Mechanism Design for Personalized Policy: A Field Experiment Incentivizing Exercise

Rebecca Dizon-Ross
University of Chicago

Ariel Zucker*
UC Santa Cruz

August 29, 2025

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 an "incentive-compatible" menu of policy choices designed to induce participants to select the variant intended for their type. 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 incentive costs relative to a one-size-fits-all benchmark. Offering choice achieves similar performance to personalizing with an extensive set of observable variables, but without the same data requirements.

*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. This material is based upon work supported by the National Science Foundation under Grant No. 1847087. We thank Rupasree Srikumar and Srinish Muthukrishnan for leading the fieldwork, and Ruoyu Chen, Katherine Daehler, Xiaomin Ju, Yadi Piao, Varun Satish, and Emily Zhang 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, Max Farrell, Meredith Fowlie, Alexander Frankel, 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. Because people's responses to policies can vary widely, tailoring a policy to individual characteristics can yield improvements beyond a one-size-fits-all approach. However, personalization can be challenging if the policymaker cannot observe each individual's type. This is especially true if individual preferences diverge from the policymaker's, which could create an incentive for people to misreport their type. This paper uses a field experiment to test whether mechanism design can 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 offer a payment to people for meeting a specific behavioral target. For instance, a workplace wellness program might pay workers for completing a set number of health activities. The ideal target for each person may vary: a low target might be most effective for workers with unhealthy lifestyles ("Low types") but may be inframarginal for those with healthy lifestyles ("High types"). 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, workers may all prefer the lower target—which offers the same reward for less effort—inducing High types to misreport. Similar issues arise in conditional cash transfer programs that provide incentives for meeting attendance targets, retirement savings programs that match savings beyond a target amount, and related settings.

Mechanism design offers a solution to this issue: design a menu of contracts for participants to choose from and make it "incentive-compatible"—that is, ensure participants have the incentive to choose the contract that aligns with the policymaker's objective. A classic mechanism for giving High types an incentive to choose higher targets is to offer 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 a second-degree price discrimination strategy where firms make it incentive-compatible for customers with a high willingness-to-pay to choose a more expensive product, such as 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.

Our experiment uses mechanism design to personalize 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, 2022a). 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, 2022b). 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). Indeed, we conducted our project in partnership with the Government of Tamil Nadu (GoTN), a southern Indian state interested in scaling up incentives for exercise among diabetics.

The specific program that we attempt to improve through personalization provides pedometers and incentives for meeting daily step targets to individuals with diabetes, hypertension, and their precursors in urban India, where both diseases have reached epidemic levels.[1] The program is promising in non-personalized form: Aggarwal, Dizon-Ross, and Zucker (2024) find 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.

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 nearly doubles the impact of incentives on walking relative to a uniform, intermediate step target that serves as our prespecified "one-size-fits-all" benchmark. While the one-size-fits-all 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 80% improvement that both the medical literature and our experimental data suggest is likely to yield meaningful health impacts. 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,

[1]It is estimated that nearly 1 in 10 adults had diabetes and 1 in 4 had hypertension in 2019 (Gupta and Ram, 2019; International Diabetes Federation, 2019), and incidence is rapidly increasing.

we cannot reject that Choice first-order stochastically dominates each of the three Fixed (non-personalized) contracts, which differ in whether the step target is 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, but avoids 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. We establish this in two parts. First, we empirically confirm the theoretical prediction that it is advantageous for the principal to 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. Second, we show empirically that participants sort in this way: 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.

We then examine the channels underlying participant sorting. Specifically, we test whether the participants who choose higher targets do so because of the higher payment levels (as in a standard economic model) or because they have nonstandard preferences that lead them to value higher targets intrinsically (e.g., a time-inconsistent demand for commitment). Our data indicate that some participants do have nonstandard preferences that may have contributed to Choice's success. However, we show that the incentive compatibility of the Choice menu—i.e., that it provided higher incentives for higher targets—was crucial for its performance.

Our final set of results benchmarks Choice against personalization based on observable characteristics, or tags. Two potential challenges with this approach are that, first, participants may manipulate their observable characteristics to access a more generous policy variant, and second, key predictors of type may be unavailable to policymakers. We compare Choice with several tagging strategies. The first—machine learning-based personalization using easily observable and hard-to-manipulate demographic and health characteristics—is designed to mitigate both concerns, but fails to increase steps relative to the non-personalized contracts. We also evaluate strategies that are more susceptible to these challenges, including personalization based on steps (where participants know the assignment rule and can potentially manipulate) and machine learning-based personalization using *all* available baseline variables (including those that are easy to manipulate or challenging to observe such as baseline steps). We find that both of these approaches perform similarly to Choice, in part

because there is limited manipulation of baseline steps when they are used to assign step targets, but both strategies require more extensive data than Choice. Choice's minimal data requirements may be particularly valuable in developing countries.

While the above analysis follows the mechanism design literature in focusing on payment costs when comparing the costs of different strategies, we also perform a cost-benefit analysis that incorporates design and implementation costs. After accounting for such costs, our best estimates of the benefits of Choice (as well as of tagging using steps measured with the potential for manipulation) still significantly outweigh the costs relative to Fixed Medium, even at small program scales. However, the estimated benefits of tagging based on all available baseline variables only outweigh the costs at large program scales, due to the costs of both generating experimental data to train the machine learning algorithm at the design stage and collecting extensive individual-level data at the implementation stage.

Overall, 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 (see Varian 1989 for a summary). 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 quantity-based incentive contracts to pay workers of differing ability. While existing empirical work has 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—works in other contexts is limited. In addition, most existing papers use observational data and structural methods, with Abubakari et al. (2024) a recent and notable exception.[2] Our contribution is to provide experimental evidence on the power of screening contracts, showing that they can be used to personalize incentives, demonstrating the channels for their effectiveness, and benchmarking them against other personalization methods.

We also tie to several other related literatures on targeting using choice (i.e., self-selection) and observables. First, a literature examines whether allowing participants to choose fi-

[2]Abubakari et al. (2024) show that non-linear pricing can help policymakers sell cleaner cooking fuel. Levitt et al. (2016) provide an additional experimental test of second-degree price discrimination, for an online gaming firm selling in-game content. They find no effect on profits, most notably because the menu they test was not designed well given their customer base's demand elasticities.

nancially dominated commitment contracts—which a rational agent would never choose—increases effort (e.g., Ashraf et al., 2006; Bai et al., 2021; 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 incentive for even rational agents to sort in a way that benefits the principal and find positive results.

Second, two papers test the impact of allowing participants to choose from a menu of non-dominated incentive schemes (Adjerid et al., 2022; Woerner et al., 2024). Adjerid et al. (2022) test a menu that is not designed to improve effectiveness for the principal, and they find that, as predicted, allowing participants to choose reduces effectiveness.[3] While Woerner et al. (2024) design their menu to improve effectiveness, they depart from the simple quantity-based price discrimination framework of Maskin and Riley (1984), instead offering a choice between contracts with dynamic streak-based incentives and more standard time-separable contracts that pay separately for each period. A key focus of their work is establishing the theoretical conditions for this type of menu to increase the targeted behavior; a second focus is explaining why choice does not work empirically in their setting. In contrast, we empirically test the classic Maskin and Riley (1984) framework using a menu composed of simple, time-separable contracts. We show that, consistent with this theory, such menus are effective empirically.

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., 2023; Ito et al., 2023; Jack, 2013), while another examines targeting on observable characteristics (e.g., 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 personalization of prices and policies based on *observables*—the analog of third-degree price discrimination. Johnson and Lipscomb (2017) and Dubé and Misra (2023) evaluate observable-based assignment for sanitation and ZipRecruiter services, respectively. Caria et al. (2024), Kasy and Sautmann (2021) and Kasy and Teytelboym (2023) examine how to efficiently learn and apply the rules for assigning treatments based on observables. In contrast, we focus on assignments based on *choices*, and benchmark this strategy against observable-based assignment.

[3]Adjerid et al. (2022) allows participants to choose between incentives that pay for success and "gain-loss" incentives that include higher payment for success but penalties for failure. They test a prediction that people choose the contract in which they perform worse.

2 Conceptual Framework and Treatment Group Design

To fix ideas, we first map standard models of second-degree price discrimination (following Maskin and Riley 1984) and third-degree price discrimination to the problem of a policymaker designing incentives to increase walking. We then show how we use key insights from these models to design mechanisms to personalize incentives for walking (steps taken).

2.1 Conceptual Framework

We assume that steps s improve health.[4] The health improvements yield private benefits $b(s)$, such as reduced morbidity and mortality, lower health care costs, and higher earnings. They also generate public fiscal externalities $g(s)$, such as reduced public health care costs and increased tax revenue from additional labor supply. Steps also have private costs $c(s;\theta)$, which include effort costs and the opportunity cost of time, and vary by participant type θ.

Types For simplicity, we consider two types, θ^H and θ^L. We assume that High types have a lower marginal cost of steps than Low types and that net private costs $c(s;\theta^j) - b(s)$ are convex in steps (yielding a single-crossing property). Without incentives, participants of type j choose steps s^j to minimize net private costs. This implies that High types take more steps than Low types, and thus that s^j is a sufficient statistic for type.

2.1.1 Personalizing Incentives

We assume that the policymaker knows the cost and benefits functions for each type. She designs incentive contracts to increase steps, aiming to maximize "principal surplus": total public fiscal externalities $g(s)$ net of incentive costs.[5]

To do so, she designs "step target contracts": participants with contract $\langle T, W \rangle$ receive a payment of W if their steps exceed the step target T. We focus on step target contracts, which are the type of contract used in Maskin and Riley (1984),[6] are the most common type of walking incentive contract in practice, can reduce payments for inframarginal steps, are simple to understand, and embed a salient daily goal which may improve performance (Mitchell et al., 2020). See Appendix B.1 for more discussion of this decision.

[4] Section 6.2 summarizes evidence on this point. While we model health as a function of steps alone, the model nests one in which health is also produced by other behaviors (e.g., diet) as long as income from the incentive payments does not directly impact these behaviors and compensatory responses do not fully undo the health benefits of steps. Both conditions are reasonable: incentive payments are small, and experimental work finds that walking interventions improve health despite any compensatory response.

[5] This objective follows Maskin and Riley (1984) and aligns with how governments and insurers often approach incentives. Were the policymaker's aim to maximize welfare under a budget constraint, the model yields a similar takeaway: the policymaker sets higher step target for High types than Low types, and, with imperfect information, makes the menu steeper to satisfy incentive-compatibility constraints.

[6] Following the mechanism design literature, we assume that although the principal cannot observe each individual's type, she knows the net cost functions for each type. Under this assumption, the principal can design personalized step target contracts that perform better than contracts that are linear in steps and equally well as linear payments after a target. If net cost functions are uncertain, however, alternative contract structures may have advantages over step target contracts.

Full-Information Contracts If the policymaker can identify each participant by type (i.e., has "full information"), she will assign personalized contracts that maximize principal surplus from each type. The step targets s^{*L} and s^{*H} will equate the marginal social costs and marginal social benefits of steps for each type, and the payments W^{*L} and W^{*H} will equal the net costs of reaching the target for each type. Notably, these contracts, which we refer to as the *full-information contracts*, assign higher step targets to High than Low types.

Contracts with Imperfect Information If the policymaker cannot observe each participant's type, she may personalize using one of two strategies: choice (second-degree price discrimination) or tagging (third-degree price discrimination).

Choice The first strategy is to offer a menu of contracts and allow participants to choose. With choice, the menu must satisfy incentive-compatibility constraints: neither type of participant may prefer the contract designed for the other.

With standard preferences, the menu of full-information contracts $\langle s^{*j}, W^{*j} \rangle$ is not incentive-compatible. High types will prefer the Low types' contract because W^{*L} exceeds their net cost of meeting the Low target, leaving them with positive surplus, while W^{*H} is exactly their net cost of meeting the High target. This is particularly clear in the special case where $W^{*L} = W^{*H}$, as the Low type's contract offers the same payment for less effort.

To ensure incentive compatibility, the principal must adjust the full-information contracts to make the menu steeper; specifically, an incentive-compatible menu requires $W^H > W^L$. We use the term *incentive-compatible choice* to describe choice menus with $W^H > W^L$, the analog of second-degree price discrimination. A range of incentive-compatible choice menus can outperform the optimal single contract from the principal's perspective.

Tagging A second way to personalize is to assign contracts based on observable proxies of θ. The challenges are that these proxies may not perfectly correlate with type, can be costly to measure, and High types might manipulate them to avoid assignment to a contract that is not incentive-compatible. The relative performance of choice and tagging thus depends on the quality, manipulability, and measurement cost of proxies for θ.

One-Size-Fits-All Contract If restricted to a single step target contract, the principal will choose the contract that maximizes the average of $g(s)$ less incentive costs across the distribution of types.

Nonstandard preferences The model above assumes that participant preferences for contracts depend only on the same net cost function that determines daily steps. However, in reality, other factors like demand for commitment may also influence preferences for contracts. These factors may loosen incentive-compatibility constraints, improving the potential performance of both choice and tagging from the principal's perspective. For ex-

ample, a menu with the full-information contracts may become implementable. We test experimentally whether nonstandard preferences play a role in contract choice.

2.2 Designing the Choice Menu and Benchmarks

Our experiment aims to evaluate the performance of an incentive-compatible choice menu relative to a single contract and benchmark the choice menu against tagging on observables. We now describe how we designed each mechanism in advance, using data from an evaluation conducted among a similar population (Aggarwal et al., 2024) as well as a small pilot conducted before launching the experiment.

Our design process broadly followed the three-stage approach to personalization described in Section 2.1.1: (1) define types, (2) select full-information contracts, and (3) design mechanisms to assign contracts in the absence of full information about type. However, since we lacked the net cost and externality functions that Section 2.1.1 assumed were known to the principal, we made several practical accommodations. First, we chose the full-information contracts based on a simple model of how steps respond to incentive contracts by type rather than by modeling the net walking cost function for each type.[7] In doing so, we constrained the contract space to a region where existing empirical data from Aggarwal et al. (2024) could inform our model of how steps respond to incentives. Second, we assumed that the externality $g(s)$ takes a linear functional form in s and selected contracts that maximized principal surplus across a range of per-step externality values.[8] Third, to understand the incentive-compatibility constraints, instead of using net walking cost functions to infer preferences, we collected direct survey data on contract preferences by type.

Stage 1: Define Types First, we defined three participant types based on baseline walking levels in the absence of incentives; these are the types for which we would design full-information contracts. We set the cutoffs using the terciles of the baseline walking distribution among the population in Aggarwal et al. (2024).

Stage 2: Select Full-Information Contracts We next selected the full-information contracts: that is, the contracts that the principal would assign to each type of participant if type were observable. Given the absence of precise data on the fiscal externality and net cost functions, we made the following practical restrictions, previewed above:

1. *Limiting the contract space:* Since we had data from Aggarwal et al. (2024) on walking behavior under contracts that paid 20 INR, we restricted attention to contracts paying 20 INR. This avoided uncertain extrapolation of walking behavior to untested payment

[7] To maximize $g(s)$ net of payments for each type, one can either model net cost functions for each type to solve for T and W or, as we do, model how contracts shift s and payments directly by type.

[8] While we assumed a linear externality for design purposes, recent work discussed in Section 6.2 suggests it is likely concave. A designer with additional information on its shape could incorporate this in the design phase. Moreover, we allow for concavity when evaluating Choice (see Section 2.3).

levels.[9] Furthermore, we chose among round-number step targets (multiples of 1,000) in order to ease communication and increase salience to participants.

2. *Assumption of a linear externality:* The linear functional form restriction is common to many second-degree price discrimination models, including Maskin and Riley (1984). It simplifies principal surplus to average steps, multiplied by the per-step externality, less average incentive payments.

As detailed in Appendix B.2, we used Aggarwal et al. (2024) data to model the steps and payments for each type under any given 20-INR incentive contract. We then chose personalized round-number step targets for each type within our restricted contract space. Ideally, we would have selected targets to maximize principal surplus, but that would require specifying the per-step externality, which is unknown. Instead, we maximized average steps, a simpler approach that yields similar targets when the externality is large relative to the payment, which appears to be the case in our setting (Section 6.2). Moreover, when limiting to round-number targets, as we do, the two methods can coincide. Indeed, our model suggests that, among round-number step target contracts paying 20 INR, our chosen contracts will maximize principal surplus from each type as long as the per-step externality is sufficiently large (at least 0.4 INR per 100 steps, which Section 6.2 suggests is conservative here).

The estimated full-information contracts, shown in Table B.1, assign step targets of 10,000, 12,000, and 14,000 steps for the Low, Medium, and High types, respectively, all with payments of 20 INR (roughly 0.29 USD).

Stage 3: Assignment Mechanisms In the final stage, we designed mechanisms to assign contracts during the experiment in the absence of full information about participant types.

Choice Since our full-information contracts are not incentive-compatible, our final step was to adjust them into an incentive-compatible menu. To understand the incentive-compatibility constraints for each type (i.e., preferences for contracts), we conducted a 70-person pilot study. We measured participants' types (i.e., steps without incentives) and then asked them to choose a contract from a menu. We piloted various menus; each included three contracts with the same step targets as the full-information contracts (10,000, 12,000, and 14,000), but with payments that increased with the step target at different rates.[10] Based on the pilot data, we selected an incentive-compatible menu that induced separation

[9] While this constraint may seem unnatural, it produces the same full-information contracts as imposing a budget constraint (for some budget level) under the assumptions we make to estimate the relationship between incentives and steps. See Appendix B.2 including footnote 67 for more details.

[10] While in the Maskin and Riley (1984) framework the principal adjusts both step targets and payments to satisfy Low types' participation constraints, we opted to maintain round-number step targets in the Choice menu. Adjusting targets was unnecessary in our case, as lower types' participation constraints likely already had slack under our full-information contracts due to 1) the initial rounding creating slack and, 2) each type representing a continuum, with some lower types finding the full-information target easier than others.

by type while maintaining payment levels close to 20 INR (to minimize the deviation from the estimated full-information contracts). To explore the impact of adjusting for incentive compatibility (which may be unnecessary with nonstandard preferences), we also test a non-incentive-compatible menu consisting simply of the full-information contracts.

One-Size-Fits-All Contract To choose our primary non-personalized benchmark, we used the same approach as for the full-information contracts, but applied it to the entire sample instead of individual types. As detailed in Appendix B.2, we used our model of how steps respond to contracts to select the step target contract that maximized average steps at the 20 INR payment level. This contract had a target of 12,000 steps (as in the full-information contract for medium types). Our model predicts that this is also the principal's optimal target at 20 INR, provided the per-step externality is at least 1.4 INR per 100 steps (roughly our estimate of the actual externality in Section 6.2), with smaller externalities yielding higher optimal targets.

Tagging In our model, types are defined by walking levels without a contract. We thus chose a natural tagging strategy for our experiment: we measured each participant's baseline steps and assigned the corresponding full-information contract, as shown in Table B.1.

While not in our model, other characteristics may also predict how individuals respond to different contracts (i.e., predict their types). Since we did not have sufficient data on these alternative observables to construct tagging rules *ex ante*, we use the random variation from our Fixed treatment groups to evaluate alternative approaches *ex post* through synthetic treatment analysis. We detail the three synthetic approaches we evaluate in Section 6.1.

