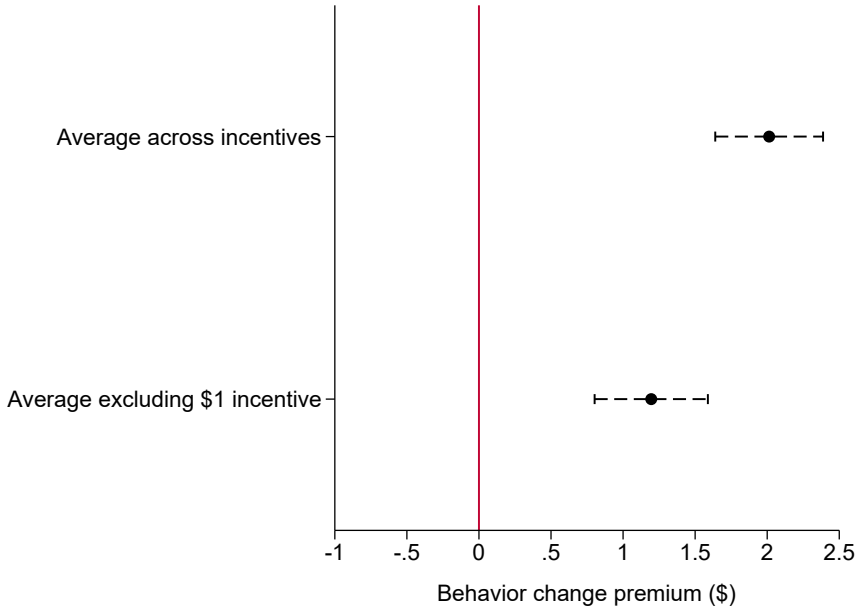
Notes: This figure presents a binned scatterplot comparing participants' WTP for piece-rate incentives to their subjective expectations of attendance under those incentives.

Figure 9: Subjective expectations of earnings and willingness to pay for piece-rate incentives



Notes: This figure compares participants' WTP for piece-rate incentives to their subjective expected earnings from the piece-rate incentives. For each incentive, subjective expected earnings are the product of the piece-rate level and participants' subjective beliefs about the number of days they would visit under that incentive.

Figure 10: Estimated average behavior change premium

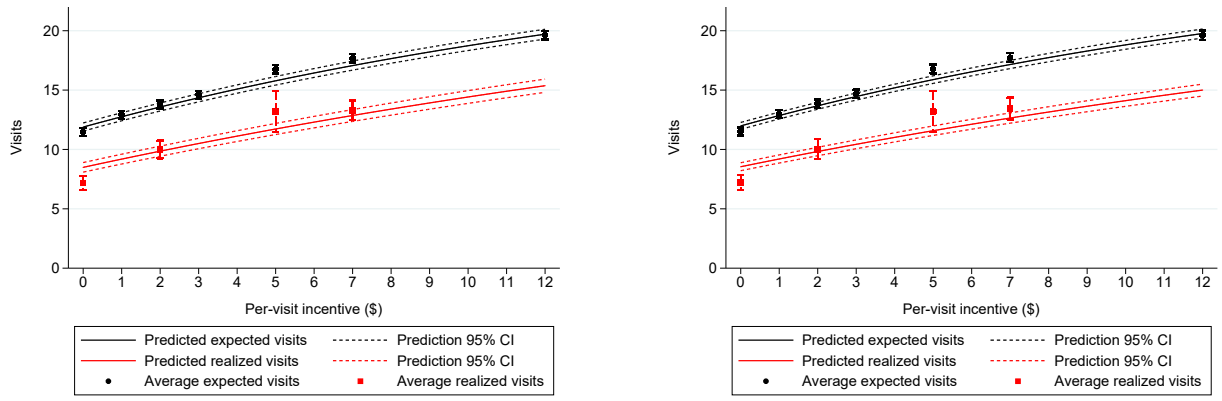


Notes: This figure shows the participants' average behavior change premium per dollar of additional incentive, as formalized in Sections 2.2 and 2.4.2. The top number averages across all incentive levels, while the bottom number reports the average excluding the $1 incentive. 95% confidence intervals are obtained from heteroskedasticity-robust standard errors.

Figure 11: Structural models' in-sample fit to participants' forecasted and realized attendance

(a) Homogenous structural parameters

(b) Heterogeneous structural parameters



Notes: These figures assess the structural models' fit to participants' subjective expectations of attendance and actual attendance. Panel (a) considers the specification in row 1 of Table 8. Panel (b) considers the structural model with eight heterogeneous types, as in Row 9 of Table 8. The empirical estimates of realized attendance and subjective expectations of attendance are as in Figure 5.

Part

Online Appendix

How are Preferences for Commitment Revealed?

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

Table of Contents

A Theory Appendix

A.1 Proof of Proposition 1

Proof. Let F_t and f_t denote the CDF and PDF, respectively, of the cost draws in period t. When the costs are distributed independently, we have

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

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

Consequently, if the terms Δ^3 and $\Delta^2\tilde{\alpha}''(p)$ are negligible,

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

Next, consider the case in which the costs are not distributed independently, but $\tilde{\beta}=1$. Here, we regard a strategy as a mapping from cost vectors $(c_1,\ldots,c_T)$ to a set of actions $(a_1,\ldots,a_T)$. The person's expected utility under piece-rate p, $V(0,p\sum a_t)$, will be differentiable in t as long as the costs are smoothly distributed. Thus, Theorem 1 of Milgrom and Segal (2002) implies that

$$\frac{d}{dp}V(0,p\sum a_t) = \tilde{\alpha}(p).$$

Proceeding as above shows that

$$V(0,(p+\Delta)\sum a_t) - V(0,p\sum a_t) = \Delta\frac{\tilde{\alpha}(p+\Delta)+\tilde{\alpha}(p)}{2} + O(\Delta^3, \Delta^2\tilde{\alpha}''(p))$$

□

Proof of the Corollary

Proof. If $p > 0$, then $w_i(p+\Delta) - w_i(p) = V_i(0,(p+\Delta))\varepsilon_{ij} - V_i(0,p)\varepsilon_{ij}$, and thus

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

If $p = 0$,

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

□

A.2 Formal results for $T = 1$

A.2.1 Behavior in absence of stochastic valuation errors or perceived social pressure

In period 1, individuals choose $a = 1$ if $\beta(p+b) - c \geq 0$, or equivalently if $c \leq \beta(p+b)$. 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 (y, ap) is

$$V(y, ap) = \beta\left[y + \int_{c\leq\tilde{\beta}(p+b)}(p+b-c)dF\right]$$

Assume $p > 0$. We call a contract $(-p, ap)$ 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 $(-p, (1-a)p)$ a commitment contract for $a = 0$ with penalty p, which we denote by $CC(p, 0)$.

A.2.2 With uncertainty about costs, quasi-hyperbolic preferences rarely generate 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 she will always choose $a = 1$. In our notation, this is a contract $(-p, p\mathbf{1}_{a=1})$ 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 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 $\Delta V(p) = V(-p, pa) - V(0, 0)$.

Proposition 2. *Suppose that $Pr(c > b)$ is positive for all realizations of c. Let $p > 0$.*

1. For a given $\bar{p} > 0$: there exist $\underline{\beta} > 0$ and $\bar{\beta} < 1$ such that $\max_{p \in [0,\bar{p}]} \Delta V(p) \leq 0$ (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 $\Delta V(p) > 0$ 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 $\max_{p \in [0,\bar{p}]} \Delta V(p) > 0$ 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), John (2019), and Schilbach (2019). 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 John (2019), who derives this non-monotonicity in the context of consumption and savings. In line with Heidhues and Kőszegi (2009) and John (2019), 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 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 $[\underline{c}, \bar{c}]$.

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 $\Delta V(p)$ is strictly increasing in $\tilde{\beta} \in [\underline{\beta}, \bar{\beta}]$. In particular, if $\underline{\beta} = 0$ and $\bar{\beta} = 1$, then $\Delta V(p)$ 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;*[40] *(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.

A.2.3 Commitment take-up with imperfect perception and demand effects

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 $\varepsilon_{ij} \equiv 1$.[41]

To characterize the new implications of the model, we begin with the observation that in the standard quasi-hyperbolic model in Section A.2.1, no individuals would ever choose commitment contracts for $a = 0$. This is simply because individuals would not choose to commit to take actions

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

[41] 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.

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. *Set $p > 0$.*

1. Assume that $\mu > 0$ or $Pr(\eta_i > \beta_i p) > 0$. Then a positive mass of individuals will choose a commitment contract for $a = 1$ with penalty p (i.e., contracts $(-p, p)$), 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. In this case, a positive mass of individuals would choose both commitment contracts for $a = 1$ and for $a = 0$.

3. Assume that $E[\tilde{\beta}_i]$ is sufficiently close to 1. Then there will be a positive correlation between demand for commitment contracts for $a = 1$ and commitment contracts for $a = 0$ if one of the following conditions holds: (i) if $\mu = 1$ and there are individual differences in η_i, (ii) if $\mu = 0$ and $Pr(\eta_i > \beta_i p) > 0$, or (iii) $\mu \in (0, 1)$ and for all individuals either $\eta_i \leq 0$ or $\eta_i > \beta_i p$.

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.[42]

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.

[42]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 $(-p, p)$, for $p \geq 0$, 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.[43] 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.

A.3 Proofs of Propositions for $T = 1$

A.3.1 Proof of Proposition 2

Proof. The perceived gains from a commitment contract are

$$\begin{aligned}\Delta V/\beta &= -p + \int_{c \leq \tilde{\beta}(p+b)} (p+b-c)dF - \int_{c \leq \tilde{\beta}b} (b-c)dF \\ &= -p(1 - F(\tilde{\beta}(b+p))) + \int_{c=\tilde{\beta}b}^{c=\tilde{\beta}(b+p)} (b-c)dF \end{aligned} \tag{7}$$

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 she will choose $a = 1$ when $c = \underline{c}$.

The commitment contract has a chance of being worthwhile when $\tilde{\beta}b < \underline{c} < \tilde{\beta}(b+p)$. In this case, if $\tilde{\beta}(b+p) < \bar{c}$ then the 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 $\tilde{\beta} \in (\underline{\beta}, \bar{\beta})$. □

[43] 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.

A.3.2 Proof of Proposition 3

Proof. We have

$$\begin{aligned}\frac{d}{d\tilde{\beta}}\Delta V/\beta &= p(b+p)f(\tilde{\beta}(b+p)) + (b+p)(b-\tilde{\beta}(b+p))f(\tilde{\beta}(b+p)) - b(b-\tilde{\beta}b)f(\tilde{\beta}b) \\ &= (1-\tilde{\beta})(b+p)^2 f(\tilde{\beta}(b+p)) - (1-\tilde{\beta})b^2 f(\tilde{\beta}b) \qquad (8)\end{aligned}$$

The expression (8) is positive if $\frac{f(\tilde{\beta}(b+p))}{f(\tilde{\beta}b)} \geq \frac{b^2}{(b+p)^2}$.

