Mechanism Design for Personalized Policy: A Field Experiment Incentivizing Exercise

Rebecca Dizon-Ross
University of Chicago

Ariel Zucker*
UC Santa Cruz

June 3, 2023

Abstract

Personalizing policies can theoretically increase their effectiveness. However, personalization is difficult when individual types are unobservable and the preferences of policymakers and individuals are not aligned, which could cause individuals to misreport their type. Mechanism design offers a strategy to overcome this issue: offer a menu of policy choices, and make it incentive-compatible for participants to choose the "right" variant. Using a field experiment that personalized incentives for exercise among 6,800 adults with diabetes and hypertension in urban India, we show that personalizing with an incentive-compatible choice menu substantially improves program performance, increasing the treatment effect of incentives on exercise by 80% without increasing program costs relative to a one-size-fits-all benchmark. Personalizing with mechanism design also performs well relative to another potential strategy for personalization: assigning policy variants based on observables.

*Dizon-Ross: University of Chicago Booth School of Business, rdr@chicagobooth.edu. Zucker: University of California, Santa Cruz, arzucker@ucsc.edu. This study was funded by the Chicago Booth School of Business, the Tata Center for Development, and the Chicago India Trust. The study protocols received approval from the IRBs of Chicago, UC Berkeley, and the Institute for Financial Management and Research (IFMR). The experiment was registered on the AEA RCT Registry. We thank Rupasree Srikumar and Srinish Muthukrishnan for leading the fieldwork, and Katherine Daehler, Emily Zhang, Varun Satish, and Ruoyu Chen for outstanding research assistance. We are grateful to Marianne Bertrand, Esther Duflo, Seema Jayachandran, and Emir Kamenica for sustained guidance and to Abhijit Banerjee, Gharad Bryan, Sydnee Caldwell, Josh Dean, Pascaline Dupas, Meredith Fowlie, Maitreesh Ghatak, Ben Golub, David Levine, Jeremy Magruder, Aprajit Mahajan, Ted Miguel, Gautam Rao, Heather Royer, Frank Schilbach, Lars Stole, and Daniel Waldinger for helpful conversations and feedback and to numerous seminar and conference participants for insightful discussions. All errors are our own.

1 Introduction

Personalizing policy is a promising approach to increase policy effectiveness. People's responses to policies can vary widely, and so tailoring a policy to fit an individual's characteristics should be more successful than adopting a uniform, one-size-fits-all approach. However, personalization can be challenging if the policymaker cannot observe each individual's type. This is especially true if the individual's preferences diverge from the policymaker's, which could give the individual an incentive to misreport their type. This paper uses a field experiment to test the use of mechanism design to overcome this principal-agent problem and effectively personalize policy.

We consider a policy that uses financial incentives to influence behavior. Such policies are increasingly common in domains such as education (e.g., Barrera-Osorio et al., 2011), savings (e.g., Gertler et al., 2019), the environment (e.g., Jayachandran et al., 2017), and preventive health (e.g., Carrera et al., 2020; Jones et al., 2019). A typical policy might set a specific behavioral target and offer a payment to those who achieve it. For instance, a workplace wellness program might pay participants for completing a set number of health activities. The ideal target for each individual may vary based on their current lifestyle. A low target might be most effective for people with unhealthy lifestyles ("low types") but may be inframarginal for those with healthy lifestyles ("high types"). Hence, to maximize the impact of the policy given its budget, the policymaker might wish to personalize the target, assigning a higher target to high types. However, with a fixed payment amount, lower targets are more generous, offering the same reward but requiring less of the participant. As a result, all participants may want the lower target, inducing high types to misreport. Similar issues arise in conditional cash transfer programs that provide incentives for hitting attendance targets, or retirement savings programs that match savings that exceed a target amount.

Mechanism design offers a solution to this issue: design a menu of contracts for participants to choose from and make it "incentive-compatible" for them to choose the contract that the policymaker wants them to. To do so, the policymaker can give high types an incentive to choose higher targets by offering a higher payment level for the high target (e.g., Maskin and Riley, 1984). This way, high types will find it in their best interest to *choose* the high target, while low types, who have a higher marginal cost of meeting the high relative to the low target, will opt for the low target. This strategy is analogous to second-degree price discrimination, whereby firms make it incentive-compatible for customers with a high willingness-to-pay to choose a more expensive product by degrading the quality of the less expensive product (e.g., Mussa and Rosen, 1978). Decreasing the payment associated with the low target to dissuade high types from choosing it is similar to decreasing the quality of the less expensive product in the standard firms-and-customers example.

Although a large theory literature suggests that personalizing incentives and other policies through mechanism design could enhance their effectiveness relative to a one-size-fits-all approach (see Varian 1989 for a summary), there is limited empirical evidence supporting this claim. This paper aims to address this gap.

Our experiment personalizes a policy that encourages exercise. The goal of this type of policy is to reduce the impact of chronic lifestyle diseases such as diabetes and hypertension. These diseases are exploding policy problems worldwide, causing significant mortality, morbidity, and lost productivity (World Health Organization, 2022). Lack of physical activity is a major contributor to these conditions (Myers, 2008; Warburton et al., 2006). Promoting exercise and healthy lifestyles is widely recognized as crucial to addressing the health and economic consequences of these diseases (World Health Organization, 2009). Motivated by the negative externalities of physical inactivity and poor lifestyle, policymakers and insurers worldwide are increasingly offering incentives for exercise and other healthy behaviors (e.g., Baicker et al., 2010; Mitchell et al., 2020).

The specific program that we attempt to improve through personalization offers pedometers and incentives for meeting daily step targets to individuals with diabetes, hypertension, and their precursors in urban India. The program is promising in non-personalized form: Aggarwal, Dizon-Ross, and Zucker (2020) finds that providing incentives for walking 10,000 steps daily to diabetics and prediabetics in India substantially increases exercise and decreases health risk. However, the program has the potential to be improved with personalization, as more than half of the program payments are for inframarginal behavior. Personalizing the step target by giving higher targets to higher walkers could greatly improve the cost-effectiveness of the program, that is, the exercise and health gains achieved relative to the payout.

We personalize the program by allowing some participants to choose their incentive contracts from an incentive-compatible menu where contracts with lower step targets offer lower payments. Our experiment randomly assigns participants either to this treatment group, which we call the Choice group; one of three Fixed groups that each received a uniform (not personalized) step target; or a Monitoring group that received a pedometer but no incentives. Our design also includes several supplementary treatment groups that allow us to explore mechanisms and benchmark the effect of Choice against personalization based on observables (an analog of third-degree price discrimination).

Our headline result is that Choice almost doubles the effectiveness of the incentive policy relative to a uniform, intermediate step target that serves as our pre-specified "one-size-fits-all" benchmark. While one-size-fits-all (Fixed) incentives increase walking by approximately 5 minutes per day relative to monitoring with a pedometer alone, the Choice treatment

increases walking by roughly 4 additional minutes per day, an improvement of 80%. Importantly, the Choice treatment achieves this increase in walking without an increase in payments. Moreover, Choice yields gains across the full distribution of walking — in fact, Choice first-order stochastically dominates each of the three Fixed (non-personalized) contracts, which differ in whether the step target was low, intermediate, or high. The Fixed contract with a low target pushes up the bottom of the distribution of walkers but does not perform well at the top. The high Fixed target does the opposite. Choice achieves the gains of the low target at the bottom of the distribution, and of the high target at the top, without the downside of "neglecting" one part of the distribution.

Our second set of results shows that, consistent with a standard mechanism design model, the Choice menu is effective because participants sort into contracts in a way that is advantageous to the principal. Specifically, we show theoretically and empirically that a principal would rather assign higher step targets to participants who walk more in the absence of incentives (i.e., who have higher "baseline steps"), and lower targets to those who walk less, as higher step targets generate relatively more steps (but not more payments) from participants with higher baseline steps. We find that participants' choices align with the principal's preference. While only 10% of participants in the lowest decile of baseline steps choose the highest step target on the Choice menu, over 60% of participants in the highest decile do so.

One interesting question is whether those with higher baseline steps choose higher targets because those targets have higher payment levels, as in a standard economic model, or because they have non-standard preferences (e.g., a time-inconsistent demand for commitment) that cause them to prefer higher targets. Our data indicate that some participants do have non-standard preferences that may have contributed to Choice's success. However, we show that the incentive-compatibility of the Choice menu — i.e., the fact that it provided higher incentives for higher targets — was nevertheless crucial for its strong performance.

Our final set of results benchmarks Choice against personalization based on observable characteristics, or tags. Two challenges with this approach are that, first, participants have incentives to manipulate their observable characteristics to access the most generous policy variant (Björkegren et al., 2020), and second, many of the variables that are most predictive of types are not available in the datasets policymakers have access to (Bryan et al., 2021). We first compare Choice with a strategy designed to overcome these challenges: personalization based on hard-to-manipulate observables that health policymakers are likely to have access to, such as gender, age, and health measurements. We find that Choice significantly outperforms this strategy, which is ineffective because none of these observable characteristics have sufficient explanatory power over steps. We then compare Choice with an "optimal tag" which uses machine learning based on *all* available baseline variables, including baseline

steps, to identify the best step target for each individual. While this optimal tag is likely not achievable in practice,[1] it provides a useful "best case" benchmark. Notably, we find that Choice performs similarly to this optimal tag, with the clear advantage that Choice is easily implementable in practice.

Taken together, our results demonstrate the effectiveness of personalizing policy using mechanism design. A large theoretical literature outlines the advantages of using choice menus for personalization, and our work shows it is possible to deliver on that promise to improve policy. Similar choice-based strategies could be helpful in a broad range of policy domains, from unemployment insurance to the promotion of eco-friendly technologies.

Our work builds on the literature outlining the theory of screening contracts and, in particular, second-degree price discrimination (e.g., Maskin and Riley, 1984; Mussa and Rosen, 1978; Stiglitz, 1977). Indeed, the seminal Maskin and Riley (1984) model of quantity-based second-degree price discrimination describes our policy problem nearly exactly. While the paper describes its model in terms of a firm choosing the optimal menu of quantity-based pricing, it also discusses how the model can be interpreted as a firm choosing the optimal menu of incentive contracts to pay workers of differing ability. While many papers have investigated the effectiveness of second-degree price discrimination for firms selling goods (e.g., Leslie, 2004; Mortimer, 2007), evidence on whether this strategy — or screening contracts more broadly — work in other contexts is limited. Moreover, the existing papers almost all use observational data and structural methods for identification.[2] Our contribution is to provide experimental evidence on the power of screening contracts and second-degree price discrimination, showing that they can be used to personalize incentives, and demonstrating the channels for their effectiveness.

We also tie to several other related literatures on choice/self-selection and on targeting on observables. First, a literature examines whether allowing participants to choose financially-dominated commitment contracts — which a rational agent would never choose — increases effort (e.g., Ashraf et al., 2006; Bai et al., 2020; Huang and Linnemayr, 2019). These papers assess whether agents with self-control problems will sort in a way that benefits their own long-run objectives and find mixed results. In contrast, we examine whether the principal can design a menu that provides the financial incentives for even rational agents to sort in a way that benefits the principal and find positive results.

[1] It may not be achievable both because some of these variables are likely unavailable to policymakers and because of the potential for manipulation. That said, interestingly, we show that there is limited manipulation of observables in our experiment when we use observables to assign step targets. However, it is an open question whether limited manipulation would hold in a scaled-up version of the program.

[2] Levitt et al. (2016) provide the lone experimental test of second-degree price discrimination, for an online gaming firm selling in-game content, such as gold bars that help customers advance in the game. They find no effect on profits, most notably because the menu they test was not designed well given their customer base's demand elasticities.

Second, two papers, Adjerid et al. (2022) and Woerner et al. (2021), both test the impact of allowing participants to choose from a menu of non-dominated incentive schemes. Two key distinctions from our work are that both deviate from the simple price discrimination framework of Maskin and Riley (1984), and both find negative impacts.[3]

Third, a large literature considers targeting or selection at the *extensive* margin — that is, who gets the program. One strand examines targeting based on self-selection (e.g., Alatas et al., 2016; Beaman et al., 2014; Deshpande and Li, 2019; Finkelstein and Notowidigdo, 2019; Ito et al., 2021; Jack, 2013),[4] while another examines targeting on observable characteristics (Burlig et al., 2020; Conner et al., 2022; Kitagawa and Tetenov, 2018). In contrast, we focus on targeting on the intensive margin — that is, who gets *what* program. This focus changes the strategies the policymaker should use, making choice menus (the analog of self-selection) and tagging (the analog of targeting on observables) the appropriate toolkits. Finally, we relate to a literature studying the use of *third-degree* price discrimination strategies (i.e., personalizing prices based on observables), such as Johnson and Lipscomb (2017) for subsidies for sanitation services and Dubé and Misra (2023) for ZipRecruiter services.

2 Physical Activity and Non-Communicable Diseases

Noncommunicable diseases, which account for 74% of global deaths, disproportionately impact low- and middle-income countries (World Health Organization, 2022). In India, both hypertension and diabetes have reached epidemic levels: it is estimated that nearly 1 in 10 adults had diabetes and 1 in 4 had hypertension in 2019, with similar numbers at high risk of developing these diseases (Gupta and Ram, 2019; International Diabetes Federation, 2019).

There is widespread agreement that increasing physical activity is a critical tool in the prevention and management of diabetes and hypertension (World Health Organization, 2013). A low level of physical activity is not only a risk factor for developing both diseases in India (Bhansali et al., 2015; Kumar et al., 2022), but also leads to more rapid development of costly complications such as cardiovascular disease, stroke, and blindness (Tandon et al., 2018), and

[3]Adjerid et al. (2022) allows participants to choose between traditional incentives that pay for success and "gain-loss" incentives that include higher payment for success but penalties for failure. They test a prediction that choice will *decrease* performance due to bad selection; in contrast, we evaluate a choice menu that theory indicates should increase performance. Woerner et al. (2021) allows students to choose between two schemes to incentivize meditation: a linear scheme and a dynamic streak-based scheme that only rewards meditation when completed on three consecutive days. Dynamic schemes are more complex than the simple static targets we used, which could make it more difficult for participants to make good decisions and sort well. Accordingly, the amount of sorting on type appears substantially smaller in the Woerner et al (2022) setting than in ours, which could help explain the lower effectiveness of choice there.

[4]Some papers examine who selects into voluntary programs (e.g., Beaman et al., 2014), including voluntary incentive programs (e.g., Einav et al., 2022; Jack and Jayachandran, 2019). Others examine targeting based on auction bids or willingness to pay (Ashraf et al., 2010; Cohen and Dupas, 2010; Jack, 2013). A third set of papers evaluate how hassle costs change selection (Alatas et al., 2016; Deshpande and Li, 2019; Finkelstein and Notowidigdo, 2019).

is associated with higher mortality (Ben-Sira and Oliveira, 2007). The World Health Organization (2018) estimates that each dollar spent on programs to increase activity in lower and middle income countries is likely to generate $2.80 in cost savings.

Previous research has shown that incentives are a promising approach for increasing physical activity and decreasing the burden of diabetes and hypertension. In closely related work, Aggarwal, Dizon-Ross, and Zucker (2020) evaluates a program which offers incentives for achieving a daily step target that is structured virtually identically to the non-personalized variant of the program we examine in this paper. The program increased physical activity and decreased blood sugar and cardiovascular health risk factors among a sample of diabetics and prediabetics in India. The program was also cost effective, especially when compared to other evidence-based approaches for increasing physical activity among populations with chronic disease, most of which are prohibitively expensive (Howells et al., 2016).

Interestingly, Aggarwal et al. (2020) finds that increasing walking via incentives appears to deliver equivalent health improvements for participants who walk more at baseline and those who walk less (see the Online Supplement for details). This aligns with several medical studies finding that the health benefits of additional exercise are near-linear up to 3,000-4,000 MET minutes per week (Kyu et al., 2016; Samitz et al., 2011) — a level of activity rarely seen in urban India (Anjana et al., 2014).[5] Therefore, personalizing incentives to boost exercise throughout the entire spectrum of walking activity seems like a promising strategy to increase the health impact and cost-effectiveness of incentives, while attempting to target based on walking activity (e.g., by screening out higher walkers) seems less so. Moreover, while the present study was unable to measure health impacts due to logistical constraints, the evidence cited above suggests that the size of the health impacts from our incentive program is likely to track with the size of the exercise gains.

3 Theoretical Framework

To fix ideas and motivate our experimental treatments, in this section, we show theoretically how a policymaker can use mechanism design to personalize incentives for greater effectiveness. We apply the classic price discrimination model of Maskin and Riley (1984) to our setting with one small adaptation: while the workers in their analysis only work for payment, the participants in our setting engage in effort (i.e., walk) even without payment. We show that this modification implies that a participant's type maps one-to-one to their level of output (walking) in the absence of incentives.

We then emphasize two classic results that justify our goal of personalization and our personalization strategy. First, if the principal had complete information, they would personalize contracts based on individual participant types, giving higher targets to higher types

[5] An MET minute is the amount of energy expended during a minute while at rest.

(i.e., to those who walk more in the absence of incentives). This solution — the analogue of perfect price discrimination — is efficient. Second, the perfect price discrimination solution is not feasible with incomplete information because of incentive compatibility constraints: high types will want to imitate low types. However, the principal can still implement a second degree price discrimination strategy that outperforms a one-size-fits all approach. This strategy offers participants a choice from an incentive-compatible menu of contracts where contracts with higher step targets pay higher incentives.

We then briefly extend the basic model to allow for non-standard participant preferences (e.g., time inconsistency). We show that many results are robust to this extension and that, conditional on participants' walking costs, non-standard preferences should actually improve the performance of choice menus from the principal's perspective.

Participant Utility For simplicity, we consider steps walked during a single period we call the "contract period." Assume a participant of type θ has the following utility function for walking in the contract period:

$$u(s, y; \theta) = y - c(s; \theta), \tag{1}$$

where y is income, s are steps walked, and $c(s; \theta)$ is the net cost of walking s steps for a type θ person. For simplicity, we assume θ can take on two values, θ^H and θ^L, with $\theta^H > \theta^L$.[6]

We assume that costs are convex in s ($c''(s; \theta) > 0$). To reflect the fact that people walk in the absence of any payments for walking, we assume that the marginal cost of steps, $c'(\theta; s)$, is negative at $s = 0$ (i.e., $c'(0; \theta) < 0$). Finally, we assume that the marginal cost of steps is always strictly *lower* for higher-θ types (i.e., $c'(s; \theta^H) < c'(s; \theta^L)$), ensuring that the single crossing property holds.

Walking without a Contract In the absence of an incentive contract, participants choose steps to minimize net costs:

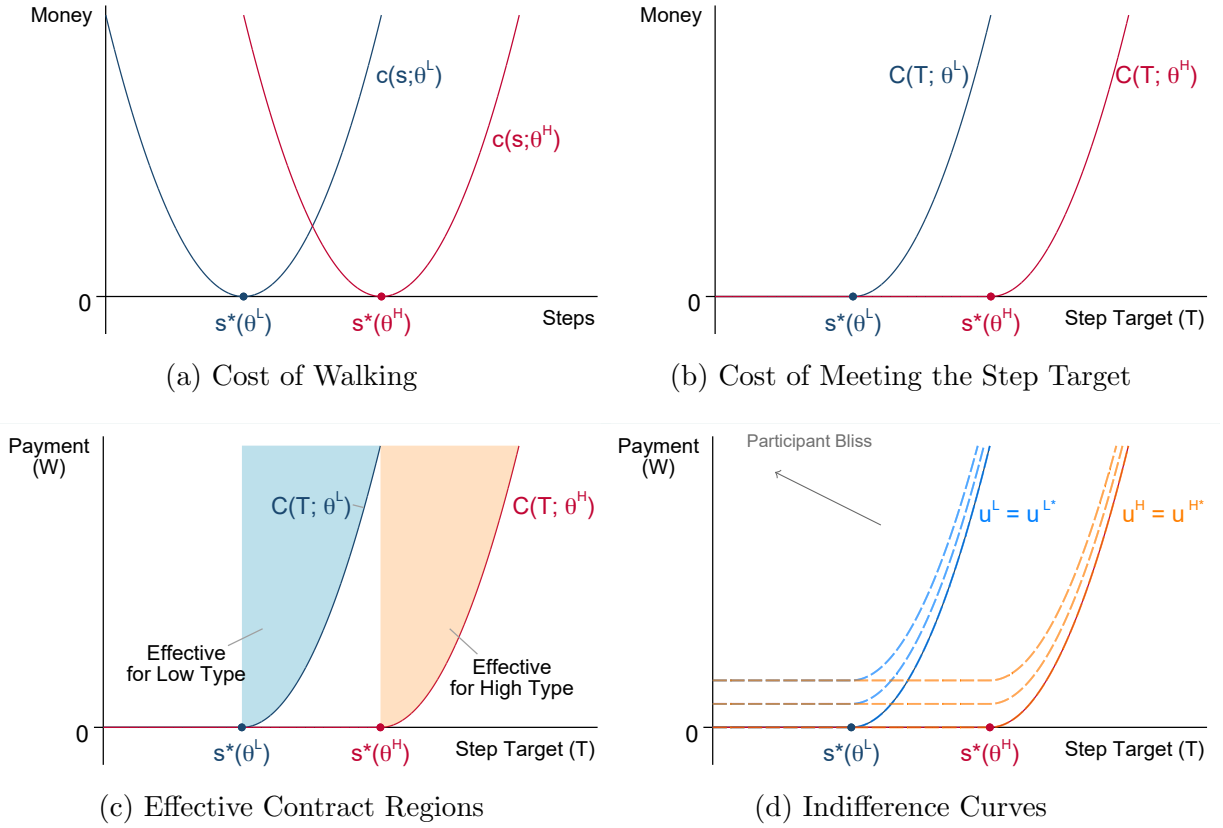
$$s^*(\theta) = \underset{s}{\operatorname{argmin}}\, c(s; \theta) \tag{2}$$

Figure 1a plots the net cost functions and chosen steps for types θ^L and θ^H. We normalize the minimum of each net cost function to be 0. As the figure shows, "high types" walk more than "low types" (i.e., $s^*(\theta^H) > s^*(\theta^L)$).

Walking with a Contract A step target contract consists of a pair of a step target level T and an incentive level W, such that a participant with contract $\langle T, W \rangle$ receives a payment of W if their steps exceed T.

[6]Many papers show that these same general results go through with more types and/or more contracts in a variety of settings. See Spence (1980) and Stole (2001) for the case with n types and n contracts, Maskin and Riley (1984), Tirole (1988), Hermalin (2005), and Varian (1989) for continuous types and continuous contracts, and Bergemann et al. (2012) and Wilson (1989) for continuous types and n contracts.

Figure 1: High Types' Cost and Indifference Curves Are Shifted to the Right of Low Types'



(a) Cost of Walking

(b) Cost of Meeting the Step Target

(c) Effective Contract Regions

(d) Indifference Curves

Notes: Panel (a) displays the cost of walking a certain number of steps, $c(s;\theta^j)$, separately for types θ^L and θ^H. Panel (b) shows the cost of meeting a set step target, $C(T;\theta^j)$, for types θ^L and θ^H. The shaded areas in Panel (c) represent the regions of step target contracts that would increase steps for types θ^L and θ^H. Panel (d) shows the indifference curve of participants of types θ^L and θ^H over step target contracts.

Figure 1b plots the additional walking costs that a participant would incur to meet a step target, T. If the step target is less than $s^*(\theta)$, then the participant will not incur any additional cost to meet the step target as they are already doing so. When $T > s^*(\theta)$, the additional cost of meeting the target is simply $c(T;\theta)$. To formally define the cost of meeting target T, $C(T;\theta)$, let $\hat{T} = \max\{T, s^*(\theta)\}$. Then $C(T;\theta) = c(\hat{T};\theta)$.

Participants will meet the step target if the payment W weakly outweighs the cost of meeting the target $C(T;\theta)$. Chosen steps under contract $\langle T, W\rangle$ are thus:

$$s^*(T, W;\theta) = \begin{cases} \hat{T} & \text{if } C(T,\theta) \leq W \\ s^*(\theta) & \text{if } C(T,\theta) > W \end{cases} \tag{3}$$

As a result, the "effective" contracts (i.e., contracts that increase steps) for a participant of type θ are those with $W \geq C(T;\theta)$, as well as $T \geq s^*(\theta)$, as depicted in Figure 1c.

Participant preferences over contracts depend on the optimized value of participant utility

under the contract. This value function for a step target contract $\langle T, W \rangle$ is:

$$V(T, W; \theta) = \begin{cases} W - C(T; \theta) & \text{if } C(T, \theta) \leq W \\ 0 & \text{if } C(T, \theta) > W \end{cases} \tag{4}$$

Equation (4) implies that participants will be indifferent between two different contracts that have the same value of $W - C(T; \theta)$. Hence, indifference curves can be expressed as $W = C(T; \theta) + K$, for some constant K, as shown in Figure 1d.

The Principal's Objective We assume that the principal derives some benefit from the steps taken by each participant, with the benefits denoted by the function $g(s)$. A natural interpretation of $g(s)$ is the financial externality to a policymaker or firm of increased exercise, e.g., through health care savings. We assume that the benefits function is increasing and weakly concave: $g'(s) > 0$ and $g''(s) \leq 0$. In the special case that the benefits function is linear, it can be fully summarized by average steps.