2.3 Comparing Approaches

We judge the success of personalization from the perspective of a policymaker whose objective is to maximize the benefits in terms of the positive externality, $g(s)$, less program costs. To do so, our primary analysis compares the average per-person benefits and costs of choice to those of our one-size-fits-all contract, with secondary analyses comparing choice to tagging. Notably, if two approaches have the same per-person costs—a hypothesis we cannot reject for our implementations of choice and the one-size-fits-all benchmark using our preferred cost measure—then the approach generating larger benefits is preferred.

Our preferred measure of program costs is incentive payments per person, consistent with the mechanism design literature. Our preferred measure of program benefits is average steps,[11] which is the measure we designed our menu to maximize and is a sufficient statistic for the fiscal externality under the assumption that $g(s)$ is linear in steps.

[11] It was infeasible to measure $g(s)$ in our setting. While we could have measured health outcomes (and assumed how they map to savings), they are statistically noisy, and comprehensive measurement is impractical as physical activity benefits every organ system and helps prevent hundreds of diseases.

We consider two alternative measures of program costs and benefits. First, we consider a non-linear externality, $g(s)$. To do so, we analyze the distribution of steps across participants under each approach. Choice first-order stochastically dominates the non-personalized benchmark, implying that average $g(s)$ is higher with choice for *any* non-decreasing $g(s)$. Second, we incorporate additional costs beyond incentive payments such as the cost of designing each mechanism. Since these costs are higher with choice than the one-size-fits-all benchmark, the preferred mechanism depends on the specific function $g(s)$. We thus do a back-of-the-envelope calculation of the per-step externality to assess the preferred mechanism, and find that offering choice is still preferred to the non-personalized benchmark.

3 Experimental Design and Data

This section first describes our sample selection, experimental timeline, and procedures. We then describe our treatment groups. Finally, we discuss the data, including potential data quality concerns such as attrition, and present baseline summary statistics.

3.1 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 self-reported diagnosis of diabetes or hypertension (either stage 1 or stage 2), or elevated blood pressure or blood sugar; low risk of injury from walking; and the ability to receive payments in the form of mobile recharges.

After screening, we contacted eligible individuals by phone, invited them to participate in a program to encourage walking, and scheduled an enrollment visit.[12] Enrollment visits were conducted on a rolling basis between May 2019 and December 2021.[13]

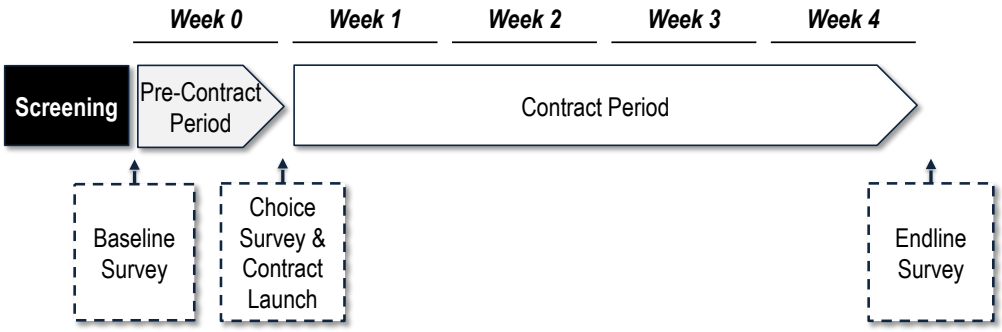
3.2 Experimental Timeline and Procedures

Figure 1 shows the experimental timeline for a participant in the study. Most treatment groups followed the same sequence of events from the enrollment visit through the end of the study. However, two supplementary treatment groups (specifically, Tag and Baseline Choice, described in Section 3.3), followed a slightly different timeline involving an earlier treatment group revelation, which was necessary to implement these treatment designs. This section describes the standard progression for all treatment groups except these two "early treatment revelation" groups. The timeline for these groups is detailed in Section 3.3.

[12]Potential enrollees were randomized into treatment groups using list randomization (stratified by median age and gender) as soon as their enrollment visits were scheduled. However, surveyors and participants were blinded to treatment group until later (as described in Section 3.2).

[13]Our experiment overlapped with two Covid-19 pandemic lockdowns: March 2020 to March 2021 and April to July 2021. We paused recruitment during lockdowns and control for lockdown days in our analyses.

Figure 1: Experimental Timeline for Sample Participant



Baseline Survey At the enrollment visit, surveyors verified the screening criteria and conducted a Baseline survey collecting health, demographic, and socioeconomic data. Surveyors launched the pre-contract period at the end of the Baseline survey.

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. While we use the pedometer data for analysis, to have real-time data during the study we also 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 told participants that we would measure their steps for six days and instructed them to walk as normal. While there were no financial rewards for meeting step targets in this period, respondents received 50 INR for wearing the pedometer and reporting steps for at least five of the six 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 2).

After the pre-contract period ended, surveyors returned for a second visit with participants.[14] They began the visit by collecting the pre-contract period pedometer data and reviewing the baseline step data with participants. Next, they conducted the Choice survey.

[14] 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 cross-randomized across treatment groups (n=2552), 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. All regressions control for whether we waited the additional week (using the "time between Baseline and Choice surveys" control). Our results are also robust to excluding those for whom we waited the extra week, with the estimated effect of Choice relative to the one-size-fits-all benchmark increasing from 420 steps in our main specification to 512 steps and the p-value<0.05 in both specifications. Regardless of second visit timing, we calculate baseline steps using the first six days following the Baseline 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 3.3.

We solicited menu choices from all participants, regardless of treatment group, to increase power and allow for heterogeneity analysis by target choice. The contract preference elicitation was "real-stakes" (i.e., not hypothetical) 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.[15]

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. See Appendix C.4 for details.

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. Imme-

[15]This held for both the Base and Flat Menus, as treatment groups received their choices on those menus (Section 3.3). For the Steep Menu, we assigned a small group (35 people) to receive their Steep Menu choices. This group is too small to examine treatment effects, and so we exclude them from all analyses.

diately 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 pedometer wearing and accurate step reporting, participants in all treatment groups received a 100 INR bonus if they wore their pedometers and accurately reported steps on 80% of contract period days, and an additional 100 INR if they did so on all days.

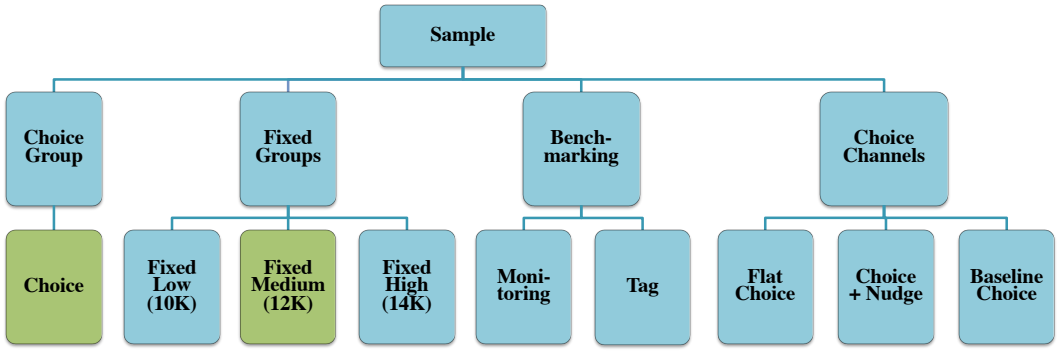
We also conducted a number of audits, both random and targeted, and suspended participants who repeatedly misreported achieving their step target.[16]

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

3.3 Treatment Groups

This section describes the treatment groups, as shown in Figure 2.

Figure 2: 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 Contract Revealed	Contract Launch	Contract Launch	Contract Launch	Contract Launch	Contract Launch	After Baseline Survey	Contract Launch	Contract Launch	After Baseline Survey
Nudge Share	0%	40%	43%	40%	18%	0%	0%	100%	0%
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. "When Contract Revealed" indicates when the participant's treatment group was revealed to them. "Nudge Share" indicates what share of the treatment group received a nudge towards a certain contract when making choices during the Choice survey. We implemented the experiment in 3 phases (see Section 3.3.5 for details). While the Nudge was cross-randomized to 60% of Fixed, Monitoring, and pooled Choice and Choice+Nudge during the initial phases, the overall treatment balance was updated in a later phase, leading to divergent Nudge shares across these groups.

[16]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.

3.3.1 Primary Treatment Groups: The Choice and Fixed Medium Groups

These groups, both designed through the process described in Section 2.2, allow us to estimate the effect of personalization using choice relative to a non-personalized approach.

Fixed Medium (12K) or "One-Size-Fits-All" Group This group received the contract that our design process suggested would maximize steps and principal surplus (provided the externality of steps meets a minimum threshold) for our full sample.

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 All participants in our Choice group were assigned a contract according to their choice from the Base Menu—the menu we created by adjusting our full-information contracts to meet incentive-compatibility constraints.

3.3.2 Other Fixed Groups

While the Fixed Medium group represents our primary prespecified 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 "full-information" contracts derived in Section 2.2.

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.

3.3.3 Benchmarking Treatment Groups

We include two treatments to benchmark the effect of Choice against other 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.[17] 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 (an early treatment revelation group) As one benchmark for the impact of personalizing with observables, we assigned participants in the Tag group to one of three contracts based on their baseline steps during the pre-contract period, using the algorithm

[17]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, Fixed Medium, and Fixed High groups.

in Table B.1.[18] These contracts had step targets of 10,000, 12,000, or 14,000 steps, each with a 20 INR payment rate, and represented our best estimate of the full-information contracts for Low, Medium, and High walkers, respectively.

Tag is one of the two treatment groups that followed a slightly shifted timeline relative to what was outlined in Section 3.2. Instead of learning their treatment assignment at the Contract Launch, they were told how their contracts would be assigned at the end of the Baseline survey, before the pre-contract period began, as indicated in the "when contract revealed" row of Figure 2. Their step targets were then assigned during the Contract Launch, based on their baseline steps. We revealed the process early because, in scaled-up versions of tagging policies, participants know that their behavior determines their contract. The Tag group was still encouraged to walk as normal during the pre-contract period.

3.3.4 Secondary Treatments: Choice Channels

We included three treatment groups to explore the channels driving the performance of Choice. The first allows us to examine the role of nonstandard preferences, the second two allow us to assess the role of incomplete information about one's own type.[19]

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. The Flat Menu contains the three estimated full-information contracts. Each has 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 targets are financially dominated.

Baseline Choice Group (an early treatment revelation 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 wearing a pedometer, making this the second group that did not follow the Section 3.2 timeline. Because treatment assignment was revealed before the Choice survey, their contract preferences in the Choice survey were hypothetical, not real-stakes, and so we exclude their Choice survey data from analysis. 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" toward a specific contract by

[18] Baseline steps were calculated as average daily steps on days with at least 200 steps. Days with fewer steps were treated as missing data, as such low counts are unlikely if someone wears the pedometer.

[19] The framework outlined in Section 2 implicitly assumes that participants have complete information about their own type; if not sorting could go awry.

informing them which contract we (the researchers) thought would maximize their steps.[20]

Nudge Cross-Randomization Our experiment also cross-randomized the same informational nudge received by the Choice + Nudge group across the Fixed and Monitoring groups. We implemented this cross-randomization for two reasons. The first was to avoid revealing treatment assignments before menu choices were made. As noted in Section 3.2, when participants from all groups except Tag and Baseline Choice made choices from the Base Menu during the Choice Survey, their treatment groups had not yet been revealed. Implementing the Nudge exclusively for the Choice + Nudge group would have thus revealed their treatment assignment to surveyors earlier than we intended. The second was to increase the statistical power for estimating the effect of the Nudge on contract choices.[21] We did not expect the Nudge to impact contract period outcomes in non-Choice groups (whose menu choices did not influence contract assignments), nor do we find evidence that it did.

Our main specifications include an indicator for being in the Choice + Nudge group, as well as an indicator for receiving the cross-randomized Nudge regardless of treatment group. We show robustness to other specifications in Appendix D.[22]

3.3.5 Implementation and Sample

We implemented the experiment in three main phases. In brief, we introduced the Baseline Choice group in phase 2, but maintained the randomization balance among existing treatments. In phase 3 (after we reached our initially preregistered sample size), we added the Flat Choice group and discontinued both the Choice + Nudge group and the associated Nudge cross-randomization. We made additional minor changes in phases 1 and 3, resulting in six subphases (detailed in Appendix C.1). All analyses control for the subphase of the experiment in which participants were enrolled.

We exclude participants who withdrew or were found ineligible prior to the end of the

[20]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 B.1.

[21]We sized the Nudge cross-randomization share for a minimum detectable effect (MDE) of the Nudge on contract choice of 5-7 percentage points at 80% power and 5% significance.

[22]For example, our *ex ante* plan was to pool the Choice and Choice + Nudge groups when analyzing step outcomes (and thereby gain statistical power), and we show the pooled comparison in column 5 of Table D.1. However, our main specifications depart from this plan. This change is in line with recent work pointing out that weighted-average effects in cross-randomized designs can be difficult to interpret and are often "neither of primary academic interest nor policy-relevant" (Muralidharan et al., 2023). The pooled results are indeed difficult to interpret in our case: the Nudge treatment led to unexpected behavior, with certain types of people (those with medium to high baseline steps) actually *less* likely to choose the contract that we recommended to them (see the Online Supplement for details). Assessing Choice and Choice + Nudge separately allows us to estimate the impact of each of these the two (relatively different) interventions separately. Moreover, we have power to do so: for example, we calculated a MDE of 420 daily steps between the Choice and the Fixed Medium groups for our final sample size, and 570 daily steps for our initially preregistered sample size (results from this sample are in Table 3 column 5).

Choice survey from all analyses, leaving a final analysis sample of 6,882 individuals.[23] The sample represents 35% of the screened, eligible population. Table A.2 shows the share of people dropped in each stage of the enrollment process.

3.4 Data

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.

3.4.1 Baseline Survey, Choice Survey, and Baseline Step Data

Baseline and Choice Surveys The Baseline survey, conducted at the first household visit, contains information on respondents' 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 respondents' pedometers 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.[24] 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.3. 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.

3.4.2 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.[25]

[23] All groups except the early treatment revelation groups (Tag and Baseline Choice), were treated identically before the Choice survey, so differential selection into this sample is not a concern outside of these two groups. We empirically rule out significant differential selection among these two groups in Table A.1.

[24] We winsorize steps at the 99th percentile. As described in footnote 18, 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.

[25] 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

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 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. Of the 1797 individuals we audited, 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 intentionally over-report (a behavior which also appears to have been rare).[26]

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 prespecified 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 1.5 percentage points (pp) 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.[27,28]

not do because of logistical constraints). Our results are robust to using actual payments instead.

[26]The rate at which pedometer data confirms participants' self-reports of meeting their step targets is similar and statistically indistinguishable among Monitoring (88.7%) and Incentives (86.2%) participants, suggesting that most discrepancies were likely mistakes.

[27]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 (column 1 of Table A.3), and the Monitoring group has 1.5pp lower missing data on a given day (column 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.

[28]As discussed in Section 3.3.5, 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

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 large 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.

3.5 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 (80 USD), slightly above the median for an urban household in Tamil Nadu (Ministry of Labour and Unemployment, 2016).

Measures of participants' health, shown in Panel B, show that the sample had high rates of chronic disease. 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, and our measurements suggest that 62% of the sample had more severe stage 2 hypertension at baseline. The average BMI of 26 kg/m^2 is in the obese range for people in India (Misra et al., 2009).[29] 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 the 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. While our primary comparison excludes these treatments, we address this imbalance (and improve precision) using the double-selection Lasso method of Belloni et al. (2014) to select controls that predict either treatment assignment or the outcome of interest in each of our regressions.

4 Choice Relative to Non-Personalized Incentives

This section empirically examines the impacts of Choice on the effectiveness of incentives, adopting the perspective of a principal who values the benefits of steps relative to the payment costs (see Section 6.3 for comparisons that take into account other potential costs). To establish a benchmark for the improvements from Choice, Section 4.1 briefly summarizes the effect of non-personalized (Fixed) incentives on average steps. Section 4.2 then compares

Choice survey and instead include everyone who was present at the Baseline survey.

[29] In India, normal BMI is considered 18.0–22.9 kg/m^2, overweight 23.0–24.9 kg/m^2, and obese >25 kg/m^2.

Choice to its primary prespecified comparison group, Fixed Medium, using our preferred measure of benefits (average steps, consistent with a linear externality of steps). Section 4.3 compares Choice with other non-personalized benchmarks, again focusing on average steps. Finally, to account for the possibility of a nonlinear externality of steps, we examine the effect of Choice on the full distribution of steps.

To compare average outcomes across treatment groups, we estimate the following least squares regression equation[30]:

$$y_{it} = \alpha + \beta \times \text{Choice}_i + \mathbf{Treat}_i'\boldsymbol{\delta} + \boldsymbol{X}_i'\gamma + \boldsymbol{X}_{it}'\lambda + \boldsymbol{Z}_i'\mu + \tau_{m(t)} + \varepsilon_{it}. \tag{1}$$

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). We omit the Fixed Medium so that the β coefficient represents our primary comparison (as prespecified in our AEA registry): Choice relative to Fixed Medium.

$\boldsymbol{X}_i$ and $\boldsymbol{X}_{it}$ are individual and day-level controls selected from the covariates listed in column 1 of Table A.5 using the double-selection Lasso method of Belloni et al. (2014). $\boldsymbol{Z}_i$ are experimental controls: fixed effects for the experiment phase, the randomly assigned length of time between the Baseline and Choice surveys (described in footnote 14) and whether the participant received the cross-randomized Nudge.[31] $\tau_{m(t)}$ are year-month fixed effects. Standard errors are clustered at the participant level.

4.1 Benchmark: Average Impacts of Fixed Incentives

As a benchmark for the potential improvement due to Choice, we briefly summarize the effect of non-personalized (Fixed) incentives relative to Monitoring. Estimates from equation 1 show that the Fixed Low, Fixed Medium, and Fixed High groups all walk more than the Monitoring group, with treatment effects ranging from 528–704 steps.[32] These increases are

[30] We use OLS with double-Lasso-selected covariates for regression analysis throughout the paper. While our primary outcome, daily steps, is a strictly positive count variable, the conditional mean of the daily step counts is large enough to be well approximated with a linear model.

[31] The Nudge dummy 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 all but the Choice groups, and the Choice + Nudge coefficient represents the impact of Choice among those receiving the Nudge. Assuming the Nudge impact 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 aligns with our expectation of a constant null effect of the Nudge on steps for non-Choice groups, which we confirm empirically: the impact of the Nudge on steps in the non-Choice groups is small and insignificant (column 1 of Table D.1). Moreover, relaxing this assumption does not change our results. The fully interacted model, which allows the Nudge effect to vary across each group, yields a nearly identical Choice coefficient (column 3 of Table D.1).

[32] 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 0.067, 0.112, and 0.044 for the Fixed Low,

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.[33]

Although the impacts of the three Fixed groups are similar and statistically indistinguishable, this similarity does not stem from participants ignoring their step targets. Figure A.1(a) shows that daily steps in each group bunch just above the randomly-assigned step target. The importance of step targets for walking suggests that personalizing the step target could in fact affect behavior. We explore this next.