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 $\tilde{\beta}$. □

A.3.3 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 fourth 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,

$$\begin{aligned}
\int_{c=\beta b}^{b}(b-c)f(c)dc &= \int_{c=\beta b}^{b}(b-c)f(b)dc + \int_{c=\beta b}^{b}(b-c)(f(c)-f(b))dc \\
&= \frac{\gamma^2}{2}f(b) + \int_{c=\beta b}^{b}(b-c)(f(c)-f(b))dc \\
&\leq \frac{\gamma^2}{2}f(b) + \int_{c=\beta b}^{b}\gamma(f(c)-f(b))dc \\
&= \frac{\gamma^2}{2}f(b) + (\mu-\gamma f(b))\gamma \\
&= \gamma\mu - \frac{\gamma^2}{2}f(b) \qquad (9)
\end{aligned}$$

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,

$$\begin{aligned}
\int_{c\geq b}(b-c)f(c)dc &= \int_{c=b}^{b+\gamma}(b-c)f(c)dc + \int_{c\geq b+\gamma}(b-c)f(c)dc \\
&\leq \int_{c=b}^{b+\gamma}(b-c)f(c)dc - \int_{c\geq b+\gamma}\gamma f(c)dc \\
&= \int_{c=b}^{b+\gamma}(b-c)f(c)dc - \gamma(1-F(b+\gamma)) \\
&= \int_{c=b}^{b+\gamma}(b-c)f(c)dc - \gamma\left[(1-F(b)-(F(b+\gamma))-F(b))\right] \\
&\leq \int_{c=b}^{b+\gamma}(b-c)f(c)dc - \gamma\left(\mu - \int_{c=b}^{b+\gamma}f(c)dc\right) \\
&= \int_{c=b}^{b+\gamma}(b+\gamma-c)f(c)dc - \gamma\mu \\
&\leq \int_{c=b}^{b+\gamma}(b+\gamma-c)f(b)dc - \gamma\mu \\
&= \frac{\gamma^2}{2}f(b) - \gamma\mu \qquad (10)
\end{aligned}$$

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 (9) and (10) 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 (10) 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

$$\int_{c=\tilde{\beta} b}^{b}(b-c)f(c)dc \leq \gamma\mu/2. \tag{11}$$

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

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

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

$$\begin{aligned}
-\int_{x=b}^{x=b+\gamma}(b-x)f(c)dc &= \int_{x=b}^{x=b+\gamma}(x-b)\frac{b^2}{x^2}f(b)dx \\
&= b^2 f(b)\int_{x=b}^{x=b+\gamma}\left(\frac{1}{x}-\frac{b}{x^2}\right)dx \\
&= b^2 f(b)\left[\ln(x)+\frac{b}{x}\right]_{x=b}^{b+\gamma} \\
&= b^2 f(b)\left[\ln(b+\gamma)+\frac{b}{b+\gamma}-\ln(b)-1\right] \\
&= b^2 f(b)\left[\ln(1+k)-\frac{k}{1+k}\right] \\
&\geq b^2 f(b)\left[k-\frac{k^2}{2}-\frac{k}{1+k}\right] \qquad (12) \\
&= b^2 f(b)\left[\frac{k+k^2-k}{1+k}-\frac{k^2}{2}\right] \\
&= b^2 f(b)\left[\frac{k^2}{1+k}-\frac{k^2}{2}\right] \\
&= f(b)\left[\frac{\gamma^2}{1+k}-\frac{\gamma^2}{2}\right] \\
&= f(b)\left[\frac{\gamma^2(1-k)}{2(1+k)}\right] \\
&= \frac{\tilde{\beta}\gamma^2}{2(1+k)}f(b) \\
&= \frac{1}{2}\tilde{\beta}\gamma\frac{\gamma}{b+\gamma}bf(b) \\
&\geq \frac{\tilde{\beta}\gamma}{2}\frac{\mu}{\tilde{\beta}} \\
&= \gamma\mu/2 \qquad (13)
\end{aligned}$$

To obtain (12), 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 (11) and (13) 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. □

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 $\tilde{\beta}$. 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 $CC(p,0)$. □

A.3.4 Proof of Proposition 5

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

$$\left[\int_{c=0}^{\tilde{\beta}_i(p+b)} (p+b-c)dF(c) - \int_{c=0}^{\tilde{\beta}_i b} (b-c)dF(c)\right]\varepsilon_{ij} \geq p - \eta_i/\beta_i \tag{14}$$

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

$$\left[\int_{c\geq\tilde{\beta}_i(b-p)} p dF(c) + \int_{c=0}^{\tilde{\beta}_i(b-p)} (b-c)dF(c) - \int_{c=0}^{\tilde{\beta}_i b} (b-c)dF(c)\right]\varepsilon_{ij} \geq p - \eta_i/\beta_i \tag{15}$$

Since $\int_{c=0}^{\tilde{\beta}_i(p+b)}(p+b-c)dF(c) - \int_{c=0}^{\tilde{\beta}_i b}(b-c)dF(c) > 0$, condition (14) will be satisfied if either $\eta_i > \beta_i p$, or if the individual is prone to stochastic valuation errors and the draw ε_{ij} is sufficiently high.

Similarly, (15) will hold in several cases. First, suppose that the left-hand-side of (15) is positive, which will be the case for $\tilde{\beta}$ sufficiently close to 1, since this expression is positive for $\tilde{\beta}_i = 1$. In this case, the inequality holds if either $\eta_i > \beta_i p$, or if the individual is prone to stochastic valuation errors

and the draw ε_{ij} is sufficiently high. Second, suppose that the left-hand-side of (15) is negative. There then exists $\bar{\eta}$ such that the individual chooses $CC(p,0)$ if $\eta_i \geq \bar{\eta}$ and $\varepsilon_{ij} \equiv 1$. Alternatively, if the individual is prone to valuation errors, she may choose $CC(p,0)$ for low draws of ε_{ij}.

To prove part 3, first suppose that $\tilde{\beta}_i = 1$ for all individuals. In this case, the propensity to choose either contract is strictly increasing in η_i. Thus, if either $\mu = 0$ or $\mu = 1$, there will be a positive correlation in take-up of contracts. Next, consider $\mu \in (0,1)$. Among individuals with $\eta_i \leq 0$, then only individuals prone to stochastic valuation errors will take up either contract with positive probability. Among individuals with $\eta_i > p\beta_i$, everyone will take up the contracts with probability 1, irrespective of whether they are prone to stochastic valuation errors or not. This establishes a positive correlation in take-up of contracts for $E[\tilde{\beta}_i] = 1$. By continuity, the positive correlation holds. □

A.3.5 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 $\tilde{\beta}_i$. □

A.4 Generalizations to the dynamic case

We now consider a dynamic environment in which the individual can choose $a_t \in \{0,1\}$ 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.

A.4.1 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 $f(c)$ 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 $\tilde{\beta}$. □

A.4.2 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.

A.5 Generalizations to other environments

A.5.1 Summary of generalizations

Continuous choice We continue to explore the robustness of our Section 2.3 results about the undesirability of commitment contracts in Appendix A.5.2. 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 A.5.2 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 $\tilde{\beta}$.[44]

Other models Finally, in Appendix A.5.3, 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,[45] though uncertainty about future costs erodes the benefits of those contracts as well.

A.5.2 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 the future: If there is a positive probability of states of the world in which the period 0 self would

[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.

[45] 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.

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 $\tilde{\beta}$.

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

An important implication of 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 she does not fail to meet the contract. Both the benefits—which derive from behavior change—and the costs—which derive from 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,

$$\begin{aligned}\frac{d}{dp}((w+p)x^* - \theta C(x^*) - Xp) &= \frac{dx^*}{dp}(w+p-\theta C'(x^*)) - (X-x^*) \\ &= \frac{dx^*}{dp}(w+p-\tilde{\beta}(w+p)) - (X-x^*) \\ &= (1-\tilde{\beta})(w+p)\frac{dx^*}{dp} - (X-x^*) \\ &= (1-\tilde{\beta})x^*\varepsilon - (X-x^*) \\ &= ((1-\tilde{\beta})\varepsilon + 1)x^* - X \end{aligned} \tag{16}$$

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

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

Next, taking the derivative of (16) 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 $\tilde{\beta} \in [0, 1]$. □

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,

$$\begin{aligned}\frac{d}{dp}(u(x^* + \theta(Y - (r+p)x^* - pX))) &= \frac{dx^*}{dp}(u'(x^*) - \theta(r+p)) - \theta(X - x^*) \\ &= \frac{dx^*}{dp}\left(\frac{1}{\tilde{\beta}}\theta(r+p) - \theta(r+p)\right) - \theta(X - x^*) \\ &= (1/\tilde{\beta} - 1)\theta(r+p)\frac{dx^*}{dp} - \theta(X - x^*) \\ &= (1/\tilde{\beta} - 1)\theta x^*\varepsilon - \theta(X - x^*) \\ &= \theta((1/\tilde{\beta} - 1)\varepsilon + 1)x^* - \theta X \end{aligned} \tag{17}$$

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 (17) 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 $\tilde{\beta} \in [0, 1]$. □

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

$$\tilde{\beta}(wx(\theta^\dagger) - p) - \theta^\dagger C(x(\theta^\dagger)) = \tilde{\beta}wX - \theta^\dagger C(X) \tag{18}$$

where x satisfies $\theta^\dagger C'(x) = \tilde{\beta} w$. Differentiating (18) with respect to $\tilde{\beta}$ using the condition $\theta^\dagger C'(x) - \tilde{\beta} w = 0$ 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 (18) 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 (18) 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 $\frac{f'(\theta)}{f(\theta)} \geq -1/\theta$. □

A.5.3 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 $\int_{c \leq b/(1+\gamma)} (b - (1 + \gamma)c) dF(c)$. The impact of the restriction is thus

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

The inequality follows because $\gamma c \geq b - c$ iff $c \geq b/(1+\gamma)$. To get a quantitative sense of this, suppose that c is uniform on $[0, \bar{c}]$, and normalize $b = 1$. Then $E[c|c > 1] - 1 = \frac{\bar{c}-1}{2}$ and $E[c|c < b] = b/2$. Then the gains are negative if $\gamma(1/2)(1/\bar{c}) \leq \frac{\bar{c}-1}{\bar{c}} \frac{\bar{c}-1}{2}$, or if $\gamma \leq (\bar{c} - 1)^2$. For example, suppose that

$\gamma = 0.3$, which is equivalent to weighting delayed benefits relative to costs by a factor of $\beta = 0.77$. In this case, the gains from full commitment are negative if $\bar{c} > 1.55$. Compared to the uniform costs case in the present focus model, this implies that binding commitment contracts are more desirable for individuals with costly self-control, for a given "weight" on delayed benefits versus immediate costs.

B Further results and robustness tests for reduced-form results

B.1 Additional results on the behavior change premium

Table A1: Correlation between the behavior change premium and expected behavior change

	Behavior change premium			
	Excl. $1	Excl. $1	$1 only	$1 only
	(1)	(2)	(3)	(4)
Expected behavior change	1.51***	1.52***	0.36***	0.36***
	(0.13)	(0.13)	(0.08)	(0.08)
Constant	0.10		5.58***	
	(0.22)		(0.35)	
Dep. var. mean:	1.20	1.20	6.10	6.10
	(0.15)	(0.15)	(0.38)	(0.38)
Wave FEs	No	Yes	No	Yes
N	6,240	6,240	1,248	1,248
Clusters	1,248	1,248	1,248	1,248

Notes: This table reports the association between the estimated behavior change premium at each piece-rate incentive level and the expected behavior change in visits per dollar increase in the piece-rate incentive. Each column presents coefficient estimates from OLS regressions with heteroskedasticity-robust standard errors in parentheses. In columns 1 and 2, all incentive levels except the $1 incentive are included, while columns 3 and 4 are restricted to the $1 incentive. Regressions in columns 2 and 4 include wave fixed effects and omit the constant term. *** denote statistics that are statistically significantly different from 0 at the 1% level.

B.2 Additional results for Section 6.2

Here we show that the results in Table 4 in the main text are robust to splitting the sample by those in the information control group and those receiving the enhanced information treatment. We find here that there is no significant correlation for the control group and the point estimates are actually negative. There is a somewhat stronger correlation between the measured behavior change premium and the take-up of "more" commitments for those who received the enhanced information intervention.