The principal's objective is to choose a step target T^j and incentive level W^j for each type of participant $j \in \{L, H\}$ in order to maximize the benefits of the additional steps to the principal net of payments to participants:

$$\max_{T^j, W^j} \quad g(\hat{T}^j) - W^j. \tag{5}$$

The principal maximizes this objective subject to constraints discussed below.

We restrict attention to "step function" contracts that reward participants for exceeding a target, which are the most common type of walking incentive contract. Moreover, Appendix B.1 shows that these contracts strictly outperform linear payments (and weakly outperform linear payments after a target) from the principal's perspective.

Full Information As a benchmark for the solution the principal hopes to attain, we begin by assuming that the principal has full information about each participant's value of θ. We then explore the more realistic case where the policymaker does not know θ.

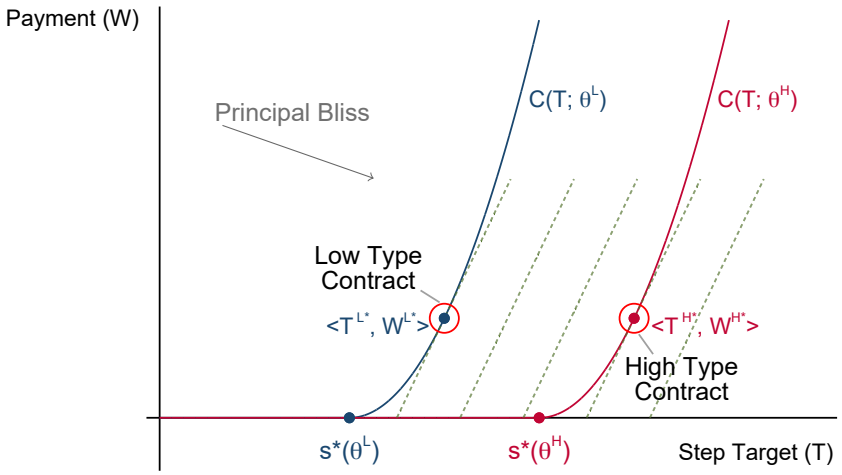
With full information on each participant's θ, the principal maximizes equation (5) subject to a "participation constraint" guaranteeing that type j will achieve T^j steps, i.e., that the contract is effective:

$$W^j \geq C(T^j, \theta^j) \tag{6}$$

We can solve for the principal's optimal contract choice for type j by overlaying the principal's indifference curves on the participation constraints (i.e., cost curves) for each type j, as in Figure 2. Specifically, Equation (5) implies that the principal's indifference curves over contracts $\langle T, W \rangle$ have the same shape as their benefit function, $g(T)$, and hence will

be increasing, weakly concave functions, as shown with the green dashed lines.[7] Given the direction of principal bliss, the participation constraints bind with equality, and so the principal chooses the contract on each cost curve that places them on their rightmost indifference curve. One can see the solutions for each type j indicated in Figure 2 as $\langle T^{j*}, W^{j*} \rangle$. The principal chooses a higher step target for high types than for low types.

Figure 2: The Full Information Solution Assigns a Higher Step Target to High Types



Notes: Figure displays the principal's full information solution: $\langle T^{L*}, W^{L*} \rangle$ for low types and $\langle T^{H*}, W^{H*} \rangle$ for high types. The straight dashed lines represent the principal's indifference curves over contracts.

As with first degree price discrimination, the principal's optimal choice in the full information case is efficient, maximizing the joint surplus of steps to the principal and participant. To see this, substitute equation (6) into equation (5) to see that the principal's maximand coincides with social surplus: $g(T) - c(T, \theta)$. We summarize our discussion as follows:

Result 1. *(Full Information) Let $\langle T^{L*}, W^{L*} \rangle$ and $\langle T^{H*}, W^{H*} \rangle$ be the contracts assigned to low and high types, respectively, by the principal in the full information case. Then $T^{H*} > T^{L*}$: the principal chooses higher step targets for high types than low types.*[8] *Moreover, the principal's chosen contracts are efficient.*

Imperfect Information and Choice Now we assume that the principal has no information about participants' types, instead offering contracts and allowing participants their

[7] Technically, policymaker indifference curves over contracts for type j set $W^j = g(\hat{T}^j) + K$, for some constant K, and hence will be specific to the type j and equal to $W^j = \begin{cases} g(T) - K & \text{if } T > s^*(\theta^j) \\ g(s^*(\theta^j)) - K & \text{if } T \leq s^*(\theta^j). \end{cases}$ For visual simplicity, we display the policymaker indifference curves as $W = g(T) - K$ in the figures (i.e., we do not show the flat part of the curve that varies by type), since the solution for each type j will always have $T^j > s^*(\theta^j)$ and so will always be on the portion of the policymaker indifference cuve where $W^j = g(T) - K$.

[8] The optimal T is characterized by $C'(T, \theta) = g'(T)$. Since $c'(s; \theta^H) < c'(s; \theta^L)$, then $C'(T, \theta) = g'(T)$ will hold at higher values of T when $\theta = \theta^H$ than when $\theta = \theta^L$, given our assumptions on $C(\cdot)$ and $g(\cdot)$.

choice. In this case, the principal's solution must also satisfy incentive compatibility constraints that neither type of participant would prefer the contract designed for the other:

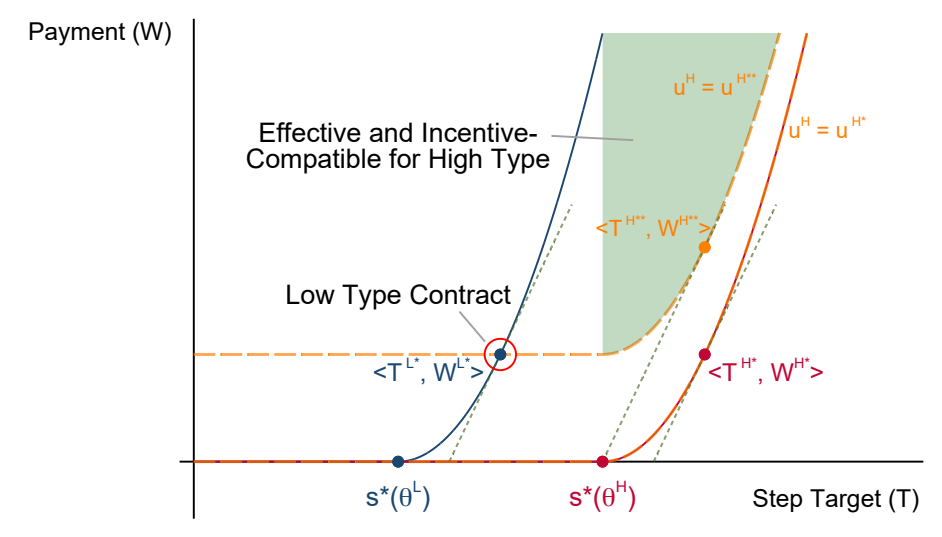
$$V(T^L, W^L; \theta^L) \geq V(T^H, W^H; \theta^L) \quad \text{(IC - L)}$$

$$V(T^H, W^H; \theta^H) \geq V(T^L, W^L; \theta^H) \quad \text{(IC - H)}$$

As is common in mechanism design, only the high type's incentive compatibility constraint (IC - H) will bind; the principal's challenge is that high types want to imitate low types.

Figure 3 shows that the full information solution is not implementable with imperfect information, since it is not incentive-compatible for high types. The contract designed for high types, $\langle T^{H*}, W^{H*} \rangle$, puts high types on the indifference curve labeled $u^H = u^{H*}$. However, the contract designed for low types, $\langle T^{L*}, W^{L*} \rangle$, would give high types higher utility, putting them on the indifference curve labeled $u^H = u^{H**}$.

Figure 3: Incentive-Compatible Menus Pay More to High Types Than Low Types



Notes: Figure shows that the full information solution ($\langle T^{L*}, W^{L*} \rangle$ and $\langle T^{H*}, W^{H*} \rangle$) is not incentive-compatible and shows an example of a contract menu that would be incentive-compatible ($\langle T^{L*}, W^{L*} \rangle$ and $\langle T^{H**}, W^{H**} \rangle$). The curves labeled $u^H = u^{H*}$ and $u^H = u^{H**}$ represent the high types' indifference curves. The shaded region indicates contracts that would both increase the high type's steps and be incentive-compatible for the high type when combined with $\langle T^{L*}, W^{L*} \rangle$.

The intuition for the result that high types prefer the low types' full-information contract is especially clear in the special case, shown in Figure 3, where both contracts have the same payment amount ($W^{L*} = W^{H*}$). In this case, high types would of course prefer to imitate low types to get the same payment for less effort. However, the result holds regardless of the relationship between W^{L*} and W^{H*}, since high types prefer any contract on the low types' cost curve to any contract on their own cost curve, as apparent in Figure 1d.

Instead, to satisfy the incentive compatibility constraint, the principal must choose a contract menu where $W^H > W^L$ (and where $W^H > C(T^H, \theta^H)$). For example, in Figure 3, using the contract $\langle T^{H**}, W^{H**} \rangle$ for high types satisfies the incentive-compatibility constraint when paired with contract $\langle T^{L*}, W^{L*} \rangle$ for low types, since the $\langle T^{H**}, W^{H**} \rangle$ contract gives high types as much utility as the $\langle T^{L*}, W^{L*} \rangle$ contract. This contract menu is both efficient and can outperform a single contract. Interestingly, it does not represent a solution to the principal's problem as formulated in equation (5), although it would solve the problem of a more efficiency-minded principal whose goal is to choose the payment-minimizing contract menu among the set of efficient contract menus (see Appendix B.2 for more detail). Moreover, the choice menu that solves the principal's problem from equation (5) will similarly induce sorting by paying $W^H > W^L$, and will outperform the optimal single contract.[9]

Result 2. *(Imperfect Information: Incentive-Compatible Choice) The principal's full information solution, $\langle T^{L*}, W^{L*} \rangle$ and $\langle T^{H*}, W^{H*} \rangle$, is not implementable as a choice menu. In addition, choice menus where $W^{L*} = W^{H*}$ are also not implementable. In both cases, high types would choose the contract designed for the low types. To induce sorting, the principal must offer a choice menu where $W^H > W^L$. Doing so can induce separation and improve the principal's utility relative to offering a single contract to everyone.*

We hereafter use the term "incentive-compatible choice" to refer to choice menus where $W^H > W^L$, the analog of second-degree price discrimination in our setting.

Non-Standard Utility Functions The above analysis assumes that participants are standard, rational agents. In reality, participants could have behavioral biases or non-standard preferences that may affect the performance of personalization. Appendix B.4 provides a framework that nests several behavioral biases that participants could have, including time-inconsistency and pride from having a higher step target. The framework suggests the main results hold with this adjustment. Specifically, the optimal full information solution still assigns higher step targets to high types than low types (Result 1), and incentive-compatible choice menus that pay $W^H > W^L$ can still outperform a one-size-fits-all approach (Result 2).

The main adjustment to the results is that behavioral biases can allow the principal to implement a wider range of choice menus, thus improving the performance of choice menus for the principal. For example, if the behavioral biases/non-standard preference elements are sufficiently strong, there is potential that the full information solution will be implementable with incomplete information. A choice menu with $W^{L*} = W^{H*}$ may also be implementable.

[9] As in Maskin and Riley (1984), the principal can do better than $\langle T^{H**}, W^{H**} \rangle$ by introducing inefficiency at the bottom, as shown in Figure B.1 and proved in Appendix B.3.

Implications for Experiment Our experiment evaluates an incentive-compatible choice menu relative to a single contract. While we cannot implement the optimal menu since we do not know the shape or distribution of participants' cost curves or preferences, we follow general screening principles by using a menu that (a) has $W^{H*} > W^{L*}$, and (b) appeared *ex ante* likely to separate types, based on piloting. We also assess the extent to which behavioral biases contribute to the incentive-compatible choice menu's success by comparing the degree of separation created by a "flat" choice menu with $W^{H*} = W^{L*}$ relative to our main choice menu that has $W^{H*} > W^{L*}$, as well as the relative performance of assigning participants based on those two respective menus.

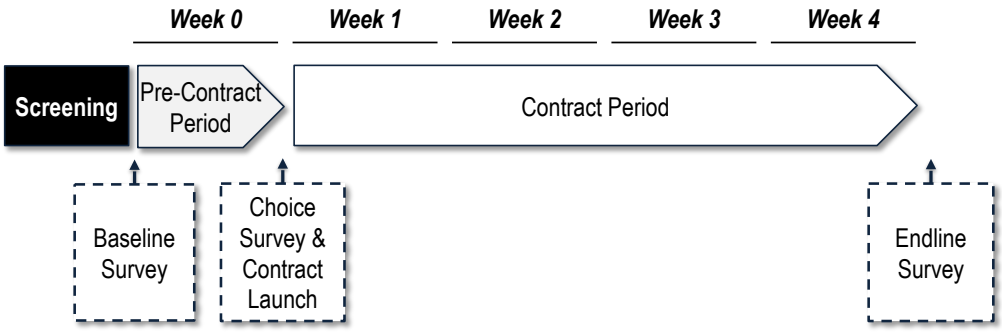
4 Experimental Design

This experiment uses an incentive-compatible choice menu to personalize a program offering incentives for meeting daily step targets. We recruited adults living with, or at high risk for developing, diabetes or hypertension and assigned them to various treatment groups. Each treatment group received an incentive contract specifying their step target and payment level, but the method of assigning contracts varied across groups. Some groups chose their contracts from a menu, some received non-personalized contracts, and some received contracts personalized based on observables, allowing us to identify the effect of choice and the channels for its efficacy.

4.1 Experimental Timeline and Procedures

The experimental timeline for a typical participant is shown in Figure 4. Nearly all participants in the experiment followed this timeline; we document deviations in Section 4.2.

Figure 4: Experimental Timeline for Sample Participant



4.1.1 Screening and Sample Selection

We recruited our sample through a series of public screening camps in the city of Coimbatore in the Indian state of Tamil Nadu. To enroll diverse groups, we held the camps in

locations ranging from markets to religious institutions. During the camps, surveyors took basic anthropometric measurements and conducted a brief eligibility survey. Our eligibility criteria, listed in Appendix C.2, included a diagnosis for diabetes or hypertension, or elevated blood pressure or blood sugar; low risk of injury from regular walking; and the ability to receive payments in the form of mobile top-ups or recharges.

After screening, we contacted eligible individuals by phone, invited them to participate in a program to encourage walking, and scheduled an enrollment visit.[10] Enrollment was conducted on a rolling basis between May 2019 and December 2021.[11]

4.1.2 Baseline Survey and Pre-Contract Period

At the enrollment visit, surveyors verified the screening criteria and conducted a Baseline survey before launching the pre-contract period.

Baseline Survey At the Baseline survey, we asked participants for basic demographic and socioeconomic information. We also verified the screening criteria for eligibility.

Pre-Contract Period Following the Baseline survey, surveyors launched the pre-contract period. This period was designed to measure baseline walking and familiarize participants with study procedures. We gave all participants pedometers for the duration of the study to measure their steps. The step data were collected by syncing the pedometers with a central database. Because syncing requires an internet connection, which most participants did not have, pedometer step data were not available in real time. Instead, we asked participants to report their daily step count to an automated calling system which called them every evening and prompted them to enter the number of steps recorded on their pedometer.

When launching the pre-contract period, surveyors explained to participants that we wanted to measure their steps for six days and instructed them to walk as normal. While there were no financial rewards for achieving step targets in this period, respondents received 50 INR for wearing the pedometer and reporting steps for at least five of the six pre-contract period days.

The pedometer data from these six days, which we refer to as the "baseline step" data, provide a measure of a person's type (θ from Section 3).

[10] Potential enrollees were randomized into treatment groups using list randomization (stratified by median age and gender) as soon as their enrollment visits were scheduled. Surveyors and participants were blinded to treatment group until later (as described in the remainder of Section 4.1).

[11] Our experiment overlapped with two Covid-19 pandemic lockdowns, the first from March 2020 to March 2021 and the second from April to July 2021. We paused recruitment during these lockdowns, and include a control for whether a day was a lockdown day in our analyses.

4.1.3 Choice Survey, Contract Launch, and Contract Period

After the pre-contract period ended, surveyors returned for a second visit with participants.[12] They began the visit by syncing the pre-contract period data from the pedometers and reviewing it with participants. Next, they conducted the Choice survey.

Choice Survey The goal of the Choice survey was to elicit participants' preferences over three contract menus, summarized in Table 1: the Base Menu, Flat Menu, and Steep Menu.

Table 1: Contract Menus

Contract Menu	Payment Levels (INR)		
	Low (10K) Step Target	Med (12K) Step Target	High (14K) Step Target
Steep	10	15	20
Base	16	18	20
Flat	20	20	20

Notes: Figure shows the payment levels used for each contract on the three different contract menus. Each menu contained three contracts, one with a 10,000 step target, one with a 12,000, and one with a 14,000.

The Base Menu was the menu used to assign contracts to our main Choice group. We included the other two menus to examine the sensitivity of choices to payment levels and for use in supplementary treatment groups, as described in Section 4.2.

We solicited menu choices from all participants, not just those who were ultimately assigned one of their choices, to increase power and allow for heterogeneity analysis by target choice. We were able to elicit contract preferences in a "real-stakes" (i.e., not hypothetical) way for all participants since we gathered preferences while participants and surveyors were still blinded to treatment group assignments. Thus, we informed all participants that there was a positive probability that their choices would be implemented.[13] Appendix C.3 contains

[12] We randomized the timing of the second visit to explore the effect of experience with the pedometer on choices, which we examine in the Online Supplement. For a subset of participants, we added a week to the typical six days between the Baseline survey and the second visit, giving these participants an additional week to walk and learn with their pedometers. We control for whether we waited the additional week (a "time between Baseline and Choice surveys" control) when estimating treatment effects. Our results are also robust to excluding those for whom we waited the extra week. Specifically, the effect of Choice relative to the one-size-fits-all benchmark goes from 414 steps, p-value<0.05, in our main specification, to 507 steps, p-value<0.05, if we exclude those for whom we waited the extra week. Regardless of second visit timing, we calculate baseline steps using the first six days following the Baseline survey.

[13] This was true for both the Base and Flat Menus because we had treatment groups which received their choices on those menus, as described in Section 4.2. To make it true for the Steep Menu, we assigned a very small portion of the sample, 35 people total, to receive their Steep Menu choices. This group is not of sufficient size to examine treatment effects, and so we exclude them from all analyses.

more details about the instructions and order of elicitation.

Because of the importance of the Base Menu, most participants made choices on the Base Menu first; however, to examine order effects, we randomized whether the Flat Menu or Base Menu was first for a short period of time.

Contract Launch Immediately after the Choice survey, surveyors told participants their treatment group assignments and the details on how their contract was assigned (e.g., by choice or lottery). Surveyors then walked participants through the details of their incentive contract, including their step target and payment level.

Contract Period The contract period lasted four weeks. During this period, all incentive groups received payments if they reported achieving their daily step target through the automated step-reporting system. We delivered incentive payments as mobile recharges (credits to the participant's mobile phone account). Incentives were delivered at a weekly frequency, along with weekly text messages summarizing walking behavior and total payments. Immediately after reporting steps, participants also received text-messages confirming their step report and payment earned, and congratulating them if they had met their target.

To encourage participants to wear their pedometers and accurately report their steps, we paid a 100 INR bonus if participants wore their pedometers and accurately reported steps on 80% of the contract period days, and an additional 100 INR if they did so on all 28 days. We also conducted a number of audits, both random and targeted, and suspended participants who repeatedly misreported achieving their step target.[14]

At the end of the four-week contract period, surveyors returned to conduct an Endline survey, sync the pedometers, and pay the bonuses for accurate reporting and pedometer wearing.

4.2 Treatment Groups

This section describes the treatment groups, as shown in Figure 5.

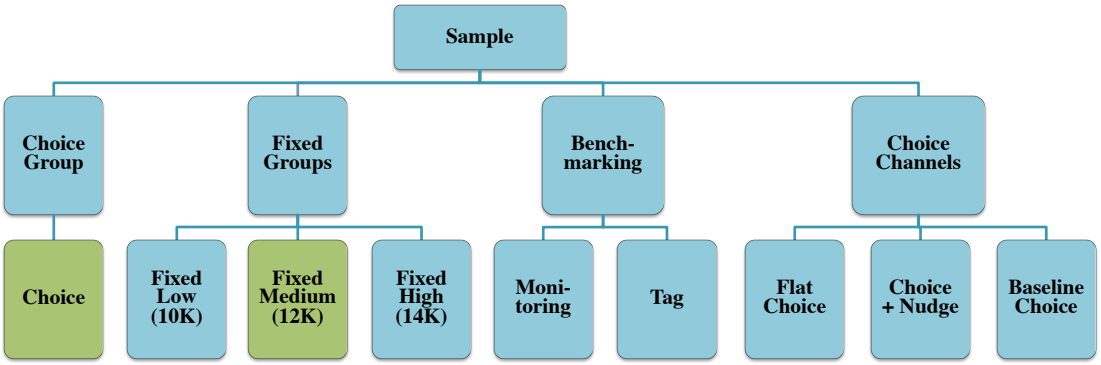
4.2.1 Main Treatment Groups: the Choice and Fixed Medium Groups

Our two primary treatment groups — the Choice and Fixed Medium groups — allow us to estimate the effect of personalization using choice relative to a non-personalized approach.

To identify the full potential of personalization, we would ideally compare the optimal choice menu with the optimal single or "one-size-fits-all" contract (where "optimal" means maximizing the principal's benefits net of payments, as in Section 3). Designing optimal

[14]We targeted audits at participants whose step reporting appeared suspicious and temporarily suspended those who were found to be over-reporting steps. We then re-audited those with temporary suspensions and permanently terminated their contracts if they were found to be over-reporting a second time.

Figure 5: Experimental Design



Payment Amount (INR)									
Low (10K) Target	16	20	-	-	0	20	20	16	16
Med (12K) Target	18	-	20	-	0	20	20	18	18
High (14K) Target	20	-	-	20	0	20	20	20	20
When Assigned	Contract Launch	Contract Launch	Contract Launch	Contract Launch	Contract Launch	Baseline	Contract Launch	Contract Launch	Baseline
Nudge Share	None	Some	Some	Some	Some	None	None	All	None
Sample Size	970	826	1,274	847	207	990	509	540	719

Notes: This figure compares the different treatment groups. "Payment Amount" shows the incentive paid for compliance with each step target in each treatment. "Nudge Share" indicates what share of the treatment group received a nudge towards a certain contract when making choices during the Choice survey.

contract menus requires information we did not have when designing our study (e.g., the cost curves for each type of participant). We approximated this strategy as best we could using data from the Aggarwal et al. (2020) evaluation and a simple model. Appendix C.4 describes the design process in more detail.

Fixed Medium (12K) or "One-Size-Fits-All" Group We first used the existing data and model to create a one-size-fits-all contract. This contract uses the step target that we estimated would maximize average steps across the sample for a 20 INR payment rate (about 0.33 USD, the same payment rate used in Aggarwal et al. 2020).

All participants in our Fixed Medium group were assigned a contract paying 20 INR for each day of compliance with a 12,000 step target.

Choice Group We used the same model and existing data to estimate the three step targets that would each maximize steps for a tercile of our sample (with terciles defined by baseline steps) for the same 20 INR payment rate. The middle tercile's target was the same as the one-size-fits all target (12,000 steps per day); the bottom and top terciles' were 10,000 and 14,000 steps per day, respectively. To construct an incentive-compatible menu that used these three step targets, we conducted a small pilot study to choose payment levels for each target that were near 20 INR and induced separation by baseline walking. This process

yielded the Base Menu shown in Table 1.

All participants in our *Choice* group were assigned a contract according to their choice from the Base Menu.

4.2.2 Other Fixed Groups

While the Fixed Medium group represents our primary pre-specified comparison group for Choice (Dizon-Ross and Zucker, 2020), it is useful to compare Choice to other non-personalized benchmarks. To facilitate these comparisons, we include two additional Fixed groups in the design which, together with the Fixed Medium group, receive the three contracts that our model suggested would each maximize steps for a tercile of the sample at the 20 INR payment rate:

Fixed Low (10K) Group All participants in our Fixed Low group were assigned a contract paying 20 INR for each day of compliance with a 10,000 step target.

Fixed High (14K) Group All participants in our Fixed High group were assigned a contract paying 20 INR for each day of compliance with a 14,000 step target.

4.2.3 Benchmarking Treatment Groups

We include two treatments in our design that allow us to benchmark the effect of Choice against other treatment effects.

Monitoring Group This group received pedometers but no incentives, allowing us to establish the treatment effect of non-personalized incentives relative to a no-incentive control. The group was treated identically to the incentivized groups save for not receiving incentives. For example, Monitoring participants were verbally encouraged to meet a step target.[15] When other groups received congratulatory texts that confirmed payment upon reaching their targets, this group also received congratulatory texts, with no mention of payments.