4.2 Main Results: Average Impacts of Choice Relative to Fixed Medium

We now estimate the impact of Choice relative to our prespecified one-size-fits-all comparison group (Fixed Medium) using two metrics: average steps, our preferred benefits measure, and average payments, our preferred cost measure.

Impact on Average Steps The difference in average steps between Choice and Fixed Medium is captured by the coefficient on Choice in equation 1, which is shown in Table 2 and plotted in Panel (a) of Figure 3.

Choice substantially increases average steps relative to Fixed Medium. While the Medium target increases daily steps by 528 steps relative to Monitoring alone, or roughly 5 minutes of brisk walking, the Choice treatment increases walking by an additional 420 steps (significant at the 5% level) or 4 minutes—an increase of roughly 80%. Section 6.2 presents evidence that this additional walking is consistent with meaningful downstream impacts on health and health care spending.

Columns 2–6 of Table 3 show that Choice's treatment effect relative to Fixed Medium is robust to alternative specifications, namely, omitting the additional control variables, controlling for actual baseline steps, not winsorizing the outcome variable, 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), and using the "one-at-a-time" estimator from Goldsmith-Pinkham et al. (2024) to mitigate potential concerns about bias from simultaneously estimating multiple treatment effects in one equation, respectively. 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 also all substantial, ranging from 62% to 106%.

Impact on Average Payments In contrast, Figure 3(b) and Table A.6 show that Choice does not significantly increase payments, with the coefficient insignificant and the point

Medium, and High groups, respectively, and 0.057 when all three Fixed groups are pooled.

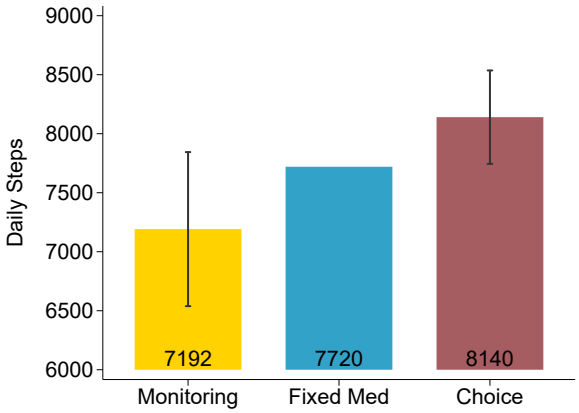
[33]We convert steps to minutes of brisk walking using a conversion rate of 100 steps per minute in order to contextualize effect sizes. In practice, participants likely walked at a mix of speeds.

Table 2: Treatment Effects on Steps, Relative to Fixed Medium

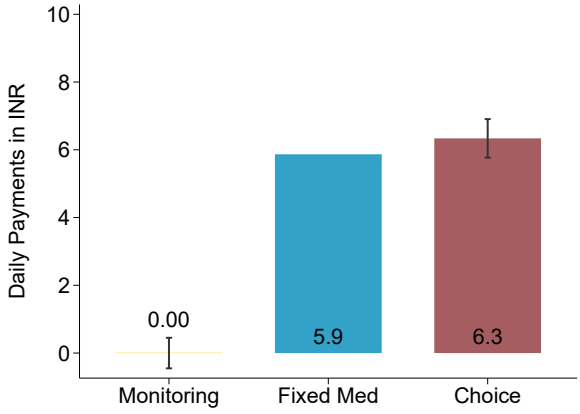
Omitted Group:	Fixed Medium
Dependent Variable:	Daily Steps
	(1)
Choice	420**
	[202]
Fixed Low (10K)	90
	[185]
Fixed High (14K)	176
	[208]
Tag	455**
	[205]
Flat Choice	104
	[252]
Baseline Choice	342
	[225]
Choice + Nudge	82
	[239]
Monitoring	-528
	[333]
Fixed Medium (12K) Mean	7,720
p-value vs Choice	
Fixed Low	0.115
Fixed High	0.282
Tag	0.867
Flat Choice	0.199
Baseline Choice	0.724
Choice + Nudge	0.234
Monitoring	0.005
p-value vs Monitoring	
Fixed Low	0.067
Fixed High	0.044
Tag	0.004
Flat Choice	0.083
Baseline Choice	0.013
Choice + Nudge	0.110
p-value Fixed High vs Fixed Low	0.694
# 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 experiment phase, time between Baseline and Choice surveys, receiving the Nudge, year-month fixed effects, and the following additional controls selected by double-Lasso from the controls shown in column 1 of Table A.5: age, mental health index, dummy for missing BMI, average predicted baseline steps, average predicted baseline steps decile 4, dummy for Sunday, dummy for first week of contract period, dummy for fourth week of contract period, dummy for day during covid lockdown. Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.

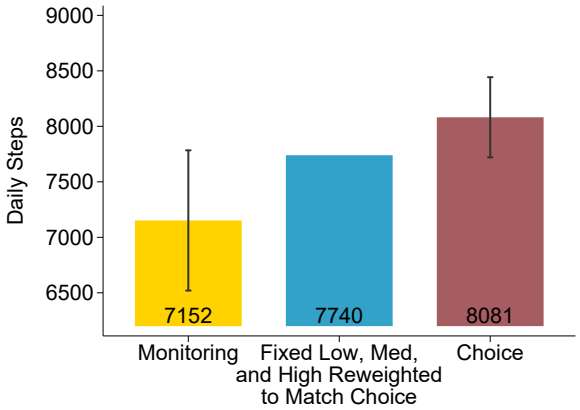
Figure 3: The Impact of Choice on Steps and Payments



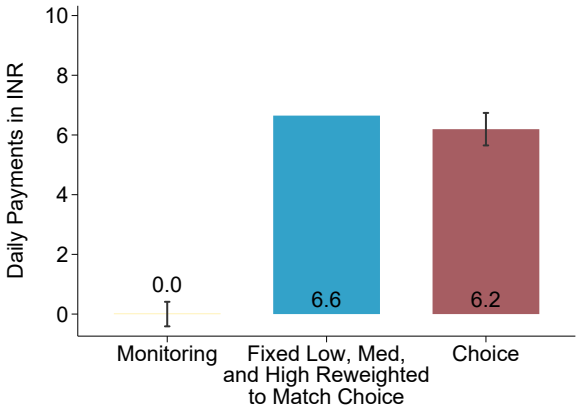
(a) Daily Steps: Choice vs. Fixed Medium



(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, columns 1 and 2, respectively.

estimate suggesting a mere 8% change.[34]

Comparing the costs and benefits of Choice relative to the Fixed Medium group, we find that Choice increases the treatment effect on average steps by 80% without significantly raising average payments. As a result, with a positive linear externality of steps, principals

[34]If we use reported steps instead of actual steps to calculate payments, the point estimate remains virtually unchanged, going from 0.47 to 0.49, although the p-value decreases to 0.097.

Table 3: Robustness of Choice Treatment Effect Estimates

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	420**	438**	384**	450**	551*	518**
	[202]	[210]	[176]	[207]	[296]	[204]
Fixed Med effect	528	414	445	529	890	583
Choice effect as % Med effect	80	106	86	85	62	89
# Observations	172,961	172,961	130,571	172,961	101,328	56,760
# Individuals	6,384	6,384	4,825	6,384	3,713	2,102
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: This table shows robustness of the estimated treatment effect of Choice from the specification shown in Table 2 (and replicated here in column 1) to alternative specifications. For brevity, only the Choice coefficient estimates from each regression are displayed; see Table A.8 for all coefficient estimates.
Columns 2–3 include alternative controls. All columns control for experiment phase, time between Baseline and Choice surveys, and receiving the Nudge ("Experimental" controls, or z_i in equation 1). Our base specification in column 1 additionally controls for a vector of controls selected by double-Lasso from the list of controls in column 1 of Table A.5 (selected controls are listed in the notes to Table A.8), 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. The selected controls are: age, average baseline steps, dummy for Sunday, dummy for first week of contract period, dummy for fourth week of contract period, dummy for day during covid lockdown. Column 4 uses non-winsorized steps as the dependent variable. Column 5 limits to experiment 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. Additional controls in these three columns are selected by double-Lasso. The selected controls are: Column 4: age, mental health index, dummy for missing BMI, average predicted baseline steps, average predicted baseline steps decile 4, dummy for Sunday, dummy for first week of contract period, dummy for fourth week of contract period, dummy for day during covid lockdown; Column 5: age, average predicted baseline steps, dummy for Sunday, dummy for Friday, dummy for first week of contract period, dummy for fourth week of contract period; Column 6: age, dummy for missing diastolic blood pressure, average predicted baseline steps, dummy for Sunday, dummy for first week of contract period, dummy for fourth week of contract period. 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 column 3 since baseline steps are endogenous in 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%.

will generally prefer Choice to a uniform (12K) target.[35]

4.3 Average Impacts of Choice Relative to Other Fixed Benchmarks

Reweighted Fixed While the Fixed Medium group was our prespecified benchmark for Choice, it is not the only non-personalized benchmark of interest. One useful benchmark, which we call the "Reweighted Fixed" group, randomly assigns 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.2). 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 3(c) compares average steps in the Choice group and the Reweighted Fixed benchmark graphically.[36] Choice increases daily walking by 342 steps more than the Reweighted Fixed group (p-value = 0.064)—an increase of roughly 58% in the treatment effect relative to Monitoring. This large increase in steps is achieved without increasing payments, as shown in Figure 3(d). Hence, even conditional on the mix of step targets, Choice substantially improves performance relative to an unpersonalized approach.[37]

Other Fixed Groups Figure 4 displays a scatter plot of average steps versus average payments in Choice and the Fixed groups. The arrow indicates the direction of principal bliss: higher steps and lower payments. While our experiment was not powered to compare Choice with Fixed Low and High, we interpret the point estimates as suggestive.

Regardless of the size of the linear per-step externality, the principal should prefer Choice not just to the Medium target, as already shown, but also to the Low target. Choice generates

[35]Taking the coefficients at face value, Choice's payment cost per extra 100 steps induced is just 0.11 INR or 0.0016 USD. The principal prefers Choice if their value of steps exceeds this amount, which is an order of magnitude below our median estimate of the externality (1.3 INR per 100 steps, see Section 6.2). It is also an order of magnitude below Fixed Medium's cost of generating steps relative to Monitoring (1.04 INR per 100 steps). This can be interpreted as the linear externality required to justify offering non-personalized incentives, suggesting that if incentives make sense, so does personalization.

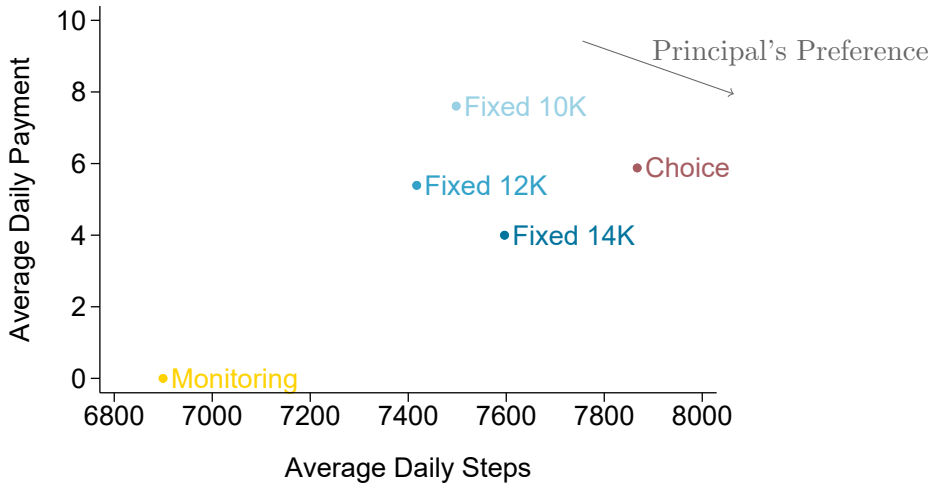
[36]Specifically, we estimate the following equation using weighted regression (results in Table A.7):

$$y_{itk} = \alpha + \beta_1 \times \text{Choice}_i + \beta_2 \times \text{Monitoring}_i + \boldsymbol{X}_i'\gamma + \boldsymbol{X}_{it}'\lambda + \mu_k + \varepsilon_{it}, \tag{2}$$

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 1. To obtain the same step target balance in the Reweighted Fixed group as the Choice group, we weight each Reweighted Fixed observation by $\frac{c_{sk}}{f_{sk}}$, where f_{sk} and c_{sk} are the respective fractions of the pooled Fixed and Choice groups assigned to step target $s \in \{Low, Med, High\}$ in experiment phase k. (Monitoring and Choice observations have a weight of 1.)

[37]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. Since payments for a given step target are weakly lower in the contracts used in Choice, conditioning on payment levels in addition to step targets would likely increase the treatment effect of Choice relative to unpersonalized incentives on steps (but might bring average payments closer together).

Figure 4: Average Steps and Payments, by Treatment Group



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.

more steps than the Low target (p-value $= 0.115$) for less payment (p-value < 0.01).

Whether the principal prefers Choice to the High target, however, depends on the size of the externality. Choice generates 244 more average daily steps (p-value $= 0.282$), but also pays out 1.9 INR more per day (p-value < 0.01). These estimates suggest the principal prefers Choice as long as the linear per-step externality is at least 0.8 INR per 100 steps (1.9/244×100). We show in Section 6.2 that this is far below our median estimate of 1.3 INR per 100 steps.

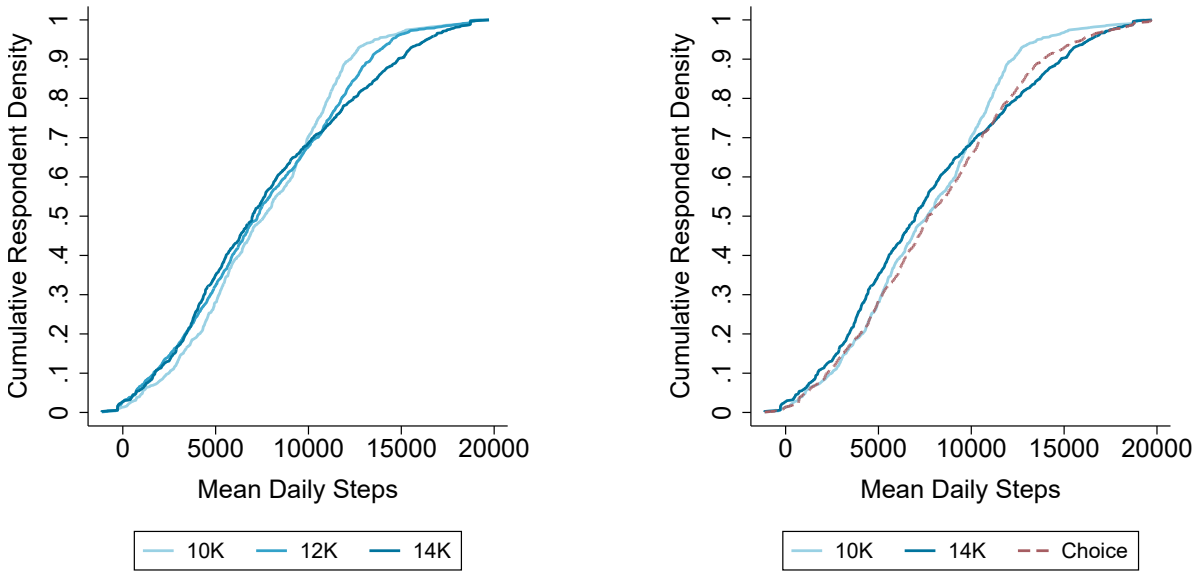
4.4 Distributional Impacts of Choice

Our initial comparisons across treatments follow Section 2.1 in assuming that the benefits of steps to the principal (i.e., the externality) is linear. To judge the performance of Choice allowing for a nonlinear externality, we next 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.[38]

Figure 5(a) shows that no one Fixed target first-order stochastically dominates the others. Fixed Low shifts the bottom of the step distribution to the right relative to the other targets (p-value < 0.05 relative to High at the 25th and 50th percentiles), while Fixed High has the largest impacts at the top (p-value < 0.01 relative to Low at the 75th percentile). Barrett and Donald (2003) tests for first-order stochastic dominance (FOSD) also reject that any single Fixed target dominates both of the others.

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

Figure 5: The Distributions of Steps under Choice and Fixed Incentives



(a) Fixed Groups

(b) Choice vs. Fixed

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.

In contrast, we cannot reject that steps under Choice first-order stochastically dominate steps under each Fixed target. Choice nearly traces the outer envelope of the Fixed target CDFs, as shown in Figure 5(b). Barrett and Donald tests fail to reject the null of FOSD when comparing Choice with each Fixed group (p-values 0.730 for Fixed Low, 0.990 for Fixed Medium, and 0.170 for Fixed High).[39] Choice performs as well as Fixed Low at the bottom of the distribution but significantly outperforms it at the top, with the difference significant from roughly the 70th percentile upwards. Analogously, Choice performs similarly to Fixed High at the top of the distribution (with a brief crossover), but significantly outperforms it at the bottom.[40] To interpret the magnitude of the differences, Table A.9 presents quantile treatment effects of the three Fixed treatments relative to Choice (the omitted group). The treatment effects of Choice relative to Monitoring at the 25th and 50th percentiles are roughly 2.5 times as large as those of Fixed High.[41]

[39] As reference for the power of the test, Barrett and Donald tests strongly reject the nulls that Fixed Low, Fixed Medium, or Fixed High dominate Choice; p-values <0.001, 0.035, and <0.001, respectively.

[40] While the Choice and Fixed High CDFs cross, Fixed High's CDF is significantly above Choice's for only around 5% of the distribution (the 85th to 90th percentile). In contrast, Choice's CDF is significantly above Fixed High's for nearly 50% of the distribution (roughly the 20th to the 65th percentile). Due to this difference in the ranges of dominance, the Barrett and Donald test does not reject the null that Choice FOSD Fixed High, although it comes closer to doing so than for the other Fixed treatments (p-value 0.170).

[41] Since Choice is omitted, its treatment effect relative to Monitoring is the negative of the Monitoring

Because average payments in Choice are lower than in Fixed Low and indistinguishable from Fixed Medium, the fact that Choice's step distribution first-order stochastically dominates those of Fixed Low and Medium implies that Choice's benefits outweigh its payment costs whenever the externality is positive—regardless of its shape. The comparison with the High target is ambiguous, as the High target also pays out less than Choice, and it is not completely clear that Choice FOSD High. However, the fact that Choice substantially increases the lower quantiles of the distribution relative to the High target means that, if the benefits of steps are concave, principals are likely to prefer Choice.

4.5 Summary of Results on the Effectiveness of Choice

In this section, we showed that, considering payment costs only, personalization using incentive-compatible choice significantly improves the effectiveness of incentives. Compared to the one-size-fits-all (Fixed Medium) benchmark, Choice increases average steps by roughly 80% and shifts the entire distribution of steps to the right, but does not meaningfully raise costs, making it preferred for nearly any positive linear or nonlinear externality. Choice is also preferred to Fixed Low for any positive externality, as steps under Choice FOSD steps under the Low target while costs are lower. Finally, Choice is preferred to Fixed High for linear externalities at least 0.8 INR per 100 steps (below our estimates of the externality in our setting), and because it particularly raises steps at the lower end of the step distribution relative to Fixed High, it is preferred for even smaller average externalities if they are concave.