Table A2: Correlation between the behavior change premium and take-up of "more" contracts

(a) Information control group

	Take-up of "more" visits contract			
	8+ visits	12+ visits	16+ visits	Pooled
	(1)	(2)	(3)	(4)
Behavior change premium	–0.040	–0.013	–0.036	–0.028
(z-score)	(0.025)	(0.024)	(0.029)	(0.022)
Dep. var. mean:	0.65	0.52	0.36	0.51
	(0.02)	(0.02)	(0.02)	(0.01)
Wave FEs	Yes	Yes	Yes	Yes
Contract FEs	No	No	No	Yes
N	429	622	429	1,480
Clusters	429	622	429	622

(b) Information treatment group

	Take-up of "more" visits contract			
	8+ visits	12+ visits	16+ visits	Pooled
	(1)	(2)	(3)	(4)
Behavior change premium	0.035***	0.041***	0.055***	0.044***
(z-score)	(0.013)	(0.013)	(0.014)	(0.012)
Dep. var. mean:	0.62	0.47	0.31	0.47
	(0.03)	(0.02)	(0.03)	(0.02)
Wave FEs	Yes	Yes	Yes	Yes
Contract FEs	No	No	No	Yes
N	246	452	246	944
Clusters	246	452	246	452

Notes: This table performs analysis identical to that of Table 4 in the body of the paper, but split by information control versus information treatment groups. *** denotes statistics that are statistically significantly different from 0 at the 1% level.

B.3 Additional results for Section 6.3

Here we show that the patterns of take-up for "more" and "fewer" commitment contracts, and in particular the positive correlation between those two decisions, holds when we split the sample separately into information control and enhanced information treatment groups. Note that the positive correlation between the two types is present in both samples, but slightly weaker for those receiving the enhanced information treatment.

Table A3: Take-up of "more" and "fewer" commitment contracts

(a) Information control group

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	0.65	0.36	0.88	0.49	0.23***	0.13***
12 visits	0.52	0.33	0.72	0.45	0.20***	0.13***
16 visits	0.36	0.31	0.56	0.48	0.20***	0.17***

(b) Information treatment group

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	0.62	0.30	0.89	0.43	0.27***	0.13***
12 visits	0.47	0.29	0.62	0.38	0.15***	0.09***
16 visits	0.31	0.22	0.47	0.34	0.16***	0.12***

Notes: This table performs analysis identical to that of Table 6 in the body of the paper, but split by information control versus information treatment groups. *** denotes statistics that are statistically significantly different from 0 at the 1% level.

Table A4: Association between the behavior change premium and take-up of "more" but not "fewer" contracts

	Take-up of "more" visits contract only			
	8+ visits (1)	12+ visits (2)	16+ visits (3)	Pooled (4)
Behavior change premium (z-score)	0.004 (0.015)	0.009 (0.014)	0.013 (0.014)	0.009 (0.013)
Dep. var. mean:	0.34 (0.02)	0.28 (0.01)	0.19 (0.01)	0.27 (0.01)
Wave FEs	Yes	Yes	Yes	Yes
Contract FEs	No	No	No	Yes
N	849	1,248	849	2,946
Clusters	849	1,248	849	1,248

Notes: This table reports the association between the estimated average behavior change premium (calculated excluding the $1 incentive) and an indicator for whether a participant took up a "more" but not a "fewer" commitment contract. Each column reports coefficient estimates from OLS regressions. Dependent variable means with standard errors in parentheses are also reported. In columns 1, 2, and 3, the dependent variables are indicators for taking up a "more" but not "fewer" contract at each threshold of 8, 12, and 16 visits, respectively. Column 4 is a regression that pools over columns 1-3. Standard errors are heteroskedasticity-robust in columns 1-3, and are clustered at the subject level in column 4. *,** denote statistics that are statistically significantly different from 0 at the 10% and 5% levels respectively.

B.4 Additional results for Section 6.4.1

Here we provide a few additional results showing that measures that are positively correlated with the take-up of "more" commitments tend to be negatively correlated with the take-up of "fewer" commitments. These results bolster the arguments in Section 6.4.1 that participants were not simply confusing "fewer" contracts for "more" contracts.

Table A5: 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"	0.12***		0.14***	–0.09***		–0.13***
	(0.02)		(0.02)	(0.03)		(0.03)
Commit to "fewer"		–0.05*	–0.08***		0.17***	0.20***
		(0.03)	(0.02)		(0.03)	(0.03)
N	399	399	399	399	399	399
"More" − "Fewer"			0.22***			-0.34***
			(0.03)			(0.05)

Notes: This table reports the association between the take-up of "more" and "fewer" commitment contracts (with a threshold of 12 visits) and subjective beliefs about the probability of success if exogenously assigned the contract. Each column presents coefficient estimates and heteroskedastic-consistent standard errors in parentheses from separate OLS regressions. Columns 1-3 display associations with participants' subjective expectations of following through on the "12 or more attendances" contract, with the subjective expectations coded on a scale of 0 to 1. Columns 4-6 display associations with participants' subjective expectations of following through on the "fewer than 12 attendances" contract, with the subjective expectations coded on a scale of of 0 to 1. The sample consists of participants in wave 3, the only wave in 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 A6: Other correlates of commitment contract take-up

	Expected attendance	Past attendance	Goal attendance
	(1)	(2)	(3)
Chose "more contract"	1.94***	1.31***	2.56***
	(0.21)	(0.22)	(0.22)
Chose "fewer" contract	–0.87***	–1.94***	–1.03***
	(0.23)	(0.23)	(0.25)
N	2,946	2,946	2,946
"More" − "Fewer"	2.81***	3.25***	3.59***
	(0.34)	(0.35)	(0.36)

Notes: This table presents results from three stacked OLS regressions that study how the three dependent variables in columns 1-3 relate to people's decision to take up the "more" contracts and the "fewer" contracts. Since participants were asked about multiple commitment contracts in waves 1 and 2, each participant contributes three observations to the regressions in these two waves. Heteroskedasticity-robust standard errors are reported in parentheses. *** denotes statistics that are statistically significantly different from 0 at the 1% level.

B.5 Additional results for Section 6.4.3

Here we present additional results that highlight that the patterns of selecting "more" and "fewer" commitment contracts are not limited to participants for whom the contract was unlikely to be binding. For each visit threshold we identify participants whose self-reported subjective expectations for gym visits in the absence of incentives were at least two or four visits below the threshold. For these individuals, the "more" contract would likely be significantly binding. Similarly, we identify participants whose subjective expectations for gym visits in the absence of incentives were at least one or three more than the threshold, which implies 2 or 4 more than the limit for compliance with the "fewer" contract. The tables show that the take-up of both types of contracts is similar if we limit to those for whom they were more likely to be binding (Table A7). Moreover, the correlation between the take-up of "more" and "fewer" contracts is similar as we limit to those for whom one of the contract types was more likely to be binding (Table A8).

Table A7: Take-up rate by expected attendance

		Chose "more"	Chose "more"		Chose "fewer"	Chose "fewer"
	Chose "more"	given exp. att.	given exp. att.	Chose "fewer"	given exp. att.	given exp. att.
	contract	$\leq r-2$	$\leq r-4$	contract	$\geq r+1$	$\geq r+3$
Threshold (r)	(1)	(2)	(3)	(4)	(5)	(6)
8 visits	0.64	0.62	0.63	0.34	0.31	0.29
12 visits	0.49	0.39	0.35	0.31	0.30	0.29
16 visits	0.32	0.24	0.23	0.27	0.31	0.32

Notes: Each column reports the take-up rate of a "more" or "fewer" commitment contract with a given visits threshold $r \in \{8, 12, 16\}$. In columns 2, 3, 5, and 6, the samples are restricted to participants whose subjective expectations of gym attendance in the absence of incentives are $\leq r-2$ (column 2), $\leq r-4$ (column 3), $\geq r+1$ (column 5), or $\geq r+3$ (column 6), respectively.

Table A8: Correlation of "more" and "fewer" take-up by expected attendance

Threshold (r)	All (1)	Exp. att. $\leq r-2$ (2)	Exp. att. $\leq r-4$ (3)	Exp. att. $\geq r+1$ (4)	Exp. att. $\geq r+3$ (5)	Exp. att. ≤ 6 (6)	Exp. att. ≥ 17 (7)
8 visits	0.37***	0.39***	0.46***	0.37***	0.38***	0.39***	0.41***
12 visits	0.24***	0.23***	0.27***	0.31***	0.27***	0.29***	0.32***
16 visits	0.23***	0.22***	0.22***	0.33***	0.33***	0.25**	0.33***

Notes: Each column reports the correlation between the take-up of "more" and "fewer" commitment contracts with a given visits threshold, with the sample limited in columns 2-7 by participants' attendance expectations in the absence of incentives. *** denotes differences that are statistically significantly different from 0 at the 1% level.

C Structural estimation appendix

C.1 Details on GMM estimation of parameters

Let $\xi = (\beta, \tilde{\beta}, b, \lambda)$ denote the vector of parameters that we are seeking to estimate. Let $\tilde{\alpha}_i(p)$ denote an individual's forecasted visits as a function of piece-rate incentive p, and let a_i denote actual visits. Let p_i denote the piece-rate incentive assigned to individual i. We have three sets of moment conditions.

The first set of moment conditions corresponds to forecasted attendance:

$$E\left[\left(28\left(1-e^{-\lambda(\tilde{\beta}(b+p))}\right)-\alpha_i(p)\right)p^n\right]=0$$

for all $p \in P^F = \{0,1,2,3,5,7,12\}$, and all $n \in \{0,1,2\}$. The set P^F is the set of all incentives for which we elicited forecasts. We use $1, p, p^2$ as the instruments for the forecasted attendance equation, and our results are virtually unchanged for smaller and higher n.

The second set of moment conditions corresponds to actual attendance:

$$E\left[\left(28\left(1-e^{-\lambda(\beta(b+p_i))}\right)-a_i\right)p_i^n\right]=0$$

for all $n \in \{0,1,2\}$.

The third set of moment conditions corresponds to the behavior change premium:

$$E\left[(1-\tilde{\beta})(b+(p_k+p_{k+1})/2)\frac{\tilde{\alpha}_i(p+\Delta_k)-\tilde{\alpha}_i(p)}{\Delta_k}-\left(\frac{w_i(p+\Delta_k)-w_i(p)}{\Delta_k}-\frac{\tilde{\alpha}_i(p+\Delta_k)+\tilde{\alpha}_i(p)}{2}\right)\right]=0$$

where p_k and p_{k+1} are one of six pairs of adjacent incentives from the set P^F, and $\Delta_k = p_{k+1} - p_k$.

Letting $\hat{\xi}$ denote the parameter estimates, the GMM estimator chooses the parameter $\hat{\xi}$ that minimizes

$$\left(m(\xi)-m(\hat{\xi})\right)' W \left(m(\xi)-m(\hat{\xi})\right),$$

where $m(\xi)$ are the theoretical moments, $m(\hat{\xi})$ 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 the two-step estimator outlined in Hall (2005). In the first step, we set W equal to the identity matrix,[46] and use this to solve the moment conditions for $\hat{\xi}$, which we denote $\hat{\xi}_1$. Since $\hat{\xi}_1$ 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

$$\left(m(\xi) - m(\hat{\xi})\right)' \hat{W} \left(m(\xi) - m(\hat{\xi})\right)$$

using $\hat{W} = \hat{S}^{-1}$, which gives the optimal $\hat{\xi}$ (Hansen, 1982).

C.2 Implications of heterogeneity for our parameter estimates

Consider a first-order, linear approximation to person i's expected linear attendance, $A_i(p) = \lambda_i^0 + \lambda_i^1 \beta_i (b_i + p)$. The forecasted attendance curve is given by $\tilde{A}_i(p) = \lambda_i^0 + \lambda_i^1 \tilde{\beta}_i (b_i + p)$, and the desired attendance curve is given by $A_i^*(p) = \lambda_i^0 + \lambda_i^1 (b_i + p)$. The behavior change premium is then given by

$$BCP_i(p, \Delta) = (1 - \tilde{\beta}_i)(b_i + p + \Delta/2)\lambda_i^1 \tilde{\beta}_i$$

We show that we can recover $E[\beta_i]$, $E[\tilde{\beta}_i]$ and $E[b_i]$ from the population averages $\bar{A}(p)$, $\bar{\tilde{A}}(p)$, and $\overline{BCP_i(p, \Delta)}$. In other words, if one assumes that the aggregate forecasted and realized attendance curves and the behavior change premium are generated by a representative agent, the parameters ascribed to that representative agent in fact correspond to the average parameters in the population.