Tag Group We also benchmark the performance of Choice against personalization based on observables. Specifically, we assign participants in our Tag group to one of three contracts based on their measured baseline steps during the pre-contract period, with the algorithm mapping baseline steps to contracts shown in Table C.2.[16] The three contracts had step targets of 10,000, 12,000, or 14,000 steps, all with a 20 INR payment rate. As detailed in Appendix C.4, we based the algorithm on the same model we used to design the choice menu.

[15]The targets were randomized between 10,000, 12,000, or 14,000 steps in the same proportion as participants were assigned to the Fixed Low, Medium, and High Target groups.

[16]Specifically, we assigned step targets based on average daily steps taken on days that participants recorded at least 200 steps. It would be extremely unlikely for a person wearing a pedometer consistently to record fewer than 200 steps, so we considered such days as missing data.

Notably, the timeline for participants in the Tag group deviated from the base timeline outlined in Section 4.1. Instead of revealing Tag participants' treatment assignment at the Contract Launch, Surveyors told the Tag group how their contracts would be assigned at the end of the Baseline survey, before the pre-contract period began. The Tag group's actual targets were then assigned during the Contract Launch, based on their baseline steps. We informed the Tag group how their targets would be assigned before the pre-contract period because, in scaled-up versions of tagging policies, participants would know that their behavior determines their contract. The Tag group was still encouraged to walk as normal during the pre-contract period.

4.2.4 Choice Channels

We include three treatment groups to explore the channels driving the performance of Choice. The first allows us to examine the role of non-standard preferences.

Flat Choice Group In this group, participants chose their contracts from the Flat Menu shown in Table 1, which is not incentive compatible for those with standard preferences. Specifically, the Flat Menu contains three contracts, each with a different step target (10,000, 12,000, and 14,000), but all with the same payment rate (20 INR), such that the contracts with higher step targest are financially dominated.

Implicit in our Section 3 model was the assumption that participants have complete information about their own type; if not (which is plausible), sorting could go awry. We include two treatments to assess the role of incomplete information about one's own type.

Baseline Choice Group To explore the role of learned information about type, in this group, participants selected their contract from the Base Menu at the end of the Baseline survey, before they had a chance to wear a pedometer. The earlier revelation of treatment assignment for the Baseline Choice group means that their contract preferences collected in the Choice survey, after the pre-contract period, were hypothetical, not real-stakes. Hence, we exclude their Choice survey data from analyses. The same is true for the Tag group.

Choice + Nudge Group We included this group to investigate the possibility that participants did not know how to sort across contracts. Like the Choice group, members of this group selected their contracts from the Base Menu during the Choice survey. However, prior to making their selection, we gave these participants a "nudge" towards a specific contract by informing them which contract we (the researchers) thought would maximize their steps.[17]

Since choices took place while participants were still blinded to treatment assignment,

[17]The recommendation was based on baseline steps, with the mapping from baseline steps to our recommended step target the same as in the Tag group and shown in Table C.2.

implementing the Nudge exclusively for this group would have unintentionally revealed their treatment assignment to surveyors. To avoid this, we cross-randomized the Nudge across the Fixed and Monitoring groups (i.e., the other groups that made real-stakes choices during the Choice survey). We did not design the Nudge to affect contract period outcomes for these non-Choice groups (since their menu choice did not impact their contract assignment) and find no evidence that it does. Our main specifications include an indicator for being in the Choice + Nudge group, as well as an indicator for receiving the Nudge regardless of treatment group. We show robustness to other specifications in Appendix D.[18]

4.2.5 Experimental Design and Sample

As described in Appendix Section C.1, we implemented the experiment in three phases. In each phase, we tweaked the design slightly in order to answer additional research questions. All analyses include a dummy which controls for the phase of the experiment in which participants were enrolled.

We exclude participants who withdrew or were found ineligible prior to the end of the Choice survey from all analyses, leaving a final analysis sample of 6,882 individuals.[19] The sample represents 35% of the screened, eligible population. See Table A.2 for the share of people dropped in each stage of the enrollment process.

5 Data and Summary Statistics

We employ four sources of data in our analysis: (1) the Baseline survey; (2) the Choice survey; (3) the baseline step data; and (4) step data from the contract period. Section 5.1 describes datasets (1)–(3). Section 5.2 describes dataset (4) and addresses potential data quality concerns such as attrition. Section 5.3 summarizes the baseline characteristics of our sample and the balance across treatment groups.

[18]The effect of the Nudge is insignificant for non-Choice participants. Note that our main analysis does not follow our *ex ante* analysis plan to pool the Choice and Choice + Nudge groups for the sake of statistical power, although we show that specification in column 5 in Table D.1. Our primary reason for this deviation is that the literature has subsequently raised concerns about presenting weighted-average effects in cross-randomized designs, given they can be difficult to interpret and are often "neither of primary academic interest nor policy-relevant" (Muralidharan et al., 2019). This proved true in our case: the Nudge treatment behaved unexpectedly, with certain types of people (specifically: those with medium to high baseline steps) actually *less* likely to choose the contract that we recommended to them, such that the pooled results are indeed difficult to interpret (see the Online Supplement for details). Moreover, the addition of phase 3 to the experiment gave us sufficient power to look at the Choice treatment on its own, which represents a cleaner test of our academic and policy-relevant questions.

[19] Prior to the Choice survey, our primary treatment groups were treated identically, and thus for our primary comparison we are not concerned about differential selection into this sample. While in theory there could be selection of the Tag or Baseline Choice group into this sample, as they were treated differently from the other groups before the Choice survey, in practice, Table A.1 shows that there is no differential selection into this sample between the Tag, Baseline Choice, and other groups.

5.1 Survey and Baseline Step Data

Baseline and Choice Surveys The Baseline survey, conducted at the first household visit, contains information on respondent health, socioeconomics, and demographics. The Choice survey, conducted during the second household visit, contains data on respondents' preferred contracts from the three contract menus shown in Table 1.

Baseline Steps Baseline step data consist of daily step counts recorded on the respondent's pedometer during the six-day pre-contract period. We hereafter use the term "baseline steps" to mean the individual-level average of these daily step counts.[20] We use baseline steps as a measure of types for analyzing sorting across contracts. While baseline steps could also be used as a baseline control in some comparisons, it is potentially endogenous to treatment in the Baseline Choice and Tag groups, who were informed of their treatments before the baseline step data were measured. This concern is particularly severe for the Tag group, who may have adjusted their baseline steps to affect their contract assignment.

To control for walking levels at baseline, we construct a Lasso prediction of baseline steps based on Baseline survey variables as described in Appendix C.2. For consistency across our various analyses, we use this predicted baseline step measure to control for baseline walking in our main specifications, even those that do not include the Tag or Baseline Choice groups. We also show that our main results are robust to controlling for actual baseline steps.

5.2 Contract-Period Steps and Potential Data Quality Concerns

The time-series of daily steps recorded on participants' pedometers during the contract period is the source of our primary outcomes. To measure the outcome of walking, we use the daily steps recorded on each participant's pedometer, winsorized at the 99th percentile (we also show robustness to using unwinsorized steps). To measure payments, we use the daily step data to infer how much a participant earned on each day according to their contract.[21]

We now address three potential concerns with these data.

Cheating A first potential concern is that participants might have "cheated" in order to increase their pedometer step counts without actually walking. We believe this concern is relatively muted, for two reasons. First, we monitored for what we saw as the most worrisome type of potential cheating: sharing the pedometer with another, potentially more

[20] We winsorize steps at the 99th percentile. As described in footnote 16, to implement the Tag treatment, we calculated baseline steps by averaging across the days where the pedometer recorded at least 200 steps. For consistency, we use the same measure of baseline steps in our analyses.

[21] This measure differs from actual payments since it depends on actual instead of reported steps. We use this measure because a scaled-up policy would likely deliver payments based on actual steps (which we could not do because of logistical constraints). Our results are robust to using actual payments instead.

active, individual. Specifically, we visited participants unannounced at their homes and workplaces, and checked if the pedometer was with them or someone else, and then synced the pedometer data to check for over-reporting. During the 1,797 audits we conducted, we witnessed only two examples of pedometer sharing. Second, the program design dulled the incentive for falsifying pedometer data. Incentive payments were based on self-reports through the phone system rather than through real-time monitoring of the pedometers. The incentive to falsify pedometer data was thus substantially less than if the payments were based on the pedometer step counts themselves. An easier way to cheat was simply to over-report (a behavior which, in practice, also appears to have been rare).

Attrition / Missing Pedometer Data A second potential concern is attrition/missing data from the pedometers. For 7% of people in the analysis sample, we have no pedometer data at all, either because they withdrew immediately after the Contract Launch (5% of people) or because of other reasons such as losing the pedometer (3% of people). In addition, among people for whom we have some pedometer data, their data is missing for an additional 3% of days, due to reasons such as sync issues. Columns 1 and 2 of Table A.3 show that both of these sources of missing data are balanced between Choice (the omitted group) and most other groups, most notably the pre-specified comparison Fixed Medium (12K) group. However, we do have one minor imbalance that is significant at the 5% level: the share of individuals missing data on a given day during the contract period is 1.5pp lower in the Tag group than the Choice group (column 2). This difference is small in magnitude, and we present Lee bounds to account for it in the table notes of Table A.3.[22,23]

Failure to Wear Pedometers. A final potential concern is that participants may not wear their pedometers every day. Our bonus payments for pedometer wearing were designed to counter this issue. Accordingly, participants wore their pedometers on a high share of days — 83%, on average. Importantly, pedometer-wearing rates are balanced across treatment groups, as shown in Table A.3 column 3. We include all daily step data in our analysis, including from days with 0 steps, although our results are robust to excluding the 0's.

[22]In addition, two of the 24 tests relative to Choice presented in Table A.3 are significant at the 10% level, as would be expected due to chance. Specifically, the Baseline Choice group has 2.4pp more people missing their full contract period data (col 1 of Table A.3), and the Monitoring group has 1.5pp lower missing data on a given day (col 2). Both differences are small and are not in our primary treatment groups. We present Lee bounds accounting for each in the Table A.3 notes.

[23]As discussed in Section 4.2.5, the Table A.3 attrition (and all of our) analyses condition on being in the analysis sample which was present through the end of the Choice survey. Since the Baseline Choice and Tag groups were treated differently before that point, one might be concerned that they would have differential attrition before that point. However, Table A.1 shows that that is not the case. Accordingly, the Table A.3 results for those groups are similar if we do not condition on being in the sample through the end of the Choice survey and instead include everyone who was present at the Baseline survey.

5.3 Summary Statistics and Balance Checks

Characteristics of our full analysis sample are in column 1 of Table A.4. As shown in Panel A, the average age was 49. 37% of the sample were female, and 58% had completed some secondary education. The average monthly income per capita was just over 5500 INR, making an incentive payment of 20 INR equivalent to 11% of average daily per capita income.

Measures of participants' health, which are shown in Panel B, show that the sample had high rates of chronic disease and disease risk. 31% of the sample had been diagnosed with diabetes and 32% with hypertension. Average blood pressure and BMI levels are both extremely high. The average blood pressure measurement of 138/92 mm Hg exceeds the hypertension cutoff of 130/80 mm Hg or greater. The average BMI of 26 kg/m^2 is in the obese range for Indians. During the pre-contract period (when there were no step target incentives), participants walked an average of 7,230 steps per day, which is very similar to average steps taken by Fitbit pedometer users across India (Dube, 2020).

Columns 3 through 9 of Table A.4 show that baseline characteristics are balanced across treatment groups. Omnibus tests of balance across all covariates fail to reject the null that each of the treatment groups has the same baseline characteristics as the Choice group or the Fixed Medium group (Bruhn and Mckenzie, 2009), with one exception. There is significant ($p<0.05$) imbalance between the Fixed High and Fixed Medium groups. This is not our primary comparison (our primary test is Choice vs. Fixed Medium), and we address the imbalance by using the post-double-selection Lasso method of Belloni et al. (2014) to choose controls for our treatment effects regressions, as described below.

6 Results

This section empirically examines the impacts of Choice on the effectiveness of incentives. We first document the effect of Fixed targets on average steps as a benchmark for the potential improvements due to choice. We then examine the effect of Choice, first on average steps and average payments, and then on the distribution of steps.

Much of our analysis centers on the following regression equation comparing our various treatment groups:

$$y_{it} = \alpha + \beta \times \text{Choice}_i + \mathbf{Treat}_i'\boldsymbol{\delta} + \boldsymbol{X}_i'\gamma + \boldsymbol{X}_{it}'\lambda + \boldsymbol{Z}_i'\mu + \tau_{m(t)} + \varepsilon_{it}. \tag{7}$$

where i represents a participant and t represents a date. The outcome y_{it} is individual i's steps on day t during the contract period. Choice_i is an indicator for being assigned to the Choice group. $\mathbf{Treat}_i$ is a vector of indicator variables for assignment to the other treatment groups (Fixed Low, Fixed High, Monitoring, Tag, Flat Choice, Baseline Choice, Choice +

Nudge), with Fixed Medium as the omitted so that the β coefficient represents our primary comparison of interest: Choice relative to Fixed Medium.

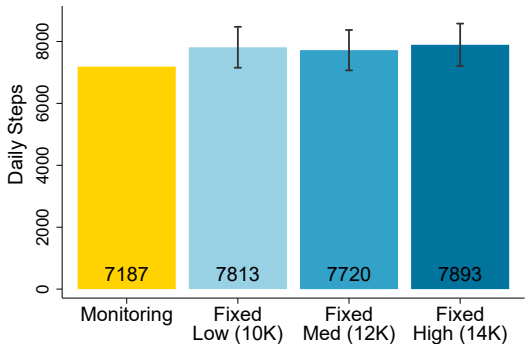
While we present many tests, our AEA registry specified only one as a primary hypothesis: the test of $\beta = 0$. (The other tests provide a fuller understanding of the effect of Choice relative to non-personalized incentives and of the channels for Choice's impact.)

$\boldsymbol{X}_i$ and $\boldsymbol{X}_{it}$ are vectors of individual and day-level controls selected from the covariates listed in column 1 of Table A.5 using the post-double-selection Lasso method of Belloni et al. (2014). $\boldsymbol{Z}_i$ are experimental controls, namely, fixed effects for the experimental phase, the length of time between the Baseline and Choice surveys, and whether the participant received the cross-randomized Nudge.[24] $\tau_{m(t)}$ are year-month fixed effects. Standard errors are clustered at the participant level. We present the results in Table 2 and highlight the main comparisons of interest graphically as we proceed through the discussion.

6.1 Fixed Incentives Increase Steps

We begin by examining the effect of non-personalized incentives — that is, the Fixed Low, Fixed Medium, and Fixed High contracts — relative to Monitoring. Figure 6 uses the coefficients from Table 2 to depict these comparisons graphically.

Figure 6: Incentives for Fixed Step Targets Similarly Increase Average Steps



Notes: The figure displays average contract-period steps in each Fixed group and the Monitoring group. The 95% confidence interval bars are relative to Monitoring and come from the regression in Table 2.

[24]This dummy, $Nudge_i$, is equal to 1 regardless of the participant's main treatment assignment. Since we include a Choice + Nudge regressor, the Nudge coefficient identifies the effect of the Nudge in the non-Choice groups. Assuming the effect of the Nudge is homogeneous across the non-Choice groups, the Choice coefficient can be interpreted as the effect of Choice relative to the no-Nudge Fixed Medium group (and likewise for the other coefficients). This assumption is in line with our expectation that the effect of the Nudge for non-Choice groups would be null (as supported by an insignificant Nudge coefficient in column 1 of Table D.1). However, we also allow for the possibility that the Nudge effect would vary by non-Choice group by estimating a fully interacted model (col 3 of Table D.1) and find a nearly identical Choice coefficient.

Table 2: Choice Has Higher Treatment Effect than Fixed Targets

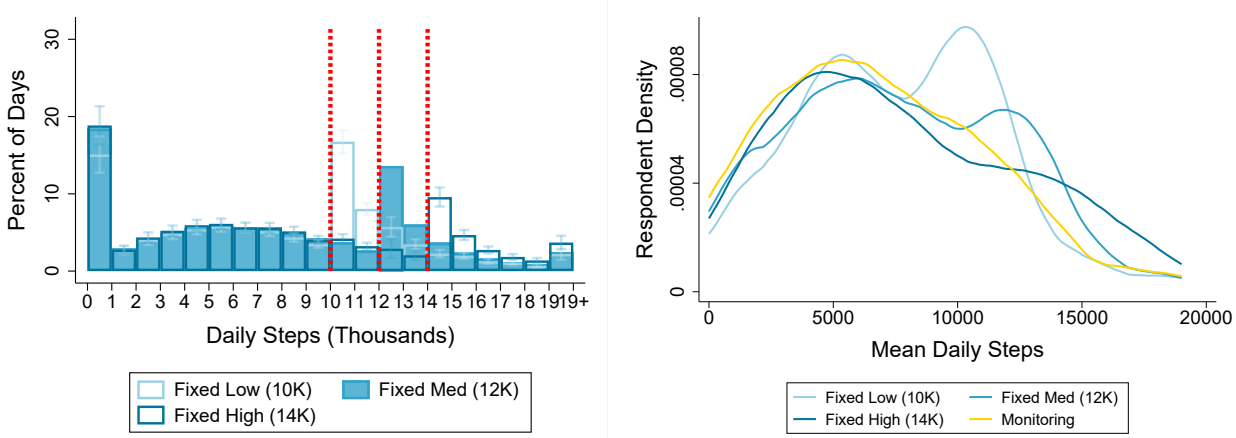
Omitted Group:	Fixed Medium
Dependent Variable:	Daily Steps
	(1)
Choice	414**
	[202]
Fixed Low (10K)	93
	[185]
Fixed High (14K)	173
	[208]
Tag	463**
	[205]
Flat Choice	98
	[252]
Baseline Choice	343
	[225]
Choice + Nudge	80
	[239]
Monitoring	-533
	[332]
Fixed Medium (12K) Mean	7,720
p-value vs Choice	
Fixed Low	0.123
Fixed High	0.287
Tag	0.817
Flat Choice	0.199
Baseline Choice	0.748
Choice + Nudge	0.239
Monitoring	0.005
p-value vs Monitoring	
Fixed Low	0.064
Fixed High	0.044
Tag	0.004
Flat Choice	0.084
Baseline Choice	0.012
Choice + Nudge	0.109
p-value Fixed High vs Fixed Low	0.711
# Observations	172,961
# Individuals	6,384

Notes: Sample sizes: Choice: 892; Fixed Low: 778; Fixed Medium: 1,210; Fixed High: 796; Tag: 928; Flat Choice: 439; Baseline Choice: 631; Choice + Nudge: 523; Monitoring: 187. The dependent variable is daily steps measured using the contract-period pedometer data. The omitted category is the Fixed Medium group. We control for experimental phase, time between Baseline and Choice surveys, receiving the Nudge treatment, year-month fixed effects, and a vector of controls selected by double-Lasso from the controls shown in col 1 of Table A.5. Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.

All three step targets have positive impacts on daily walking, ranging from 533–706 steps. While our power for comparisons with the Monitoring group is somewhat limited due to the fact that that group is small, the p-values for equality with Monitoring are 0.064, 0.109, and 0.044 for the Fixed Low, Medium, and High groups, respectively, and 0.055 when all three Fixed groups are pooled. Moreover, these estimates are all meaningful in size, equivalent to approximately 5–7 additional minutes of brisk walking, on average, each day — roughly a 7–10% increase relative to the Monitoring group.[25]

The impacts of the three Fixed groups are similar and statistically indistinguishable, but this does not stem from participants ignoring their step targets. Figure 7(a) shows that daily steps bunch steeply just above the randomly-assigned step target. Consequently, the distributions of average individual-level steps over the contract period differ markedly across the Fixed groups, as shown in Figure 7(b). Compared to the Monitoring group, the High target barely moves the bottom of the distribution, but has strong effects at the top, while the opposite is true for the Low target. The importance of step targets for walking suggests that personalizing the step target could in fact affect behavior. We explore this next.

Figure 7: Step Targets Matter



(a) Daily Steps Bunch Just Above Step Targets (b) Average Daily Steps Bunch Near Step Targets

Notes: Panel (a) displays histograms of daily steps during the contract period. The vertical red lines are drawn at each of the three step targets. The 95% confidence interval bars are drawn relative to the Fixed Medium group and use the same controls as column 1 of Table 2. Panel (b) displays kernel density plots of individual-average daily steps across the contract period.

6.2 Relative to Fixed Incentives, Choice Increases Performance

This section analyzes the effects of using incentive-compatible choice to personalize the step target. We begin by taking the perspective of a principal who values all steps equally

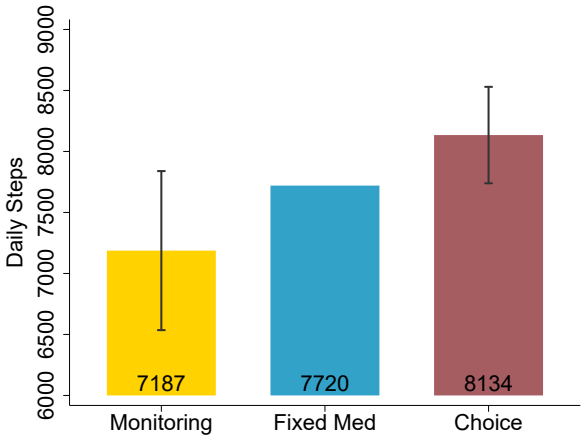
[25]We estimate minutes of brisk walking a day using a conversion rate of 100 steps per minute.

(i.e., has a linear benefits function) and examine the effect of choice on average steps relative to average payments. We then consider a principal with a nonlinear benefits function and examine the effect of choice on the full distribution of steps.

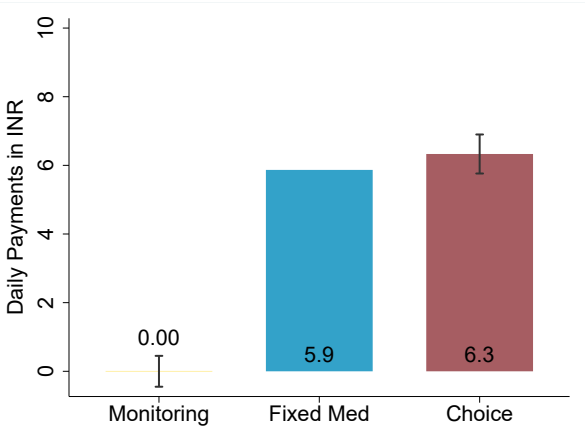
6.2.1 Average Impacts of Choice

We first compare average steps in our Choice group with the pre-specified one-size-fits-all comparison group (the Fixed Medium group). The Choice coefficient from Table 2 captures this comparison, which we also depict graphically in Panel (a) of Figure 8.

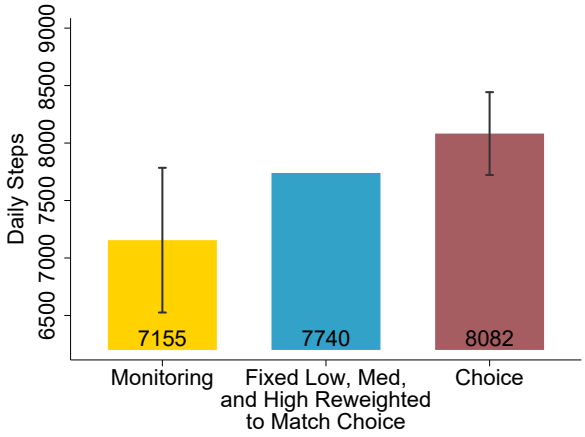
Figure 8: Choice Increases Walking with Small Impacts on Incentive Payments



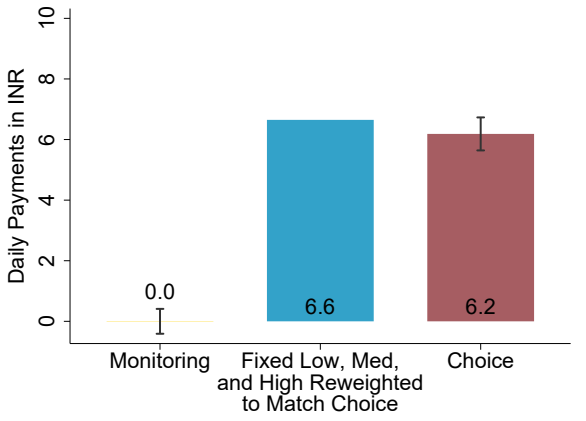
(a) Daily Steps: Choice vs. Fixed Medium



(b) Daily Payments: Choice vs. Fixed Medium



(c) Daily Steps: Choice vs. Reweighted Fixed



(d) Daily Payments: Choice vs. Reweighted Fixed

Notes: Figures show the impact of Choice on average contract-period steps (panels (a) and (c)) and payments (panels (b) and (d)). In panels (a) and (b), 95% confidence intervals shown relative to Fixed Medium and come from the regressions in Table 2 and A.6, respectively. In panels (c) and (d), 95% confidence intervals shown relative to the "Reweighted Fixed" group (i.e., the Fixed groups reweighted in the proportion that their targets appear in the Choice group) and come from the regressions in Table A.7, cols 1 and 2, respectively.