5 Channels for Choice's Impact

Classic mechanism design frameworks, such as Maskin and Riley (1984) (Section 2), highlight two main channels for Choice's effectiveness: (1) the principal prefers to assign higher targets to higher types, and (2) the Choice menu sorts higher types into higher targets. We provide evidence for both channels. We also investigate what underlies (2)—that is, why higher types choose higher targets. While some participants exhibit nonstandard preferences, choosing higher targets even when financially dominated, the incentive compatibility of our Base Menu, which offers higher payments for higher targets, is crucial for inducing this sorting. Finally, we find no evidence that information frictions about one's own type hinder effective sorting in Choice.

Further from the standard mechanism design 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.

5.1 Heterogeneity in Step Target Impacts by Type

We first examine whether higher step targets are more effective for those with higher baseline walking. Among participants in the Fixed groups, we regress daily steps and payments

coefficient, and Fixed High's relative to Monitoring is the Fixed High coefficient minus Monitoring's.

on the randomly-assigned step target, baseline steps, and their interaction.

The results with steps as the outcome, shown in column 1 of Table A.10, show that the interaction term is positive and significant: higher step targets generate more steps from higher baseline walkers. To better understand the magnitudes, Figure A.3 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.

In contrast, when payments are the outcome, there is no statistically significant or meaningful heterogeneity in step target effects by baseline steps (column 2 of Table A.10). High step targets are generally less expensive than low step targets, and no less so for high walkers.

Hence, principals who value average steps relative to payments should prefer higher targets for higher walkers: relative to lower walkers, the higher targets generate more steps for higher walkers without higher payments. Moreover, the substantial heterogeneity just demonstrated in the effects of step targets by baseline steps could explain Choice's effectiveness if participants sort by baseline steps when selecting targets. We examine this next.

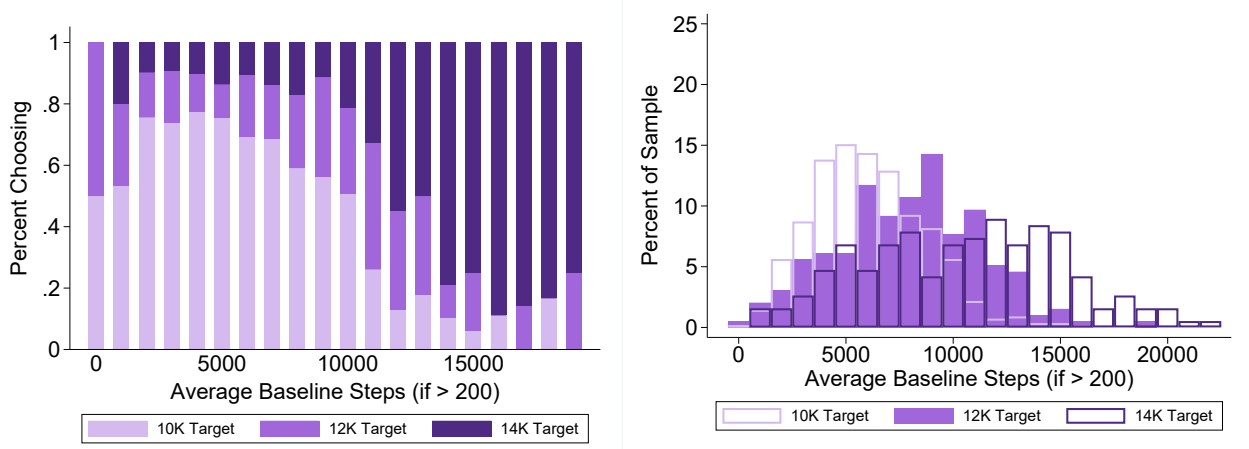
5.2 Sorting by Type

Figure 6 shows that participants in the Choice group sort across contracts by type. Figure 6(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 6(b). The correlation between choices and baseline steps is highly statistically significant (Table A.11, column 1).

While baseline steps are a sufficient statistic for type in our unidimensional Section 2 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;[42] in fact, the correlation between the predicted

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

Figure 6: Sorting by Type on the Choice Menu



(a) Chosen Step Targets by Type (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).

treatment effects and baseline steps is 0.59. However, there are other important predictors, such as health measurements and age (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 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 6(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 2 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.4 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.

5.3 Prevalence of Nonstandard 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.2 shows that 33% of participants still choose Medium and High targets. It appears that nonstandard factors, such as pride or demand for commitment (e.g., Ashraf et al., 2006), may be influencing choices.[43] This raises an important question: did High types only sort into higher targets because of nonstandard factors, or was the incentive compatibility of the Choice menu also critical?[44]

5.4 Sorting and Incentive Compatibility

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.5 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.[45] Five pp fewer participants choose the High target on the Flat Menu than the Base Menu, off of a base of 18%.

[43] 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 89% of participants answered both questions correctly.

[44] Nonstandard preferences could cause sorting by baseline steps even if not correlated with baseline steps. For example, even if all participants have a time-inconsistent demand for commitment, a higher target only serves as an effective commitment device for those with sufficiently high baseline steps.

[45] Recall that we randomized choice order for a short period to explore choice order effects. Choice order appears to matter: the difference between Flat and Base Menu choices is over 5 times larger for first than second choices, though the p-value is 0.151 due to the small sample for which we randomized order.

The implications of the shift towards lower targets depend on which participants shift. Panels B and C of Figure A.5 show the results separately for those with above-median 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 5.1 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 420 more steps on average, daily, than the Fixed Medium group, the Flat Choice group only walks 104 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 our Base Menu was important for its success.

5.5 Information Frictions and Choice

In the standard model, respondents understand their own 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. We briefly summarize our results here and offer more detail in the Online Supplement.

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 pre-contract period makes Choice simpler and cheaper to implement. Second, the Choice + Nudge group that received information about which target we (the principal) thought might be best has 338 *fewer* steps than the main Choice group, although the difference is not statistically significant (p-value 0.234, Table 2). Several other studies also find evidence of informational nudges backfiring (e.g., Beshears et al., 2015; Byrne et al., 2018). In our case, part of the negative impact appears to stem from participants with medium-to-high baseline steps becoming *less* likely

to choose the recommended target, which is sometimes referred to as a "boomerang effect."[46]

5.6 Summary of Channels for Choice's Effectiveness

We show that the Choice treatment is effective because it sorts participants based on their types, with the incentive compatibility of the menu improving sorting. We also find that some people prefer higher step targets even when they are financially dominated.

6 Policy Implications: Benchmarking and Cost-Effectiveness

This section examines factors that are helpful for judging the policy relevance of Choice. We begin by comparing Choice with another strategy for personalization, tagging on observables, focused on the impacts on average steps and average payments. From this perspective, Choice outperforms the most scalable tagging approach and performs similarly to tagging based on measured baseline steps or on an extensive set of observables.

Next, we present evidence that increasing steps through personalization will lead to health improvements and healthcare savings externalities. In order to provide context on the costs a policymaker would be willing to incur for steps, we provide conservative estimates of the magnitude of these benefits.

Finally, with our benefit estimates in hand, we extend our cost benefit comparison of Choice with non-personalized incentives and tagging to incorporate design and implementation costs. While these costs make Choice more expensive than non-personalized incentives, these costs appear to be outweighed by the savings from health benefits. On the other hand, while tagging entails additional implementation costs it does not increase steps relative to choice, emphasizing the cost competitiveness of Choice when individual data on which to tag are costly to collect.

6.1 Benchmarking Choice against Tagging on Observables

We now benchmark Choice's average steps and payments against tagging based on observables. In addition to the algorithm implemented in the Tag group, which assigned step targets based on potentially manipulated baseline steps, we use the Fixed groups to construct three other tagging algorithms a policymaker might consider:

1. *Policy Variables:* We use the policy tree machine learning procedure of Athey and Wager (2021) in our Fixed groups to estimate which step target would be best for each participant given a set of observables that health policymakers in a developing country setting would plausibly have access to and that are challenging to manipulate. See Appendix C.5 for details. 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).

[46]The rest may stem from participants following the nudge even when they had private information about a better target or feeling pressured to comply and resenting the loss of autonomy.

2. *"Unmanipulated" Steps:* To consider tagging based on unmanipulated steps, we assign targets to the Fixed groups based on their baseline steps, which they had no incentive to manipulate, using the Table B.1 algorithm (the same as is used in the Tag group). While not implementable, this tag allows us to isolate the effect of manipulation.[47]
3. *All Variables:* We again use the policy tree algorithm in our Fixed groups, but now include a larger set of variables including all policy variables, baseline steps, self-reported measures of wealth, and more. Like Unmanipulated Steps, this tag is not implementable even if the policymaker had the ability to survey participants with our baseline instrument, as some variables (such as baseline steps) are easily manipulable.

To compare each of these three tagging algorithms with Choice, we create three synthetic treatment groups composed of all Fixed group participants randomly assigned the step target the respective algorithm would have chosen for them. We compare each synthetic treatment group to Choice using regressions of the following form:

$$\begin{aligned} y_{it} =& \alpha + \beta_1 \times \text{Synthetic Tag}_i + \beta_2 \times \text{Tag}_i + \beta_3 \times \text{Fixed Medium}_i \\ &+ \boldsymbol{X}_i'\gamma + \boldsymbol{X}_{it}'\lambda + \boldsymbol{Z}_i'\mu + \tau_{m(t)} + \varepsilon_{it}. \end{aligned} \tag{3}$$

Synthetic Tag represents a dummy for being in the relevant synthetic treatment group (Policy Variables, Unmanipulated Steps, or All Variables).[48] The omitted group is Choice. Tag and Fixed Medium are dummies for being in those treatment groups, with Fixed Medium included as a benchmark. All other variables are defined as in equation 1.

Results Columns I and II of Figure 7 show the results for steps and payments, respectively. We show Gaussian confidence intervals that condition on the synthetic tag assignments for all regressions. We also show bootstrapped confidence intervals for the Policy and All Variables tags which account for noise in the creation of these tag algorithms from data.

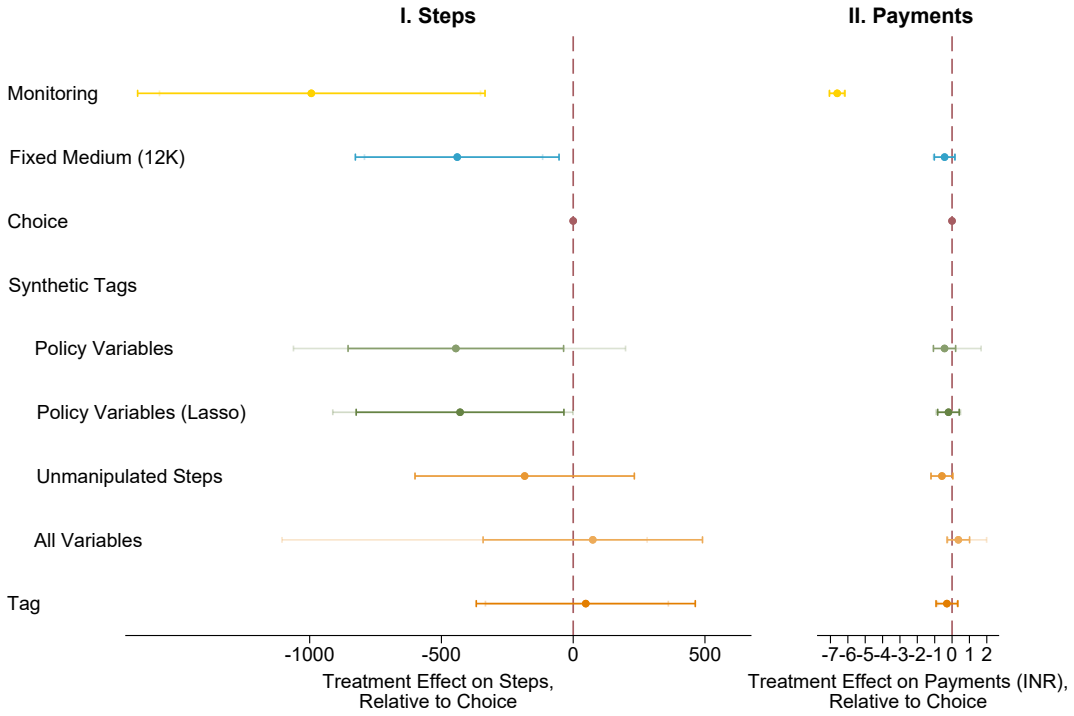
Personalizing using Policy Variables generates significantly fewer steps than Choice (Gaussian p-value 0.015; bootstrapped 0.255), with no significant difference in payments. In fact, it performs nearly identically to (and statistically indistinguishably from) the one-size-fits-all benchmark.[49] The coefficient estimates suggest that, relative to the Policy Variables tag, Choice generates additional steps for only 0.10 INR per 100 steps.

[47] An alternative approach is to machine learn the algorithm based on unmanipulated steps. In Appendix B.3, we show that approach yields statistically indistinguishable but numerically slightly worse results for the synthetic tag relative to Choice. This Appendix also validates the Table B.1 algorithm.

[48] Since step target assignment was random in the Fixed groups, each Synthetic Tag group represents a random segment of the population. However, because we assigned more Fixed target participants to the Medium target than the other targets, Medium target participants are over-represented. To adjust for this, the regression weights observations by the inverse probability of assignment within the Fixed groups.

[49] To assess robustness of this result to the machine learning procedure, we estimate another tag using the same predictor variables but a simpler Lasso-based prediction procedure (described in Appendix C.5); the results, shown in Figure 7 with the "Policy variables (Lasso)" Synthetic Tag, are similar.

Figure 7: Treatment Effects of Various Tagging Strategies Relative to Choice



Notes: 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 experiment 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 = 0.015) and Policy Variables (Lasso) (Gaussian p-value = 0.014) Synthetic Tag groups. Controls are the same as in Table 2.

We find that tagging is more effective using predictors that are manipulable and challenging to collect. The Unmanipulated Steps tag closes over half of the gap with Choice in steps.[50] Tagging with All Variables performs similarly to Choice, with similar and statistically indistinguishable impacts on steps and payments.

While the potential for manipulation is theoretically a downside of such tags, interestingly, in our experiment manipulation did not appear to harm the performance of personalizing based on observables. Steps in the Tag group, which received targets based on manipulated steps, were somewhat *higher* than those in the Unmanipulated Steps Synthetic Tag group, though the difference is not significant (p-value 0.259). Tag also yields similar, statistically indistinguishable steps and payments to Choice. By comparing baseline steps in Tag to those in groups with no incentive to manipulate them, Figure A.6 suggests that Tag performs well

[50]The Unmanipulated Steps group also incurs significantly lower costs than Choice. Taking our coefficients at face value, the cost of each additional 100 steps induced by Choice, relative to this tag, is 0.31 INR.

because there is limited manipulation, perhaps reflecting significant practical or health costs of reducing step counts. Moreover, the manipulation is on net *upwards*. Since all step target contracts in the Tag treatment pay the same amount (20 INR), higher targets are financially dominated, and upwards manipulation suggests nonstandard preferences.

The lack of problematic manipulation is perhaps surprising given evidence that people manipulate the observables that affect decision rules in other contexts (e.g., Banerjee et al., 2020; Björkegren et al., 2024; Gonzalez-Lira and Mobarak, 2019). Understanding when manipulation is more likely is thus an important question for future work, as manipulation could substantially influence the relative performance of the Choice and Tag approaches.

In summary, these results suggest that, comparing steps and payments, Choice significantly outperforms tagging based on readily available characteristics (Policy Variables), and performs just as well as both tagging based on baseline steps (Tag) and all characteristics that our surveys collected (All Variables) while eliminating the need to gather data on personal characteristics.

6.2 The Health Benefits of Steps

The benefit of additional steps from Choice stems from health improvements and reduced public healthcare costs. This section presents evidence that these impacts are large. We summarize extensive research showing that physical activity, including step interventions, improves health and reduces health care costs. We also present estimates from our experiment on the marginal returns to steps for key cardiovascular health indicators, which further support the conclusion that Choice will improve health. Finally, we present back-of-the-envelope estimates of the externality's magnitude in our setting.

Literature on Health Impacts of Exercise Strong experimental and observational evidence suggests that increased physical activity, especially walking, benefits those with hypertension and diabetes. Health outcomes continue to improve with activity for activity levels beyond those of our participants, including those in Choice. In addition, the fact that walking interventions improve health suggests that compensatory behaviors are typically not large enough to offset the health benefits of the additional steps.[51]

For example, a recent meta-analysis of 126 randomized controlled trials of exercise interventions among diabetics shows that walking is one of the three most effective methods for improving blood sugar control (Gallardo-Gómez et al., 2024), and that the returns to activity, while decreasing, are positive even for interventions increasing energy expenditure by three to five times more than Choice.[52] A smaller study of adults with diabetes and

[51] More directly, Aggarwal et al. (2024) find that steps induced by a similar step target incentives program improve health without any evidence of negative compensatory changes to diet, smoking, or drinking.

[52] Author's calculations. Relative to Monitoring, Choice-based step target incentives lead to an additional 947 steps per day, which translates to an intervention dose between 226 and 332 MET minutes per week

prediabetes (J. del Pozo-Cruz et al., 2022) finds that all-cause mortality declines with steps up to a level reached by only a third of Choice participants, with the evidence inconclusive beyond that.[53] Recent work using accelerometers to precisely measure steps among diabetics finds continued mortality reductions with additional activity even at the highest activity levels (Cao et al., 2024). Among people with hypertension, experimental work also shows that physical activity and walking reduce blood pressure (e.g., Lee et al., 2021), and a non-experimental study of 40,000 hypertensives finds large mortality decreases up to the 90th percentile of physical activity (B. del Pozo Cruz et al., 2022).[54]

Evidence on Health Benefits of Step Target Incentives While activity is widely recognized to improve health, little evidence exists on the returns to the steps induced by step target incentive programs in particular. Health outcome data from our experiment help address this gap. Note that we did not design our experiment to measure impacts on health outcomes, nor did we prespecify any health measures as primary or secondary outcomes; however, we did collect health measures that provide suggestive evidence on the health returns to steps. To maximize statistical power, we estimate the marginal health returns to steps using an instrumental variables strategy that leverages all the variation across our treatments. We regress each health outcome on average daily intervention period steps, instrumented with incentive treatment indicators. Table A.14 reports results for random blood sugar (RBS),[55] blood pressure (BP), BMI, and waist circumference.

Despite the small sample for which we collected RBS, we see fairly large reductions in RBS (7.2 mg/dl, p-value 0.081) for every additional 1,000 daily steps. The reduction is larger for those with higher baseline RBS (Panel B): 12.5 mg/dl per 1,000 daily steps (p-value 0.081). A back-of-the-envelope calculation suggests that the 420 daily steps induced by Choice in this subsample would close 10–20% of the gap in blood sugar control between diabetic and normal levels. These estimates mirror those from our earlier study of a related step target incentive program (Aggarwal et al., 2024), where a parallel strategy shows that

(depending on the intensity of the additional steps). In comparison, Gallardo-Gómez et al. (2024) estimate that HbA1c control is maximized through an intervention dose of 1,100 MET minutes per week.