We make the following assumptions:

Assumption 1. *The parameters $\tilde{\beta}_i, b_i, \lambda_i^1$ are mutually independent.*

Assumption 2. *The parameters $\beta_i, b_i, \lambda_i^1$ are mutually independent.*

Assumption 3. *Terms of order $E[(1 - \tilde{\beta}_i)^2]$ are negligible.*

Proof. Without loss of generality, consider two values of p, p_1 and $p_2 = p_1 + 1$. Let $\bar{\tilde{A}}^{-1}$ denote the inverse of $\bar{\tilde{A}}(p)$, which is also approximately linear, by assumption. We then have

$$E[\tilde{A}_i(p_2) - \tilde{A}_i(p_1)] = E[\tilde{\beta}_i]E[\lambda_i^1] \tag{19}$$

$$E[A_i(p_2) - A_i(p_1)] = E[\beta_i]E[\lambda_i^1] \tag{20}$$

$$\bar{\tilde{A}}^{-1}(0) = -E[b_i] \tag{21}$$

Since the left-hand-side of all three equations above is known, we can solve for $E[\tilde{\beta}_i]E[\lambda_i^1]$, $E[\beta_i]E[\lambda_i^1]$, $E[b_i]$.

[46] 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.

Next, note that

$$\begin{aligned}E[BCP_i(p,\Delta)] &= E[(1-\tilde{\beta}_i)(b_i+p+\Delta/2)\lambda_i^1\tilde{\beta}_i] \\ &= E[(1-\tilde{\beta}_i)(b_i+p+\Delta/2)]E[\tilde{\beta}_i]E[\lambda_i^1] + O\left(E[(1-\tilde{\beta}_i)^2]\right) \\ &= E[1-\tilde{\beta}_i]\left(E[b_i]E[\tilde{\beta}_i]E[\lambda_i^1] + (p+\Delta/2)E[\tilde{\beta}_i]E[\lambda_i^1]\right) \\ &+ O\left(E[(1-\tilde{\beta}_i)^2]\right)\end{aligned}$$

Since $E[b_i]E[\tilde{\beta}_i]E[\lambda_i^1]$ and $E[\tilde{\beta}_i]E[\lambda_i^1]$ are identified from the system of equations (19)-(21), we can therefore solve for $E[1-\tilde{\beta}_i]$ given a value of $E[BCP_i(p,\Delta)]$ for a pair of (p,Δ). Given a value of $E[\tilde{\beta}_i]$, equation (19) then identifies $E[\lambda_i^1]$, and given the value of $E[\lambda_i^1]$, equation (20) then identifies $E[\beta_i]$. □

C.3 Details on equilibrium strategies, value functions, and simulated behavior

C.3.1 Equilibrium value functions and strategies

We let f denote the probability density function (PDF) of a random variable given by $\underline{c}+X$, where X is distributed exponentially with rate parameter λ. We let F denote the cumulative distribution function (CDF). The exponential distribution provides closed-form solutions for both the conditional expectation and the CDF.

$$\int_{c=\underline{c}}^{x} cf(c)dc = \underline{c} + \frac{1}{\lambda}\left(1-e^{-\lambda(x-\underline{c})}\right) - (x-\underline{c})e^{-\lambda(x-\underline{c})} \tag{22}$$

$$F(x) = 1-e^{-\lambda(x-\underline{c})} \tag{23}$$

Let $h_t = \sum_{j=1}^{t-1} a_t$ denote the period-t history summarizing a person's total attendance in periods $1,\ldots,t-1$. Given a contract $\mathcal{C}$, we let $W_t^*(\mathcal{C},h_t;\beta,\tilde{\beta})$ denote a person's expected utility using the period $t-1$ information set and the long-run criterion. Let $W_t(\mathcal{C},h_t;\beta,\tilde{\beta})$ denote a person's forecast of the expected utility (normalized by β), which may differ from W_t^* if $\tilde{\beta}\neq\beta$. When $\mathcal{C}$ is a linear piece-rate incentive of p per attendance,

$$W_t^*(\mathcal{C},h_t;\beta,\tilde{\beta}) = (T-t)\cdot\int_{c=0}^{\max(\beta(b+p)-\underline{c},0)}(b-c-\underline{c})f(c)dc$$

$$W_t(\mathcal{C},h_t;\beta,\tilde{\beta}) = (T-t)\cdot\int_{c=0}^{\max(\tilde{\beta}(b+p)-\underline{c},0)}(b-c-\underline{c})f(c)dc$$

and in each period a person chooses to attend the gym if and only if $\beta(b+p)\geq c_t$. We now characterize W_t^* and W_t when $\mathcal{C}$ is a contract where participants lose p if they don't attend at least g times. We start with the sophisticated case where $\beta=\tilde{\beta}$. In period T,

$$W_T^*(h_t) = \begin{cases} \int_{c=\underline{c}}^{\beta b}(b-c)f(c)dc & \text{if } h_T \geq g \\ \int_{c=\underline{c}}^{\beta(b+p)}(b-c)f(c)dc - (1-F(\beta(b+p)))p & \text{if } h_T = g-1 \\ \int_{c=\underline{c}}^{\beta b}(b-c)f(c)dc - p & \text{if } h_T < g-1 \end{cases}$$

Now set $\Delta W_{t+1}^*(h) := W_{t+1}^*(h+1) - W_{t+1}^*(h)$. Then a person chooses to attend the gym in period t if and only if $\beta(b + \Delta W_{t+1}^*(h_t)) \geq c_t$. For $t < T$, we have the following recursion on the value functions:

$$W_t^*(h_t) = \int_{c=\underline{c}}^{\beta(b+\Delta W_{t+1}^*(h_t))}(b + W_{t+1}^*(h_t+1) - c)f(c)dc + \int_{c=\beta(b+\Delta W_{t+1}^*(h_t))}^{\infty} W_{t+1}^*(h_t)f(c)dc. \quad (24)$$

Note that (22) and (23) imply that the expression in (24) above has a closed-form solution for W_t given a value function W_{t+1}.

Next, note that $W_t(\mathcal{C}, h_t; \beta, \tilde{\beta}) = W_t^*(\mathcal{C}, h_t; \tilde{\beta}, \tilde{\beta})$, meaning that subjective expectations of utility of partial naifs are immediately implied by the recursion for sophisticates. In period T,

$$W_T(\mathcal{C}, h_t; \beta, \tilde{\beta}) = \begin{cases} \int_{c=\underline{c}}^{\tilde{\beta} b}(b-c)f(c)dc & \text{if } h_T \geq g \\ \int_{c=\underline{c}}^{\tilde{\beta}(b+p)}(b-c)f(c)dc - (1-F(\beta(b+p)))p & \text{if } h_T = g-1 \\ \int_{c=\underline{c}}^{\tilde{\beta} b}(b-c)f(c)dc - p & \text{if } h_T < g-1 \end{cases}$$

while $W_T^*(\mathcal{C}, h_T; \beta, \tilde{\beta}) = W_T^*(\mathcal{C}, h_T; \tilde{\beta}, \tilde{\beta})$. Set $\Delta W_{t+1}(h) := W_{t+1}(h+1) - W_{t+1}(h)$. In period t, a person chooses to attend the gym if and only if $\beta(b + \Delta W_{t+1}(h_t)) \geq c_t$. For $t < T$, we have the following recursion on the value functions:

$$W_t^*(\mathcal{C}, h_t; \beta, \tilde{\beta}) = \int_{c=\underline{c}}^{\beta(b+\Delta W_{t+1}(h_t))}(b + W_{t+1}^*(h_t+1) - c)f(c)dc + \int_{c=\beta(b+\Delta W_{t+1}(h_t))}^{\infty} W_{t+1}^*(h_t)f(c)dc. \quad (25)$$

A person's incremental gain from the contract is given by $W_0^*(\mathcal{C}, h_t; \beta, \tilde{\beta}) - W_0^*(\emptyset, h_t; \beta, \tilde{\beta})$, where $\emptyset$ denotes the absence of a contract.

C.3.2 Simulating the impacts of contracts on behavior

Under a piece-rate incentive of p per attendance, a person attends in period t if and only if $\beta(b+p) \geq c_t$, and thus the impact of a piece-rate incentive on behavior is simply $F(\beta(b+p)) - F(\beta b)$, for which an analytic solution is given by (23). An analytic solution does not exist for the impacts of commitment contracts. We thus study the effects using simulation methods.

Specifically, we simulate attendance under a commitment contract over 10,000 draws of a T-period cost vector $(c_1, c_2, \ldots, c_T)$, where each c_t is an independent draw from the exponential distribution with CDF F. In each draw, a person's behavior in each period can be computed recursively by "forward induction"—i.e., first computing behavior in period $t = 1$, then $t = 2$, and so forth. In

particular, in period 1, a person chooses $a_1 = 1$ if

$$c_1 \leq \beta \left[b + W_2(\mathcal{C}, 1; \beta, \tilde{\beta}) - V_2(\mathcal{C}, 0; \beta, \tilde{\beta}) \right].$$

For periods $t > 1$, a person chooses $a_t = 1$ if

$$c_t \leq \beta \left[b + W_{t+1}(\mathcal{C}, h_t + 1; \beta, \tilde{\beta}) - V_{t+1}(\mathcal{C}, h_t; \beta, \tilde{\beta}) \right].$$

C.3.3 Optimal piece-rate incentives for efficient behavior change

Consider a set J of types indexed by j, and having a share μ_j in the population. The efficiency of behavior change under a piece-rate incentive p is given by

$$W^E = T \cdot \left[\sum_{j \in J} \mu_j \int_{c=b_j}^{c=b_j+p} (b_j - c) f_j(c) dc \right]$$

The first-order condition is

$$\sum_j \mu_j \beta_j (b_j(1 - \beta_j) - \beta_j p) f(\beta_j(b_j + p)) = 0$$

which implies that the optimal incentive must satisfy

$$p = \frac{\sum_{j \in J} \mu_j (1 - \beta_j) b_j \beta_j f_j(\beta_j(b_j + p))}{\sum_j \mu_j \beta_j^2 f_j(\beta_j(b_j + p))}.$$

For example, under homogeneity, the optimal value of p is simply $(1 - \beta)b/\beta$. We verify numerically that there is a unique value of p satisfying the condition above in the heterogeneous cases that we study.

C.4 Welfare effects of other commitment contracts

Table A9: Estimated welfare effects of piece-rates and commitment contracts

		(1) Avg. Δ in attendance	(2) Δ Agent surplus	(3) Δ Health benefits	(4) Δ Attendance costs	(5) Δ Social Surplus
1	8+ visits contract	0.77	−$5.09	$6.41	$6.14	$0.27
2	Linear incentive, $p = \$1.21$	0.77	$14.42	$8.18	$5.26	$2.93
3	16+ visits contract	1.43	−$3.40	$15.00	$12.05	$2.94
4	Linear incentive, $p = \$2.24$	1.43	$27.75	$14.77	$9.70	$5.06

Notes: Analogous to Table 10, this table reports the estimated effects of four different incentive schemes, averaged over the full population. There are eight heterogeneous types in all rows, analogous to Table 10. In rows 1 and 2, we assume that there are eight types of individuals, corresponding to eight subgroups: below- or above-median past attendance, crossed with receiving either the enhanced information treatment or no information treatment, crossed with choosing the 8+ commitment contract. In rows 3 and 4, we assume that there are eight types of individuals, corresponding to eight subgroups: below- or above-median past attendance, crossed with receiving either the enhanced information treatment or no information treatment, crossed with choosing the 16+ commitment contract.