Choice substantially increases the impact of incentives relative to the Fixed Medium group. While the Medium target increases daily steps by 533 steps relative to Monitoring alone, or roughly 5 minutes of brisk walking, the Choice treatment increases walking by an additional 414 steps (significant at the 5% level) or 4 minutes — an increase of roughly 80%.

Table 3 shows that Choice's advantage over the Fixed Medium group is robust to alternative specifications, namely, omitting the additional control variables (col 2), controlling for actual baseline steps (col 3), not winsorizing the outcome variable (col 4), limiting to the first two phases of the experiment (as we originally designed our experiment to detect Choice's impact in the Phase 1 and 2 samples; col 5), and using the "one-at-a-time" estimator from Goldsmith-Pinkham et al. (2022) to mitigate potential concerns about bias from simultaneously estimating multiple treatment effects in one equation. In all specifications, the magnitude of the difference between the Choice and Fixed Medium groups remains large and significant at at least the 10% level. The estimates of the percentage increase in the treatment effect due to choice are all substantial, ranging from 60% (col 5) to 106% (col 2).

So far, the discussion has focused on the effect of choice on the incentivized behavior (steps), but of course its effect on payments matters as well. Importantly, Figure 8(b) and Table A.6 show that Choice achieves the 80% increase in average steps without meaningfully increasing payments. The change in payments is not statistically significant, and the point estimate suggests a mere 8% change.[26] Thus, because Choice substantially increases the treatment effect on steps but only minimally affects payments, principals who value average steps should prefer Choice to assigning everyone a uniform Medium (12K) Target.

While the Fixed Medium group was our pre-specified benchmark for Choice, it is not the only non-personalized benchmark of interest. One useful benchmark, which we call the "Reweighted Fixed" group, is to consider randomly assigning participants to step targets with the randomization probabilities set to match the probabilities with which each step target appears in the Choice group (which are 58%, 21%, and 20% for the Low, Medium, and High targets respectively, as shown in Figure A.1). While it may be unlikely that policymakers would randomize step targets in practice, this benchmark allows us to hold the mix of step targets constant when comparing Choice with an unpersonalized approach.

Figure 8(c) compares the Choice group and the Reweighted Fixed benchmark graphically.[27] Choice increases daily walking by 343 steps more than the Reweighted Fixed group

[26] If we use reported steps instead of actual steps to calculate payments, the point estimate remains virtually unchanged, going from 0.46 to 0.49, although the p-value decreases to 0.097.

[27] Specifically, we estimate the following regression equation using weighted regression:

$$y_{itk} = \alpha + \beta_1 \times \text{Choice}_i + \beta_2 \times \text{Monitoring}_i + \boldsymbol{X}_i'\gamma + \boldsymbol{X}_{it}'\lambda + \mu_k + \varepsilon_{it}, \tag{8}$$

where the omitted group is the "Reweighted Fixed" group (i.e., the pooled Fixed Low, Fixed Medium, and Fixed High groups) and all variables are defined as in equation 7. To create the same step target balance in

Table 3: The Improvement Due to Choice is Robust across Specifications

Omitted Group:	Fixed Medium					
Dep Variable:	Daily Steps					
Robustness to:		Controls		Dep Var	Sample	
	Base Spec	Basic	Actual Steps	Non-Winsorized	Phases 1 & 2	Choice & 12K Only
	(1)	(2)	(3)	(4)	(5)	(6)
Choice	414**	436**	383**	444**	501*	518**
	[202]	[210]	[176]	[207]	[275]	[203]
Fixed Med effect	533	411	444	534	838	585
Choice effect as % Med effect	78	106	86	83	60	89
# Observations	172,961	172,961	130,571	172,961	109,380	172,961
# Individuals	6,384	6,384	4,825	6,384	4,008	6,384
Controls						
Predicted Steps	Yes	No	No	Yes	Yes	Yes
Steps	No	No	Yes	No	No	No
Demographics	Yes	No	Yes	Yes	Yes	Yes
Year-Month FEs	Yes	No	Yes	Yes	Yes	Yes
Experimental	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table shows robustness of the specification shown in Table 2 (and replicated here in column 1). For brevity, this table only displays the Choice coefficient from the regressions; see Table A.8 for all coefficient estimates.
Columns 2-3 show robustness to different controls. All columns control for experiment phase, time between Baseline and Choice surveys, and the Nudge treatment (the "Experimental" controls, equivalent to z_i in equation 7). Our base specification in Column 1 additionally controls from a vector of controls selected by double-Lasso from the list of controls in column 1 of Table A.5, which includes both predicted baseline steps (panel C of Table A.5, the "Predicted Steps" control) and other controls (panels A, B, and E of Table A.5, the "Demographics" control), in addition to year-month fixed effects. Column 2 omits these additional controls. Column 3 includes the same control specification as in column 1 except that it uses actual baseline steps (panel D of Table A.5) rather than predicted steps in the vector of controls that Lasso can select from, as listed in Table A.5 column 2. Column 4 shows robustness to using non-winsorized steps as our dependent variable. Column 5 limites to experimental phases 1 and 2. Column 6 limits to only the Choice and Fixed Medium groups. The Fixed Medium effect in this column comes from a separate regression that only includes Fixed Medium and Monitoring.
While only the Choice and Fixed Medium results are shown here, the sample for columns 1-5 includes the Monitoring, Tag, Choice, Flat Choice, Fixed, Baseline Choice, and Choice + Nudge groups (the Tag and Baseline Choice groups are omitted from col 3 since baseline steps are endogenous for those groups). The omitted category is the Fixed Medium group.
Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.

(p-value = 0.062)—an increase of roughly 59% in the treatment effect relative to Monitoring. This large increase in steps is achieved while actually marginally decreasing payments, as shown in Figure 8(d) (p-value = 0.096). Hence, even conditional on the mix of step targets,

the Reweighted Fixed group as the Choice group, we assign each Reweighted Fixed observation a weight of $\frac{c_{sk}}{f_{sk}}$, where, f_{sk} and c_{sk} represent the fractions of the pooled Fixed and Choice groups, respectively, assigned to step target s ($s \in \{Low, Med, High\}$) in experiment phase k. (All Monitoring and Choice observations simply have a weight of 1.) Table A.7 shows the results.

Choice substantially improves performance relative to an unpersonalized approach.[28]

In addition to comparing Choice to the Fixed Medium and Reweighted Fixed groups, we can also compare it to the other Fixed groups. While we did not power our experiment to test for these differences, we interpret the point estimates as suggestive. Figure 9 displays a scatter plot of average steps versus average payments in the different treatment groups. The arrow indicates the direction of principal bliss: higher steps and lower payments. Regardless of the value of steps to the principal, the principal should prefer Choice not just to the Medium target, as already shown, but also to the Low target. Choice generates more steps than the Low target (p-value = 0.123) for less payment (p-value < 0.01). Whether the principal prefers Choice to the High target depends, however, on the principal's specific value of steps, as Choice generates more steps (p-value = 0.287) but also higher payments (p-value < 0.01). Thus, the higher the marginal value of steps to the principal, the more likely they are to prefer Choice. Moreover, recall that the High target does particularly poorly at the bottom of the distribution (as shown in Figure 7(b)). We show next that Choice performs better at the bottom of the distribution, which means that principals who particularly value steps among lower walkers are particulary likely to prefer Choice.

6.2.2 Distributional Impacts of Choice

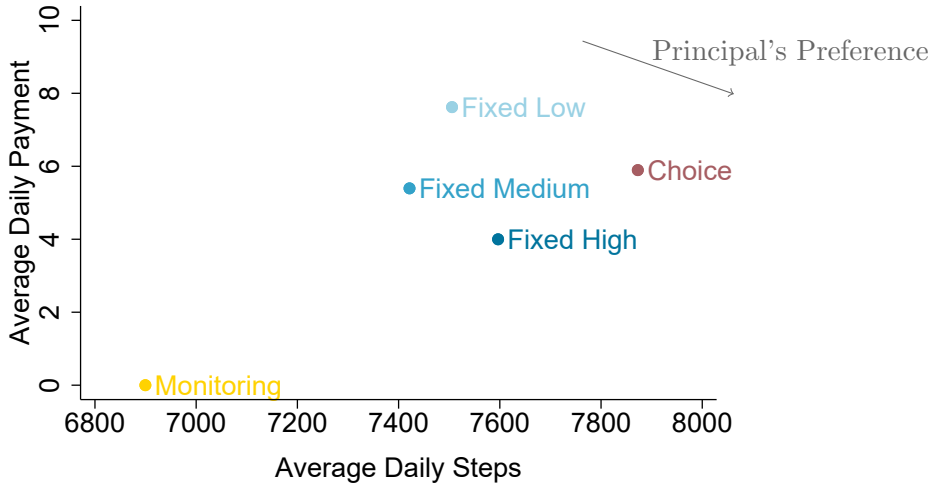
To assess Choice from the perspective of a policymaker with a nonlinear value of steps, we now assess the impact of Choice on the cumulative distribution function (CDF) of steps. We begin by comparing the CDFs of average individual-level contract-period steps across the Fixed groups.[29] Figure 10(a) shows that no one Fixed target dominates another, with the Low and High targets both having upsides and downsides relative to the Medium target. Fixed Low has the highest performance at the bottom of the distribution (p-value < 0.05 relative to High at the 25th and 50th percentiles of the distribution), but Fixed High has the highest performance at the top (p-value < 0.01 relative to Low at the 75th percentile). Somers' D tests confirm that no Fixed target first-order stochastically dominates another.

In contrast, as shown in Figure 10(b), Choice first-order stochastically dominates every one of the fixed targets, as it nearly traces the outer envelope of their CDFs. Somers' D tests confirm the first-order stochastic dominance (p-values 0.065, 0.016, and 0.040 for comparisons with the Fixed 10K, 12K, and 14K groups, respectively). Specifically, Choice performs similarly to the Low target at the bottom of the distribution, outperforming the High target at both the 25th and 50th percentiles of the distribution (p-value < 0.05). Analogously,

[28] Since the contracts used in the Choice menu have slightly different payment levels than those used in the Fixed groups, this analysis does not condition on the mix of *contracts*, only the mix of step targets. However, for a given step target, payments are weakly lower in the contracts used in the Choice menu, so conditioning on the payment levels (in addition to step targets) is likely to make the Choice group look even better for the outcome of steps (but likely make the average payments more equivalent).

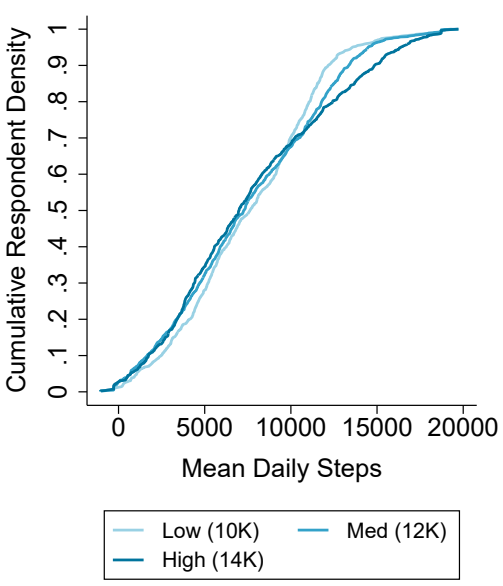
[29] We residualize individual-level steps on experiment phase dummies to ensure orthogonality to treatment.

Figure 9: Choice Cost-Effectively Generates Steps

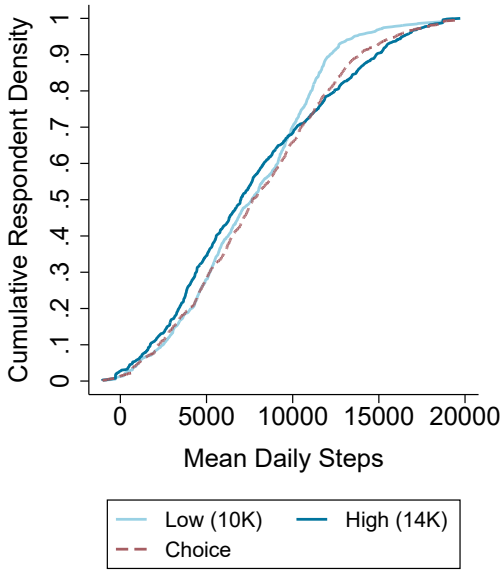


Notes: The figure plots average daily steps against average daily payments in several treatment groups. For consistency with the regression estimates, average daily steps and average daily payments are each residualized using the same double-Lasso-selected controls as in Table 2 and Table A.6, respectively.

Figure 10: Choice First-Order Stochastically Dominates Each Fixed Target



(a) No Fixed Target Dominates the Others



(b) Choice Achieves the Outer Envelope of the Fixed Targets

Notes: The figures display CDFs of average individual-level steps in the contract period, by treatment group. To ensure orthogonality to treatment, average steps have been residualized on a control for experiment phase. Panel (a) shows the three Fixed groups only, while panel (b) brings in the Choice group. We omit the Fixed Medium line from panel (b) for visual clarity, since it is always between the Fixed Low and Fixed High lines.

it performs similarly to the High target at the top of the distribution, outperforming the Low at the 75th percentile of the distribution (p-value < 0.01). By sorting participants into the targets appropriate for them, Choice achieves the upsides of the more extreme targets without their downsides.

Because payments in Choice are weakly less than in Fixed Low and Fixed Medium, the fact that steps in Choice first-order stochastically dominate steps in Fixed Low and Medium means that any policymaker — regardless of their benefits function of steps — should prefer Choice to uniform Low and Medium targets. The relationship with the High target is more ambiguous given that the High target also pays out less than the Choice group. However, Choice's better performance at the bottom of the distribution means that principals who particularly value steps among low walkers are particularly likely to prefer Choice to a uniform High target.[30]

The magnitude of Choice's advantage over Fixed High at the bottom of the distribution is large. Table A.9 presents quantile treatment effects of the three Fixed treatments relative to Choice (the omitted group). Choice's treatment effects at the 25th and 50th percentiles of the distribution are roughly 2.5 times as large as Fixed High's.

6.2.3 Summary of Results on the Effectiveness of Choice

In this section, we showed that personalization using incentive-compatible choice substantially improves the effectiveness of incentives. Relative to an intermediate one-size-fits-all benchmark, Choice increases average steps by roughly 80% without increasing costs. Choice also outperforms the Low target, increasing steps while decreasing costs, while substantially increasing steps relative to the High target at the bottom of the distribution. Moreover, Choice first-order stochastically dominates each of the Fixed targets, suggesting that it would likely be preferred by a range of principals.

7 Choice: Channels and Benchmarking

This section examines the channels through which Choice increases walking and benchmarks Choice against personalization on observables.

7.1 Channels for the Effectiveness of Choice

The Maskin and Riley (1984) framework, presented in Section 3, suggests two main channels for the effectiveness of Choice: (1) higher targets should be more effective for higher types (i.e., those who walk more at baseline; Result 1), and (2) the Choice menu should sort people into targets by type (Result 2). We begin by providing evidence that

[30] While some evidence suggests that the health benefits of exercise are linear, other evidence suggests that it could be concave (Loprinzi, 2015). That said, technically what the principal cares about is not the shape of total health benefits but the shape of the health externality, on which there is less evidence.

Choice is in fact effective because of channels (1) and (2). We also show that, although there are non-standard preferences at play, the incentive-compatibility of our Base Menu is crucial for Choice's success. Finally, we briefly examine whether information frictions about one's own type hinder effective sorting in Choice and find no evidence that they do.

Further from the standard model, an alternate theory is that choice operates not by sorting but through creating autonomy effects from being allowed to choose. We examine this possibility in the Online Supplement and find no evidence for it.

7.1.1 Higher Step Targets Are More Effective for Higher Walkers

We first examine whether higher step targets work better for those with higher baseline walking. Among participants in the Fixed groups, we run the following regression:

$$y_{it} = \alpha + \beta_1 \times \text{Step Target}_i \times y_i^{BL} + \beta_2 \times y_i^{BL} + \beta_3 \times \text{Step Target}_i + \boldsymbol{X}_i'\gamma + \boldsymbol{X}_{it}'\lambda + \boldsymbol{Z}_i'\mu + \tau_{m(t)} + \varepsilon_{it}, \tag{9}$$

where y_i^{BL} is baseline steps (in thousands), Step Target_i is a continuous measure of the step target assigned to participant i (in thousands), and y_{it} are daily steps. The remainder of the variables are defined as in equation (7). The coefficient of interest, β_1, represents the additional increase in daily contract-period steps from increasing the step target by 1,000 steps for those whose baseline steps are 1,000 steps higher.

The results, shown in column 1 of Table A.10, show that β_1 is positive and significant, confirming that higher step targets generate more walking from higher baseline walkers. To better understand the magnitudes, Figure A.2 displays the treatment effects on steps of each Fixed group relative to Monitoring separately for each tercile of the baseline step distribution. For those in the top tercile, the effect of being in Fixed High instead of Fixed Low is nearly 1,200 steps greater than for those in the bottom tercile — a large difference, roughly twice the size of the average effect of Fixed incentives.

Column 2 of Table A.10 presents the results using daily payments as the outcome variable y_{it}. There is no statistically significant or meaningful heterogeneity in the payments by step target for higher walkers. High step targets are generally less expensive than low step targets (Figure 9), and no less so for high walkers. Hence, principals should prefer higher targets for higher walkers, as they generate substantially more steps without higher payments.

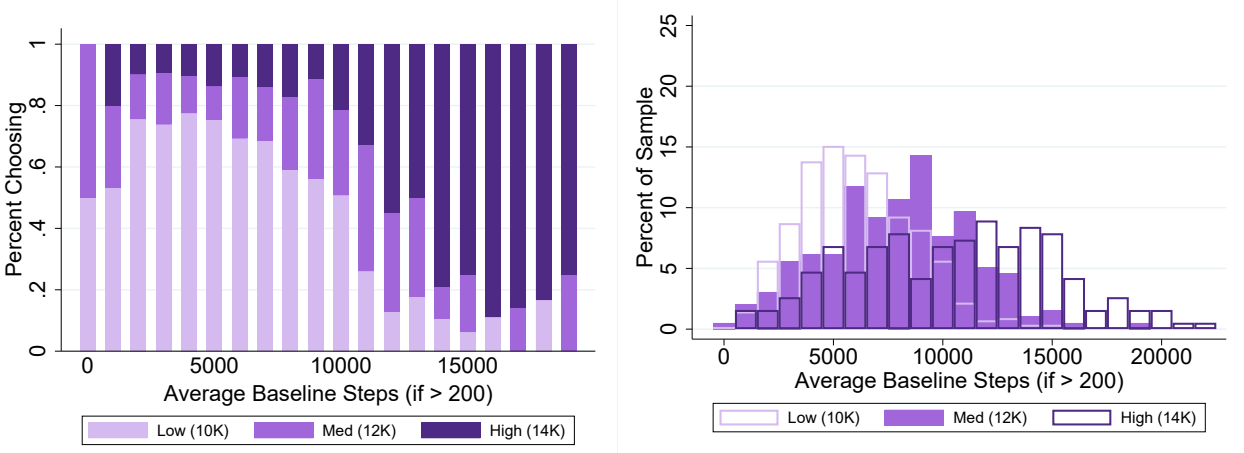
The large magnitude of the heterogeneity in the treatment effects of step targets by baseline steps could explain the large improvement from Choice if participants sort based on baseline steps when choosing step targets from the Base Menu. We examine this next.

7.1.2 Choice Sorts Participants by Type

Figure 11 shows strong evidence that, consistent with Result 2, participants in the Choice group sort across contracts based on their type. Figure 11(a) shows that lower walkers are more likely to choose lower step targets, and higher walkers are more likely to choose higher

step targets. While 80% of walkers with baseline steps in the bottom quintile choose the Low Target, only 20% of walkers in the top quintile do. Put another way, the distribution of baseline steps is markedly different among the participants who choose (and are then assigned to) the Low, Medium, and High targets, as shown in Figure 11(b). The correlation between choices and baseline steps is highly statistically significant (Table A.11, col 1).

Figure 11: Choice Sorts Participants by Baseline Walking



(a) Chosen Step Targets by Baseline Steps

(b) Distributions of Baseline Steps by Chosen Step Target

Notes: Panel (a) show the fraction of the Choice group that chose the Low, Medium, and High target on the Base Menu, by bins of baseline steps. Panel (b) shows the resulting distributions of baseline steps among Choice group participants who chose each step target (Low, Medium, and High).

While baseline steps are a sufficient statistic for type in our unidimensional Section 3 model, outside the model, there could be other factors that could also impact individuals' treatment effects from different targets (i.e., their true "types"). For example, employed people may have less capacity than unemployed people to reach the High target relative to the Low. To explore whether participants sort based on these other factors as well, we follow the methodology of Athey et al. (2019) and estimate a causal forest in our Fixed groups to predict each individual's treatment effect from assignment to the High relative to the Low step target, based on a large set of observables (including baseline steps; see Appendix C.5 for details). The causal forest selects baseline steps as the most important predictor of treatment effect heterogeneity;[31] in fact, the correlation between the predicted treatment effects and baseline steps is 0.59. However, there are other important predictors, such as age and health measurements (see Table A.12 for the list). Column 2 of Table A.11 shows that participants' choices correlate significantly with their predicted treatment effects. However, if we control for baseline steps, column 3 shows that predicted treatment effects do not have

[31]Importance indicates how frequently the trees in the causal forest split on each variable.

any additional positive predictive power over choices. The primary observable characteristic on which participants sort appears to be baseline steps.

However, there also appear to be unobservable factors that influence choices. As seen in Figure 11(a), some people who walked little at baseline choose high targets. While these participants might be making mistakes, they could also have better information about their own true type than their baseline steps alone. After all, even within the context of our unidimensional Section 3 model, an individual's true type maps 1:1 with their counterfactual *contract period* steps in the absence of incentives, of which baseline steps may be an imperfect measure (e.g., because of a temporal shock such as a pre-contract period injury).

If baseline measurements are, in fact, poor type measures for some people, choices can provide supplementary information about type. Figure A.3 provides evidence that this is the case. Specifically, in the Monitoring group, contract period steps represent a perfect measure of type (i.e., contract period steps without incentives). Since the Choice survey measured menu choices from the Monitoring group, we can show that participants with higher chosen targets have higher types (i.e., higher contract period steps), even conditional on baseline steps and predicted treatment effects. This suggests that choices capture unobservable information about type and that allowing people to choose their contracts may help overcome the noise that arises when personalizing based on (noisy) baseline observables.

We also use the Fixed groups to provide a final piece of evidence that participants sort by type. Table A.13 shows that participants who chose higher step targets have more positive treatment effects from being randomly assigned to higher (rather than lower) step targets.

Thus, we have shown that the two main mechanisms for the effectiveness of Choice from the Maskin and Riley (1984) framework hold in our setting.

7.1.3 Some Participants Have Non-Standard Preferences

Embedded in the Maskin and Riley (1984) framework is also the idea that higher types only choose higher targets because of the higher payment rates associated with them. However, this final implication does not appear to hold in our setting. On the Flat Menu, where there is no financial incentive to choose higher targets, Figure A.1 shows that 33% of participants still choose Medium and High targets. It appears that non-standard factors, such as pride or demand for commitment (e.g., Ashraf et al., 2006), may be influencing choices.[32] This raises an important question: did high types only sort into higher targets because of non-standard factors, or was the incentive-compatibility of the Choice menu also critical?[33]

[32]Carrera et al. (2020) provide evidence that demand for commitment contracts can also reflect confusion. We asked two questions to confirm whether participants understood that the Medium and High targets were dominated on the Flat Menu, and 90% of participants answered both questions correctly.

[33]Non-standard preferences could cause sorting by baseline steps even if those preferences were not correlated with baseline steps. For example, even if all participants have a time-inconsistent demand for commitment, a higher target would only be an effective commitment device for participants with sufficiently

7.1.4 Using Higher Payments for Higher Targets Induces Better Sorting

We now explore how the incentives to choose higher targets affects sorting and performance in Choice. We first compare the choices on the Base Menu with choices on the Flat Menu, which gave no financial incentive to choose higher targets, and on the Steep Menu, which gave stronger incentives to choose higher targets. Second, we examine the treatment effect of assigning contracts based on Flat Menu choices relative to Base Menu choices.

Choices Figure A.4 shows that participants' choices respond to the incentives to sort. Specifically, Panel A of the Figure shows the differences in the percent of participants choosing the Low, Medium, and High targets on the Flat Menu (sub-graphs I and II) and Steep Menu (sub-graph III), both relative to the Base Menu. Significantly more participants choose the Low target on the Flat Menu and the High target on the Steep Menu. The magnitudes in sub-graph I, which focuses only on first-choice menus to control for order effects, are meaningful.[34] Five pp fewer participants choose the High target on the Flat Menu than the Base Menu, off of a base of 18%.

The implications of the shift towards lower targets depends on which participants shift. Panels B and C of Figure A.4 show the results separately for those with above-median baseline steps and below-median baseline steps. The greater fraction of Low choices on the Flat Menu are entirely driven by those with above-median baseline steps — precisely those that Section 7.1.1 showed the principal does not want to move into lower targets. The differences in sorting between those with above-median and below-median steps are significant in the all choices sample at the 1% level. Hence, making the menu incentive-compatible improves sorting.