[53]Specifically, J. del Pozo-Cruz et al. (2022) find that all-cause mortality decreases with average daily steps until just over an average of 10,000 steps per day (10,177 among diabetics and 10,678 among pre-diabetics), with a statistically noisy flattening beyond that level. With Choice, only 36% of people achieve more than 10,000 steps per day on average; the remaining 64% walk in the range where J. del Pozo-Cruz et al. (2022) find clear mortality reductions for prediabetics and diabetics from additional steps.

[54]We don't know of any studies that estimate the dose-response of health outcomes to activity levels for people with hypertension in units that we can translate into steps. However, recent meta-analyses of associational studies have found declines in all-cause mortality with daily steps among the general population: Banach et al. (2023) finds mortality reductions even up to 20,000 steps per day, while Paluch et al. (2022) finds that reductions taper to negligible levels after 8,000–10,000 steps per day (sooner for adults over 60).

[55]We collected RBS at baseline and endline for the first approximately 1,500 participants we enrolled, until new rules instituted by the Indian Council on Medical Research prevented us from continuing RBS testing.

each 1,000 steps reduced RBS by 5.3 mg/dl in the full sample and by 8.6 mg/dl in the higher-RBS sample (see the Online Supplement Table F.1).

We also find substantial reductions in waist circumference (0.45 cm per 1,000 daily steps). BMI is unchanged, potentially suggesting muscle gain alongside fat loss. BP is unaffected.

Literature on Monetary Benefits of Exercise Exercise also decreases both private and public health care costs (e.g., Anokye et al., 2018; Cobiac et al., 2009; Johnson et al., 2015; Sangarapillai et al., 2021). The World Health Organization (2018) estimates that each $1 spent on programs to increase activity in lower and middle income countries generates $2.80 in cost savings.

Appendix E.1 provides back-of-the-envelope estimates of the public and private cost savings from inducing steps among diabetics in India. These estimates combine data on healthcare costs in India with studies on cost savings and complication risk reductions from exercise in similar populations.[56] Table E.1 summarizes the results. The mean and median estimates of the public cost savings, corresponding to the linear externality our Section 2 principal maximizes, are both 1.3 INR per 100 steps (range: 0.3–2.4 INR), while the mean and median estimates of the value of the private health benefits range are both 2.4 per 100 steps (range: 0.5–4.5 INR).[57] These estimates are likely conservative as they do not account for the persistent impacts of step target incentives (Aggarwal et al., 2024), which would significantly amplify the benefits.[58]

6.3 Cost-Benefit Analysis of Personalization

When comparing the costs and benefits of incentive strategies so far, our cost estimates only included the direct cost of the incentive payments (e.g., Sections 4.5 and 6.1). This section incorporates design and implementation costs into the comparison of personalized versus one-size-fits-all incentives. We report costs for all personalized treatments, but focus our discussion on those with statistically significant step increases at the 5% level: Choice, Tag, and All Variables Tag.

While personalized incentives are more expensive to design and implement than a one-size-fits-all approach, our Section 6.2 estimates of the benefit of steps to the principal (i.e.,

[56]We focus on diabetics as they are the likeliest target for a scale-up by GoTN, and there is the strongest evidence for the cost savings impacts of exercise in this population. We also provide one estimate from the general population. While we estimate larger treatment effects of Choice among diabetics than in the full population, we conservatively use the full-sample estimates for cost-effectiveness calculations.

[57]Since these estimates appear to come from populations with similar baseline exercise levels to our population, the healthcare cost impact per step from our intervention is likely to be similar. Johnson et al. (2015) and Yates et al. (2014) report (unconditional) baseline pedometer counts of 6,645 and 6,245, while unconditional baseline pedometer counts in our sample are 6,800. (Two of the studies do not report baseline exercise in a manner that we can translate to our study.)

[58]The estimates assume that extra steps today have no impact on steps tomorrow. However, Aggarwal et al. (2024) finds substantial persistence: 50% as large an effect in the 3 months *after* payments end as the 3 months of payment, which increases the benefits by 50% even without further persistence.

the externality) indicate that the principal still prefers Choice and Tag to Fixed Medium even at small program scales.[59] In contrast, the All Variables tag—which is substantially more costly to design and implement than other approaches—is only preferred to Fixed Medium at large scales, and is always dominated by both Choice and Tag.

Program Scales We estimate design and implementation costs of the personalized strategies relative to Fixed Medium and then evaluate their relative cost-effectiveness at three program scales: (1) 7,000 people—the annual number of newly diagnosed diabetics in the city of Coimbatore and our experimental sample size; (2) 170,000 people—all diabetics in Coimbatore; and (3) 11.6 million people—all diabetics in Tamil Nadu, aligning with the statewide scale-up goal of our GoTN partnership.[60]

Payment Costs Column 5 of Table E.2 reports the additional payments (above Fixed Medium) needed to generate 100 additional steps for each personalization strategy. As noted in Section 4.5, Choice's payment costs are nearly the same as Fixed Medium, costing just 0.11 INR per 100 additional steps—an order of magnitude below our median externality estimate of 1.3 INR per 100 steps. Tag and All Variables have similarly low payment costs while achieving similar increases in steps (Section 6.1). Thus, considering payment costs alone, the estimated externality benefits of Choice, Tag, and All Variables all vastly exceed the costs. While our point estimates suggest that Tag generates additional steps at the lowest cost, its differences from Choice and All Variables are small and statistically insignificant.

Design Costs Design costs (column 3 of Table E.2) are minimal for Tag (no additional design costs) and low for Choice (just a small pilot to understand preferences), but high for All Variables, which required experimental data from the Fixed groups to train a machine learning model.

Incorporating both payment and design costs (as shown in columns 6–8 of Table E.2), the cost per 100 steps at the small program scale is thus higher for Choice than Tag (0.83 INR versus 0.05 INR) because of Choice's design cost, but the gap narrows at the medium scale. Across all scales, the costs of both strategies remain below the median externality benefit estimate. In contrast, the All Variables tag is far more costly at the small scale (19.66 INR per 100 steps) and is not cost-effective at that scale.

Implementation Costs Implementation costs (column 4 of Table E.2) may also increase with personalization. While Choice's implementation costs were modest, the data require-

[59]The principal in Section 2 aims to maximize the externality $g(s)$ net of program costs. A program is preferred if the externality generated by its additional steps exceeds its additional cost.

[60]We focus on diabetics here, as opposed to diabetics and hypertensives since (a) there is stronger evidence of the health impacts of walking among diabetics than hypertensives, as discussed in Section 6.2, and (b) our discussions with GoTN regarded scaling the program up for diabetics, since our first evaluation of step target incentives (Aggarwal et al., 2024) focused on diabetics only.

ments for tagging raised costs—especially for the All Variables tag, which required extensive data collection.

After incorporating these (and all other) costs (columns 9–11 of Table E.2), Choice and Tag are both preferred to Fixed Medium at all program scales. While Tag may be preferred to Choice at small scales due to its lack of fixed design costs, Choice is preferred to Tag at large scales as it lacks costly personal data requirements. In contrast, tagging using All Variables is dominated by Choice and Tag at any scale. It is not preferred to the one-size-fits-all benchmark even at medium program scales and does not become preferred until the program reaches a scale of 274,000 people.

Since implementation costs vary with program design, it is also useful to consider how they might differ in other settings. While the cost of Choice could have been further reduced even using our experimental infrastructure (e.g., by collecting choices via phone surveys), reducing the higher cost of tagging would require additional technology and/or data. For example, the cost of Tag would decrease with technology that automates pedometer data syncing. Tagging via All Variables would potentially become cost-effective in a setting with rich administrative data, such as a developed country or in the private sector.

7 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 not been frequently 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 on average steps by 80% without significantly increasing payments. Moreover, Choice is more effective than non-personalized incentives across the full distribution of steps, likely first-order stochastically dominating each Fixed contract. Choice also outperforms the most scalable tagging approach (the Policy Variables tag) while achieving step increases comparable to tagging strategies with higher data requirements (Tag and All Variables tag). 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 increases their steps most, relative to their payments. While nonstandard 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 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 adjustments in their cost function over time (e.g., due to random shocks or habit formation). Dynamic approaches to Choice could be contrasted with static approaches and with dynamic tagging approaches, as have been implemented in some apps (e.g., Kramer et al., 2020). A final direction for future work is to evaluate different processes for designing choice menus. For example, a different and more expensive approach to design a menu would be to conduct a full design experiment upfront that randomizes contract features and estimates their impacts by type, as done in Abubakari et al. (2024) to design a menu for selling clean fuel. Future work can compare different approaches to menu design to determine their relative performance and costs.

References

Abubakari, S., K. Asante, M. Daouda, B. K. Jack, D. Jack, F. Malagutti, and P. Oliva (2024). Targeting subsidies through price menus: Menu design and evidence from clean fuels. *Working paper*.

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 (2024). Designing incentives for impatient people: An RCT promoting exercise to manage diabetes. *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. *Journal of Political Economy 124*.

Anokye, N., J. Fox-Rushby, S. Sanghera, D. G. Cook, E. Limb, et al. (2018). Short-term and long-term cost-effectiveness of a pedometer-based exercise intervention in primary care: A within-trial analysis and beyond-trial modelling. *BMJ open 8*(10), e021978.

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. Handel, E. Miguel, and G. Rao (2021). Self-control and demand for preventive health: Evidence from hypertension in india. *Review of Economics and Statistics 103*(5), 835–856.

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

Banach, M., J. Lewek, S. Surma, P. E. Penson, A. Sahebkar, et al. (2023). The association between daily step count and all-cause and cardiovascular mortality: A meta-analysis. *European journal of preventive cardiology 30*(18), 1975–1985.

Banerjee, A., R. Hanna, B. A. Olken, and S. Sumarto (2020). The (lack of) distortionary effects of proxy-means tests: Results from a nationwide experiment in indonesia. *Journal of Public Economics Plus 1*, 1–9.

Barrera-Osorio, F., M. Bertrand, L. L. Linden, and F. Perez-Calle (2011). Improving the design of conditional transfer programs: Evidence from a randomized education experiment in Colombia. *American Economic Journal: Applied Economics 3*(2), 167–195.

Barrett, G. F. and S. G. Donald (2003). Consistent tests for stochastic dominance. *Econometrica 71*(1), 71–104.

Beaman, L., D. Karlan, B. Thuysbaert, and C. Udry (2023). Selection into credit markets: Evidence from agriculture in Mali. *Econometrica 91*(5), 1595–1627.

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.

Beshears, J., J. J. Choi, D. Laibson, B. C. Madrian, and K. L. Milkman (2015). The effect of providing peer information on retirement savings decisions. *The Journal of finance 70*(3), 1161–1201.

Björkegren, D., J. E. Blumenstock, and S. Knight (2024). Manipulation-robust prediction. *Unpublished manuscript*.

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.

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.

Byrne, D. P., A. L. Nauze, and L. A. Martin (2018). Tell me something I don't already know: Informedness and the impact of information programs. *Review of Economics and Statistics 100*(3), 510–527.

Cao, Z., J. Min, H. Chen, Y. Hou, H. Yang, et al. (2024). Accelerometer-derived physical activity and mortality in individuals with type 2 diabetes. *Nature Communications 15*(1), 5164.

Caria, A. S., G. Gordon, M. Kasy, S. Quinn, S. O. Shami, and A. Teytelboym (2024). An adaptive targeted field experiment: Job search assistance for refugees in Jordan. *Journal of the European Economic Association 22*(2), 781–836.

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.

Cobiac, L. J., T. Vos, and J. J. Barendregt (2009). Cost-effectiveness of interventions to promote physical activity: a modelling study. *PLoS medicine 6*(7), e1000110.

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

del Pozo Cruz, B., M. N. Ahmadi, I.-M. Lee, and E. Stamatakis (2022). Prospective associations of daily step counts and intensity with cancer and cardiovascular disease incidence and mortality and all-cause mortality. *JAMA internal medicine 182*(11), 1139–1148.

del Pozo-Cruz, J., F. Alvarez-Barbosa, D. Gallardo-Gomez, and B. del Pozo Cruz (2022). Optimal number of steps per day to prevent all-cause mortality in people with prediabetes and diabetes. *Diabetes care 45*(9), 2156–2158.

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.

Gallardo-Gómez, D., E. Salazar-Martínez, R. M. Alfonso-Rosa, J. Ramos-Munell, J. del Pozo-Cruz, et al. (2024). Optimal dose and type of physical activity to improve glycemic control in people diagnosed with type 2 diabetes: A systematic review and meta-analysis. *Diabetes Care 47*(2), 295–303.

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 (2024). Contamination bias in linear regressions. *American Economic Review 114*(12), 4015–4051.

Gonzalez-Lira, A. and A. M. Mobarak (2019). Slippery fish: Enforcing regulation under subversive adaptation. *IZA Discussion Paper*.

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

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 (2023). Selection on welfare gains: Experimental evidence from electricity plan choice. *American Economic Review 113*(11), 2937–2973.

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

Jayachandran, S., J. D. 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*, 267–273.

Johnson, S., D. Lier, A. Soprovich, C. Mundt, and J. Johnson (2015). How much will we pay to increase steps per day? Examining the cost-effectiveness of a pedometer-based lifestyle program in primary care. *Preventive Medicine Reports 2*, 645–650.

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.

Kasy, M. and A. Sautmann (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica 89*(1), 113–132.

Kasy, M. and A. Teytelboym (2023). Matching with semi-bandits. *The Econometrics Journal 26*(1), 45–66.

Khongrangjem, T., S. Phadnis, and S. Kumar (2019). Cost of illness (COI) of type-II diabetes mellitus in Shillong, Meghalaya. *International Journal of Diabetes in Developing Countries 39*, 201–205.

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

Kramer, J.-N., F. Künzler, V. Mishra, S. N. Smith, D. Kotz, et al. (2020). Which components of a smartphone walking app help users to reach personalized step goals? Results from an optimization trial. *Annals of Behavioral Medicine 54*(7), 518–528.

Lee, L. L., C. A. Mulvaney, Y. K. Y. Wong, E. S. Chan, M. C. Watson, et al. (2021). Walking for hypertension. *Cochrane Database of Systematic Reviews* (2).

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.

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

Ministry of Labour and Unemployment (2016). Report on fifth annual employment - unemployment survey (2015-16). Technical report, Labour Bureau, Government of India, Chandigarh.

Misra, A., P. Chowbey, B. Makkar, N. Vikram, J. Wasir, et al. (2009). Consensus statement

for diagnosis of obesity, abdominal obesity and the metabolic syndrome for Asian Indians and recommendations for physical activity, medical and surgical management. *Japi 57*(2), 163–70.

Mitchell, M. S., S. L. Orstad, A. Biswas, P. I. Oh, M. Jay, 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 (2023). Factorial designs, model selection, and (incorrect) inference in randomized experiments. *The Review of Economics and Statistics*, 1–44.

Mussa, M. and S. Rosen (1978). 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.

National Health Systems Resource Centre (2024). National health accounts estimates for India (2021-22). New Delhi: Ministry of Health and Family Welfare, Government of India.

Paluch, A. E., S. Bajpai, D. R. Bassett, M. R. Carnethon, U. Ekelund, et al. (2022). Daily steps and all-cause mortality: A meta-analysis of 15 international cohorts. *The Lancet Public Health 7*(3), e219–e228.

Sangarapillai, T., M. Hajizadeh, S. S. Daskalopoulou, and K. Dasgupta (2021). Cost-comparison analysis of a physician-delivered step-count prescription strategy. *CJC open 3*(8), 1043–1050.

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.

Woerner, A., G. Romagnoli, B. M. Probst, N. Bartmann, J. N. Cloughesy, and J. W. Lindemans (2024). Should individuals choose their own incentives? Evidence from a mindfulness meditation intervention. *Management Science*.

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

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

World Health Organization (2022b). Global status report on physical activity 2022.

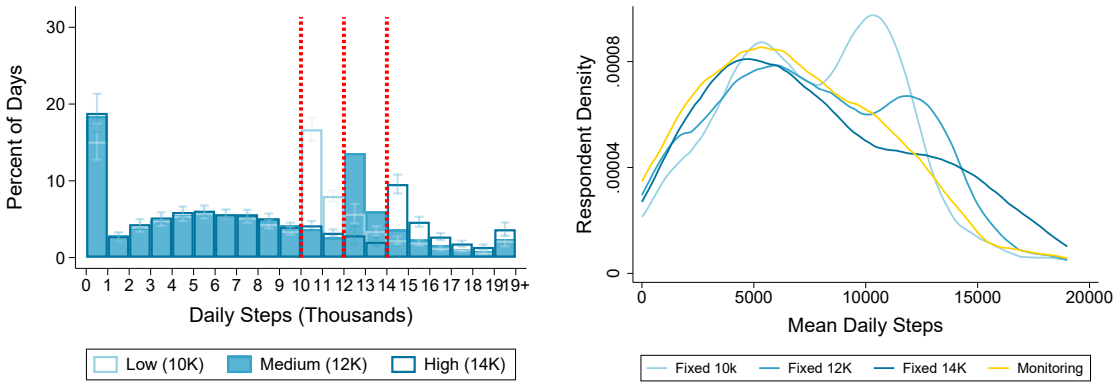
Yates, T., S. M. Haffner, P. J. Schulte, L. Thomas, K. M. Huffman, et al. (2014). Association between change in daily ambulatory activity and cardiovascular events in people with impaired glucose tolerance (navigator trial): A cohort analysis. *The Lancet 383*(9922), 1059–1066.

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 - E. 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

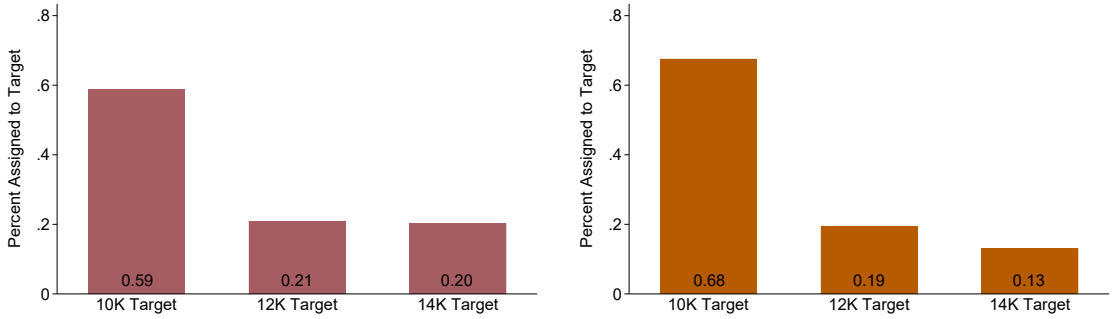
Appendix Figure A.1: Impact of Step Targets on Steps



(a) Daily Step Results (b) Average Daily Step Results

Notes: Panel (a) displays histograms of daily steps during the contract period in the Fixed groups. 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 Table 2. Panel (b) displays kernel density plots of individual-average daily steps across the contract period.