C.5 Welfare estimates for alternative specifications of heterogeneity

Table A10: Estimated welfare effects of piece-rates and commitment contracts, homogeneity

		(1) Avg. Δ in attendance	(2) Δ Agent surplus	(3) Δ Health benefits	(4) Δ Attendance costs	(5) Δ Social Surplus
1	12+ visits contract	1.51	−$3.82	$14.49	$14.86	−$0.38
2	Linear incentive, $p = \$2.15$	1.51	$26.91	$14.37	$8.67	$5.70
3	Optimal linear incentive, $p = \$7.98$	5.04	$118.61	$48.07	$36.53	$11.53
4	8+ visits contract	0.63	−$1.39	$5.81	$6.08	−$0.28
5	Linear incentive, $p = \$0.88$	0.63	$10.57	$6.13	$3.62	$2.50
6	16+ visits contract	1.64	−$3.46	$16.88	$16.69	$0.20
7	Linear incentive, $p = \$2.32$	1.64	$29.80	$15.61	$9.42	$6.19

Notes: This table reports welfare effects for the incentive schemes considered in Table 10 and A9, but under different assumptions about heterogeneity. In this table, we assume that individuals are homogeneous conditional on their choice of contract, as in row 2 of Table 9 (and its analogues for rows 4/5 and rows 6/7).

Table A11: Estimated welfare effects of piece-rates and commitment contracts, heteroegenity along past attendance (below/above median)

		(1) Avg. Δ in attendance	(2) Δ Agent surplus	(3) Δ Health benefits	(4) Δ Attendance costs	(5) Δ Social Surplus
1	12+ visits contract	1.29	−$9.10	$10.84	$9.83	$1.02
2	Linear incentive, $p = \$1.97$	1.29	$23.72	$12.67	$7.99	$4.68
3	Optimal linear incentive, $p = \$7.83$	4.61	$111.78	$45.17	$35.17	$10.00
4	8+ visits contract	0.91	−$5.07	$6.53	$6.80	−$0.27
5	Linear incentive, $p = \$1.37$	0.91	$16.27	$9.07	$5.77	$3.30
6	16+ visits contract	1.31	−$5.95	$12.60	$11.91	$0.70
7	Linear incentive, $p = \$1.98$	1.31	$24.22	$12.86	$8.15	$4.71

Notes: This table reports welfare effects for the incentive schemes considered in Table 10 and A9, but under different assumptions about heterogeneity. In this table, we make the heterogeneity assumption in row 4 of Table 9 (and its analogues for rows 4/5 and rows 6/7).

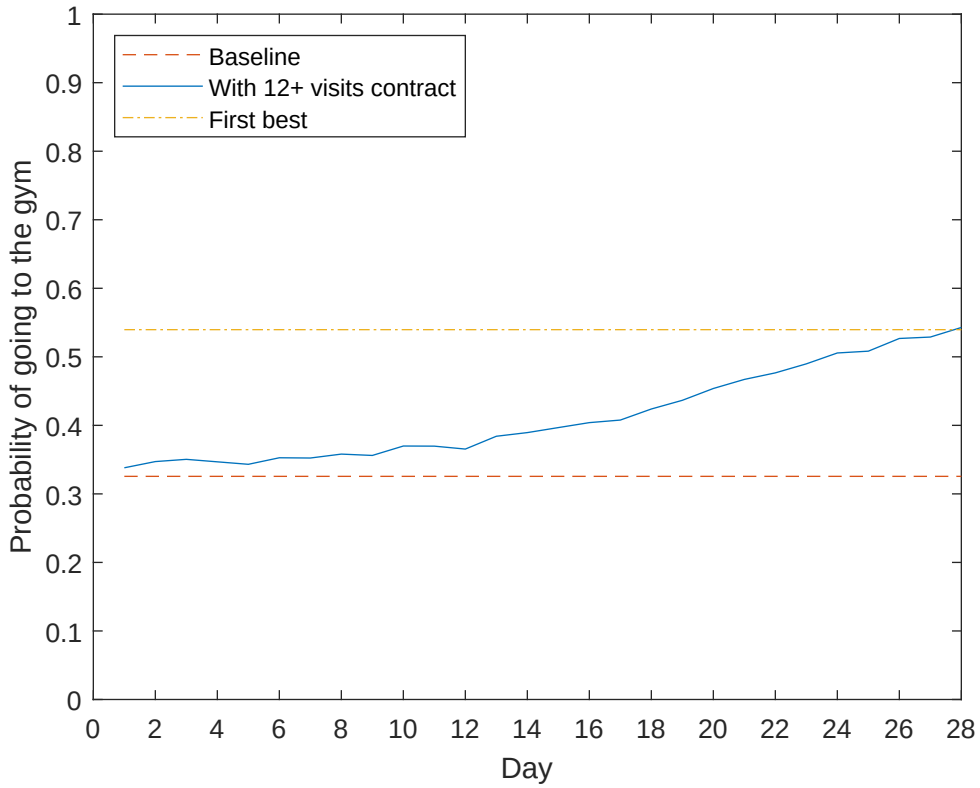
Table A12: Estimated welfare effects of piece-rates and commitment contracts, heterogeneity along past attendance (quartile)

		(1) Avg. Δ in attendance	(2) Δ Agent surplus	(3) Δ Health benefits	(4) Δ Attendance costs	(5) Δ Social Surplus
1	12+ visits contract	1.35	−$9.82	$11.04	$10.17	$0.86
2	Linear incentive, $p = \$2.15$	1.35	$25.74	$13.48	$8.65	$4.83
3	Optimal linear incentive, $p = \$7.74$	4.42	$108.70	$43.75	$34.23	$9.52
4	8+ visits contract	0.91	−$7.29	$6.64	$6.40	$0.24
5	Linear incentive, $p = \$1.43$	0.91	$16.85	$9.25	$6.04	$3.20
6	16+ visits contract	1.25	−$6.82	$11.09	$10.41	$0.68
7	Linear incentive, $p = \$1.95$	1.25	$23.52	$12.39	$8.06	$4.34

Notes: This table reports welfare effects for the incentive schemes considered in Table 10 and A9, but under different assumptions about heterogeneity. In this table, we make the heterogeneity assumption of row 5 of Table 9 (and its analogues for rows 4/5 and rows 6/7).

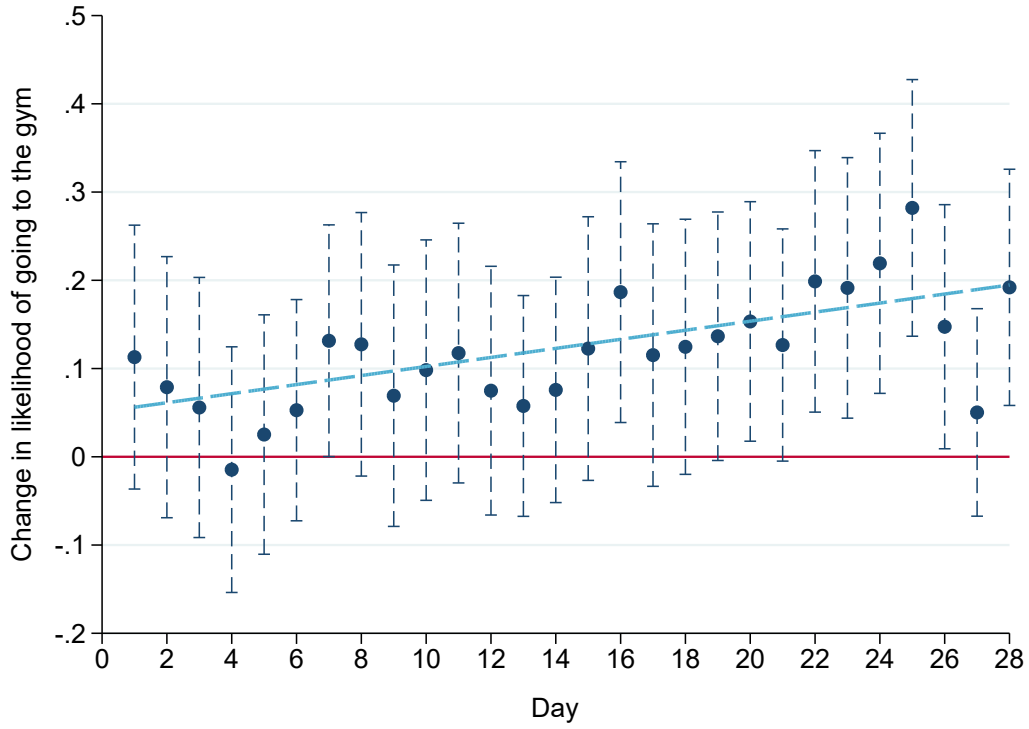
C.6 How commitment contracts affect attendance over time

Figure A1: Simulated probability of attendance each day, chose 12+ visits contract



Notes: This figure displays the simulated probability of attending the gym each day, under the heterogeneity assumptions of Table 10.

Figure A2: Change in likelihood of attendance each day, chose 12+ visits contract



Notes: This figure displays the estimated change in the likelihood of attending the gym each day from assignment to the "more" contract with a threshold of 12 visits. Estimates are obtained from an OLS regression of gym attendance on indicators for each day and their interactions with an indicator for assignment to the contract. The coefficients on the interaction terms are plotted with 95% confidence intervals, obtained from standard errors clustered at the participant level. The sample is limited to participants who wanted the contract and were exogenously assigned to either receive the contract or to receive no incentives. A line is plotted with an intercept and slope equal to the coefficients on *12+ visits contract* and *Day* × *12+ visits contract*, respectively, from the regression in Table A13.

Table A13: Daily likelihood of attendance, chose 12+ visits contract

	Likelihood of attendance
	(1)
Day	–0.005***
	(0.001)
12+ visits contract	0.051
	(0.045)
Day × 12+ visits contract	0.005**
	(0.002)
Wave FEs	Yes
N	7,336
Clusters	262

Notes: This table reports the estimated change in the likelihood of attending the gym each day by assignment to the "more" contract with a threshold of 12 visits. *Day* is an index for the day in the 4-week study period, from 1 to 28, and *12+ visits contract* is an indicator for assignment to the contract. The table presents coefficient estimates and standard errors clustered at the participant level in parentheses from an OLS regression. The sample is limited to participants who wanted the contract and were exogenously assigned to either receive it or no piece-rate incentive or contract. **,*** denote statistics that are statistically significantly different from 0 at the 5% and 1% level respectively.