Treatment Effects Our finding that sorting varied across the Flat Menu and the incentive-compatible Base Menu suggests that the treatment effects of assigning participants on the two menus may also differ. We therefore compare steps in the Flat Choice group, whose contracts depended on their Flat Menu choices, with steps in our Choice group, whose contracts depended on their Base Menu choices. As shown in Table 2, while the main Choice group walks 414 more steps on average, daily, than the Fixed Medium group, the Flat Choice group only walks 98 more steps on average than the Fixed Medium group—an improvement which is not statistically different from 0. While we cannot reject equality between the Flat Choice and Choice groups (p-value 0.199), we interpret the evidence as suggestive. Taken together with the above analysis of sorting, it appears that the incentive-compatibility of

high baseline steps. See Appendix B.4 for further discussion of contract preferences with time inconsistency.

[34] Recall that we varied choice order for a short period to explore choice order effects. The evidence suggests that choice order matters: the difference between Flat and Base Menu choices is over 5 times as large for first as second choices, although the p-value for the difference is only 0.151 due to the small sample for which we randomized order.

our Base Menu was important for its success.[35]

7.1.5 Information Frictions Do Not Appear to Impede Choice

In the standard model, respondents understand their own behavioral response type. Given the above evidence that participants sorted by type, participants must have had *some* information about their types. If they had more information, would Choice have worked better? Perhaps surprisingly, we do not find any evidence that more information would have made Choice more effective. The Online Supplement discusses our results in detail; for brevity, we just summarize them here. First, having more time with pedometers does not have much impact on choices or sorting. Sorting and walking are similar (and statistically indistinguishable) between the Baseline Choice group, which had 0 days with a pedometer before making choices, and the main Choice group, which had their pedometers for at least 6 days before making decisions. This result is notable from a policy perspective, as eliminating the phase-in period makes choice more scalable. Second, the Choice + Nudge group that received information about which target we (the principal) thought might be best has 334 *fewer* steps than the main Choice group, although the difference is not statistically significant (p-value 0.239, Table 2). This appears to reflect that the Nudge backfired in some cases, making participants with medium-to-high baseline steps *less* likely to choose the recommended target, as documented in the Online Supplement.

7.1.6 Summary of Channels for Choice's Effectiveness

We find that (1) the Choice treatment is effective because it sorts participants based on their types, and (2) the Choice menu's incentive-compatibility is important for achieving good sorting. We also find that some people prefer contracts with higher step targets, even when there is no financial incentive for such a preference.

7.2 Benchmarking Choice against Tagging on Observables

We now benchmark Choice against another potential strategy for personalization: tagging based on observables. In addition to considering the tagging algorithm used for the Tag group, we also construct other tags a policymaker could consider using our Fixed groups.

7.2.1 Constructing Synthetic Tag Groups

We consider three additional tag algorithms. To compare each potential algorithm with Choice, we construct a "Synthetic Tag" group composed of all participants in the Fixed groups who were randomly assigned the step target that the respective algorithm would have assigned them to. The algorithms we consider are:

[35]The conclusion that the Choice and Flat Choice groups have meaningfully different (but not statistically different) steps is robust to restricting attention to phase 3 of the experiment, which is the only phase in which the Flat Choice group appeared. Specifically, the Flat Choice coefficient relative to Choice becomes -309, instead of -316.

Tag Based on All Variables: We use the policy tree machine learning algorithm of Athey and Wager (2021) in our Fixed groups to estimate which step target would be best for each participant given a large set of observables including baseline steps, health measurements, wealth variables, etc. See Appendix C.5 for details. This tag may not be implementable, as some variables may be unavailable to policymakers and/or prone to manipulation.

Tag Based on Policy Variables: We again use the policy tree algorithm in our Fixed groups, but now exclude predictor variables that health policymakers in our setting do not have access to and/or that are easy to manipulate, most notably, baseline steps and the wealth variables. Column 1 of Table A.12 shows the predictors we include, which incorporate demographics (e.g., age, gender) and health measures (e.g., weight, BMI).

Tag Based on "Unmanipulated" Steps: In the Tag group, we assigned step targets based on potentially-manipulated baseline steps. To consider tagging based on unmanipulated steps instead, we assign targets to the Fixed group participants based on their baseline steps, which they had no incentive to manipulate, using the algorithm from Table C.2 (the same used in the Tag group). While not implementable, this tag allows us to isolate the effect of manipulation.[36]

7.2.2 Comparing the Tag and Synthetic Tag Groups with Choice

We compare each Synthetic Tag with Choice using a regression of the following form:

$$\begin{aligned} y_{it} =& \alpha + \beta_1 \times \text{Synthetic Tag}_i + \beta_2 \times \text{Tag}_i + \beta_3 \times \text{Fixed Medium}_i \\ &+ \boldsymbol{X}_i'\gamma + \boldsymbol{X}_{it}'\lambda + \boldsymbol{Z}_i'\mu + \tau_{m(t)} + \varepsilon_{it}. \end{aligned} \tag{10}$$

Synthetic Tag represents a dummy for being in the relevant Synthetic Tag group (All Variables, Policy Variables, or Unmanipulated Steps).[37] The omitted group is Choice. Tag and Fixed Medium are dummies for being in those treatment groups, each included in the sample for comparison.[38] All other variables are defined as in equation (7).

[36] An alternative approach is to machine-learn the algorithm based on unmanipulated steps. That approach yields statistically indistinguishable but numerically slightly worse results for Tag, and hence slightly better results for Choice. To be conservative in benchmarking Choice, we hence present the Table C.2 results.

[37] Since step target assignment was random in the Fixed groups, each Synthetic Tag group represents a randomly-selected segment of the population. However, because we assigned more Fixed target participants to the Medium target than the other targets, the people assigned to the Medium target are over-represented in the Synthetic Tag groups. To correct for the unequal probabilities of assignment to each Fixed group, the regression weights observations by the inverse of the probability of assignment to a given step target within the Fixed groups in their experimental phase.

[38] Some of the observations in the Synthetic Tag group come from the Fixed Medium group. Hence, to include both groups in the regression, we duplicate any observations that appear in both groups. All observations in the Fixed Medium group are included in the regression once with regressors Fixed Medium Target= 1 and Synthetic Tag= 0, and then the subset of those observations that are also members of the Synthetic Tag group appear a second time with regressors Fixed Medium Target= 0 and Synthetic Tag= 1. We cluster standard errors at the individual level. The results are nearly the same if we exclude the Fixed

Figure A.5 and Table A.14 show the results with steps and payments as the outcomes, respectively. Payments under Tag and Synthetic Tag never differ significantly or meaningfully from Choice and so our discussion focuses on the step results. We show Gaussian confidence intervals that condition on the synthetic tag assignments for all regressions. For the Policy and All Variables tags, which we construct based on data, we also show bootstrapped confidence intervals that account for noise in the creation of the synthetic tag assignments.

Personalizing using Policy Variables, the most scalable tag, is not effective. It generates significantly fewer steps than Choice (Gaussian p-value 0.017; bootstrapped 0.260) and performs nearly identically to the one-size-fits-all benchmark.[39] To achieve better performance, one needs to bring in other predictors — most notably baseline steps, as evidenced by the fact that the Unmanipulated Steps Synthetic Tag closes over half of the gap with Choice.

Indeed, the "best case" of personalizing using All Variables performs similarly to Choice, with numerically similar and statistically indistinguishable impacts on steps. However, the All Variables tag may be hard to implement, as the data may not be available and/or prone to manipulation. Choice has the clear advantage of being more implementable in the real world. Moreover, these results may be lower bounds on the effectiveness of Choice. While the Synthetic Tag was optimized with machine learning, we designed our Choice menu with imperfect information. An optimized Choice menu could perform even better.

Interestingly, in our experiment, the potential for manipulation did not appear to harm the performance of personalizing based on observables. The Tag group, where personalization was based on manipulated steps, has somewhat *higher* steps than the Unmanipulated Steps Synthetic Tag group, although the difference is not significant (p-value 0.257). Tag also performs statistically indistinguishably from the Choice group (p-value 0.817, Table 2). Figure A.6 presents evidence that the reason Tag performs well is that manipulation of baseline steps is relatively limited, likely reflecting a cost of manipulation. Moreover, if anything, the manipulation is on net *upwards*. Since all of the step target contracts in the Tag treatment pay the same amount (20 INR), upwards manipulation suggests non-standard preferences. However, it is unclear whether the manipulation results would hold in a scaled-up version of the program, when information about how to "game the system" might spread more widely. We view it as promising that Choice performs roughly as well without similar concerns.

Overall, we view these results as promising for Choice. Choice outperforms the most scalable version of tagging, and performs indistinguishably from the better-performing tagging options which, unlike Choice, may not be scalable in practice.

Medium group from the regression and avoid the duplication process.

[39] To assess robustness of this result to the machine learning procedure used, we also estimate another tag using the same predictor variables but a simpler Lasso-based prediction procedure (described in Appendix C.5); the results are similar, as shown in Figure A.5 with the "Policy variables (Lasso)" Synthetic Tag.

8 Conclusion

This paper highlights the power of mechanism design for personalizing incentives and policies. We focus on screening contracts, which, despite a large theoretical literature, have only been infrequently tested. Relative to a one-size-fits-all contract, we find that personalizing incentives by offering an incentive-compatible choice increases the impact of incentives by 80% without increasing payments. Moreover, Choice is more effective than non-personalized incentives across the full distribution of behavior, first-order stochastically dominating any single non-personalized contract. Choice also compares favorably against personalization based on observables, matching the performance of an optimal Tag that may be infeasible to implement in practice. As in standard mechanism design, sorting is the primary driver of Choice's efficacy: when offered an incentive-compatible menu, many participants prefer the contract that is most effective for them. While non-standard preferences appear to enhance Choice's effectiveness in our specific policy domain, we show that the incentive-compatibility of the menu is nonetheless crucial for Choice's effectiveness, suggesting that choice is likely relevant to a wide range of policy areas.

The implications of our findings are widespread. Similar incentive-compatible menus could be used for other programs incentivizing beneficial behaviors, such as schooling, R&D by firms, or the adoption of eco-friendly technologies. For example, homeowners investing in energy efficiency could choose from incentive-compatible menus of targets, trading off higher targets for higher payments. Incentive-compatible menus could also personalize other types of policies besides incentives. Take unemployment insurance as an example: incentive-compatible choice menus could enable participants to balance the duration of benefits against the payout levels, sorting based on their underlying employability.

Our results open up several potential directions for future work. A first is to test the effectiveness of incentive-compatible menus in these other policy domains (e.g., for personalizing unemployment insurance). A second is to test the effectiveness of more dynamic approaches to Choice. Our approach to Choice was (for simplicity) fundamentally static, allowing participants to choose their contracts only once. However, allowing participants to choose contracts repeatedly over time could further improve performance by allowing participants' choices to adapt to any adjustments in their cost function over time (e.g., due to random shocks, habit formation). A final direction for future work is to evaluate a more optimal choice menu. The personalization mechanism used in this experiment was designed imperfectly; we did not estimate how walking cost functions varied by type to estimate the optimal menu. While our approach substantially improved performance, future work can estimate the further gains possible from implementing a more optimal menu.

References

Adjerid, I., G. Loewenstein, R. Purta, and A. Striegel (2022). Gain-loss incentives and physical activity: the role of choice and wearable health tools. *Management Science 68*, 2642–2667.

Aggarwal, S., R. Dizon-Ross, and A. D. Zucker (2020). Incentivizing behavioral change: The role of time preferences. *NBER Working Paper*, No. 27079.

Alatas, V., A. V. Banerjee, R. Hanna, B. A. Olken, R. Purnamasari, and M. Wai-poi (2016). Self-targeting : Evidence from a field experiment in indonesia abhijit banerjee rema hanna ririn purnamasari matthew wai-poi. *Journal of Political Economy 124*.

Anjana, R. M., R. Pradeepa, A. K. Das, M. Deepa, A. Bhansali, and S. R. J. et al (2014). Physical activity and inactivity patterns in india - results from the ICMR-INDIAB study. *International Journal of Behavioral Nutrition and Physical Activity 11*, 1–11.

Ashraf, N., J. Berry, and J. M. Shapiro (2010). Can higher prices stimulate product use? evidence from a field experiment in zambia. *American Economic Review 100*, 2383–2413.

Ashraf, N., D. S. Karlan, and W. Yin (2006). Tying odysseus to the mast: Evidence from a commitment savings product in the philippines. *The Quarterly Journal of Economics 121*, 635–672. ISBN: 00206.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *Annals of Statistics 47*, 1179–1203.

Athey, S. and S. Wager (2021). Policy learning with observational data. *Econometrica 89*, 133–161.

Bai, L., B. R. Handel, E. Miguel, and G. Rao (2020). Self-control and demand for preventive health: Evidence from hypertension in india. *Review of Economics and Statistics Forthcomin*.

Baicker, K., D. Cutler, and Z. Song (2010). Workplace wellness programs can generate savings. *Health Affairs 29*, 1–8.

Barrera-Osorio, F., M. Bertrand, L. L. Linden, and F. Perez-Calle (2011). Improving the design of conditional cash transfer programs : Evidence from a randomized evaluation in colombia organ donations. *American Economic Journal: Applied Economics 3*, 167–195.

Beaman, L., D. Karlan, B. Thuysbaert, and C. Udry (2014). Self-selection into credit markets: Evidence from agriculture in mali.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*, 608–650.

Ben-Sira, D. and J. M. F. Oliveira (2007). Hypertension in aging: physical activity as primary prevention. *European Review of Aging and Physical Activity 4*, 85–89.

Bergemann, D., J. Shen, Y. Xu, and E. M. Yeh (2012). Mechanism design with limited information: the case of nonlinear pricing. pp. 1–10.

Bhansali, A., V. K. Dhandania, M. Deepa, R. M. Anjana, S. R. Joshi, and P. P. J. et al (2015). Prevalence of and risk factors for hypertension in urban and rural india: The icmr-indiab study. *Journal of Human Hypertension 29*, 204–209.

Björkegren, D., J. E. Blumenstock, and S. Knight (2020). Manipulation-proof machine learning. *arXiv preprint arXiv:2004.03865*.

Bruhn, M. and D. Mckenzie (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics 1*, 200–232.

Bryan, G. T., D. Karlan, and A. Osman (2021). Big loans to small businesses: Predicting winners and losers in an entrepreneurial lending experiment.

Burlig, F., C. Knittel, D. Rapson, M. Reguant, and C. Wolfram (2020). Machine learning from schools about energy efficiency. *Journal of the Association of Environmental and Resource Economists 7*, 1181–1217. Publisher: The University of Chicago Press Chicago, IL.

Carrera, M., H. Royer, M. Stehr, and J. Sydnor (2020). The structure of health incentives: Evidence from a field experiment. *Management Science 66*, 1783–2290.

Cohen, J. and P. Dupas (2010). Free distribution or cost-sharing? evidence from a randomized malaria prevention experiment. *The Quarterly Journal of Economics*, 1–45. Publisher: JSTOR.

Conner, P., L. Einav, A. Finkelstein, P. Persson, and H. L. Williams (2022). Targeting precision medicine: Evidence from prenatal screening.

Deshpande, M. and Y. Li (2019). Who is screened out? application costs and the targeting of disability programs. *SSRN Electronic Journal 11*, 213–248.

Dizon-Ross, R. and A. D. Zucker (2020). Targeting incentive contracts in heterogeneous populations. *AEA RCT Registry*.

Dube, A. (2020). Indians are the least active and second most sleep deprived country in the world, claims fitbit study.

Dubé, J.-P. and S. Misra (2023). Personalized pricing and consumer welfare. *Journal of Political Economy 131*(1), 131–189.

Einav, L., A. Finkelstein, Y. Ji, and N. Mahoney (2022). Voluntary regulation: Evidence from medicare payment reform. *The quarterly journal of economics 137*, 565–618.

Finkelstein, A. and M. J. Notowidigdo (2019). Take-up and targeting: Experimental evidence from snap. *The Quarterly Journal of Economics 134*, 1505–1556.

Gertler, P., S. Higgins, A. Scott, and E. Seira (2019). Increasing financial inclusion and attracting deposits through prize-linked savings. *Unpublished manuscript*.

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2022). Contamination bias in linear regressions. Technical report, National Bureau of Economic Research.

Gupta, R. and C. V. S. Ram (2019). Hypertension epidemiology in india: emerging aspects. *Current opinion in cardiology 34*, 331–341. ISBN: 0000000000000.

Hermalin, B. (2005). Lecture notes for economics.

Howells, L., B. Musaddaq, A. J. McKay, and A. Majeed (2016). Clinical impact of lifestyle interventions for the prevention of diabetes: An overview of systematic reviews. *BMJ Open 6*, 1–17.

Huang, H. and S. Linnemayr (2019). Moving the goalpost closer: Do flexible targets improve the behavioral impact of incentives?

International Diabetes Federation (2019). *IDF Diabetes Atlas* (9 ed.). International Diabetes Federation.

Ito, K., T. Ida, and M. Tanaka (2021). Selection on welfare gains: Experimental evidence from electricity plan choice.

Jack, B. K. (2013). Private information and the allocation of land use subsidies in malawi. *American Economic Journal: Applied Economics 5*, 113–135.

Jack, B. K. and S. Jayachandran (2019). Self-selection into payments for ecosystem services programs. *Proceedings of the National Academy of Sciences 116*, 5326–5333.

Jayachandran, S., J. De Laat, E. F. Lambin, C. Y. Stanton, R. Audy, and N. E. Thomas (2017). Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation. *Science 357*(6348), 267–273.

Johnson, T. and M. Lipscomb (2017). Pricing people into the market: Targeting through mechanism design. *Working paper*.

Jones, D., D. Molitor, and J. Reif (2019). What do workplace wellness programs do? evidence from the illinois workplace wellness study. *The Quarterly Journal of Economics 134*(4), 1747–1791.

Kitagawa, T. and A. Tetenov (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica 86*, 591–616.

Kumar, P., R. Patel, T. Muhammad, and S. Srivastava (2022, 1). Does engagement in frequent physical activity improve diabetes mellitus among older adults in india? a propensity score matching approach. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews 16*. Publisher: Elsevier Ltd.

Kyu, H. H., V. F. Bachman, L. T. Alexander, J. E. Mumford, A. Afshin, and K. E. et al (2016). Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events. *BMJ (Online) 354*, 1–10.

Leslie, P. (2004). Price discrimination in broadway theater. *The RAND Journal of Economics 35*, 520.

Levitt, S. D., J. A. List, S. Neckermann, and D. Nelson (2016). Quantity discounts on a virtual good: The results of a massive pricing experiment at king digital entertainment. *Proceedings of the National Academy of Sciences of the United States of America 113*, 7323–7328.

Loprinzi, P. D. (2015). Frequency of moderate-to-vigorous physical activity (mvpa) is a greater predictor of systemic inflammation than total weekly volume of mvpa: Implications for physical activity promotion. *Physiology and Behavior 141*, 46–50. Publisher: Elsevier Inc.

Maskin, E. and J. Riley (1984). Monopoly with incomplete information. *The RAND Journal of Economics 15*, 171–196.

Mitchell, M. S., S. L. Orstad, A. Biswas, P. I. Oh, M. Jay, and M. T. Pakosh et al (2020). Financial incentives for physical activity in adults: systematic review and meta-analysis. *British Journal of Sports Medicine 54*(21), 1259–1268.

Mortimer, J. H. (2007). Price discrimination, copyright law, and technological innovation: Evidence from the introduction of dvds. *Quarterly Journal of Economics 122*, 1307–1350.

Muralidharan, K., M. Romero, and K. Wüthrich (2019). Factorial designs, model selection, and (incorrect) inference in randomized experiments. Technical report, National Bureau of Economic Research.

Mussa, M. and S. Rosen (1978, 8). Monopoly and product quality. *Journal of Economic Theory 18*, 301–317.

Myers, J. (2008). The health benefits and economics of physical activity. *Current Sports Medicine Reports 7*, 314–316.

Rehill, P. (2022). Policy learning for many outcomes of interest: Combining optimal policy trees with multi-objective bayesian optimisation. *arXiv preprint arXiv:2212.06312*.

Samitz, G., M. Egger, and M. Zwahlen (2011). Domains of physical activity and all-cause mortality: Systematic review and dose-response meta-analysis of cohort studies. *International Journal of Epidemiology 40*, 1382–1400.

Spence, A. M. (1980). Multi-product quantity-dependent prices and profitability constraints. *The Review of Economic Studies 47*, 821–841.

Stiglitz, J. E. (1977). Monopoly, non-linear pricing and imperfect information: the insurance market. *The Review of Economic Studies 44*, 407–430.

Stole, L. (2001). Lectures on the theory of contracts and organizations. *Unpublished monograph*.

Tandon, N., R. M. Anjana, V. Mohan, T. Kaur, A. Afshin, and K. O. et al (2018). The increasing burden of diabetes and variations among the states of india: the global burden of disease study 1990–2016. *The Lancet Global Health 6*, e1352–e1362.

Tirole, J. (1988). *The theory of industrial organization*. MIT press.

Varian, H. R. (1989). Price discrimination. *Handbook of industrial organization 1*, 597–654.

Warburton, D. E. R., C. W. Nicol, and S. S. D. Bredin (2006). Health benefits of physical activity: The evidence. *Canadian Medical Association Journal 174*, 801–809.

Whitehead, D. and G. Russell (2004). How effective are health education programmes - resistance, reactance, rationality and risk? recommendations for effective practice. *International Journal of Nursing Studies 41*, 163–172.

Wilson, R. (1989). Efficient and competitive rationing. *Econometrica: Journal of the Econometric Society*, 1–40.

Woerner, A., G. Romagnoli, B. M. Probst, N. Bartmann, J. N. Cloughesy, and J. W. Lindemans (2021). Should individuals choose their own incentives? evidence from a mindfulness meditation intervention. *Evidence from a Mindfulness Meditation Intervention*.

World Health Organization (2009). Global health risks, mortality and burden of disease attributable to selected major risks.

World Health Organization (2013). Global action plan for the prevention and control of noncommunicable disease.

World Health Organization (2018). Saving lives, spending less: a strategic response to noncommunicable diseases.

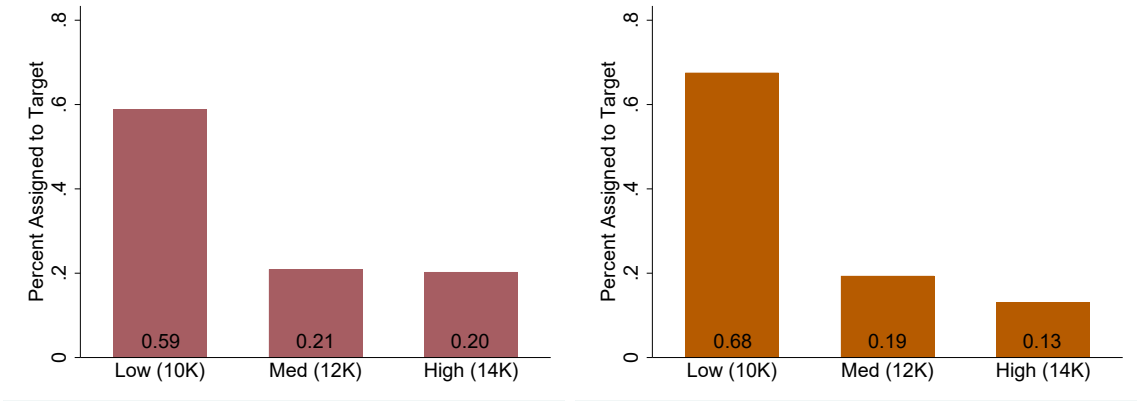
World Health Organization (2022). Global health estimates: Leading causes of death.