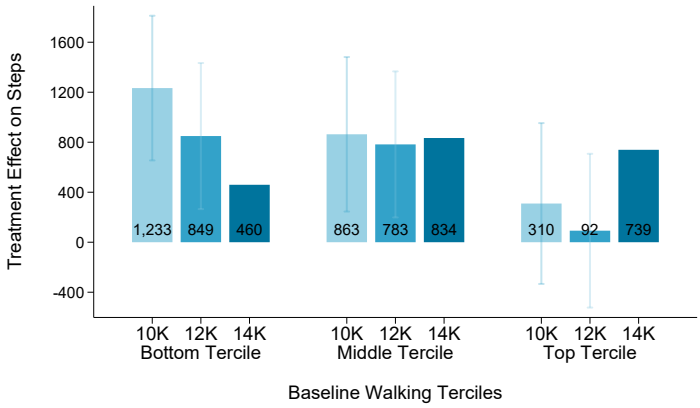
Appendix Figure A.2: Step Target Distribution in Choice



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

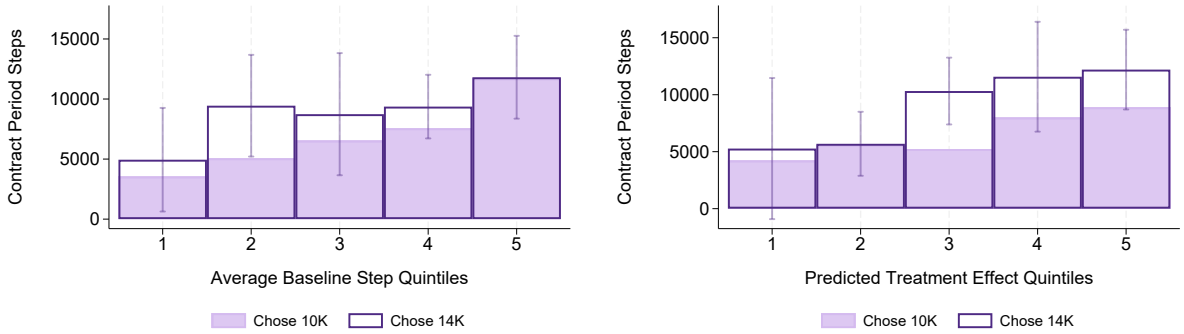
Notes: Panel (a) displays the percentage of Choice participants who chose each of the three targets from the Base Menu. Panel (b) displays the percentage of Flat Choice participants who chose each of the three targets from the Flat Menu.

Appendix Figure A.3: Heterogeneity in the Performance of Step Targets by Baseline Steps



Notes: The figure shows the treatment effects of the Fixed groups relative to the Monitoring group for each baseline step tercile (bottom tercile: $<$ 5171 steps; top: $>$ 8217 steps). The 95% confidence intervals are relative to Fixed High, controlling for the experiment phase, the time between the Baseline and Choice surveys, receiving the Nudge, year-month fixed effects, and controls selected by double-Lasso for the middle tercile from the controls in column 1 of Table A.5.

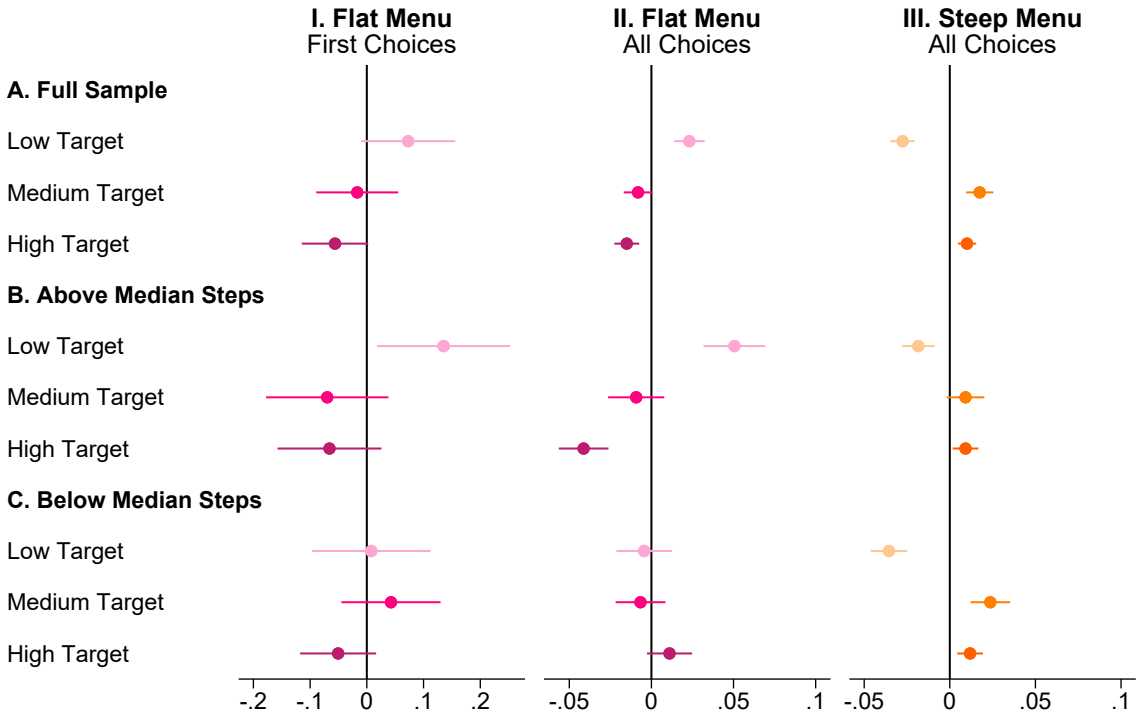
Appendix Figure A.4: Variation in Contract-Period Steps by Choices on the Choice Menu, Conditional on Baseline Steps or Predicted Treatment Effects



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

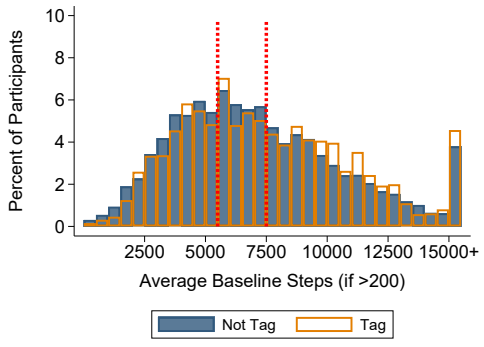
Notes: The figure shows contract-period walking in the Monitoring group, separately for those who chose the High (14K) target (shaded bars) and those who chose the Low (10K) target (outlined bars) from the Base Menu during the Choice survey. Panel (a) further splits the sample by quintiles of baseline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

Appendix Figure A.5: Chosen Step Targets on Flat and Steep Menus Relative to Base Menu

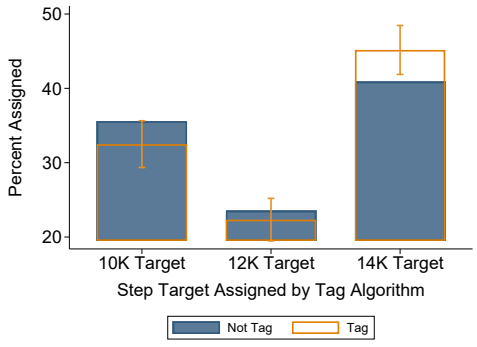


Notes: The figure shows the difference in (and 95% confidence intervals for) the fraction of participants choosing each step target on the Flat Menu (sub-graphs I and II) and the Steep Menu (sub-graph III), both compared to the Base Menu. Sub-graph I is restricted to choices from the first menu shown; 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. All regressions control for experiment phase, time between the Baseline and Choice surveys, receiving the Nudge, and controls selected by double-Lasso for the middle tercile from the controls in column 1 of Table A.5.

Appendix Figure A.6: Baseline Steps and Assigned Step Targets in Tag vs. Other Groups



(a) Baseline Steps



(b) 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 the Tag target assignment algorithm (Table B.1) had been applied to *unmanipulated* baseline steps. The confidence interval bars represent tests of equality between the likelihood individuals are assigned to each step target at the 95% confidence level. Regressions in Panel (b) include 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 the Baseline and Choice surveys, and year-month fixed effects for the date of the Baseline survey.

Appendix Table A.1: Balance in Pre-Contract-Launch Withdrawals

Omitted Group:	Not Tag or Baseline Choice	
	Withdrew Before Contract Launch	Withdrew Before Contract Period
	(1)	(2)
Tag	0.0148	0.0124
	[0.0102]	[0.0121]
Baseline Choice	0.00258	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 restricted to those who completed the Baseline survey up to the point that treatment was revealed to Tag. Controls include experiment phase, time between the Baseline and Choice surveys, and year-month fixed effects for the date of the Baseline survey. Additionally, column-specific controls are selected by double-Lasso for each column from the list of controls in column 3 of Table A.5. Robust standard errors are in brackets. 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	6,417	28%

Notes: This table reports statistics on participant dropout at each stage of the experiment design. Critically, dropout is extremely limited following contract launch in the Choice survey, when the majority of the treatment groups were assigned. The most common reasons given for withdrawing between the Baseline survey and contract launch in the Choice survey (i.e., between lines 3 and 4) are busy schedule (40%), not motivated (30%), and health issues/concerns (29%). 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 (fewer than 50) to their menu choice from the Steep Menu in order to make choices on this menu incentive-compatible. This group is omitted from all analyses; however, they are included here since they were enrolled and screened with the rest of the sample.

Appendix Table A.3: Balance in Attrition Across Treatment Groups

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	0.00444	-0.00507	-0.0191
	[0.0119]	[0.00731]	[0.0132]
Fixed Medium	0.00594	-0.00137	0.0120
	[0.0109]	[0.00671]	[0.0125]
Fixed High	0.0110	-0.00453	0.0166
	[0.0119]	[0.00727]	[0.0134]
Tag	0.00433	-0.0151**	0.00206
	[0.0110]	[0.00613]	[0.0124]
Flat Choice	0.0238	0.00691	-0.00120
	[0.0167]	[0.0104]	[0.0158]
Baseline Choice	0.0243*	-0.00227	0.00351
	[0.0142]	[0.00801]	[0.0137]
Choice + Nudge	0.0139	-0.00789	-0.00168
	[0.0133]	[0.00842]	[0.0176]
Monitoring	0.0215	-0.0152*	0.000665
	[0.0197]	[0.00897]	[0.0223]
Choice Mean	0.08	0.04	0.17
p-value vs Fixed Medium			
Fixed Low	0.872	0.535	0.008
Fixed High	0.595	0.605	0.714
Tag	0.874	0.014	0.418
Flat Chocie	0.284	0.428	0.413
Baseline Choice	0.187	0.909	0.536
Choice + Nudge	0.407	0.323	0.361
Monitoring	0.413	0.109	0.608
p-value vs Monitoring			
Fixed Low	0.383	0.260	0.380
Fixed High	0.591	0.234	0.484
Tag	0.378	0.993	0.951
Flat Choice	0.923	0.058	0.938
Baseline Choice	0.899	0.180	0.901
Choice + Nudge	0.707	0.454	0.925
p-value Fixed High vs Fixed Low	0.539	0.934	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 Choice. The dependent variable in column 1 is a person-level indicator for missing all of their contract period data. In column 2, it is a person-day level indicator for missing data on a given day, 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 not wearing the pedometer (recorded fewer than 200 steps) conditional on the pedometer data not being missing in column 2. The sample includes all treatment groups. All columns include controls for experiment phase, time between Baseline and Choice surveys, receiving the Nudge, and year-month fixed effects for either Baseline survey date (column 1) or day (columns 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 column 1) and column 1 (for columns 2 and 3). The analysis is conditioned on being in our main analysis sample that was present at the Contract Launch.
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 (individual × day level), the lower bound is 680 (standard error 399) and the upper bound is 1317 (standard error 443). For Tag vs Choice (individual × day level), the lower bound is -194 (standard error 266) and the upper bound is 200 (standard error 347). For Baseline Choice vs Choice (individual level), the lower bound is -212 (standard error 365) and the upper bound is 106 (standard error 341). We also report Lee bounds to account for any differential attrition following *Baseline survey* (instead of Choice survey) completion across the treatment groups that were revealed at Baseline (Baseline Choice and Tag) and their key comparison group (Choice, which was revealed at contract launch), all at the individual level. These are calculated for those in the sample at the end of the Baseline survey. For Tag vs Choice, the lower bound is -59 (standard error 311) and the upper bound is 58 (standard error 300). For Baseline Choice vs Choice, the lower bound is -147 (standard error 371) and the upper bound is 48 (standard error 313).
Standard errors (robust for columns 1; clustered at the individual level for columns 2 and 3) are in brackets. Significance levels: * 10%, ** 5%, *** 1%.

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 and Exercise 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	1083	7096	7123	7147	7121	7116	7163	6982	7201	7078	6882
p-values for joint orthogonality of covariates versus:												
Choice			0.971	0.747	0.147	0.376		0.425	0.115			
Fixed Med			0.558	0.160		0.022	0.147	0.461	0.599	0.112	0.536	
Monitoring				0.970	0.558	0.771	0.971	0.767	0.828	0.994	0.621	
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. Diagnosed diabetic and diagnosed hypertensive are self-reported by participants. BP refers to blood pressure, and BMI refers to 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 informed of their treatment, baseline steps exclude these groups. 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 estimated from a column-specific regression. Columns 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 covariates were selected 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.3 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 (i.e., partialled out).

Appendix Table A.6: Treatment Effects on Payments, Relative to Fixed Medium

Omitted Group:	Fixed Medium
Dependent Variable:	Daily Payments
	(1)
Choice	0.47
	[0.29]
Fixed Low	2.22***
	[0.31]
Fixed High	-1.40***
	[0.29]
Tag	0.22
	[0.31]
Flat Choice	0.96**
	[0.38]
Baseline Choice	0.39
	[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	0.401
Flat Choice	0.170
BL choice	0.784
Choice + Nudge	0.158
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 the experiment phase, the time between the Baseline and Choice survey, receiving the Nudge, and year-month fixed effects. Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.7: Treatment Effects of Choice Relative to the "Reweighted Fixed" Group

Omitted Group:	Reweighted Fixed	
Dependent Variable:	Daily Steps	Daily Payments
	(1)	(2)
Choice	342*	-0.45
	[184]	[0.28]
Monitoring	-588*	-6.35***
	[322]	[0.21]
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, and year-month fixed effects. Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.8: Robustness of Treatment Effect Estimates 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	420**	438**	384**	450**	551*	518**
	[202]	[210]	[176]	[207]	[296]	[204]
Fixed Low	90	62	232	72	91	57
	[185]	[192]	[161]	[188]	[224]	[57]
Fixed High	176	169	156	182	175	199
	[208]	[215]	[178]	[212]	[252]	[199]
Tag	455**	466**		497**	494**	539***
	[205]	[213]		[212]	[248]	[539]
Flat Choice	104	130	34	96		65
	[252]	[266]	[222]	[255]		[65]
Baseline Choice	342	381		359	742*	319
	[225]	[234]		[230]	[429]	[319]
Choice + Nudge	82	27	132	96	80	38
	[239]	[248]	[205]	[246]	[247]	[38]
Monitoring	-528	-414	-445	-529	-890*	-583*
	[333]	[348]	[281]	[340]	[496]	[-583]
Fixed Medium Mean	7,720	7,720	7,720	7,770	7,859	7,720
p-value vs Choice						
Fixed Low	0.115	0.084	0.408	0.077	0.145	
Fixed High	0.282	0.255	0.241	0.249	0.257	
Tag	0.867	0.900		0.828	0.845	
Flat Choice	0.199	0.235	0.101	0.155		
BL choice	0.724	0.806		0.692	0.674	
Choice + Nudge	0.234	0.163	0.298	0.227	0.186	
Monitoring	0.005	0.016	0.004	0.005	0.007	
p-value vs Monitoring						
Fixed Low	0.067	0.176	0.017	0.081	0.053	
Fixed High	0.044	0.109	0.040	0.047	0.040	
Tag	0.004	0.014		0.004	0.007	
Flat Choice	0.083	0.155	0.119	0.093		
Choice + Nudge	0.110	0.269	0.072	0.111	0.064	
p-value Fixed High vs Fixed Low	0.694	0.633	0.679	0.618	0.758	
# Observations	172,961	172,961	130,571	172,961	101,328	54,241
# Individuals	6,384	6,384	4,825	6,384	3,713	1,994
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: 353; Fixed 10K: 510; Fixed 12K: 922; Fixed 14K: 544; Tag: 646; Baseline Choice: 142; Choice + Nudge: 523; Monitoring: 73.
This Table is the same as Table 3, but shows coefficients for all treatment groups. The dependent variable is daily steps measured using the contract-period pedometer data. Column 1 is the same as Table 2. Columns 2–3 show robustness to different sets of controls, and column 4 to not winsorizing the outcome variable. 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. (2024) 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. Year-Month fixed effects are included in all columns other than column 2. Additional controls are selected by double-Lasso, and listed in the notes to Table 3. Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.9: Quantile Regression Results: Fixed Groups Relative to Choice

Omitted Group:	Choice		
Dependent Variable:	Individual-Average Steps		
Percentile:	25	50	75
	(1)	(2)	(3)
Fixed Low (10K)	-262	-195	-759**
	[266]	[321]	[295]
Fixed Medium (12K)	-500*	-441	-417
	[268]	[294]	[301]
Fixed High (14K)	-725***	-776**	-53
	[251]	[313]	[405]
Monitoring	-1282***	-1289**	-1425***
	[434]	[502]	[453]
Choice Quantiles	4,372	7,640	11,014
p-val Fixed Low vs. Fixed High	0.058	0.076	0.062
# 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: The 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, and year-month fixed effects for the date of the Baseline survey. In addition, since there is no double-Lasso command for quantile regression, each column includes Lasso-selected controls selected for an OLS regression with an indicator that the participant's steps were above median. Robust standard errors are in brackets. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.10: Heterogeneity in the Impacts of Step Targets by Baseline Steps

Dependent Variable:	Daily Steps	Daily Payments
	(1)	(2)
Step Target (1,000s) ×	41***	-0.01
Baseline Steps (1,000s)	[15]	[0.02]
Baseline Steps (1,000s)	137	0.99***
	[181]	[0.26]
Step Target (1,000s)	-305***	-0.86***
	[111]	[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 1,000s). The sample includes the Fixed groups only. The dependent variable is daily steps in column 1 and daily payments in column 2. 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, and year-month fixed effects. Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.11: Correlations Between Choices and Baseline Steps or Predicted Treatment Effects

Dependent Variable:	Chosen Step Target (Steps)		
	(1)	(2)	(3)
Baseline Steps	0.181***		0.209***
	[0.0123]		[0.0148]
Predicted Treatment Effect		4.752***	-2.031**
		[0.745]	[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 1,000s) of the step target chosen on the Base Menu. The sample includes only the Choice group. 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: Variables Used in Causal Forest and Their Importance Values

Variable name	Included in Policy Variable Prediction?	Importance Value
	(1)	(2)
Baseline steps	No	0.20
Mental health index	No	0.12
Age	Yes	0.11
Weight (kg)	Yes	0.09
Systolic BP	Yes	0.09
Diastolic BP	Yes	0.07
Waist circumference (cm)	Yes	0.06
BMI	Yes	0.05
Height (cm)	Yes	0.05
Diagnosed diabetic	Yes	0.02
Female	Yes	0.02
Number of smartphones owned	No	0.01
Household size	Yes	0.01
Home has running water	No	0.01
Owns home	No	0.01
Above median education level	Yes	0.01
Number of mobilephones owned	No	0.01
Dianosed hypertensive	Yes	0.01
Number of rooms in home	No	0.01
Number of scooters owned	No	0.00
Married	Yes	0.00
Participating in labor force	No	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: Heterogeneity in the Impacts of Step Targets by Chosen Step Target

Dependent Variable:	Daily Steps	Daily Payments
	(1)	(2)
Assigned Target (1,000s) ×	95***	0.08
Chosen Target (1,000s)	[36]	[0.05]
Chosen Target (1,000s)	-314	0.29
	[434]	[0.65]
Assigned Target (1,000s)	-1063***	-1.88***
	[396]	[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 1,000s) with assigned step target assignment (in 1,000s). The sample includes only the Fixed groups. 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, and year-month fixed effects. Standard errors, in brackets, are clustered at the individual level. Significance levels: * 10%, ** 5%, *** 1%.