C.7 Alternative assumptions about the cost distribution

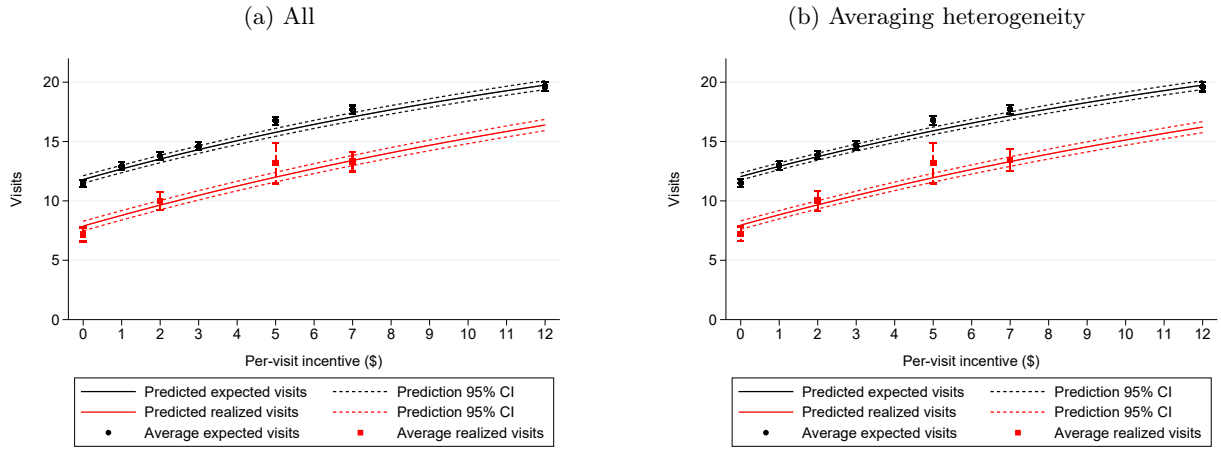
C.7.1 Minimal cost draw of $10

Table A14: Parameter estimates

		(1) $\hat{\beta}$	(2) $\hat{\tilde{\beta}}$	(3) $\hat{b}$	(4) $(1-\hat{\beta})\cdot\hat{b}$	(5) $(1-\hat{\tilde{\beta}})\cdot\hat{b}$	(6) $\frac{(1-\hat{\tilde{\beta}})}{(1-\hat{\beta})}$
1	All	0.74	0.91	20.70	5.35	1.84	0.34
	(N=1,126)	(0.72, 0.76)	(0.89, 0.93)	(20.08, 21.32)	(4.88, 5.81)	(1.41, 2.28)	(0.28, 0.41)
2	Information control	0.73	0.92	20.95	5.56	1.61	0.29
	(N=560)	(0.71, 0.76)	(0.90, 0.94)	(20.09, 21.82)	(5.01, 6.11)	(1.18, 2.05)	(0.22, 0.36)
3	Enhanced information	0.74	0.88	21.18	5.51	2.47	0.45
	treatment (N=392)	(0.70, 0.78)	(0.84, 0.93)	(20.06, 22.30)	(4.54, 6.48)	(1.50, 3.43)	(0.33, 0.57)
4	Below median past	0.68	0.89	18.31	5.84	1.95	0.33
	attendance (N=550)	(0.66, 0.71)	(0.86, 0.93)	(17.63, 18.99)	(5.21, 6.47)	(1.31, 2.60)	(0.25, 0.42)
5	Above median past	0.80	0.93	23.39	4.66	1.62	0.35
	attendance (N=576)	(0.78, 0.82)	(0.91, 0.95)	(22.28, 24.51)	(4.03, 5.29)	(1.09, 2.14)	(0.26, 0.44)
6	Chose 8+ visit	0.74	0.91	20.12	5.22	1.81	0.35
	contract (N=546)	(0.71, 0.77)	(0.88, 0.94)	(19.27, 20.97)	(4.54, 5.90)	(1.11, 2.50)	(0.24, 0.45)
7	Chose 12+ visit	0.71	0.90	20.83	5.97	2.04	0.34
	contract (N=556)	(0.69, 0.74)	(0.87, 0.93)	(19.95, 21.71)	(5.29, 6.64)	(1.36, 2.73)	(0.25, 0.43)
8	Chose 16+ visit	0.69	0.87	22.16	6.78	2.99	0.44
	contract (N=275)	(0.65, 0.73)	(0.81, 0.92)	(20.69, 23.63)	(5.64, 7.92)	(1.74, 4.24)	(0.31, 0.57)
9	Averaging heterogeneity	0.75	0.92	21.30	5.28	1.68	0.33
	(N=952)	(0.73, 0.76)	(0.90, 0.94)	(20.55, 22.04)	(4.83, 5.74)	(1.28, 2.07)	(0.27, 0.40)

Notes: This table replicates Table 8, but assumes that the distribution of cost draws is given by $10 + X$, where X is an exponentially distributed random variable.

Figure A3: Predicted and observed actual and expected attendance by incentive



Notes: This figure replicates Figure 11, but assumes that the distribution of cost draws is given by $10 + X$, where X is an exponentially distributed random variable.

Table A15: Estimated impact of 12+ contract on attendance

		(1) Δ in att.	(2) Pr(att. ≥ 12) with contract	(3) Pr(att. ≥ 12) without contract	(4) Δ in Pr(att. ≥ 12)
1	Empirical	3.51 (1.38, 5.65)	0.65 (0.52, 0.78)	0.22 (0.10, 0.35)	0.42 (0.26, 0.58)
2	Homogeneous	3.78	0.96	0.10	0.86
3	Heterogeneous by median past att., info. treatment	3.80	0.89	0.30	0.58
4	Heterogeneous by median past att.	3.99	0.91	0.30	0.61
5	Heterogeneous by quartile past att.	4.24	0.90	0.29	0.61
6	Heterogeneous by quartile past att., info. treatment	4.03	0.89	0.31	0.59

Notes: This table replicates Table 9, but assumes that the distribution of cost draws is given by $10 + X$, where X is an exponentially distributed random variable.

Table A16: Estimated welfare effects of piece-rates and commitment contracts

		(1) Avg. Δ in attendance	(2) Δ Agent surplus	(3) Δ Health benefits	(4) Δ Attendance costs	(5) Δ Social Surplus
1	12+ visits contract	1.88	−$2.01	$38.03	$35.64	$2.39
2	Linear incentive, $p = \$2.21$	1.88	$30.33	$39.19	$30.61	$8.58
3	Optimal linear incentive, $p = \$7.34$	5.56	$114.84	$115.99	$100.26	$15.74
4	8+ visits contract	1.15	−$1.06	$22.46	$21.61	$0.85
5	Linear incentive, $p = \$1.36$	1.15	$18.20	$24.27	$18.76	$5.51
6	16+ visits contract	1.76	−$1.53	$39.83	$36.81	$3.02
7	Linear incentive, $p = \$2.12$	1.76	$29.22	$37.37	$29.11	$8.26

Notes: This table replicates Tables 10 and A9, but assumes that the distribution of cost draws is given by $10 + X$, where X is an exponentially distributed random variable.

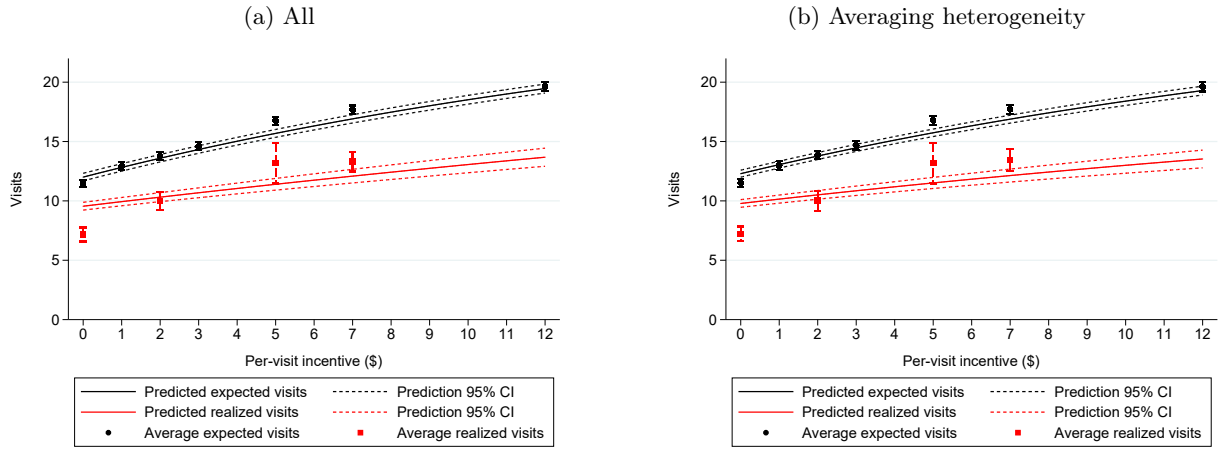
C.7.2 Minimal cost draw of -$5

Table A17: Parameter estimates

		(1) $\hat{\beta}$	(2) $\hat{\tilde{\beta}}$	(3) $\hat{b}$	(4) $(1-\hat{\beta})\cdot\hat{b}$	(5) $(1-\hat{\tilde{\beta}})\cdot\hat{b}$	(6) $\frac{(1-\hat{\tilde{\beta}})}{(1-\hat{\beta})}$
1	All	0.33	0.82	4.57	3.07	0.84	0.28
	(N=1,126)	(0.24, 0.42)	(0.74, 0.89)	(3.56, 5.58)	(2.64, 3.49)	(0.60, 1.09)	(0.19, 0.36)
2	Information control	0.30	0.83	4.96	3.47	0.85	0.25
	(N=560)	(0.20, 0.40)	(0.75, 0.90)	(3.66, 6.25)	(2.86, 4.08)	(0.59, 1.11)	(0.16, 0.33)
3	Enhanced information	0.38	0.77	4.69	2.91	1.10	0.38
	treatment (N=392)	(0.18, 0.58)	(0.57, 0.96)	(2.42, 6.95)	(2.16, 3.66)	(0.57, 1.63)	(0.17, 0.58)
4	Below median past	0.15	0.89	3.88	3.28	0.41	0.13
	attendance (N=550)	(0.03, 0.28)	(0.79, 1.00)	(2.77, 4.99)	(2.60, 3.97)	(0.07, 0.76)	(0.01, 0.24)
5	Above median past	0.52	0.83	6.75	3.27	1.17	0.36
	attendance (N=576)	(0.43, 0.60)	(0.75, 0.90)	(5.33, 8.16)	(2.78, 3.76)	(0.79, 1.56)	(0.25, 0.47)
6	Chose 8+ visit	0.33	0.85	4.54	3.03	0.69	0.23
	contract (N=546)	(0.20, 0.46)	(0.73, 0.97)	(3.17, 5.91)	(2.47, 3.60)	(0.28, 1.10)	(0.08, 0.37)
7	Chose 12+ visit	0.28	0.81	4.96	3.57	0.94	0.26
	contract (N=556)	(0.15, 0.41)	(0.69, 0.93)	(3.46, 6.46)	(2.93, 4.21)	(0.51, 1.37)	(0.13, 0.39)
8	Chose 16+ visit	0.26	0.74	5.28	3.93	1.36	0.34
	contract (N=275)	(0.04, 0.47)	(0.52, 0.97)	(2.46, 8.11)	(2.77, 5.09)	(0.66, 2.05)	(0.13, 0.56)
9	Averaging heterogeneity	0.37	0.85	5.66	3.34	0.81	0.27
	(N=952)	(0.29, 0.44)	(0.79, 0.91)	(4.68, 6.64)	(2.88, 3.79)	(0.54, 1.09)	(0.19, 0.35)

Notes: This table replicates Table 8, but assumes that the distribution of cost draws is given by $-5 + X$, where X is an exponentially distributed random variable.

Figure A4: Predicted and observed actual and expected attendance by incentive



Notes: This figure replicates Figure 11, but assumes that the distribution of cost draws is given by $-5 + X$, where X is an exponentially distributed random variable.

Table A18: Estimated impact of 12+ contract on attendance

		(1) Δ in att.	(2) Pr(att. ≥ 12) with contract	(3) Pr(att. ≥ 12) without contract	(4) Δ in Pr(att. ≥ 12)
1	Empirical	3.51 (1.38, 5.65)	0.65 (0.52, 0.78)	0.22 (0.10, 0.35)	0.42 (0.26, 0.58)
2	Homogeneous	1.57	0.78	0.33	0.45
3	Heterogeneous by median past att., info. treatment	0.64	0.58	0.41	0.17
4	Heterogeneous by median past att.	0.63	0.58	0.41	0.17
5	Heterogeneous by quartile past att.	0.69	0.57	0.39	0.18
6	Heterogeneous by quartile past att., info. treatment	0.70	0.59	0.39	0.19

Notes: This table replicates Table 9, but assumes that the distribution of cost draws is given by $-5 + X$, where X is an exponentially distributed random variable.