Appendices for Online Publication

This section contains all tables and figures labeled with an A at the beginning (e.g., Table A.1), as well as Appendices B - D. The Online Supplement is a separate document and can be found at: https://faculty.chicagobooth.edu/-/media/faculty/rebecca-dizon-ross/research/customizingincentives_onlinesupp.pdf

A Appendix Tables and Figures

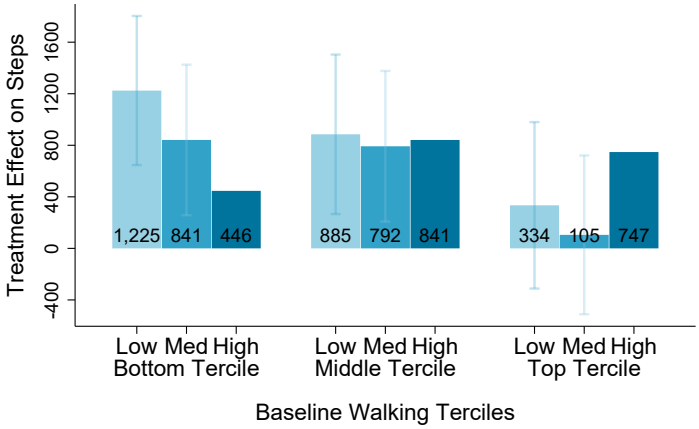
Appendix Figure A.1: Step Target Distribution in Choice



(a) Step Target Distribution in Choice (b) Step Target Distribution in Flat Choice

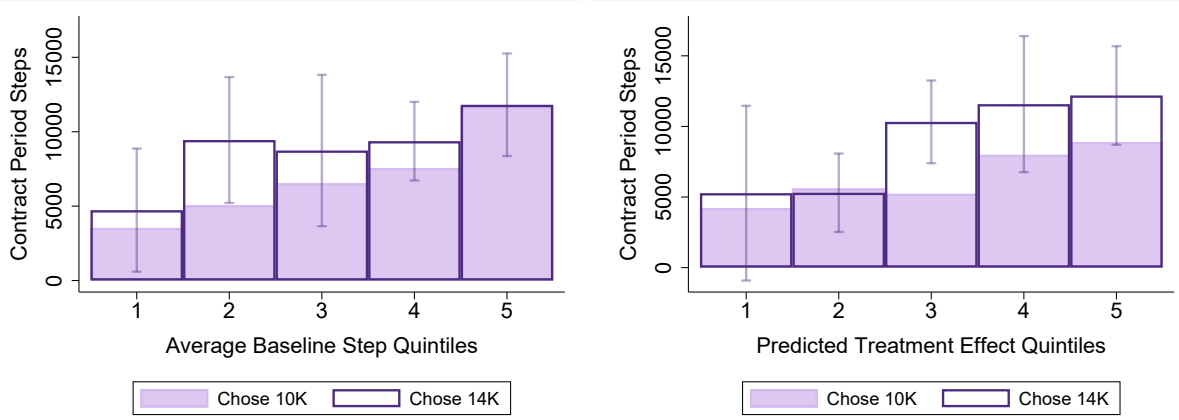
Notes: Panel (a) displays the percentage of Choice participants who chose each of the three targets from the Base Menu. Panel (b) displays the percentage of Flat Choice participants who choose each of the three targets from the Flat Menu.

Appendix Figure A.2: High Step Targets Generate More Steps from Higher Walkers



Notes: Figure shows the treatment effects of the Fixed groups relative to Monitoring for each baseline step tercile. The 95% confidence intervals are relative to Fixed High, controlling for experiment phase, time between Baseline and Choice surveys, the Nudge, year-month fixed effects, and controls selected by double-Lasso for the middle tercile from the controls in col 1 of Table A.5.

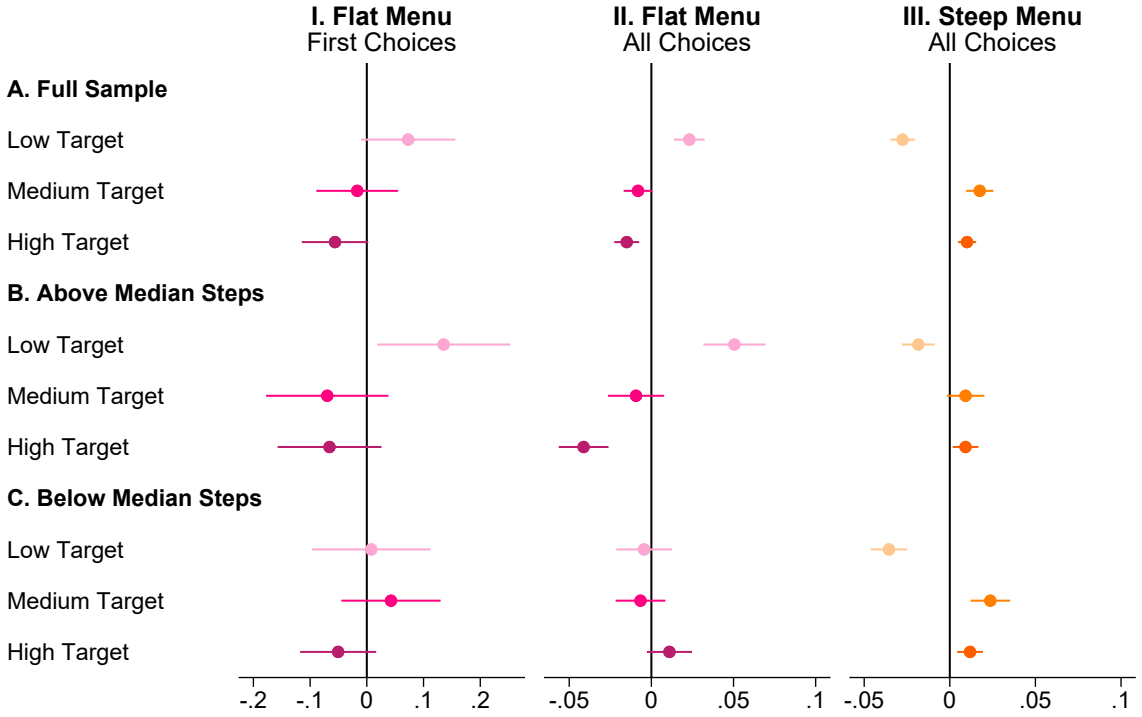
Appendix Figure A.3: In the Monitoring Group, Those Who Choose Higher Targets Have Higher Contract-Period Steps, Conditional on Baseline Steps or Predicted Treatment Effects



(a) Contract Period Steps, by Baseline Steps (b) Contract Period Steps, by Predicted TE

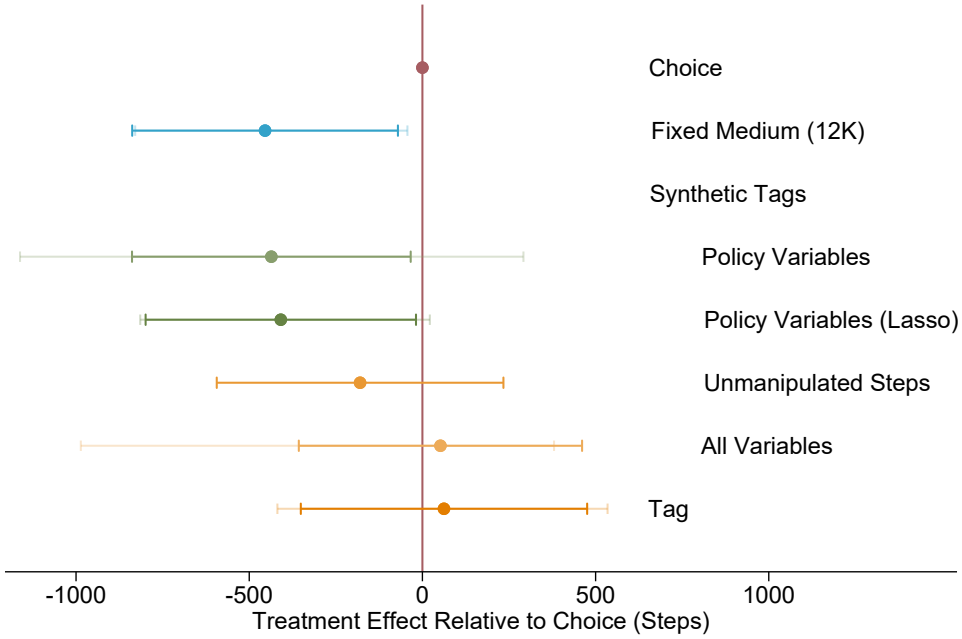
Notes: The figure shows contract period walking in the Monitoring group, separately for those who chose the High (14K) target (in the shaded bars) and those who chose the Low (10K) target (in the outlined bars) from the Base Menu during the Choice survey. Panel (a) further splits the sample by quintiles of baseline walking, while panel (b) splits it by quintiles of the predicted treatment effect of Fixed High vs Fixed Low. Confidence interval bars represent tests of equality between contract period walking among those who chose the High and Low targets, controlling for experiment phase and time between Baseline and Choice surveys.

Appendix Figure A.4: Participants Choose Lower Targets on the Flat Menu, Especially Those with Higher Baseline Steps



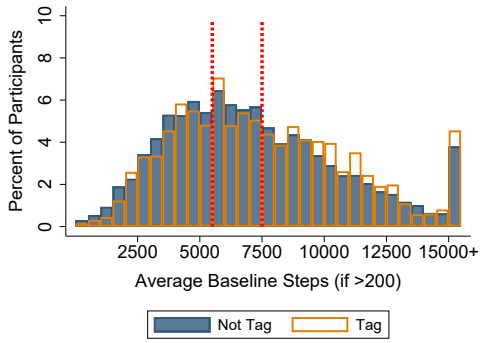
Notes: Figure shows the difference in (and 95% confidence intervals for) the fraction of participants choosing each step target on the Flat Menu (sub-graphs I and II) and the Steep Menu (sub-graph III), both compared to the Base Menu. Sub-graph I limits to choices from the first menu shown; sub-graphs II and III include the full sample. Flat Menu choices are limited to phase 3—the only phase in which choices on the menu were "incentive-compatible." The sample includes the Choice, Monitoring, Flat Choice, and Fixed groups, excluding those who received the Nudge.

Appendix Figure A.5: Choice Performs Well Relative to Personalizing With Observables



Notes: The figure displays the treatment effects of various tag assignment mechanisms relative to Choice. The Synthetic Tag groups include individuals from the Fixed groups whose randomly assigned target matches the target they would have been assigned under the respective tag mechanism. The figure displays both Gaussian (darker colored) and bootstrapped (lighter colored) 95% confidence intervals for all groups with the exception of Unmanipulated Steps (for which the tag assignment rule does not depend on data). Estimates come from a weighted regression where each Synthetic Tag observation is weighted by the inverse of the probability of assignment to a given step target within the Fixed groups in its experimental phase. (All other observations receive a weight of 1.) Choice is statistically indistinguishable from all of the tags except the Policy Variable (Gaussian p-value = 0.017) and Policy Variables (Lasso) (Gaussian p-value = 0.020) Synthetic Tag groups. Controls are the same as in Table 2. See Table A.14 for the table version of the results.

Appendix Figure A.6: Tag Group Does Not Manipulate Baseline Steps Downwards



(a) Tag Group, If Anything, Increases Baseline Steps

(b) Increased Steps Raises Step Targets

Notes: The figure shows how being assigned to the Tag group influences Baseline steps. Panel (a) shows the distribution of average baseline steps among the Tag group compared to all other groups (excluding Baseline Choice, for whom baseline steps were also endogenous to treatment). Panel (b) shows how step target assignment in the Tag group differs from how target assignment would have looked in the Not Tag group if we had applied the Tag target assignment algorithm (Table C.2) to *unmanipulated* The confidence interval bars represent tests of equality between the likelihood individuals are assigned to each step target at the 95% confidence level. Regressions includ controls selected by double-Lasso for the Medium (12K) Target from the list of potential controls in column 3 of Table A.5; the selected controls are then included in the regressions for the Low (10K) and High (14K) Targets. We also control for experiment phase, time between Baseline and Choice survey, and year-month fixed effects for the date of the Baseline survey.

Appendix Table A.1: There Is No Significant Difference in Pre-Contract-Launch Withdrawals in the Tag or Baseline Choice Groups

Omitted Group:	Not Tag or Baseline Choice	
	Withdrew Before Contract Launch	Withdrew Before Contract Period
	(1)	(2)
Tag	0.0149	0.0126
	[0.0102]	[0.0121]
Baseline Choice	0.00265	0.0144
	[0.0120]	[0.0152]
Not Tag or Baseline Choice Mean	0.11	0.19
# Individuals	7,893	7,893
Tag	1,141	1,141
Baseline Choice	831	831
Not Tag or Baseline Choice Mean	5,921	5,921

Notes: This table compares rates of withdrawal prior to contract launch between Tag, Baseline Choice, and all other groups pooled. The sample is limited to those who completed the Baseline survey up to the point that treatment was revealed to Tag. Controls include experiment phase, time between Baseline and Choice surveys, and year-month fixed effects for the date of the Baseline survey. In addition, column-specific controls are selected by double-Lasso for each column from the list of controls in Table A.5 column 3. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.2: Enrollment Statistics

Total screened: 94,421

Total eligible: 22,577

	# Individuals	% of total eligible
	(1)	(2)
Successfully contacted	19,438	86%
Interested in enrolling	13,302	59%
Completed Baseline survey	7,920	35%
Completed Choice survey up to contract launch	6,917	31%
Started contract period	6,751	30%
Completed Endline survey	6,714	30%

Notes: This table reports statistics on how many participants dropped out of the study at each stage of the experiment design. Critically, there is extremely limited dropout following contract launch in the Choice survey, when the majority of the treatment groups were assigned. Note that participants could elect to participate in the Endline survey even if they withdraw from the rest of the program.
The number of participants is slightly off from elsewhere in the paper due to the inclusion of an extra treatment group. We assigned very few people (less than 50) to their menu choice from the Steep Menu in order to make choices on this menu incentive-compatible. We omit this group from all of our analyses, however they are included here since they were enrolled and screened with the rest of the sample.

Appendix Table A.3: There Is Limited Differential Attrition

Omitted Group:	Choice		
		Missing Day-Level Data During Contract Period	
	Individual Missing Data for Full Contract Period	No Pedometer Data (e.g. Sync Issue)	Did Not Wear Pedometer
	(1)	(2)	(3)
Fixed Low	0.00475	-0.00523	-0.0195
	[0.0119]	[0.00730]	[0.0132]
Fixed Medium	0.00602	-0.00144	0.0118
	[0.0109]	[0.00671]	[0.0125]
Fixed High	0.0112	-0.00452	0.0164
	[0.0119]	[0.00726]	[0.0133]
Tag	0.00504	-0.0149**	0.00174
	[0.0110]	[0.00609]	[0.0124]
Flat Choice	0.0238	0.00709	-0.00152
	[0.0167]	[0.0104]	[0.0158]
Baseline Choice	0.0244*	-0.00207	0.00304
	[0.0142]	[0.00798]	[0.0136]
Choice + Nudge	0.0143	-0.00809	-0.00203
	[0.0133]	[0.00842]	[0.0175]
Monitoring	0.0216	-0.0151*	0.000208
	[0.0197]	[0.00897]	[0.0223]
Choice Mean	0.08	0.04	0.17
p-value vs Fixed Medium			
Fixed Low	0.893	0.524	0.007
Fixed High	0.585	0.614	0.710
Tag	0.923	0.016	0.410
Flat Chocie	0.284	0.414	0.407
Baseline Choice	0.184	0.936	0.521
Choice + Nudge	0.389	0.313	0.355
Monitoring	0.415	0.112	0.598
p-value vs Monitoring			
Fixed Low	0.391	0.271	0.380
Fixed High	0.598	0.237	0.474
Tag	0.398	0.975	0.946
Flat Choice	0.921	0.057	0.943
Baseline Choice	0.892	0.175	0.902
Choice + Nudge	0.720	0.472	0.929
p-value Fixed High vs Fixed Low	0.546	0.914	0.006
# Observations	6,882	178,752	172,961
# Individuals	6,882	6,384	6,384
Choice	970	892	892
Fixed Low	826	778	778
Fixed Medium	1,274	1,210	1,210
Fixed High	847	796	796
Tag	990	928	928
Flat Choice	509	439	439
Baseline Choice	719	631	631
Choice + Nudge	540	523	523
Monitoring	207	187	187

Notes: This table shows the causes of missing data during the contract period. The omitted group is the Choice group. The dependent variable in column 1 is a person-level indicator for if the respondent is missing all of their contract period data for any reason. In column 2, it is a person-day level indicator for if the respondent is missing data on a given day for any reason, conditional on having data from the pedometer at some point during the contract period. In column 3, it is a person-day level indicator for if the respondent did not wear their pedometer on a given day (had fewer than 200 steps), conditional on having data from the pedometer. Cols 2 and 3 cluster standard errors at the individual level. The sample includes all treatment groups. All columns include controls for experiment phase, time between Baseline and Choice surveys, the Nudge treatment, and year-month fixed effects for either Baseline survey date (in col 1) or day (in cols 2 and 3). In addition, column-specific controls are selected by double-Lasso for each column from the list of controls in Table A.5 column 3 (for col 1) and column 1 (for cols 2 and 3). The analysis conditions on being in our main analysis sample that was present at the Contract Launch. Significance levels: * 10%, ** 5%, *** 1%.

To account for the small imbalances in the table above, we also report Lee bounds for the Monitoring, Tag and Baseline Choice groups relative to Choice. For Monitoring vs Choice, the lower bound is 693 (standard error 369) and the upper bound is 1292 (standard error 450). For Tag vs Choice, the lower bound is -201 (standard error 293) and the upper bound is 210 (standard error 364). For Baseline Choice vs Choice, the lower bound is -221 (standard error 364) and the upper bound is 112 (standard error 341).

Appendix Table A.4: Baseline Summary Statistics in Full Sample and by Treatment Group

	Full Sample		Monitoring	Fixed Low	Fixed Med	Fixed High	Choice	Tag	Flat Choice	Choice + Nudge	Baseline Choice	# Obs.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Mean	SD	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Count
A. Demographics												
Age	49.38	8.77	49.22	49.24	49.38	48.87	49.75	49.43	49.67	48.62	49.99	6882
Female	0.37	0.48	0.39	0.36	0.37	0.37	0.35	0.36	0.37	0.40	0.35	6882
Married	0.91	0.28	0.91	0.92	0.90	0.92	0.91	0.93	0.91	0.91	0.91	6882
Household Size	3.74	1.51	3.71	3.82	3.75	3.81	3.69	3.72	3.64	3.88	3.60	6882
Monthly Income/Capita (INR)	5516	7302	5104	5521	5971	5165	5392	5353	6100	5148	5555	5111
Wealth Index	0.04	0.48	0.02	0.05	0.05	0.05	0.01	0.04	0.03	0.05	0.00	6882
Any Secondary Education	0.58	0.49	0.57	0.57	0.59	0.57	0.56	0.58	0.63	0.56	0.59	6882
Participating in Labor Force	0.80	0.40	0.81	0.80	0.79	0.80	0.80	0.79	0.79	0.80	0.80	6882
B. Health statistics												
Diagnosed Diabetic	0.31	0.46	0.31	0.33	0.33	0.30	0.32	0.32	0.28	0.32	0.24	6882
Diagnosed Hypertensive	0.32	0.47	0.38	0.34	0.29	0.29	0.34	0.30	0.39	0.24	0.39	6882
Diastolic BP	92	12.29	93	93	92	91	93	92	94	91	94	6840
Systolic BP	138	20.33	139	139	137	137	140	138	141	135	142	6840
BMI	26	4.59	26	26	27	27	26	26	27	26	26	6858
Weight (kg)	68	12.75	67	68	68	68	69	68	68	68	67	6870
Height (cm)	160	9.11	161	160	160	160	161	160	160	160	160	6865
Waist Circumference (cm)	95	10.31	94	95	95	95	95	95	95	94	94	6860
Mental Health Index	-0.03	0.67	0.00	-0.08	-0.05	0.00	-0.02	-0.06	0.01	-0.07	0.01	6882
Days of Exercise in Past Week	1.40	2.61	1.43	1.29	1.36	1.26	1.49	1.42	1.74	1.24	1.44	6882
Exercised Yesterday	0.23	0.42	0.22	0.21	0.23	0.19	0.24	0.24	0.28	0.19	0.24	6882
C. Baseline Walking												
Baseline Steps	7230	3636	7193	7025	7254	7296	7335		7106	7323		6792
Predicted Baseline Steps	7121	1108	7097	7122	7148	7124	7114	7166	6974	7209	7073	6882
p-values for joint orthogonality of covariates versus:												
Choice			0.955	0.652	0.155	0.281		0.350	0.117	0.267	0.107	
Fixed Med			0.556	0.173		0.021	0.155	0.439	0.598	0.079	0.555	
Monitoring				0.970	0.556	0.766	0.955	0.768	0.828	0.995	0.600	
Sample size												
Number of individuals	6,882		207	826	1,274	847	970	990	509	540	719	
Percent of sample	100.0		3.0	12.0	18.5	12.3	14.1	14.4	7.4	7.8	10.4	
Number of ind. with ped data	6,384		187	778	1,210	796	892	928	439	523	631	

Notes: This table shows summary statistics for characteristics measured at Baseline for all participants in our main analysis sample. The wealth index is the simple average of the following standardized variables: number of scooters owned, number of cars owned, number of computers owned, number of smartphones owned, number of not-smart phones owned, number of rooms in house, a home-ownership dummy, whether the home has a private water connection, and whether the participant has a bank account. BP is blood pressure, and BMI is body mass index. The mental health index is a simple average of answers to seven mental health questions from RAND's 36-Item Short Form Survey, standardized relative to the Monitoring group.

Baseline steps represent the average steps taken across the first 6 days after the Baseline survey, conditioning on days when the participant wore the pedometer (steps >200). Because baseline step data were collected after the Tag and Baseline Choice groups were told their treatment, Baseline Steps exclude the Tag group. The F-statistics test the joint orthogonality of all characteristics to treatment assignment relative to the Choice, Fixed Medium, or Monitoring group (the primary three comparison groups in our analyses), holding constant the experiment phase and time between Choice and Baseline surveys. Each F-statistic is obtained by running a column-specific regression. Cols 8 and 11 include predicted baseline steps in the regression; all other columns include baseline steps.

"Number of ind. with ped data" shows the number of participants in our analysis sample for whom we have any pedometer data during the contract period. This is lower than "number of individuals" due to a combination of participants withdrawing from the program and problems syncing steps from the pedometers. Column 1 of Table A.3 shows that whether participants have pedometer data is balanced across our main treatment groups.

Appendix Table A.5: Variables Used in Double-Lasso Selection Method

	Resp × Day Specifications		Respondent-Level Specifications	
	Base Specification Controls	Robustness to Using Actual Steps	Base Specification Controls	Robustness to Using Actual Steps
	(1)	(2)	(3)	(4)
A. Self-Reported at Baseline				
Gender	X	X	X	X
Age	X	X	X	X
Diagnosed with diabetes	X	X	X	X
Diagnosed with hypertension	X	X	X	X
Excersized yesterday	X	X	X	X
Days exercised last week	X	X	X	X
Mental health index	X	X	X	X
Household size	X	X	X	X
Household income per capita	X	X	X	X
Participating in labor force	X	X	X	X
Above median education	X	X	X	X
Married	X	X	X	X
Number of scooters owned	X	X	X	X
Number of cars owned	X	X	X	X
Number of computers owned	X	X	X	X
Number of smartphones owned	X	X	X	X
Number of mobile phones owned	X	X	X	X
Number of rooms in home	X	X	X	X
Owns home	X	X	X	X
Home has running water	X	X	X	X
Has bank account	X	X	X	X
B. Measured at Baseline				
Weight	X	X	X	X
Height	X	X	X	X
BMI	X	X	X	X
Systolic BP	X	X	X	X
Diastolic BP	X	X	X	X
Waist circumference	X	X	X	X
C. Estimated Using Baseline Variables				
Average predicted baseline steps	X		X	
Average predicted baseline steps (deciles)	X		X	
D. Measured During Pre-contract Period				
Average baseline steps (> 200)		X		X
Average baseline steps (deciles)		X		X
E. Covid and Temporal Indicators				
Day during Covid lockdown	X	X		
Contract period overlapped with Covid lockdown			X	X
Day of week	X	X		
Contract period week	X	X		
F. Other Variables				
Dummies for Missing	X	X	X	X
G. Always Included Controls				
Experiment phase	X	X	X	X
Choice survey timing	X	X	X	X
Year-Month fixed effects	X	X		
Baseline Survey year-month fixed effects			X	X

Notes: This table lists the variables from which we selected covariates using the double-Lasso selection method of Belloni et al. (2014). The variables in Panel A were self-reported at the Baseline survey, or are indices of standardized self-reported variables. The variables in Panel B were directly measured at Baseline. The variables in Panel C are predictions from a cross-validated Lasso model of pre-contract period walking (see Appendix Section C.2 for more information). The variables in Panel D are measured during the pre-contract period. The variables in panel E are a variety of temporal controls such as Covid lockdown controls. Panel F shows that we included dummies for any missing values. Panel G shows the variables that we required Lasso to select (that is, partialled out).