Appendix Table A.14: Health Impacts of Marginal Steps

	Blood sugar		Other health outcomes		
Dependent variable:	Random blood sugar	Health risk index	Mean arterial BP	BMI	Waist circum-ference
	(1)	(2)	(3)	(4)	(5)
Panel A. Full sample					
Average Steps (1000)	-7.23*	-0.027	0.10	-0.0012	-0.45*
	[4.14]	[0.017]	[0.47]	[0.059]	[0.27]
Monitoring mean	243.1	0.0	105.8	26.2	94.1
# Individuals	1,520	5,429	5,610	5,614	5,451
Panel B. Above-median blood sugar sample					
Average Steps (1000)	-12.5*	-0.089**	-0.31	-0.090	-0.73**
	[7.14]	[0.043]	[0.76]	[0.082]	[0.30]
Monitoring mean	325.5	0.6	104.0	26.4	99.4
# Individuals	765	750	763	766	753

Notes: This table shows the effect of average steps per day (in 1,000s) during the contract period on health outcomes from an IV specification. Panel A includes the full sample; Panel B includes those with above-median baseline RBS. The health risk index is the average of RBS, mean arterial BP, BMI, and waist circumference standardized (with missing RBS imputed using the average RBS in the Monitoring group). Instruments are dummies for each incentive treatment group (Choice, Tag, Fixed Low/Med/High, Flat Choice, Baseline Choice, Choice + Nudge); Monitoring is omitted. All specifications control for experiment phase, time between Baseline and Choice surveys, and Baseline survey year-month fixed effects (Panel G in Table A.5 column 3). Additional controls are selected by double-Lasso from the following list: the baseline value of the outcome, its missing dummy, the controls listed in Table A.5 column 3, as well (for column 2 only) the baseline value of all components of the health risk index and their missing dummies. Robust standard errors are reported in brackets. Significance levels: * 10%, ** 5%, *** 1%.

B Design of Contracts, Choice Menus, and Tag Mechanisms

In this appendix, we first outline our rationale for using "step target" contracts rather than alternative contract structures. We then describe the process through which we designed our choice menu, tag treatments, and fixed contracts. Finally, we provide empirical evidence that the model underlying our design process performed well in practice.

B.1 Rationale for Step Target Contracts

We discuss our rationale for employing step target contracts, first in comparison to linear contracts and then to linear contracts after a step target.

Why Step Target Instead of Linear Contracts? There are two key reasons we employ step target contracts instead of linear contracts. First, step targets facilitate sorting because they have two parameters, the step target and the wage, which allows for differentiation across contracts. These two parameters allow for menus where one contract appeals to one type and the other contract appeals to another, as shown in Maskin and Riley (1984). For example, a low-wage low-target contract may appeal to low walkers, while a high-wage high-target contract appeals to high walkers. In contrast, linear contracts have a single parameter, the piece rate, and so all types will simply prefer the linear contract with the highest piece rate. Additional design features (e.g., adding non-payment-related frictions to contracts) are thus necessary to achieve sorting by type.

Second, step target contracts can generate compliance at a lower cost than linear contracts. Specifically, for any given participant and any given linear contract, a principal can always design a step target contract that induces the same level of walking as the linear contract but at a lower cost. The reason is that step target contracts only pay the exact cost for the additional steps needed to meet the target, while linear contracts pay for all steps at the same rate as the marginal cost of the final (most expensive) step.[61] This makes step targets particularly advantageous in settings with significant inframarginal behavior like ours (average baseline steps exceed 7,000): this behavior is essentially free under step target contracts but paid at the marginal cost of the most expensive step in linear contracts. Even if the effect of the treatment on walking were 20% (a substantial treatment effect, larger than what we see), over 80% of payments in a linear contract would go to inframarginal steps.

However, the ability of a principal to match the walking level of a linear contract at a lower cost using a step target contract requires having accurate and precise information about each type's cost function. If the principal is uncertain or inaccurate, or if there is heterogeneity in cost functions within type, the principal may prefer linear contracts. The intuition is that a single linear contract can increase steps from participants with a wide range of cost functions, making them more robust to errors in the principal's estimates of participants' cost functions. In contrast, step target contracts are tailored for a specific cost function, so if the principal misjudges that function, the contract will not be tailored correctly. The contract could even entirely fail to change behavior if the principal either sets the target too high for the payment, or sets it too low (i.e., below baseline steps). Linear contracts, while costly, still increase steps when cost curves differ from expectations.

[61] To see this mathematically, let $\tilde{c}(s;\theta)$ be participant's net private cost of steps $(c(s;\theta) - b(s))$. A linear contract that pays k per step will increase steps to the level s^k where the marginal net cost of steps is k (i.e., $\tilde{c}'(s^k;\theta) = k$), and will cost ks^k. A step target contract can increase steps to the same level s^k by paying $\tilde{c}(s^k;\theta)$. Since net step costs are convex, the step target contract pays less: $\tilde{c}(s^k;\theta) = \int_0^{s^k} \tilde{c}'(u;\theta)du < ks^k$.

Why Step Target Instead of Linear-After-Step-Target Contracts? With two parameters (step target and payment rate) linear-after-step-target contracts can naturally support sorting. They can also be as cost-effective as step target contracts,[62] while maintaining much of the robustness of linear contracts to mis-specification of the cost function.[63] However, we did not use linear-after-step-target contracts for three reasons, two behavioral and one logistical. First, our piloting and previous data collection suggested that participants struggle to understand these contracts.[64] Likely as a result, existing data suggested that there is less sorting by type across such contracts than across step target contracts.[65] Simple step target contracts, in contrast, are easier for participants to understand and resulted in more separation across types. Second, evidence identifies a realistic daily goal as a key component of effective physical activity incentives (Mitchell et al., 2020), as it may improve the performance of inattentive participants. With linear-after-a-target contracts, the salient number is the target, which is not what the designer wants the participants to hit (they want the participants to walk beyond the target).

Finally, to select full-information contracts, we employed a model of how participants' walking would respond to different step target contracts (rather than a more primitive model of net walking costs) based on data from a previous evaluation of step target contracts (as described next). In contrast, modeling participant responses to linear-after-step-target contracts would have required costly experimentation in the design phase.

B.2 Selecting the Full-Information and One-Size-Fits-All Contracts

This section describes how we selected the "full-information" contracts—that is, the contracts that the policymaker would assign to each participant type (low, medium, and high walkers) if type were known—as well as the one-size-fits-all contract.

[62] For a given net cost function, the cost-minimizing step target and linear-after-step-target contracts pay the same amount to generate a given number of steps. A step target contract can generate s^k steps by paying net costs $\tilde{c}(s^k;\theta)$. A linear-after-step-target contract generating s^k steps must pay at least this amount to satisfy the participation constraint, and can do so by setting the payment rate $k = \tilde{c}'(s^k;\theta)$ and step target $\hat{T} = s^k - \frac{\tilde{c}(s^k;\theta)}{k}$).

[63] For example, the policymaker can reduce the risk of choosing too high a target (and hence eliciting no effort) by setting a conservative target and then paying participants linearly after that.

[64] During the pilot phase of Aggarwal et al. (2024), in which we used a form of linear-after-a-target contract (see footnote 65 for details), our field team struggled to ensure participants fully understood the linear-after-a-target contract. To address this, we invested significant time testing different explanation strategies, incorporating visual aids. Despite these efforts, even with the clearest explanation and visuals we could develop, 8–12% of participants answered basic understanding questions about the contract incorrectly, compared with just 0–1% for linear contracts (the comparison contract in that project).

[65] Data from Aggarwal et al. (2024) suggest limited correlation of choices between linear-after-a-target contracts and type (proxied with baseline steps), ranging from 0.03 to 0.06 across choices. In contrast, the correlations between contract choice and baseline steps in this study are an order of magnitude higher (>0.3). A caveat is that Aggarwal et al. (2024) offered a different type of linear-after-a-target contracts than we would have considered for this experiment: they featured linear payments after achieving a step target on a target number of *days* rather than for a target number of *steps*. Specifically, all contracts had a 10,000 daily step target, but paid participants per day of walking 10,000 steps *only* if the participant did so on at least a target number of days per week—e.g., at least 4, 5, or 0 (the last of which is a linear contract over days). In contrast, the linear-after-a-target contracts we would have implemented in this setting would have paid linearly for steps taken above a target within a day. However, we were concerned about similar challenges to understanding and sorting for linear-after-a-target contracts.

Previous Evaluation Our design process used data from Aggarwal et al. (2024), an existing 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. (2024).[66]

Full-Information Contracts for Each Type As mentioned in Section 2.2, we used terciles of the baseline step distribution among participants in Aggarwal et al. (2024) to define three discrete types. We then used a simple model of how steps respond to contracts (estimated using Aggarwal et al. 2024 data) to select the three round-number step targets that would maximize average steps for each type given the 20 INR payment level; importantly, our model also implies these step targets would maximize principal surplus at the 20 INR payment level as long as the externality is sufficiently large (at least 0.4 INR / 100 steps). We first describe the model and how we used it to estimate step-maximizing targets, and then show the evidence that these targets are surplus-maximizing.

Modeling the Response to Step Target Contracts To estimate the relationship between steps and contracts, we first assumed that each person's net cost curve is a horizontally shifted version of the others'. This implies that, for a given payment level, the treatment effect of a step target contract on average steps in the contract period is a function *only* of the gap between a participant's baseline steps (which is uniquely determined by their net cost curve) and the step target. Thus, we can use the heterogeneous treatment effects of a 10,000 step target contract paying 20 INR offered in Aggarwal et al. (2024) to estimate the treatment effect of any step target contract paying 20 INR on a person with any baseline step level. For example, the treatment effect of a 10,000 step target for participants with 5,000 baseline steps would be the same as the treatment effect of a 12,000 step target for participants with 7,000 baseline steps. Moreover, the step-maximizing target for a given participant at a given payment level will equal baseline steps plus a constant.[67]

We then non-parametrically estimated heterogeneous treatment effects of the 10,000 step target contract from Aggarwal et al. (2024), relative to a Monitoring group (i.e., a group that did not receive incentives), according to participants' baseline steps. Figure B.1a shows the estimated treatment effects for participants binned into 1,000-step-width bins. The function has a roughly inverted U shape in baseline steps, with a peak among participants who walked 4,000–5,000 steps at baseline.[68] Under our modeling assumptions, this suggests that the step-maximizing target for each participant for a 20 INR payment rate is around 5,000–6,000 steps higher than baseline steps, and that the treatment effect of a target b steps above a participant's baseline would be the same as the treatment effect observed in Figure

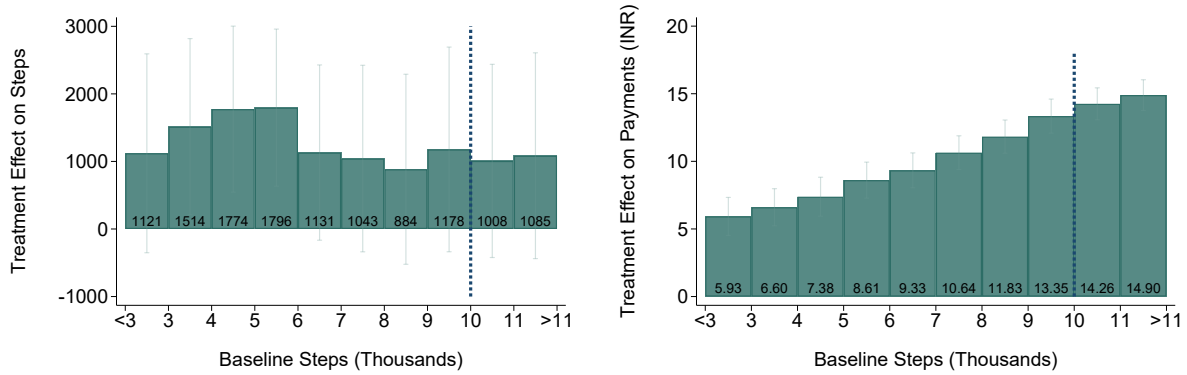
[66]The primary differences are that we shortened the contract period from twelve to four weeks and that we offered multiple step targets instead of only one. Aggarwal et al. (2024) also evaluated more complicated contracts, such as the linear-after-a-target number of days discussed in Appendix B.1. We exclude data from participants offered these contracts when estimating treatment effects.

[67]Interestingly, paired with our linear externality assumption, this model of net costs also implies that the contracts that globally maximize principal surplus for each type, searching across the full contract space of payments and step targets, would all feature the same payment level but different step targets. This in turn implies that there exists a budget such that maximizing principal surplus subject to this budget constraint would yield the same full-information contracts that we select.

[68]The estimated treatment effect in this bin is significantly larger than the treatment effects in all other bins combined or in the surrounding three bins (p-values 0.046 and 0.036, respectively, with p-value relative to the surrounding two bins 0.110.

B.1a among participants with $10,000 - b$ steps.

Appendix Figure B.1: Treatment Effects from Aggarwal et al. (2024), by Baseline Steps



(a) Treatment Effects on Steps

(b) Treatment Effects on Payments

Notes: Data from Aggarwal et al. (2024). The figures display the treatment effects of the 10,000-step-target-incentives on daily steps walked and payments during the contract period by participants' average baseline steps. Participants are grouped into bins of 1,000 baseline steps. Participants with more than 15,000 baseline steps are excluded. We estimate the treatment effect of incentives relative to the Monitoring group (whose steps were monitored but who received no incentive) by baseline step bins, controlling for baseline step bin fixed effects and all standard controls used in regressions in Aggarwal et al. (2024). The treatment effects of each step bin shown in the figures are smoothed by averaging the treatment effect of the bin itself with those of its immediate neighboring bins (one on each side).

Selecting Full-Information Contracts We next used the treatment effect estimates to select the step-maximizing target for each of the three types, assuming that the baseline step distributions in each type would closely resemble those in Aggarwal et al. (2024). Specifically, we used Figure B.1a to estimate average contract-period steps for each type under a set of round-number step targets, and chose the step-maximizing target for each type among these.[69] This process yielded full-information contracts paying 20 INR with step targets of 10,000, 12,000, and 14,000 steps for the Low, Medium, and High types, respectively, as shown in Table B.1.

Appendix Table B.1: Full-Information Contracts

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

[69]Since we estimated these treatment effects only in 1,000-step-width bins of baseline steps (e.g., 3,000–4,000 steps; 4,000–5,000 steps), we apply the same treatment effect estimates for any participant whose baseline steps fall within the same 1,000-step bin. Consequently, we can only estimate the treatment effects of step targets rounded to the nearest 1,000. For each type, we searched among the five (rounded) targets in the range from 10,000 through 14,000.

Estimating Principal Surplus From the Full-Information Contracts We assume in Section 2.2 that the externality of steps is linear. Thus, the principal aims to maximize expected average steps, multiplied by the per-step externality, less expected average payments. Having already estimated average steps under each contract, to estimate principal surplus, we next estimated expected average payments for each contract.

We use the same assumption of horizontally-shifted cost curves to estimate a model of average payments to each participant under a given contract, in a manner analogous to our model of average steps. Specifically, Figure B.1b shows average payments of the 10,000-step-target, 20-INR incentive contract from Aggarwal et al. (2024) as a function of participants' baseline steps. Our modeling assumptions imply that the average payments made to a participant under a step target that is b steps above their baseline steps would be the same as the payments observed in Figure B.1b among participants with $10,000 - b$ steps.

We can then calculate expected principal surplus for each type under each target, for any assumption of the per-step externality. The full-information contracts assigned to each type, though initially selected to maximize steps, also maximize the principal's surplus from that type (in the explored contract space) if the per-step externality is at least 0.2 INR/100 steps, 0.4 INR/100 steps, and > 0 INR/100 steps, for the low, medium, and high types, respectively.

Selecting the One-Size-Fits-All Contract We also used the same model to select our one-size-fits-all contract. Now we aimed to choose the average-step-maximizing contract for the full sample (instead of for each type), again at the 20 INR payment rate. Assuming again that the type distribution would mirror Aggarwal et al. (2024), the model implied that the average-step-maximizing target for the full sample at the 20 INR payment rate would be 12,000. Hence, we used this contract as our one-size-fits-all contract. Moreover, the model implies that this target also maximizes principal surplus in the explored contract space as long as the per-step externality is at least 1.4 INR per 100 steps.

B.3 Validating our Tag (Full-Information Contract) Algorithm

We now provide evidence that the set of full-information contracts (shown in Table B.1) that we developed to assign participants, based on their baseline steps, to targets that would increase their step counts does in fact accomplish its goal. We begin with reduced form evidence and then turn to evidence from machine learning.

B.3.1 Reduced Form Evidence

Our full-information algorithm suggested that the 10K, 12K, and 14K step targets would generate the most steps for the low, medium, and high walkers, respectively. We can use evidence from the Fixed groups to provide evidence that this is the case. Figure A.3, which shows the performance of each of the Fixed groups in each of those groups, shows that this aligns with the data. While all targets perform similarly in the middle group, in the bottom group, the low 10K target generates the most steps (p-value 0.023), while in the top group, the high 14K target generates the most steps (p-value 0.047).[70] We can also directly test whether participants in the Fixed groups walked more steps if they were randomly assigned

[70]This figure cuts the sample at the terciles of the baseline step distribution, as it aims to show the patterns of the step targets across the distribution of steps. While the cut points based on terciles are slightly different than the cut-points between categories in our Tag algorithm, they are very similar: $<$5171, 5171$-$8217, and $>$ 8217 for the terciles versus $<$ 5500, 5500–7500, and $>$ 7500 for the bins used in our algorithm). The figure using the cut-points from our Tag algorithm is very similar, with the same ordering of bars in each bin and

to the contract that was the same as their full-information contract assignment (e.g., the 10K target if they were a low walker or the 14K target if they were a high walker). Conditional on fixed effects for the randomly assigned target and for their full-information contract, we find that being assigned to the contract our algorithm said would be best for them increases average steps by 273 (p-value 0.068). This is a meaningful increase, equal to roughly 50% of the treatment effect of Fixed Medium.

B.3.2 Policy tree evidence

We can also directly compare our Tag algorithm for mapping from steps to step maximizing targets (Table B.1) to an algorithm created by machine learning, shown in Table B.2. Specifically, we use the policy tree machine learning algorithm of Athey and Wager (2021) in our Fixed groups to determine which step target maximizes steps for each individual based on their baseline steps.