Table A19: Estimated welfare effects of piece-rates and commitment contracts

		(1) Avg. Δ in attendance	(2) Δ Agent surplus	(3) Δ Health benefits	(4) Δ Attendance costs	(5) Δ Social Surplus
1	12+ visits contract	0.32	−$16.27	$1.85	$1.65	$0.20
2	Linear incentive, $p = \$0.86$	0.32	$9.51	$1.94	$1.08	$0.86
3	Optimal linear incentive, $p = \$6.12$	2.09	$75.70	$12.73	$9.60	$3.12
4	8+ visits contract	0.19	−$10.25	$0.91	$0.66	$0.25
5	Linear incentive, $p = \$0.55$	0.19	$6.02	$1.28	$0.71	$0.58
6	16+ visits contract	0.50	−$11.49	$3.75	$4.19	−$0.44
7	Linear incentive, $p = \$1.41$	0.50	$15.88	$3.10	$1.82	$1.28

Notes: This table replicates Tables 10 and A9, but assumes that the distribution of cost draws is given by $-5 + X$, where X is an exponentially distributed random variable.

C.8 Dollar value of exercise from public health estimates

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

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

$$0.112 \times (\$50,000)/(2.5 \times 104) = \$21.5$$

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[47]

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

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

$$48.1/100,000 \times 9,000,000 = \$4,329$$

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

$$\$4,329/(0.75 \times 365) = \$15.81$$

Alternatively, we could assume that a QALY is worth \$50,000, use life tables to calculate the probability of survival to each age beyond 34, and calculate the present discounted value (PDV) of life remaining. Using a discount rate of 2%, we calculate \$1,431,000 for men and \$1,519,000 for women. Performing similar calculations to the ones above for men and women and then taking the weighted average based on the fraction of each gender in the sample, we obtain \$2.60 per hour of exercise. Since part of the reason for discounting is to take account of the decreasing probability of survival at higher ages, it may be appropriate to apply an even lower discount rate. If we assume a discount rate of 0% so that the decrease in the contribution of QALYs at higher ages is entirely attributable to a decreased probability of survival, the value of life remaining past age 34 increases to \$2,189,000 for men and \$2,390,000 for women, and the value of an hour of exercise increases to \$4.01.

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

D Further study details and instructions

Table A20: 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 A21: Study demographics by wave

	Wave 1	Wave 2	Wave 3	Total
Female	0.66	0.61	0.57	0.61
	(0.47)	(0.49)	(0.50)	(0.49)
Age[a]	30.93	34.55	34.38	33.51
	(12.61)	(15.29)	(15.70)	(14.82)
Student, full-time	0.64	0.54	0.55	0.57
	(0.48)	(0.50)	(0.50)	(0.50)
Working, full- or part-time	0.50	0.60	0.59	0.57
	(0.50)	(0.49)	(0.49)	(0.50)
Married	0.25	0.28	0.27	0.27
	(0.44)	(0.45)	(0.45)	(0.44)
Advanced degree[b]	0.41	0.48	0.47	0.46
	(0.49)	(0.50)	(0.50)	(0.50)
Household income[a]	45,804	58,502	58,527	55,139
	(40,574)	(48,248)	(49,722)	(47,121)
Visits in the past 4 weeks, recorded	7.04	7.63	5.89	6.91
	(5.86)	(6.12)	(5.36)	(5.86)
N	340	509	399	1,248

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 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.

D.1 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 A5.

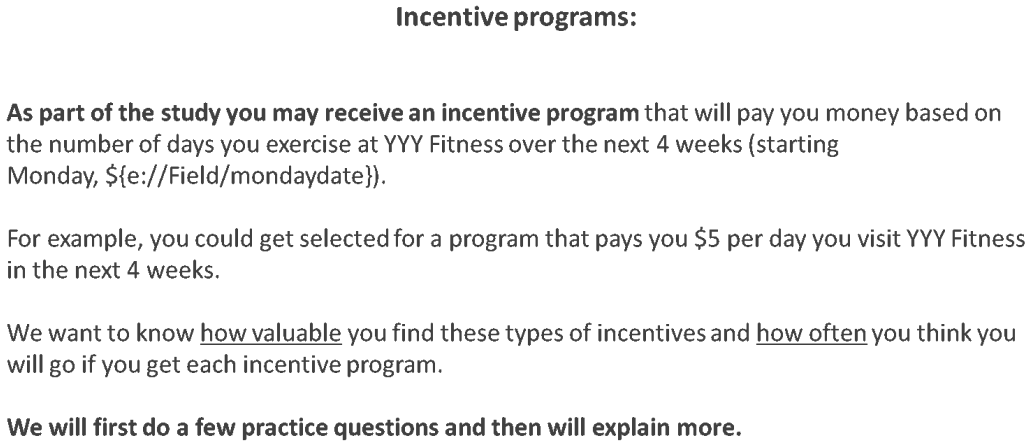


Figure A5: Introduction to willingness to pay section of the study

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 A6.

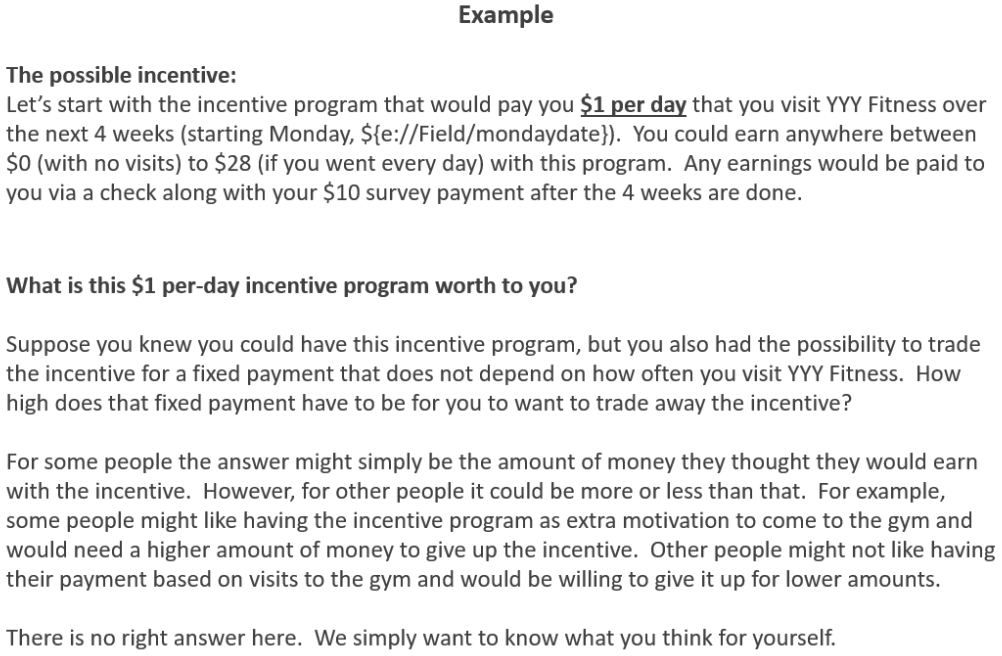


Figure A6: Explanation of willingness to pay for $1 incentive program

Since participants may not have been familiar with the idea of willingness to pay, we presented them with rows 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 A7 and A8.

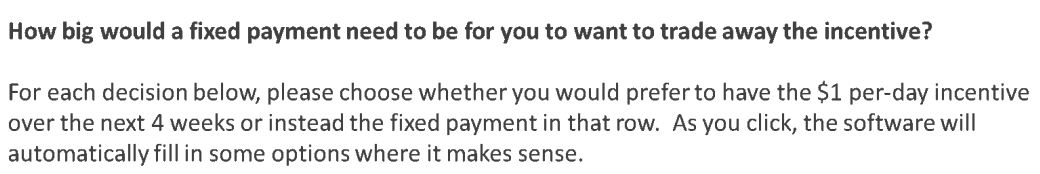


Figure A7: Instructions for decision table

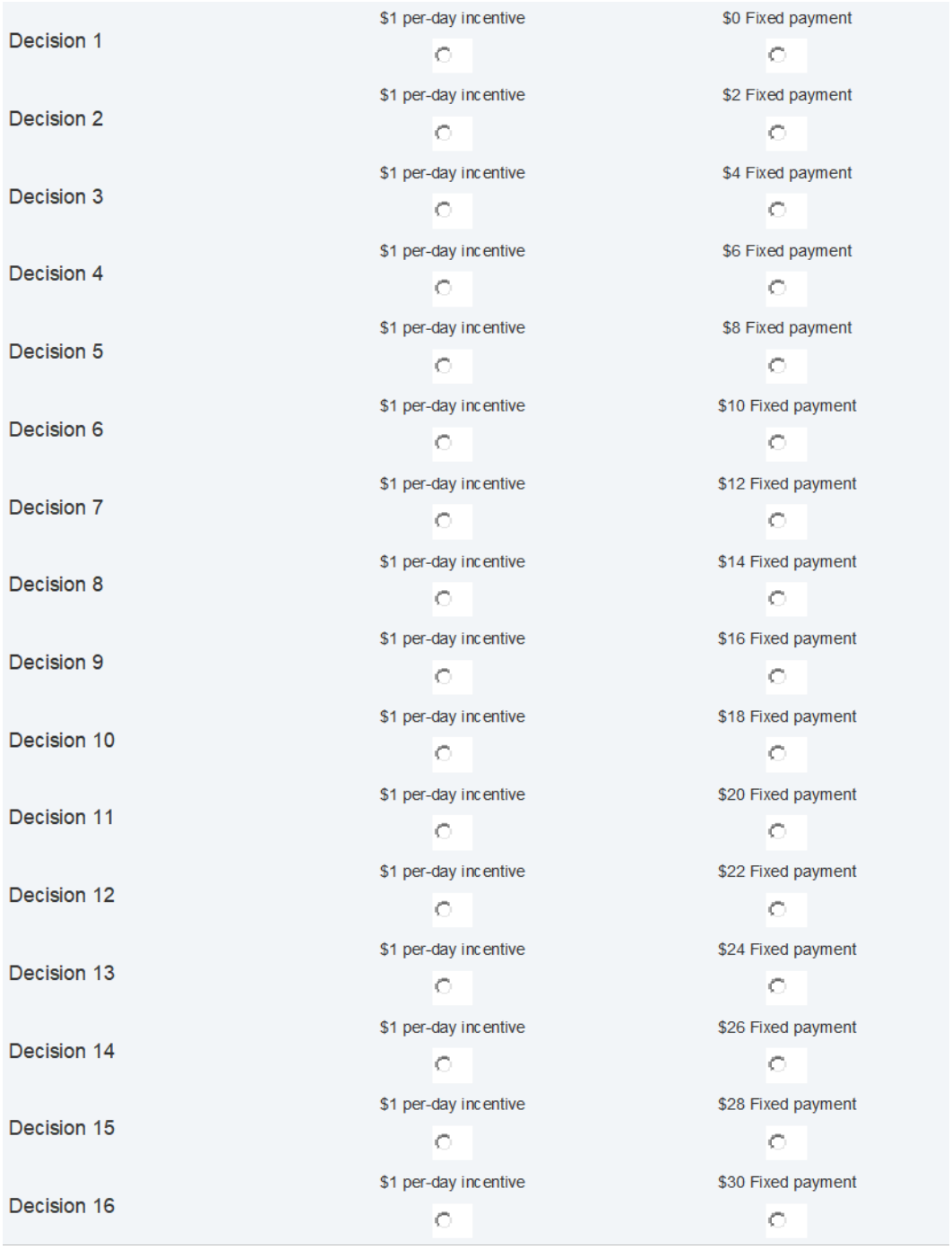


Figure A8: Decision table

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 A9 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.