Appendix Table A.6: Choice Does Not Meaningfully Increase Payments

Omitted Group:	Fixed Medium
Dependent Variable:	Daily Payments
	(1)
Choice	0.46
	[0.29]
Fixed Low	2.23***
	[0.31]
Fixed High	-1.41***
	[0.29]
Tag	0.21
	[0.31]
Flat Choice	0.96**
	[0.38]
Baseline Choice	0.38
	[0.32]
Choice + Nudge	-0.10
	[0.35]
Monitoring	-5.47***
	[0.23]
Fixed Medium Mean	5.87
p-value vs Choice	
Fixed Low	0.000
Fixed High	0.000
Tag	0.397
Flat Choice	0.171
BL choice	0.785
Choice + Nudge	0.164
Monitoring	0.000
p-value vs Monitoring	
Fixed Low	0.000
Fixed High	0.000
Tag	0.000
Flat Choice	0.000
Choice + Nudge	0.000
p-value Fixed High vs Fixed Low	0.000
# Observations	190,420
# Individuals	6,801
Choice	957
Fixed Low	819
Fixed Medium	1,263
Fixed High	840
Tag	983
Flat Choice	496
BL Choice	701
Choice + Nudge	540
Monitoring	202

Notes: The dependent variable is daily payments. The sample includes the Monitoring, Tag, Choice, Flat Choice, Fixed, and Baseline Choice groups. The omitted category is the Fixed Medium group. Controls are selected by double-Lasso from the controls shown in column 1 of Table A.5.We also control for experiment phase, time between Baseline and Choice survey, receiving the Nudge treatment, and year-month fixed effects. Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.7: Choice Increases Steps and Decreases Payments Relative to the "Reweighted Fixed" Group

Omitted Group:	Reweighted Fixed	
Dependent Variable:	Daily Steps	Daily Payments
	(1)	(2)
Choice	342.8*	-0.463*
	[183.8]	[0.278]
Monitoring	-584.6*	-6.367***
	[321.0]	[0.209]
Reweighted Fixed Mean	7,740	6.65
# Observations	104,600	114,263
# Individuals	3,863	4,081
Reweighted Fixed	2,784	2,922
Choice	892	957
Monitoring	187	202

Notes: The dependent variable in column 1 is daily steps measured using the contract-period pedometer data. In column 2, it is daily payments during the contract period. The sample includes the Choice and Monitoring groups, along with the Fixed Low, Medium, and High groups reweighted in the proportion realized by the Choice group ("Reweighted Fixed" group). Specifically, each Fixed group observation receives a weight of $\frac{c_{sk}}{f_{sk}}$, where f_{sk} and c_{sk} are the fractions of the pooled Fixed and Choice groups, respectively, assigned to step target s ($s \in \{Low, Med, High\}$) in experiment phase k. (All Monitoring and Choice observations simply have a weight of 1.) Controls are selected by double-Lasso from the list of controls shown in column 1 of A.5 separately for each column.We also control for experiment phase, time between Baseline and Choice survey, receiving the Nudge treatment, and year-month fixed effects. Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.8: Choice's Improvement Robust across Specifications

Omitted Group:	Fixed Medium					
Dependent Variable:	Daily Steps					
Robustness to:		Controls		Dep Var	Sample	
	Base Spec	Basic	Actual Steps	Non-Winsorized	Phases 1 & 2	One at a Time
	(1)	(2)	(3)	(4)	(5)	(6)
Choice	414**	436**	383**	444**	501*	518**
	[202]	[210]	[176]	[207]	[275]	[203]
Fixed Low	93	61	237	74	78	61
	[185]	[191]	[161]	[187]	[216]	[61]
Fixed High	173	171	156	179	180	195
	[208]	[215]	[178]	[212]	[244]	[195]
Tag	463**	484**		504**	474**	538***
	[205]	[213]		[212]	[241]	[538]
Flat Choice	98	133	34	89		63
	[252]	[266]	[222]	[254]		[63]
Baseline Choice	343	387*		360	517	321
	[225]	[234]		[230]	[368]	[321]
Choice + Nudge	80	25	134	94	76	43
	[239]	[247]	[205]	[247]	[245]	[43]
Monitoring	-533	-411	-444	-534	-838*	-585*
	[332]	[347]	[281]	[339]	[463]	[-585]
Fixed Medium Mean	7,720	7,720	7,720	7,770	7,895	7,720
p-value vs Choice						
Fixed Low	0.123	0.084	0.425	0.083	0.148	
Fixed High	0.287	0.262	0.243	0.253	0.302	
Tag	0.817	0.825		0.782	0.924	
Flat Choice	0.199	0.243	0.102	0.154		
BL choice	0.748	0.832		0.714	0.967	
Choice + Nudge	0.239	0.164	0.303	0.233	0.211	
Monitoring	0.005	0.016	0.004	0.005	0.007	
p-value vs Monitoring						
Fixed Low	0.064	0.180	0.017	0.077	0.053	
Fixed High	0.044	0.110	0.041	0.046	0.036	
Tag	0.004	0.012		0.003	0.006	
Flat Choice	0.084	0.155	0.119	0.093		
Choice + Nudge	0.109	0.274	0.071	0.109	0.064	
p-value Fixed High vs Fixed Low	0.711	0.621	0.660	0.635	0.695	
# Observations	172,961	172,961	130,571	172,961	109,380	172,961
# Individuals	6,384	6,384	4,825	6,384	4,008	6,384
Controls						
Predicted Steps	Yes	No	No	Yes	Yes	Yes
Steps	No	No	Yes	No	No	No
Demographics	Yes	No	Yes	Yes	Yes	Yes
Year-Month FEs	Yes	No	Yes	Yes	Yes	Yes
Experimental	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Treatment group sample sizes, columns 1, 2, 4, and 6: Choice: 892; Fixed 10K: 778; Fixed 12K: 1,210; Fixed 14K: 796; Tag: 928; Flat CHoice: 439; Baseline Choice: 631; Choice + Nudge: 523; Monitoring: 187. Column 3 is the same as column 1 but excludes the Tag and Baseline Choice groups. Column 5: Choice: 415; Fixed 10K: 552; Fixed 12K: 979; Fixed 14K: 576; Tag: 677; Baseline Choice: 207; Choice + Nudge: 523; Monitoring: 79.
The dependent variable is daily steps measured using the contract-period pedometer data. Column 1 is the same as column 1 of table 2. Columns 2–3 and 4 show robustness to different sets of controls and to not winsorizing the outcome variable, respectively. Columns 5–6 show robustness to different samples. Column 5 is limited to those who were enrolled during Phase 1 or 2 of our experiment, excluding those from Phase 3 who were enrolled after we had met our enrollment target specified in our AEA registry. Column 6 shows robustness to using the "one-at-a-time" estimator from Goldsmith-Pinkham et al. (2022) which simply re-estimates the effect of each treatment relative to Fixed Medium in a sample that only includes those two groups. The sample includes the Fixed, Monitoring, Choice, Tag, Flat Choice, Choice + Nudge, and Baseline Choice groups. The omitted category in all columns is the Fixed Medium group. All columns include controls for experiment phase, time between Baseline and Choice surveys, and receiving the Nudge treatment. Year-Month fixed effects are included in all columns other than col 2. Additional controls are selected by double-Lasso from the list of controls shown in column 1 of Table A.5, except for column 3 which selects from the list of controls shown in column 2 of Table A.5. Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.9: The Low and High Targets Perform Worse than Choice at the Top and Bottom of the Distribution, Respectively

Omitted Group:	Choice		
Dependent Variable:	Individual-Average Steps		
Percentile:	25	50	75
	(1)	(2)	(3)
Fixed Low (10K)	-269	-207	-711**
	[267]	[309]	[293]
Fixed Medium (12K)	-514*	-410	-341
	[264]	[296]	[296]
Fixed High (14K)	-717***	-778**	-38
	[250]	[315]	[420]
Monitoring	-1238***	-1298***	-1411***
	[409]	[469]	[525]
Choice Quantiles	4,372	7,640	11,014
p-val Fixed Low vs. Fixed High	0.068	0.069	0.091
# Individuals	3,863	3,863	3,863
Fixed Low	778	778	778
Fixed Medium	1,210	1,210	1,210
Fixed High	796	796	796
Monitoring	187	187	187
Choice	892	892	892

Notes: Table shows quantile regressions of individual-level contract period steps averaged across the contract period. The sample includes all three Fixed target groups, along with Monitoring and Choice (the omitted group). All columns control for experiment phase, time between Baseline and Choice surveys, receiving the Nudge treatment, and year-month fixed effects for the date of the Baseline survey. In addition, since there is no double-Lass command for quantile regression, each column includes Lasso-selected controls selected for an OLS regression with an indicator that the participant's steps were at least the 25th percentile, 50th percentile, or 75th percentile (for columns 1, 2, and 3, respectively). Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.10: Higher Step Targets Increase Steps (But Not Payments) More for Higher Walkers

Dependent Variable:	Daily Steps	Daily Payments
	(1)	(2)
Step Target (1,000s) ×	41.0***	-0.013
Baseline Steps (1,000s)	[14.9]	[0.021]
Baseline Steps (1,000s)	136.2	0.99***
	[181.1]	[0.26]
Step Target (1,000s)	-306.9***	-0.87***
	[111.2]	[0.16]
# Observations	75,520	81,811
# Individuals	2,784	2,922
Fixed Low	778	819
Fixed Medium	1,210	1,263
Fixed High	796	840

Notes: This table shows the interaction of baseline steps (in 1000s) with assigned step target assignment (in 1000s). The dependent variable in column 1 is daily steps and the dependent variable in column 2 is daily payments. The sample includes the Fixed Target groups only. Controls are selected separately for each column by double-Lasso from the list of controls in Table A.5 column 2 (with the exception of average pre-contract period steps (deciles), which are excluded). We also control for experiment phase, time between Baseline and Choice surveys, receiving the Nudge treatment, and year-month fixed effects. Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.11: Both Baseline Steps and Predicted Treatment Effects Predict Choices

Dependent Variable:	Chosen Step Target (Steps)		
	(1)	(2)	(3)
Baseline Steps	0.182***		0.210***
	[0.0123]		[0.0147]
Predicted Treatment Effect		4.757***	-2.060**
		[0.746]	[0.824]
# Individuals	970	948	948

Notes: This table shows the correlation between choices on the Base Menu and both baseline walking and predicted treatment effects. Predicted treatment effects are the predicted effect of the 14K target relative to the 10K target, as generated by the causal forest methodology of Athey et al. (2019). The dependent variable is a continuous measure (in 1000s) of the step target chosen on the Base Menu. The sample is the Choice group only. All columns control for experiment phase, time between Baseline and Choice surveys, and year-month fixed effects for the date of the Baseline survey. Robust standard errors are in brackets. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.12: Baseline Steps Are the Most Important Predictor in the Causal Forest

Variable name	Included in Policy Variable Prediction?	Importance Value
	(1)	(2)
Baseline steps	No	0.17
Age	Yes	0.11
Systolic BP	Yes	0.10
Weight (kg)	Yes	0.09
Mental health index	No	0.08
Diastolic BP	Yes	0.07
Waist circumference (cm)	Yes	0.05
Height (cm)	Yes	0.05
BMI	Yes	0.04
Female	Yes	0.02
Above median education level	Yes	0.01
Number of smartphones owned	No	0.01
Diagnosed diabetic	Yes	0.01
Owns home	No	0.01
Home has running water	No	0.01
Number of rooms in home	No	0.01
Number of mobilephones owned	No	0.01
Household size	Yes	0.01
Dianosed hypertensive	Yes	0.00
Number of scooters owned	No	0.00
Participating in labor force	No	0.00
Married	Yes	0.00
Number of cars owned	No	0.00
Number of computers owned	No	0.00
Has bank account	No	0.00
Mobile balance	No	0.00

Notes: This table shows the list of variables used in the multi-arm causal forest for predicting the optimal treatment for each participant. The importance value indicates how frequently the trees in the causal forest split on each variable. The list includes all variables from panels A, B, and D in Table A.5.

Appendix Table A.13: Higher Step Targets Increase Steps More for Those Who Chose Higher Step Targets on the Choice Menu

Dependent Variable:	Daily Steps	Daily Payments
	(1)	(2)
Assigned Target (1,000s) ×	94.7***	0.076
Chosen Target (1,000s)	[36.2]	[0.052]
Chosen Target (1,000s)	-304.9	0.39
	[433.5]	[0.64]
Assigned Target (1,000s)	-1055.8***	-1.80***
	[395.1]	[0.58]
# Observations	75,520	81,811
# Individuals	2,784	2,922
Fixed Low	778	819
Fixed Medium	1,210	1,263
Fixed High	796	840

Notes: This table shows the interaction of chosen step targets (in 1000s) with assigned step target assignment (in 1000s). Chosen step targets are the respondent's choice on the Base Menu. Controls are selected separately for each column by double-Lasso from the list of controls in Table A.5 column 3. We also control for experiment phase, time between Baseline and Choice surveys, receiving the Nudge treatment, and year-month fixed effects for the date of the Baseline survey. Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.14: Choice Performs Well Relative to Tagging

Omitted Group:	Choice			
Dependent Variable:	Daily Steps			
Synthetic Tag Type:	Policy Variables	Policy Variables (Lasso)	Unmanipulated Steps	All Variables
	(1)	(2)	(3)	(4)
Synthetic Tag	-436**	-409**	-180	52
	[-838, -34]	[-799, -19]	[-594, 234]	[-357, 461]
	[-1162, 292]	[-815, 21]		[-986, 380]
Tag	45	69	59	62
	[-368, 458]	[-344, 482]	[-354, 472]	[-351, 476]
	[-406, 504]	[-403, 505]		[-419, 535]
Fixed Medium (12K)	-472**	-450**	-458**	-454**
	[-855, -88]	[-833, -66]	[-842, -75]	[-838, -71]
	[-785, -64]	[-822, -25]		[-829, -43]
Choice Mean	7851	7851	7851	7851
p-value Synthetic Tag vs Fixed Medium (12K)	0.425	0.409	0.076	0.004
p-value Synthetic Tag vs Tag	0.018	0.014	0.257	0.961
# Observations	112,103	116,621	111,648	111,257
# Individuals	4,116	4,285	4,101	4,092
Synthetic Tag	928	1,097	913	904
Tag	925	925	925	925
Fixed Medium (12K)	1,197	1,197	1,197	1,197
Choice	880	880	880	880

Notes: This table shows the results from regressions in Figure A.5, along with 95% confidence intervals based on Gaussian and bootstrapped standard errors. Gaussian confidence intervals are displayed above and bootstrapped ones are displayed below. Significance stars and p-values are based on Gaussian standard errors. Bootstrap is not performed for the Unmanipulated Steps tag, as its tag assignment rule is based on fixed cutoffs and does not depend upon data. Estimates come from a weighted regression where each Synthetic Tag observation is weighted by the inverse of the probability of assignment to a given step target within the Fixed groups in its experimental phase. (All other observations receive a weight of 1.) Coefficients for Tag, Fixed Medium and Monitoring included in Figure A.5 are from the last column. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.15: Payments With Tagging Do Not Differ Meaningfully From Choice

Omitted Group:	Choice			
Dependent Variable:	Daily Payments			
Synthetic Tag Type:	Policy Variables	Policy Variables (Lasso)	Unmanipulated Steps	All Variables
	(1)	(2)	(3)	(4)
Synthetic Tag	-0.32	-0.18	-0.57*	0.34
	[-0.95, 0.31]	[-0.80, 0.43]	[-1.19, 0.06]	[-0.30, 0.98]
	[-0.68, 1.71]	[-0.89, 0.43]		[-0.26, 2.03]
Tag	-0.31	-0.23	-0.27	-0.29
	[-0.94, 0.31]	[-0.86, 0.39]	[-0.90, 0.35]	[-0.92, 0.33]
	[-1.00, 0.37]	[-0.99, 0.46]		[-0.99, 0.37]
Fixed Medium (12K)	-0.48	-0.40	-0.44	-0.46
	[-1.07, 0.11]	[-0.99, 0.19]	[-1.04, 0.15]	[-1.05, 0.13]
	[-1.03, 0.23]	[-0.93, 0.24]		[-1.04, 0.21]
Choice Mean	6.58	6.58	6.58	6.58
p-value Synthetic Tag vs Fixed Medium (12K)	0.301	0.229	0.343	0.006
p-value Synthetic Tag vs Tag	0.987	0.872	0.362	0.059
# Observations	112,103	116,621	111,648	111,257
# Individuals	4,116	4,285	4,101	4,092
Synthetic Tag	928	1,097	913	904
Tag	925	925	925	925
Fixed Medium (12K)	1,197	1,197	1,197	1,197
Choice	880	880	880	880

Notes: This table is the same as Table A.14 except that it uses payments instead of steps as the outcome variable.

B Theory Appendix

B.1 Alternative Payment Functions

This section demonstrates the rationale for a principal to pay participants for achieving a step target. For simplicity, we consider the case where principals have full information on participant type. We show that step target contracts are as or more effective than:

1. Payment functions that increase linearly with steps: $W(s) = ks$
2. Linear payments after a step target: $W(s,T) = \begin{cases} 0, & \text{if } s < T. \\ k(s-T), & \text{if } s \geq T. \end{cases}$
3. n-Step function payments: $W(s) = \sum_{i=1}^{n} b_i \mathbb{1}\{s \geq T_i\}$

Step target contracts are more effective than linear contracts because they cost less. To see this, imagine the principal offers a participant of type θ a linear contract that pays k per step. This contract will increase steps to the level s^k where the marginal cost of steps is k (i.e., where $c'(s^k;\theta) = k$), and will cost ks^k. A step target contract, on the other hand, can increase steps to the same level s^k by paying $c(s^k;\theta)$. Since step costs are convex, the needed payment under the step target contract is smaller than under the linear contract: $c(s^k;\theta) = \int_0^{s^k} c'(u;\theta)du < ks^k$. Hence any linear contract is dominated by a step target contract.

Step targets are as effective as linear payments after a step target because they cost the same amount to achieve any level of steps. As before, a step target contract can achieve steps s^k by paying $c(s^k;\theta)$. A linear payment after a step target can achieve the same steps for the same payment level (to do so, it can set $k = c'(s^k;\theta)$ and $T = s^k - \frac{c(s^k;\theta)}{k}$), but cannot pay less than this amount. If it did, the participant would just choose to walk $s^*(\theta)$.

While step target contracts dominate linear contracts, there is no benefit to adding multiple step targets. To see this, note that a participant will walk up to a target T_i if $W(T_i) \geq c(T_i;\theta)$ and $b_i \geq c(T_i;\theta) - c(T_{i-1};\theta)$. Thus, the minimum condition to induce a participant to achieve the highest target $T_n = s^e$ in a multiple-target contract is to set its marginal payment at $b_n = c(s^e) - b_{n-1}$ and its total payment at $W_n = c(s^e)$. However, the principal can induce the participant to achieve the same s^e with a single step target contract offering the same total payment, $c(s^e)$. Thus, the optimal contract with multiple step targets is equivalent (in steps and cost) to the optimal contract with only one step target.

B.2 An Alternative Principal Objective

In this section, we briefly consider an alternate objective function that a principal might adopt instead of equation (5) of Section 3. Some principals may care about efficiency and only want to assign participants to efficient contracts. Because there are multiple efficient contracts for participants of each type $j \in \{L, H\}$ (specifically, any contract with target T^{j*} and payment level $\geq W^{j*}$ will be efficient) an efficiency-minded principal might choose from among the set of efficient contracts to meet some other objective, such as minimizing expected payments. That is, taking λ as the share of high-type participants, a principal might try to solve the following problem:

$$\min_{T^j, W^j} \lambda W^H + (1-\lambda) W^L \tag{11}$$

subject to the participation constraint (equation (6)), the incentive compatibility constraints (IC−L and IC−H), and a constraint that the solution be efficient:

$$\langle T^j, W^j \rangle \in \arg\max_{T^j, W^j} g(\hat{T}^j) - C(T^j, \theta^j). \tag{12}$$

The solution to this problem is the $\langle T^{L*}, W^{L*} \rangle$, $\langle T^{H**}, W^{H**} \rangle$ menu shown in Figure 3.

B.3 Inefficiency of the Principal's Imperfect Information Solution

In this section, we replicate the standard result that the principal's utility-maximizing contract menu under imperfect information will sort the high types into an efficient contract and the low types into inefficient contracts, and that this contract menu improves the principal's utility relative to a one-size-fits-all or single contract solution.

Our starting point is the model in Section 3. The principal cannot observe participant type but is aware of the share of high-type participants, which we denote as $\lambda \in (0, 1)$. The principal's problem is thus to maximize the benefits of additional steps net of payments subject to each type's incentive compatibility and participation constraints:

$$\max_{T^H, W^H, T^L, W^L} \lambda(g(\hat{T}^H) - W^H) + (1 - \lambda)(g(\hat{T}^L) - W^L)$$

where $\hat{T}^j = max(T, s^*(\theta^j))$, such that

$$C(T^L, \theta^L) \leq W^L \tag{PC - L}$$
$$C(T^H, \theta^H) \leq W^H \tag{PC - H}$$
$$W^L - C(T^L, \theta^L) \geq W^H - C(T^H, \theta^L) \tag{IC - L}$$
$$W^H - C(T^H, \theta^H) \geq W^L - C(T^L, \theta^H) \tag{IC - H}$$

Let $\langle T^{L***}, W^{L***} \rangle$ and $\langle T^{H***}, W^{H***} \rangle$ denote the optimal contracts for the low and high types. First note that the high-type IC constraint binds at the optimum (i.e., $W^{H***} - C(T^{H***}, \theta^H) = W^{L***} - C(T^{L***}, \theta^H)$). If it did not, the principal could choose another high-type contract with ε higher step target or lower payment that the high type would prefer to the low-type contract. Moreover, to extract the maximum surplus from the high types, W^L must be set as low as possible. This implies that the low-type PC constraint binds at the optimum (i.e., $C(T^{L***}, \theta^L) = W^{L***}$). The constrained maximization problem can then be rewritten as an unconstrained one:

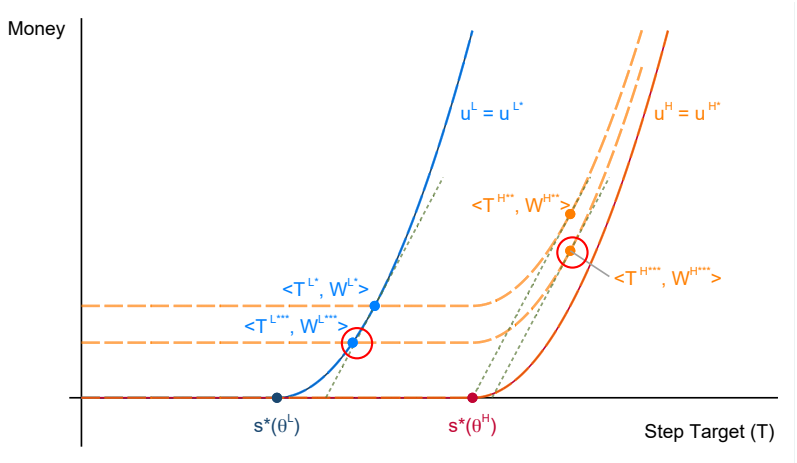
$$\max_{T^H, T^L} \lambda \left(g(\hat{T}^H) - C(T^H, \theta^H) - C(T^L, \theta^L) + C(T^L, \theta^H) \right) + (1 - \lambda) \left(g(\hat{T}^L) - C(T^L, \theta^L) \right)$$

The first-order conditions with respect to T^H and T^L are

$$\lambda g'(\hat{T}^H) = \lambda C'(T^H, \theta^H) \tag{FOC1}$$
$$(1 - \lambda) g'(\hat{T}^L) = C'(T^L, \theta^L) - \lambda C'(T^L, \theta^H) \tag{FOC2}$$

FOC1 and FOC2 yield the standard "efficiency at the top, inefficiency at the bottom" result (illustrated in Figure B.1).[40] FOC1 and FOC2 also imply that the optimal high-type

[40]The solution to FOC1 is at the tangency between the principal's benefit function and the participant's cost curve, so the optimal high-type step target will be efficient. To see inefficiency at the bottom, first note



Appendix Figure B.1: The Second-Degree Price Discrimination Solution Is Not Efficient

Notes: This figure shows how the second-degree price discrimination solution (represented by $\langle T^{L***}, W^{L***}\rangle$ and $\langle T^{H***}, W^{H***}\rangle$) entails an inefficiently low amount of effort from low types, as $\langle T^{L***}, W^{L***}\rangle$ is not tangent to the principal's indifference curve (shown in the dotted line). The contract pair $\langle T^{L*}, W^{L*}\rangle$ and $\langle T^{H**}, W^{H**}\rangle$ would be an efficient incentive-compatible approach.

and low-type contracts would differ as long as $\lambda \notin \{0, 1\}$, which in turn implies that the optimal menu dominates the optimal one-size-fits-all contract.[41]

B.4 Extending the Model: Non-Standard Utility functions

In the standard model, participant preferences over contracts are determined solely by their payments and cost of steps under that contract (equation (1)). However, in the real world, peoples' preferences over contracts may be influenced by other factors such as pride in having higher targets, wanting commitment from a higher target, or incorrect forecasts of walking costs. We flexibly allow for these forces through a non-standard utility function where utility depends not only on money and steps, but also directly on the step contract $\langle T, W\rangle$ itself:

$$u(y, s, T, W; \theta) = y - c(s; \theta) + b(T, W; \theta). \tag{13}$$

We assume that the non-standard term, $b(T, W; \theta)$, is weakly increasing in T ($\frac{\partial b}{\partial T} \geq 0$) in the range of contracts for which a given type of participant will comply and meet the step target (i.e., for all contracts such that $C(T(\theta); \theta) \leq W$).

We hold the cost function $c(s; \theta)$ constant when we introduce non-standard preferences. Thus we are considering the implications of non-standard preferences for a participant's preferences over contracts conditional on their walking choices under different incentive levels.