Appendix Table B.2: Policytree Assignment Algorithm

Baseline Steps	Assigned Step Target
<4,650	10,000 steps
4,650–5,650	14,000 steps
5,650–7,350	12,000 steps
>7,350	14,000 steps

Notes: This table shows the results from using the policy tree machine learning algorithm of Athey and Wager (2021) in our Fixed groups to determine which step target will maximize steps for each individual based on their baseline steps. To estimate the policy tree itself, we use the policy_tree method of the policytree package in R. All parameters take default values except for tree.depth, where, to improve interpretability and reduce overfitting, we show results for a tree depth of 2 (the cross-validated tree depth).

The algorithms themselves are relatively similar; e.g., the cutoff for assignment to the 14,000 step target instead of the 12,000 step target is 7,350 for policy tree versus 7,500 for ours. However, the policy tree algorithm is not monotonic and assigns some low-step participants to the highest step target—which is inconsistent with standard theory and so may be a data anomaly. Indeed, using the same "synthetic group" method used in Section 6.1 of the paper, we actually estimate that assignment based on our algorithm would outperform assignment based on the policy tree algorithm, although not significantly. We estimate that assigning participants based on our algorithm would increase steps by 269, bringing it 60% of the way from Fixed Medium to Choice. In contrast, we estimate that assignment based on the policy tree algorithm shown in Table B.2 would achieve only a 43 step gain relative to Fixed Medium, bringing it just 10% of the way from Fixed Medium to Choice.[71]

the p-values for the low vs other targets in the low bin 0.084 (instead of 0.023) and for the high vs other targets in the high bin 0.136 (instead of 0.047).

[71]Note that these policy tree results use the cross-validated tree depth. To check the robustness of this result to the depth, we also ran the policy tree with all depths from 2–5 (where 5 is the maximum depth recommended for the method given our sample size). The depth that yields the highest estimate of impact according to the synthetic group method is depth 5, which may be overfitting, given it cuts the sample into $2^5 = 32$ groups based on baseline steps—and is hence highly non-monotonic—while the cross-validated depth only cut it into $2^2 = 4$. That said, even that version does not outperform our algorithm, achieving a gain of 244 steps relative to Fixed Medium, bringing it roughly 54% of the way from Fixed Medium to Choice. Thus, overall, we find no evidence that the policy tree algorithm would outperform our own.

C Appendices Describing Experimental Design and Analysis

C.1 Description of Experiment Phases

In this section, we describe the three phases (and six subphases) of the experiment mentioned in Section 3. We preregistered the design elements introduced in each phase in the AEA registry (Dizon-Ross and Zucker, 2020). Table C.1 summarizes the design changes.

Treatment Group Changes We introduced seven treatments in phase 1: Choice, Tag, the three Fixed Groups, Monitoring, and the Choice + Nudge group (the Nudge was also cross-randomized to 60% of the Monitoring and Fixed groups, the same fraction who received it among the pooled Choice and Choice + Nudge groups). In phase 2, we introduced the *Baseline Choice* group, but did not adjust the randomization balance of the initial treatments. In phase 3, we eliminated the Nudge (defining subphase 3a), and then introduced the *Flat Choice* group and changed the treatment balance among the initial treatment groups (defining subphase 3b), increasing the relative size of Choice and Monitoring but decreasing the relative size of Fixed Medium.[72]

Changes to Choice Survey Timing We changed another element of our design within the main experiment phases: the Choice survey timing. As described in footnote 14, we randomly added an additional week to the typical six days between the Baseline survey and the second visit for some participants. We introduced this variation shortly before phase 2 (defining subphase 1b), cross-randomizing the additional week to 93% of participants in all treatment groups. We maintained this 93% cross-randomization rate through phase 2, but adjusted it early in phase 3 (defining subphase 3c).

All analyses control for a 6-level categorical variable ("experiment phase") representing the subphases, which are summarized in Table C.1. We also control for each participants' Choice survey timing and for whether participants received the Nudge.

Appendix Table C.1: Phases of the Experiment

Sub. Phase	Start Date	Treatment Groups							Nudge Cross-Randomization: Share	Additional Time between Baseline and Choice: Share
		Choice	Tag	Fixed	Monitoring	Flat Choice	Baseline Choice	Choice + Nudge		
Phase 1a	May 15, 2019	X	X	X†	X†			X‡	60%	0%
Phase 1b	Oct 31, 2019	X	X	X†	X†			X‡	60%	93%
Phase 2	Dec 9, 2019	X	X	X†	X†		X	X‡	60%	93%
Phase 3a	Jan 28, 2020§	X	X	X	X		X		0%	93%
Phase 3b	Jan 28, 2020	X	X	X	X	X	X		0%	93%
Phase 3c	Feb 18, 2020	X	X	X	X	X	X		0%	25%

Notes: § indicates that the start date of phase 3a is different from others—the start date of phase 3a refers to the date of the Choice survey, whereas the other start dates refer to the date of the Baseline survey. X indicates that a treatment group was included in the design in a given phase; † indicates that some fraction of the treatment group in the given phase was cross-randomized to receive the Nudge (with the share given in the "Nudge Cross-Randomization: Share" column); ‡ indicates that all participants in the treatment group received the Nudge. The choice timing was cross-randomized across all treatments.

[72] For logistical reasons, we eliminated the Nudge based on the timing of participants' Choice survey but changed the other treatments based on the timing of participants' Baseline survey. We lump both changes into phase 3 (rather labeling the second as "phase 4") since both went into effect on the same date. Subphase 3a includes the small set of participants who had completed Baseline but not Choice on this date.

C.2 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 and without unlimited calling;[73] 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 6,532 eligible 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).

C.3 Prediction of 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 Lasso regression coefficients to create individual-level predictions of baseline steps in all groups, including Tag and Baseline Choice.

C.4 Choice Survey: Scripts and Order

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

During 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.5 Causal Forest Estimation and Synthetic Tag Construction

C.5.1 Causal Forest Estimates for Sorting Analysis

Among participants in the Fixed groups, we use the multi arm causal forest method implemented by the grf package in R to predict the treatment effect of the High (14K) relative

[73] We exclude individuals with unlimited calling plans because they are less likely to respond to incentives provided as mobile recharges.

to Low (10K) target (our predictor variables are listed in Table A.12).[74] 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. We used a *multi-arm* causal forest to be consistent with the machine-learning procedure used to estimate the best step target for each participant (which we describe next), which requires a multi-arm causal forest. Results from a single-arm causal forest are similar.

C.5.2 Policy Tree Assignments for All Variables and Policy Variables Synthetic Tags

To estimate the step-maximizing target for each participant, 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 sets of predictors:

- *All Variables Synthetic Tag*: The variables used in the causal forest estimation described in Appendix Section C.5.1.
- *Policy Variables Synthetic Tag*: The variables above but excluding (a) baseline steps, and (b) all wealth variables (see column 1 of Table A.12 for the specific variables excluded).

To estimate the policy tree, we used the hybrid_policy_tree method of the policytree package in R. All parameters take default values except tree.depth, where we show results for a depth of 5 (the maximum depth for which the Athey and Wager (2021) results hold given our sample size). Results for depths 2–4 perform similarly (or worse). By comparing Synthetic Tags with the best performing tree depth to Choice, we present conservative estimates of the relative performance of Choice.

C.5.3 Constructing a Simpler Tag with Lasso

To assess the robustness of the Policy Variables results to a simpler process, we use a cross-validated Lasso regression to predict steps with the same variables as the main Policy Variables tag. We then apply the Tag algorithm in Table B.1 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

[74] We include 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). We exclude Panel C, predicted baseline steps, since we use actual baseline steps, and Panel E, time indicators, which we think a policymaker would be unlikely to use for prediction.

Appendix Table D.1: Robustness 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	420**	480**	435**	430**	
	[202]	[194]	[218]	[219]	
Choice + Nudge	82	-4		56	
	[239]	[223]		[262]	
Choice or Choice + Nudge					285*
					[167]
Fixed Low	90	89	23	27	87
	[185]	[185]	[242]	[242]	[185]
Fixed Low × Nudge				105	
				[375]	
Fixed High	176	177	310	329	172
	[208]	[208]	[264]	[264]	[208]
Fixed High × Nudge				-323	
				[426]	
Monitoring	-528	-503	-577	-566	-559*
	[333]	[331]	[372]	[371]	[332]
Monitoring × Nudge				294	
				[864]	
Nudge	-180			-122	-284*
	[179]			[257]	[155]
Fixed Medium Mean	7,720	7,720	7,631	7,720	7,720
p-value vs Choice					
Fixed Low	0.115	0.053	0.070	0.076	
Fixed High	0.282	0.174	0.619	0.686	
Choice + Nudge	0.234	0.050		0.272	
Monitoring	0.005	0.003	0.005	0.006	
p-values for the significance of the Nudge in Fixed Low, Fixed High, and Monitoring groups					
Nudge + Fixed Low × Nudge				0.955	
Nudge + Fixed High × Nudge				0.216	
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				0.338	
Monitoring × Nudge vs					
Fixed High × Nudge				0.492	
Monitoring × Nudge vs					
Fixed Low × Nudge				0.829	
# 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 Table 2. Column 2 is the same as column 1, but excludes the control for receiving the Nudge. Column 3 excludes all participants who received the Nudge. Column 4 interacts a control for receiving the Nudge 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, 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%.

by showing a "fully interacted" model. Specifically, the 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. Column 4 also shows that the Nudge is insignificant in each of the non-Choice groups, and 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. The pooled coefficient is fairly large (nearly 300 steps) and significant at the 10% level. However, it is smaller than the effect of Choice alone, reflecting the fact that the Nudge backfired for certain types of participants as shown in the Online Supplement.

E Cost-Effectiveness

E.1 Back-of-the-Envelope Estimates of the Financial Value of Steps

This section provides details on our back-of-the-envelope estimates of the financial benefits of exercise among people with diabetes. We first present estimates of the private and public healthcare costs among people with diabetes in India. We then present four estimates of the private and public cost savings from exercise in this population, where public cost savings correspond to the fiscal externality to a government policymaker.

E.1.1 Estimates of Healthcare Costs

Khongrangjem et al. (2019) estimates the total cost of illness borne by diabetic patients (i.e., the total private costs) in India, including direct costs such as expenses on medications and procedures, as well as indirect costs measured as productivity losses. The median monthly cost of illness per diabetic patient is estimated at 5,307 INR per month, which corresponds to a daily cost of 176.91 INR (5307 INR / 30 days), of which 125.46 INR (roughly 70%) are direct costs and the remainder indirect.

While we could not find direct estimates of the public healthcare costs per patient, the Government of India estimates that, for every 1 INR of private household expenditure on health, the government spends roughly 0.768 INR (National Health Systems Resource Centre, 2024). We apply this factor to the 125.46 INR *direct* private benefit estimate from the previous paragraph; we assume there are no public benefits from the indirect benefits (the 30% productivity benefits), which is conservative as the government would also in reality experience some fiscal benefit from the indirect productivity benefits via the tax system. We thus estimate that the average daily government cost per day per diabetic is 96.37 INR (125.46 INR direct private cost × 0.768 INR public costs for every 1 INR in private costs).

E.1.2 Estimates of Healthcare Savings From Steps

We searched the literature for any studies that estimated the percent change in cost or in cardiovascular disease events from steps in similar populations to ours. For each of the four studies we found, we combine the study's estimate with the total healthcare cost estimate from Appendix E.1.1 to create a back-of-the-envelope estimate of the healthcare cost savings (i.e., benefits) from steps. These estimates are shown in Table E.1, and the underlying studies and calculations are described further in Section I of the Online Supplement. Our estimates of the public cost savings (or public externality) per 100 steps walked range from 0.3–2.12 INR, and our median estimate is 1.30.

Appendix Table E.1: Estimates of the Cost Savings from Exercise

Study	Study Description	Key Estimate From Study	Implied Cost Savings Per 100 Steps Walked	
			Private Cost Savings	Public Cost Savings
(1)	(2)	(3)	(4)	(5)
A. Studies Estimating Cost Savings From Steps				
Johnson et al. (2015)	Experiment estimating the impact of pedometer-based walking intervention among diabetics over 6 months on health costs	20.24% reduction in avg health cost over 6 months from an increase in average daily steps of 919 over that period	3.90	2.12
Anokye et al. (2018)	Experiment estimating the impact of pedometer-based walking intervention among 45–75 year olds over one year on health costs	16.6% reduction in avg health cost in 1 year from an increase in average daily steps of 660 over that year	4.46	2.43
Di Loreto et al. (2005)	Observational study comparing the change in physical activity over 2 years (relative to baseline) in response to an intervention with the change in health care costs in 2 years (relative to baseline) among diabetics	6.9% cost savings from every additional 1 mile (2252 steps) walked per day	0.54	0.30
B. Studies Estimating Cardiovascular Events Prevented By Steps				
Yates et al. (2014)	Observational study comparing the cardiovascular event risk over a 5 year period with a baseline measure of walking	0.5% reduction in CVD event risk for each additional 100 steps per day	0.88	0.48
Median Estimates			2.39	1.30

Notes: In each row, we estimate the implied cost savings in columns 4 and 5 by multiplying estimates of the daily total private and public costs by the percentage reduction in cost per 100 steps shown in Column 3. The daily total healthcare cost estimates for diabetics are described in Section E.1.1; the estimates are 176.91 INR for private costs (Khongrangjem et al., 2019) and 96.37 INR for public costs (176.91 INR $\times$ 0.7681 $\times$ 0.7092). As an example, in Row 1, column 4 is calculated as 20.24% $\times$ 176.91 INR $\times \frac{100}{919} = 3.90$ INR. Similarly, Column 5 is calculated as 20.24% $\times$ 96.37 INR $\times \frac{100}{919} = 2.12$ INR.

E.2 Design and Implementation Costs of Personalization

This section briefly describes the design and implementation costs of our personalization treatments *over-and-above* those of the Fixed Medium treatment, and then incorporates these into the estimated cost of each treatment relative to Fixed Medium.

Design Costs: The fixed design costs, in column 3 of Table E.2, include the following:[75]

- *Choice* and *Baseline Choice*: The cost of the 70-person pilot to gather preference data to

[75]Note that, although we used the Aggarwal et al. (2024) data to design all treatments, we do not include the costs of gathering or analyzing these data as we calculate design costs *relative* to the design costs of Fixed Medium for which we also used the Aggarwal et al. (2024) data. Morevoer, the Aggarwal et al. (2024) data already existed before we began the design process for this experiment.

design the incentive-compatible menu (see Section 3.2)

- *Synthetic Tags that use Machine Learning* (*Policy Variables* and *All Variables*): This includes the cost of conducting our experiment in the Fixed arms, as developing these tags relied on data from the Fixed Arms.[76]
- *Tag* and *Synthetic Tag - Unmanipulated Steps*: No additional costs since we developed these using the same approach used to develop Fixed Medium (see Appendix B.2).

Implementation Costs Column 4 of Table E.2 shows the additional per-person implementation costs by treatment group, relative to the cost of implementing the one-size-fits-all approach (Fixed Medium), for the treatments as we implemented them in our experiment. While we did not measure these costs directly during the experiment, our field team estimated them *ex post*. Specifically, we estimate the cost of the following implementation activities:

- *Choice:* 5 minutes at the Choice survey visit to elicit contract choices from the Base Menu.
- *Synthetic Tag - Policy Variables*: 15 minutes at the Baseline survey visit to measure the policy variables and calculate assigned target.
- *Synthetic Tag - All Variables*: 25 minutes at the Choice survey visit to (a) sync pedometers to measure steps and (b) measure additional variables needed to implement the tag and calculate assigned target.[77]
- *Synthetic Tag - Unmanipulated Steps:* 10 minutes at the Choice survey visit to sync steps and calculate assigned step targets.
- *Tag Group*: 5 minutes at the Baseline survey visit to explain the Tag algorithm and 10 minutes at the Choice survey visit to sync steps and calculate assigned step targets.
- *Baseline Choice*: 5 minutes at the Baseline survey visit to elicit contract choices from the Base Menu.

[76]To estimate this cost, our field team built a budget based on our actual experimental cost, projecting the cost of a smaller experiment limited to the Fixed arms and with fewer measurements (e.g., excluding health measurements). Despite these reductions, the experiment remains expensive due to the large sample sizes needed to detect small differences across the Fixed treatments.

[77]In the current study, all variables but steps were measured at Baseline; however, since it is cheaper to measure everything at once and baseline steps could only be measured after the precontract period, we budget to measure all variables at the Choice visit.

Appendix Table E.2: Costs and Benefits of Personalization Relative to Fixed Medium

	Additional Benefits	Additional Costs			Additional Costs (INR)/ Additional 100 Steps						
					Payments Only	Payments + Design Costs			Payments + Design + Implementation Costs		
	Steps (per person)	Payments (INR/person)	Design cost (INR 1,000)	Implemen -tation cost (INR/person)		7K people	170K people	11.6M people	7K people	170K people	11.6M people
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Choice	11,760	13	593	21	0.11	0.83	0.14	0.11	1.01	0.32	0.29
Tag Group	12,740	6	0	57	0.05	0.05	0.05	0.05	0.50	0.50	0.50
Synthetic Tags											
Policy Variables	-812	1	18,946	53	NA	NA	NA	NA	NA	NA	NA
Unmanipulated Steps	6,692	-3	0	36	-0.04	-0.04	-0.04	-0.04	0.50	0.50	0.50
All Variables	13,888	24	18,946	88	0.17	19.66	0.97	0.18	20.29	1.61	0.81
Baseline Choice	9,576	11	593	21	0.11	1.00	0.15	0.11	1.22	0.37	0.33

Notes: This table shows the benefit-cost analysis of different treatment strategies. Columns 1–4 report results relative to the Fixed Medium group. Columns 1 and 2 are the total per-person treatment effects on steps and payments over the 28-day contract period; estimates represent the daily estimates from Tables 2 and A.6, respectively, each multiplied by 28. Column 3 shows the actual fixed costs we paid to develop each approach, relative to the cost of developing the Fixed Medium approach, as described in Section E.2. Column 4 shows the additional per-person implementation costs by treatment group, relative to the cost of implementing the One-Size-Fits-All approach, for implementing the treatments as in our experiment, as described in Section E.2. Columns 5–11 report costs per additional 100 steps, for the treatments that increased steps above the One-Size-Fits-All group only. The seven columns include different cost components when calculating the costs, with column 5 including only the total payments per person (column 2), columns 6–8 including both total payments and fixed cost (columns 2 and 3), and columns 9–11 including all additional costs (columns 2–4). Columns 9–11 are calculated using the formula: (Col 2 × Number of participants + Col 3 × 1000 + Col 4 × Number of participants) / (Col 1 / 100 × Number of participants); columns 5–8 omit the cost components described above. Columns 6 and 9 assume 7,000 participants, roughly the annual newly diagnosed diabetics in Coimbatore as well as the size of our experimental sample. Columns 7 and 10 assume 170,000 participants, the estimated number of diabetics in Coimbatore. Columns 8 and 11 assume 11,600,000 participants, the estimated number of diabetics in Tamil Nadu. We display NA's in columns 5–11 when the treatment does not generate more steps than Fixed Medium.