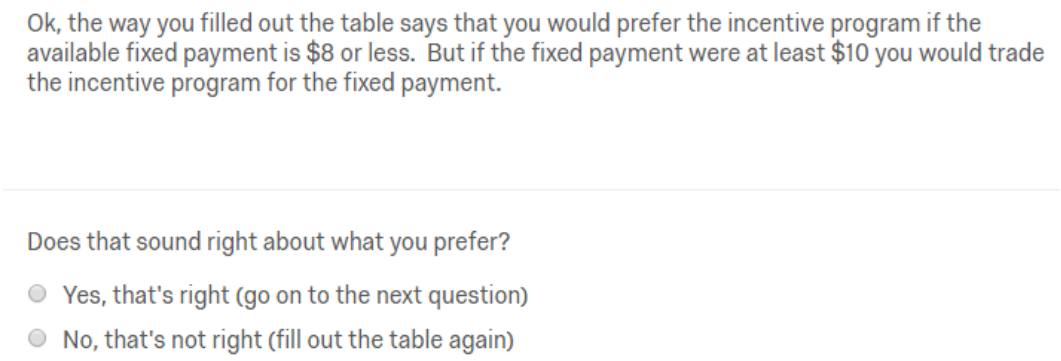


Figure A9: Comprehension check for table

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 A10.

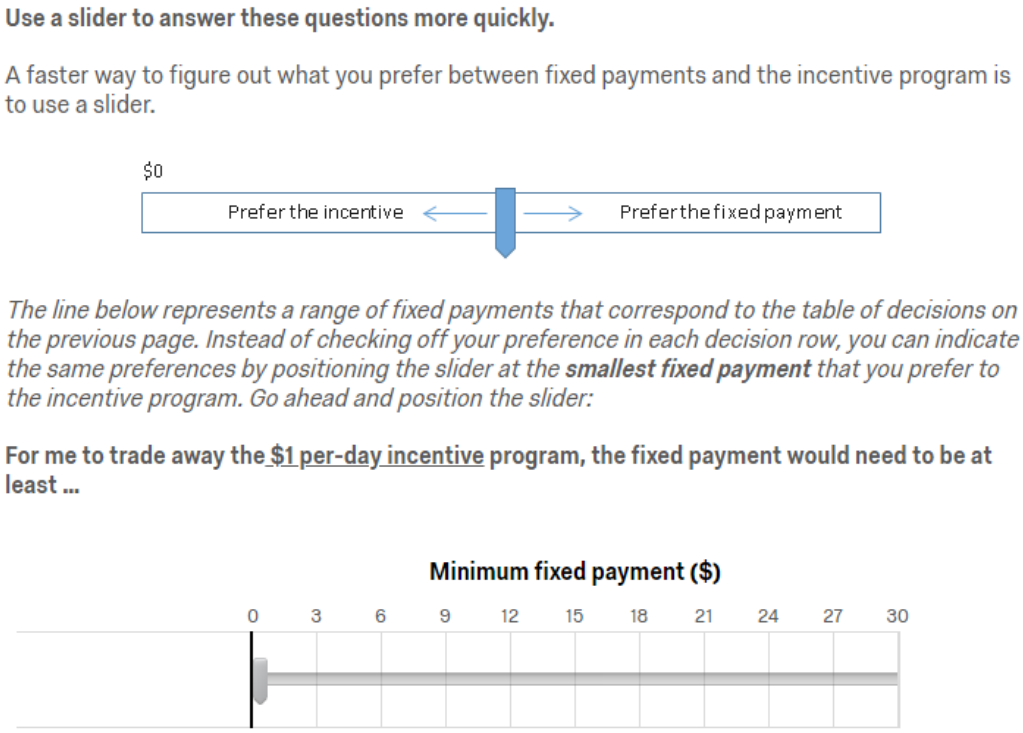


Figure A10: Slider for WTP for $1 per day incentive program

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 A11.

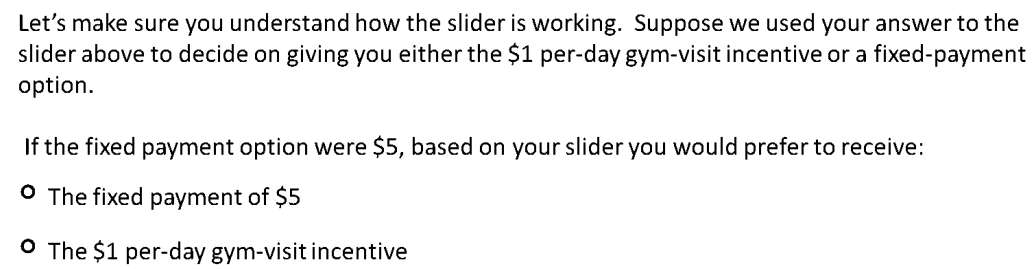


Figure A11: Comprehension check for slider

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 A12) and given the chance to try again.

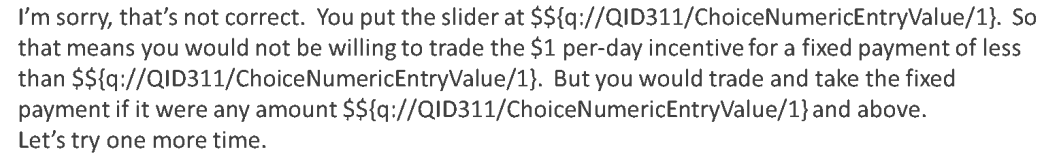


Figure A12: Explanation of incorrect answer

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 A13. After being presented with some answers to frequently asked questions (see Figure A14), 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 A15 and A16. 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 A13: Explanation of incentive program selection

Frequently asked questions:

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

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

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

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

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

Figure A14: 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 A15: Instructions for incentive program questions

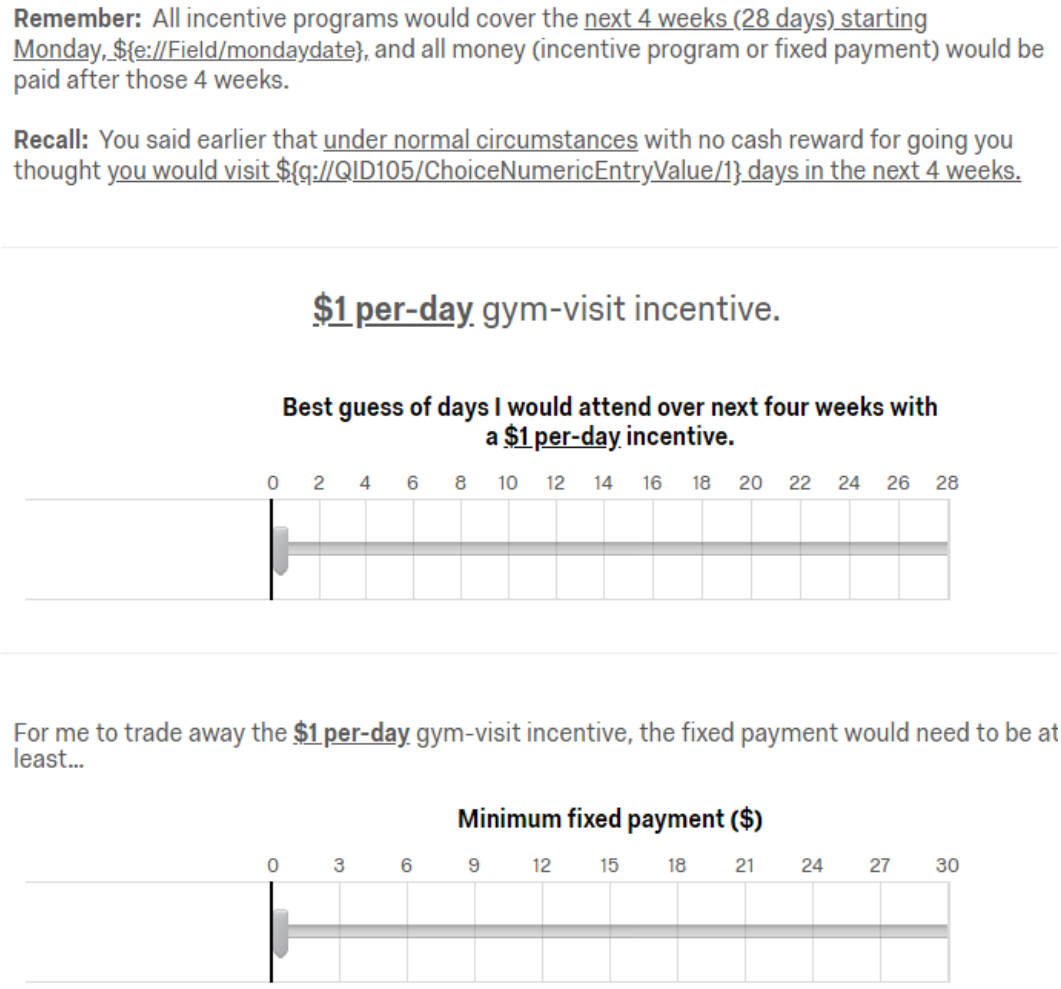


Figure A16: Page for $1 per day gym visit incentive

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 A17 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.

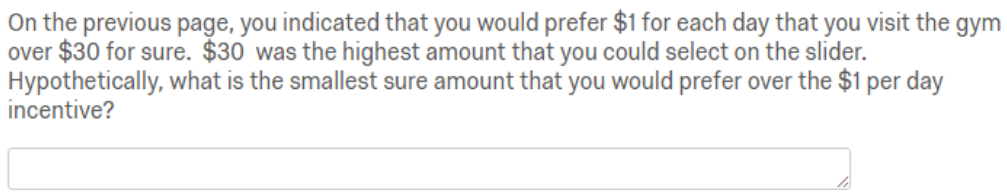


Figure A17: Fill-in-the-blank for off-slider WTP

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 A18 is for a participant whose choices revealed that she would prefer the fixed payment that was drawn to the incentive program that was drawn.

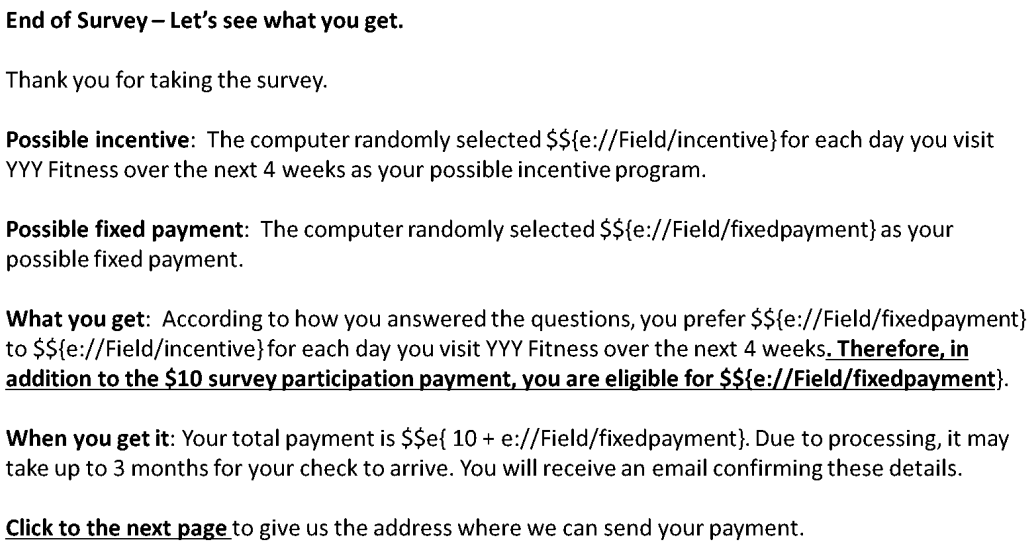


Figure A18: End of online component announcement of fixed payment or incentive program