---

that, as long as $(1-\lambda) > 0$, the principal will always choose $T^L > s^*(\theta^L)$. This is because the derivative of the principal's unconstrained objective function is strictly positive at $T^L = s^*(\theta^L)$. As a result, at the optimum, $C'(T^L, \theta^H) < C'(T^L, \theta^L)$ (which holds for all $T^L > s^*(\theta^L)$ given our definition of types). So, the right hand side of FOC2, $C'(T^L, \theta^L) - \lambda C'(T^L, \theta^H)$, is strictly greater than $(1-\lambda)C'(T^L, \theta^L)$. Combined with the left hand side of FOC2, this implies that $(1-\lambda)g'(T^L) = C'(T^L, \theta^L) - \lambda C'(T^L, \theta^H) > (1-\lambda)C'(T^L, \theta^L)$, which implies that $g'(T^L) > C'(T^L, \theta^L)$. This indicates inefficiency, as efficiency requires $g'(T^L) = C'(T^L, \theta^L)$.

[41] To see this, suppose $T^H = T^L = T$. FOC1 imposes $g'(T) = C'(T, \theta^H)$. Then FOC2 would impose $(1-\lambda)g'(T) = C'(T, \theta^L) - \lambda C'(T, \theta^H) = C'(T, \theta^L) - \lambda g'(T)$, which implies $g'(T) = C'(T, \theta^L)$. This is a contradiction, as by the above argument, the optimal step target is always above $s^*(\theta^L)$, so $C'(T, \theta^L) > C'(T, \theta^H)$ by our definition of types.

The Participant's Problem We analyze the participant's problem under the utility function in equation (13) in two stages: the choice of contract and the choice of steps given the contract. Working backwards, we have:

Stage 2: Participant chooses steps to maximize utility given their incentive contract. Because $\langle T, W \rangle$ is not a choice variable at this stage – only s and y are – then the addition of the $b(T, W; \theta)$ term does not affect the participant's optimization relative to our earlier analysis. As before, the participant chooses steps to solve equation (1).

Stage 1: Participant chooses their incentive contract. In the standard model from Section 3, with the standard utility function from equation (1), participant's preferences over contracts are simply equal to their value function from stage 2, $V(T, W; \theta)$, which is shown in equation (4). In contrast, with the non-standard utility function in equation (13), the $b(T, W; \theta)$ term will affect preferences – and incentive-compatibility – in stage 1. Participant's preferences over contracts can thus be expressed as:

$$\tilde{V}(T, W; \theta) = V(T, W; \theta) + b(T, W; \theta). \tag{14}$$

A number of behavioral factors could underly the $b(T, W; \theta)$ term. First, people may want to show off or take pride in a high step target. Alternatively, the $b(T, W; \theta)$ term could be microfounded by either sophisticated or naive time-inconsistant preferences (we show this in the Online Supplement). Since walking has costs now but health benefits that are realized in the future, participants who are time-inconsistent and sophisticated know they will walk less in stage 2 than their stage 1 selves would prefer. As a result, they could prefer a higher step target to commit their stage 2 selves to walk more. Alternatively, people who are time-inconsistent and naive overestimate in stage 1 how much they will walk in stage 2 (because they underestimate their future net cost of walking). As a result, they could also place higher values than the standard model predicts on contracts with higher targets.

Because $\frac{\partial b}{\partial T} \geq 0$, the amount of payment needed to maintain a given utility may decrease in T (i.e., the slope of the indifference curve can be negative in T), as opposed to always weakly increasing with T. This captures the idea that participants might actually *prefer* higher targets in certain regions. Figure B.2 shows how indifference curves might look with non-standard preferences, showing three examples: non-pecuniary benefits of higher targets (e.g., pride); sophisticated time-inconsistency; and naive time-inconsistency.

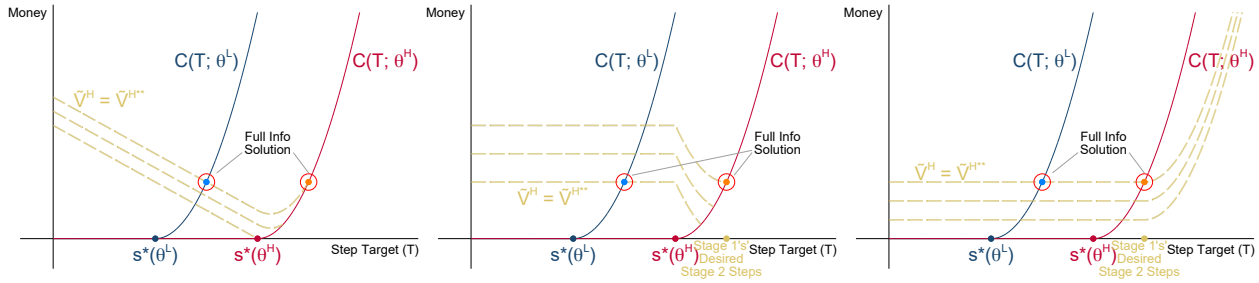
The Principal's Problem The principal's problem under full information is still as expressed in equations (5) and (6). Because the principal only values steps, not participant utility,[42] the principal values contracts based on the walking they induce in stage 2. Hence, the non-standard utility function does not change the principal's full information solution.

Interestingly, the principal's full information solution may be implementable with imperfect information. This is because the high type's incentive compatibility constraint is relaxed under non-standard preferences. As a result, high types may not prefer the low-type contract to the high-type contract, as shown in Figure B.2.

Moreover, for any choice of contract for the low types, a broader range of contracts for the high types will induce separation. Figure B.3 depicts this: the lighter shaded region would be implementable with both standard and non-standard preferences, whereas the darker

[42] We thus assume that the principal is not trying to correct any internalities.

Appendix Figure B.2: With Non-Standard Preferences, the Full Information Solution May Be Incentive-Compatible

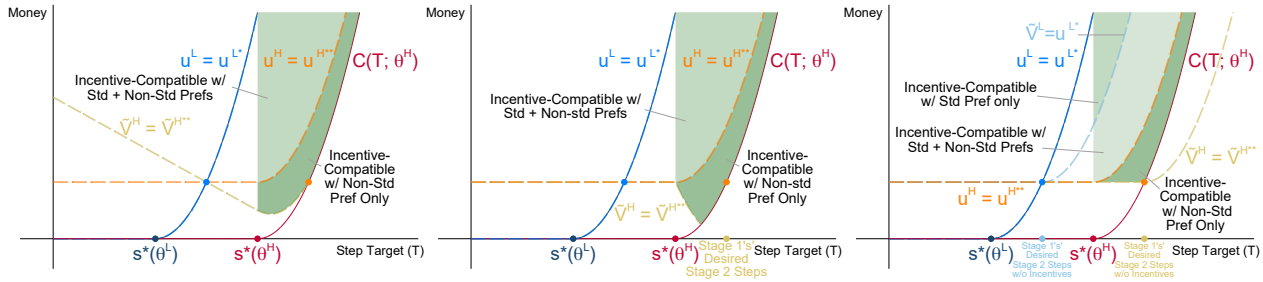


(a) Non-Pecuniary Benefits (b) Time-Inconsistency: Soph. (c) Time-Inconsistency: Naive

Notes: Figures shows the non-standard indifference curves for the high-types as the dashed curves. Stage 1's desired stage-2 steps in the two time-inconsistency cases are the steps that the stage 1 self would prefer her stage 2 (future) self to take.

shaded region is only implementable with non-standard preferences.[43] As a result, principals do better with non-standard preferences because they have a wider range of contracts to choose from. Importantly, the dark region contains much of the participant's cost curve, which is the area where the principal's preferred contracts lie. The figure also demonstrates that $W^{L*} < W^{H*}$ is no longer necessary to achieve separation.

Appendix Figure B.3: With Non-Standard Preferences, a Broader Range of Effective Mechanisms May Induce Separation



(a) Non-Pecuniary Benefits (b) Time-Inconsistency: Soph. (c) Time-Inconsistency: Naive

Notes: Figures show the regions of effective incentive-compatible contracts under standard and non-standard preferences. The dashed curves in a lighter color, labeled $\tilde{V}^H = \tilde{V}^{H**}$, are the non-standard indifference curves for the high-types. Stage 1's desired stage-2 steps in the two time-inconsistency cases are the steps that the stage 1 self would prefer her stage 2 (future) self to take.

Note that this discussion highlights the benefits of non-standard preferences from relaxing the high type's incentive compatibility constraint. However, non-standard preferences also open up the potential that the low-type's incentive compatibility constraint will bind; for example, people who are time-inconsistent and partially naive might prefer a contract with a step target that they then do not follow through with. This may prevent implementation of some contract menus. We summarize our discussion with the following result.

[43] In the naive time-inconsistency case, there is also a region that is only implementable with standard preferences, but this region is undesirable to the principal as it is far from the high type's cost curve.

Result 3. *(Non-Standard Preferences) When contract valuations locally increase in the step target (as may result from time-inconsistency or pride from the chosen target), the principal's full information solution, $\langle T^{L*}, W^{L*} \rangle$ and $\langle T^{H*}, W^{H*} \rangle$, may be implementable through a Choice menu. In addition, contract assignments with $W^{L*} = W^{H*}$ are also potentially implementable. Moreover, a broader range of contract menus are implementable, thus weakly increasing the utility of the principal relative to the standard case.*

C Appendix to the Experimental Design

C.1 Description of the 3 Phases of the Experiment

As discussed in Section 4 and shown in Table C.1, we implemented the experiment in three phases. We pre-registered the additional design elements of Phase 2 and Phase 3 in our AEA registry (Dizon-Ross and Zucker, 2020). In this section, we first describe the treatment group changes implemented in phases 2 and 3, followed by other more minor changes.

Treatment Group Changes In Phase 2 of the experiment, we introduced the *Baseline Choice* group. We layered this treatment into the randomization without changing the balance (i.e., randomization percentages) of the remaining treatments.

We began Phase 3 of the experiment only after reaching our pre-registered target sample sizes. We introduced the *Flat Choice* group in this phase. We also changed the treatment balance among the remaining treatment groups: we increased the relative size of the Fixed Low and Fixed High groups, and eliminated the Choice + Nudge group.

Other Changes As described in footnote 12, we introduced cross-randomized variation in the Choice survey timing in Phase 2. For some participants, we waited an additional week after the pre-contract period to schedule the Choice survey. (In Phase 1, we never waited the extra week). We continued this cross-randomized variation in Phase 3 but adjusted the balance. For logistical reasons, the adjustments to the Choice survey timing went in place roughly one month before the treatment group changes.

All analyses control for a 5-level categorical variable, which we call "experiment phase," for the treatment randomization phase and cross-randomization regime in which the person was enrolled. We also control for the timing of their Choice survey.

Appendix Table C.1: Phases of the Experiment

Rand. Phase	Start Date	Start Date of Phase-in Change	Treatment Groups							13-day phase-in share
			Choice	Tag	Fixed	Monitoring	Flat Choice	Baseline Choice	Choice + Nudge	
Phase 1	May 15, 2019	-	X	X	X	X			X	0%
Phase 2	Dec 9, 2019	Oct 31, 2019	X	X	X	X	X	X	X	56.5%
Phase 3	Jan 28, 2020	Feb 18, 2020	X	X	X	X	X	X		13.4%

C.2 Details on Experimental Procedures

This section describes our eligibility criteria and predicted baseline steps measure.

Eligibility Criteria The initial full list of eligibility criteria was: diabetic or elevated random blood sugar (> 140 mg/dL); 30–65 years old; physically capable of walking 30 minutes; literate in Tamil; not pregnant; not on insulin; have a prepaid mobile number used solely by them, without unlimited calling; reside in Coimbatore; not have blindness, kidney disease, type 1 diabetes, or foot ulcers; and not have had major medical events such as stroke or heart attack. Due to a rule change at the Indian Council of Medical Research mid-study, we were only able to collect random blood sugar from the first 3,300 respondents. We therefore adjusted the first eligibility criterion to include non-diabetic individuals with a hypertension diagnosis, elevated blood pressure (systolic blood pressure > 120 or diastolic blood pressure > 80 mm Hg), or slightly lower elevated blood sugar (> 135 mg/dL).

Predicted Baseline Steps To construct our measure of predicted baseline steps, we implement a cross-validated Lasso regression among all groups except Tag and Baseline Choice, regressing baseline steps on the baseline characteristics listed in Panels A, B and F of Table A.5. We then use the predictive coefficients from the Lasso regression to create individual-level predictions of baseline steps in all groups, including Tag and Baseline Choice.

C.3 Choice Survey: Scripts and Order

This appendix section provides more detail on the order in which the menus were presented during the Choice survey, as well as the stakes associated with the choices, by experimental phase.

During Experiment Phases 1 and 2, only the Base Menu and Steep Menu choices were real-stakes (i.e., had a positive probability of being implemented); the Flat Menu was hypothetical, and so we exclude the Phase 1 and 2 Flat Menu choices from analysis. The Base Menu was presented first, followed by the Steep Menu, and then the Flat Menu. Study participants were instructed to take the first two menus seriously since each choice had a positive probability of being implemented; however we emphasized that the probability of being assigned the Base Menu choice was relatively large and that the likelihood of being assigned the Steep Menu choice was relatively small.

During Phase 3, all three menus had a positive probability of implementation (i.e., were real-stakes, not hypothetical). For the majority of Phase 3, we asked the Base Menu first, followed by the Flat Menu and then the Steep Menu. For a small portion of Phase 3, in order to examine choice order effects, we randomized the order of the Base Menu and Flat Choice Menu (the Steep Menu was always last). Irrespective of the order of the Base and Choice Menus, we emphasized to participants that the first two choices had relatively large probabilities of being implemented while the likelihood of being assigned the Steep Menu choice was relatively small.

In all phases, respondents were presented with a visual aid for each menu to clarify the choice being presented.

C.4 Designing the Fixed, Choice, and Tag Treatments

In this section, we provide more detail about the process for designing the Choice and Fixed groups described in Section 4.2.1. We also describe how we designed the Tag algorithm.

Previous Evaluation Our design process used the results from the previous Aggarwal et al. (2020) evaluation of a similar incentive program. This program paid participants 20 INR for achieving a daily 10,000 step target. The details of the present study's setting,

recruitment, and procedures closely follow Aggarwal et al. (2020). The primary differences are that we shortened the contract period from twelve to four weeks and that we offered multiple step targets instead of only one.

Choosing the One-size-fits-all Contract To construct the one-size-fits-all contract, we aimed to choose the step target that would have the largest treatment effect on contract-period steps across our sample at a given payment rate.[44] To do so, we modeled the step-maximizing target for an individual as a function of their baseline steps. Specifically, we assumed that, for a given payment level, the step-maximizing target was simply baseline steps plus some constant and that the treatment effect of any target was a quadratic function of baseline steps (that decreases as the target moves away from the step-maximizing target). To estimate the quadratic function and the constant, we used a linear regression to model the treatment effect of the 10,000 step target from Aggarwal et al. (2020) as a quadratic function of baseline steps. The model implied that the treatment effect of a 10,000 step target that paid 20 INR was maximized for people walking 4,500 daily baseline steps. Therefore, we estimated that each individual's step-maximizing target for a 20 INR payment rate would be their average daily steps plus 5,500.

We then used this model to select our one-size-fits-all target. Doing so required an assumption on the distribution of baseline steps in the current study, and so we assumed that distribution would closely resemble Aggarwal et al. (2020). Together, our quadratic model and distributional assumptions implied that the average step-maximizing target for the 20 INR payment rate would be 12,000. Hence, we used a contract with a 12,000 step target and 20 INR payment rate as our one-size-fits all contract.

Selecting the Three Step Targets to Use in the Base Menu To construct the Choice menu, we aimed to choose three step targets that would each have the largest treatment effect on contract-period steps for a tercile of our sample (based on baseline steps) for a given payment rate. Using the same quadratic model and distributional assumptions we used to develop our one-size-fits-all target, we chose three round-number step targets that we predicted would maximize steps for a 20 INR payment rate. These targets were 10,000, 12,000, and 14,000.

Choosing the Payment Levels for the Base Menu To choose the payment levels for each step target on the Base Menu, we conducted a small pilot study. Pilot participants were given a pedometer for six days, and then asked which contract they would prefer among various menus. All menus offered the same three targets (10,000, 12,000, and 14,000), but the menus used different payment levels for each target. Based on our piloting, we chose payment levels that induced separation while maintaining levels close to the 20 INR payment level at which the targets were chosen.

Designing the Tag Algorithm Our design process for the Choice Menu already required estimating which three targets would each maximize steps for one third of our sample, based on baseline steps. Our Tag Algorithm simply assigns participants to those three targets

[44]Note that the step-maximizing target for a given payment level is not the same as the "optimal" target for that payment level, even for a principal who cares about maximizing steps for a given payout. This is because the payout is not equal to the payment level; rather, it equals the payment level times the compliance rate. Moving to the optimal target would have involved further assumptions about the principal's objective and the shape of compliance with target and type, which we chose not to make for simplicity.

based on which baseline step bin they fall in, with the mapping as in Table C.2.

Appendix Table C.2: Tag Assignment Algorithm

Baseline Steps	Assigned Step Target
<5,500	10,000 steps
5,500-7,500	12,000 steps
>7,500	14,000 steps

C.5 Causal Forest Estimation and Synthetic Tag Construction

C.5.1 Constructing the Causal Forest Estimates Used to Analyze Sorting

Among the participants in our Fixed groups, we use the multi_arm_causal_forest method implemented by the grf package in R to make predictions for the treatment effects of the High (14K) relative to Low (10K) target. The set of predictor variables we used is listed in Table A.12.[45] All parameters used for the training are default values except min.node.size, whose value is selected based on cross-validation results from the causal_forest method in the same package. The reason that we used a *multi-arm* causal forest to obtain the treatment effect was to maintain consistency with the machine-learning procedure used to estimate the best step target for each participant, which we describe next and which requires the use of a multi-arm causal forest. Results from a single-arm causal forest comparing the Fixed High and Fixed Low groups are similar.

C.5.2 Constructing the Policy Tree Assignments Underlying the All Variables and Policy Variables Synthetic Tags

To estimate the best step target assignment for each individual, we use the policy tree machine learning algorithm of Athey and Wager (2021) in our Fixed groups. The output of this algorithm is a step target assignment for each individual calculated based on a minimum-regret criterion. To avoid overfitting, we use a leave-one-out procedure to estimate the policy tree. Specifically, we predict the step target assignment for each individual using the policy tree algorithm estimated with every other individual in the sample.

The policy tree algorithm takes as input a multi-arm causal forest, which we estimate the same way as described in Section C.5.1, using one of the following list of predictors:

- **All Variables Synthetic Tag**: We use all variables used to estimate the causal forest described above in Appendix Section C.5.1.
- **Policy Variables Synthetic Tag**: We start from the same set of variables used in the All Variables Synthetic Tag and then exclude (a) baseline steps, and (b) all wealth variables (see column 1 of Table A.12 for the specific variables excluded).

[45]This list includes all variables from Sections A, B, and D of Table A.5, except household income per capita (since it was often missing) and self-reported activity levels (since we included actual activity levels instead – panel D). We exclude Panel C, predicted baseline steps, since we use actual baseline steps instead (panel D), and exclude Panel E, time indicators, since those would not be used for prediction.

To estimate the policy tree itself, we used the hybrid_policy_tree method of the policytree package in R.[46] All parameters take default values except tree.depth, where, to allow for greater flexibility, we show results for a tree depth of 5 (the highest tree depth for which the Athey and Wager (2021) results hold given our sample size). We have also verified that all of the results are similar (or worse) for tree depths 2-4 (to be conservative in considering the effect of Choice relative to tagging strategies, we are comfortable not showing the worse results – which are worse for tagging and hence better for Choice.)

C.5.3 Constructing a Simpler Tag with Lasso

To assess the robustness of the results with the policy variables to a simpler process that avoids machine learning of heterogeneous treatment effects (a complicated process), we predict steps using a cross-validated Lasso regression using the same set of predictor variables used in the main Policy Variables tag. To construct assignments, we then apply the same tag algorithm in Table C.2, which was designed to be used for actual steps, to participants' *predicted* steps.

D Nudge Robustness

This section shows that the estimated impact of Choice is robust to various ways of controlling for the Nudge. For reference, column 1 of Table D.1 replicates our main specification from Table 2, where the Nudge variable controls for the effect of receiving the Nudge in the non-Choice groups. The specification in Column 2 omits the control for the Nudge; the effect of Choice is similar, as the Nudge had negligible impacts in the non-Choice groups. Column 3 demonstrates that the estimates are robust to simply excluding all participants who received the Nudge, regardless of treatment group assignment, from the regression. This shows that the Nudge is not driving any of our main estimates. Column 4 relaxes the assumption made in our base specification that the effect of the Nudge was uniform across all non-Choice groups by showing a "fully interacted" model. Specifically, the column 4 specification controls for the interaction terms between the Nudge and each other treatment group (e.g., Fixed High × Nudge). The estimated effect of Choice remains very similar to our main specification. In addition, column 4 shows that the Nudge is insignificant in each of the non-Choice groups, and that we cannot reject the hypothesis that the Nudge effect is the same across each of the non-Choice groups (i.e., we cannot reject the assumption used in our base specification). Across columns 1-4, our main Choice coefficient remains large and significant at the 5% level. Finally, column 5 pools the Choice and Choice + Nudge groups together, testing for their difference from the Fixed Medium group. We still find that the pooled coefficient is large, nearly 300 steps, and significant at the 10% level. However, it is smaller, reflecting the fact that the Nudge backfired for certain types of participants, as shown in the Online Supplement.

[46] We preferred the hybrid_policy_tree method over the regular policy_tree method for shorter runtime when fitting the tree using our leave-one-out method, as the regular policy_tree method is expected to take over 2 years to produce the predictions we present. Rehill (2022) showed that the expected regret of policies predicted by hybrid policy trees comes close to the regret from regular policy trees.

Appendix Table D.1: Main Result Robust to Various Ways of Controlling for the Nudge

Omitted Group:	Fixed Medium				
Dependent Variable:	Daily Steps				
	Base Spec	No Nudge Control	No Nudge Sample	Fully Interacted	Pooling Choice & Choice + Nudge
	(1)	(2)	(3)	(4)	(5)
Choice	414**	473**	435**	430**	
	[202]	[194]	[218]	[219]	
Choice + Nudge	80	-4		55	
	[239]	[223]		[262]	
Choice or Choice + Nudge					281*
					[167]
Fixed Low	93	92	30	33	89
	[185]	[185]	[241]	[242]	[185]
Fixed Low × Nudge				100	
				[375]	
Fixed High	173	174	308	326	169
	[208]	[208]	[264]	[264]	[208]
Fixed High × Nudge				-322	
				[425]	
Monitoring	-533	-509	-576	-572	-563*
	[332]	[331]	[371]	[370]	[331]
Monitoring × Nudge				305	
				[863]	
Nudge	-174			-133	-276*
	[177]			[255]	[153]
Fixed Medium Mean	7,720	7,720	7,631	7,720	7,720
p-value vs Choice					
Fixed Low	0.123	0.058	0.075	0.081	
Fixed High	0.287	0.178	0.615	0.680	
Choice + Nudge	0.239	0.053		0.272	
Monitoring	0.005	0.003	0.005	0.005	
p-values for the significance of the Nudge in Fixed Low, Fixed High, and Monitoring groups					
Nudge + Fixed Low × Nudge				0.909	
Nudge + Fixed High × Nudge				0.202	
Nudge + Monitoring × Nudge				0.837	
p-values for the difference in the Nudge effect across non-Choice groups					
Fixed Low × Nudge vs					
Fixed High × Nudge				0.345	
Monitoring × Nudge vs					
Fixed High × Nudge				0.484	
Monitoring × Nudge vs					
Fixed Low × Nudge				0.814	
# Observations	172,961	172,961	125,217	118,923	172,961
# Individuals	6,384	6,384	4,635	4,386	6,384

Notes: Treatment group sample sizes, columns 1–4: Choice: 892; Fixed Low: 778; Fixed Medium: 1,210; Fixed High: 796; Tag: 928; Flat Choice: 439; Baseline Choice: 631; Choice + Nudge: 523; Monitoring: 187. Columns 5–6: Choice: 892; Fixed Low: 454; Fixed Medium: 671; Fixed High: 468; Tag: 928; Flat Choice: 439; Baseline Choice: 631; Monitoring: 152.
The dependent variable is daily steps in the contract period. Column 1 is the same as column 1 of table 2. Column 2 is the same as column 1, but excludes the control for receiving the Nudge treatment. Column 3 excludes all participants who received the Nudge treatment. Column 4 interacts a control for receiving the Nudge treatment with each treatment group. Note that "Pooled Choice and Choice + Nudge" is logically equivalent to "Choice × Nudge." Column 5 shows robustness to pooling the Choice and Choice + Nudge groups into a single pooled group. The sample includes the Fixed, Monitoring, Choice, Choice + Nudge, Tag, Flat Choice, and Baseline Choice groups in columns 1, 2, and 5; columns 3 and 4 exclude the Flat Choice, Baseline Choice, and Tag groups because the Nudge treatment was not assigned in these groups. We control for Tag, Flat Choice, and Baseline Choice in columns 1, 2, and 5 but exclude their coefficients from the table for simplicity. The omitted category in all columns is the Fixed Medium group. All columns control for experiment phase, time between Baseline and Choice surveys, receiving the Nudge treatment, and year-month fixed effects. Additional controls are selected individually for each column by double-Lasso from the list of controls shown in column 1 of Table A.5. Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.