# Healthcare Appointments as Commitment Devices

Laura Derksen, Jason T. Kerwin,
Natalia Ordaz Reynoso, and Olivier Sterck*

May 27, 2024

**Abstract**

We show that ordinary appointments can act as effective substitutes for hard commitment devices and increase demand for a critical healthcare service, particularly among those with self-control problems. We show this using an experiment that randomly offered HIV testing appointments and hard commitment devices to high-risk men in Malawi. Appointments more than double testing rates, with effects concentrated among those who demand commitment. In contrast, most men who take up hard commitments lose their investments. Appointments overcome commitment problems without the potential drawback of commitment failure, and have the potential to increase demand for healthcare in the developing world.

*Keywords*: Appointments, Commitment Devices, Self-Control, Health, HIV
*JEL Classification*: D81, I15, O12

---

*Derksen: Ragnar Frisch Centre for Economic Research and Department of Management, University of Toronto (laura.derksen@frisch.uio.no); Kerwin: Department of Economics, University of Washington, IZA, and J-PAL (jkerwin@umn.edu); Ordaz Reynoso: AXA Research Lab on Gender Equality, Dondena Research Centre, Bocconi University (ordaz008@umn.edu); Sterck: IOB, University of Antwerp and Department of International Development, University of Oxford (olivier.sterck@uantwerpen.be). We thank Susan Watkins for her extensive guidance and mentorship on this project. We are grateful for insightful comments from Manuela Angelucci, Liang Bai, Leah Bevis, Lasse Brune, Gharad Bryan, Eric Chyn, Jon de Quidt, Stefano DellaVigna, Kate Dovel, Jessica Gallant, Erick Gong, Anett John, Maggie McConnell, Rachael Meager, Mauricio Romero, and Simone Schaner, and from seminar participants at PacDev, the Canadian Economics Association, BU, the KDI DIME conference, the EGROW Foundation, Maryland, and Rutgers, and from MPC, IAS, and the Applied Economics department at the University of Minnesota. We also thank the editor, Sule Alan, and three anonymous referees for their detailed and constructive comments. This project would not have been possible without the hard work of our fantastic fieldwork supervisor, Abdul Chilungo, the efforts of our wonderful team of survey interviewers and HDAs, and the participation of the men who generously shared their time with us to take part in the study. We thank Dina O'Brien, Ethan Sansom, Tim White, and Heather Wong for their excellent research assistance. Connaught, GATE and SRCHSS at the University of Toronto and CIFAP at the University of Minnesota provided grant funding for this research. Kerwin gratefully acknowledges support from the Minnesota Population Center (P2C HD041023) funded through a grant from the Eunice Kennedy Shriver National Institute for Child Health and Human Development (NICHD), and from a UMN Institute for Advanced Study Residential Fellowship. This study was reviewed and approved by IRBs in Malawi (National Health Sciences Research Committee, protocol # 19/03/2268) and at the University of Minnesota (protocol # STUDY00005587) and the University of Toronto (protocol # 00037491), and is registered with the AEA RCT Registry under registration number AEARCTR-0004295. All errors and omissions are our own.

# 1 Introduction

Health behaviors are plagued by commitment problems. The decision to seek healthcare, exercise more, or to eat a healthier diet requires people to incur short-term costs for long-term gains and to follow through with plans. This makes health behaviors prone to behavioral biases (Kessler and Zhang 2015), particularly self-control problems (DellaVigna and Malmendier 2006). One potential solution to these challenges is to use "hard" commitment devices with financial penalties (Laibson 1997; Gul and Pesendorfer 2001; Halpern, Asch, and Volpp 2012; Laibson 2015) or strict enforcement (see for example Dupas and Robinson 2013, Kim et al. 2019, and Sadoff, Samek, and Sprenger 2020). Hard commitment devices can encourage healthier behaviors in settings ranging from smoking (Giné, Karlan, and Zinman 2010) and drinking (Schilbach 2019) to gym attendance (Royer, Stehr, and Sydnor 2015).[1] They have also been used at scale in health policy, for example, to encourage adherence to tuberculosis treatment in Bangladesh (Islam et al. 2002). However, hard commitment devices have an important drawback: many people who demand hard commitment devices do not follow through (John 2020). In the average study of financial commitments tied to health behaviors, 65 percent of people lose the money they staked on their own behavior (Table 1). Indeed, there is evidence that these commitment devices can actually be welfare-diminishing (Bai et al. 2021).

A soft commitment, in the form of a scheduled healthcare appointment, offers an alternative to a hard commitment device. Booking an appointment is a commitment to come in for healthcare, with no enforcement or financial penalty.[2] By combining many different nudges in a single, natural intervention, appointments address several behavioural barriers to healthcare simultaneously, possibly with fewer downsides than hard commitments because there is no financial penalty for defaulting. Yet while health appointments appear to have large impacts on healthcare demand in some settings (Salvadori et al. 2020), they have zero impact in others (Chang et al. 2021).

[1] Commitment devices have also been extensively studied in non-health settings. For example, they can be an effective way to promote savings (Ashraf, Karlan, and Yin 2006*b*, Burke, Luoto, and Perez-Arce 2018), and increase work effort (Kaur, Kremer, and Mullainathan 2015). For a review of the literature on commitment devices, see Bryan, Karlan, and Nelson (2010).

[2] Some healthcare providers do charge fees for missing appointments, but these penalties can typically be avoided by providing notice. In Malawi there are no fees for missing healthcare appointments.

In this paper we show that healthcare appointments can act as a highly effective substitute for hard commitment devices, with particularly large effects for those who *demand* commitment due to self-control problems. We compare the effectiveness of healthcare appointments and hard commitment devices, and examine heterogeneity in their effectiveness by the demand for commitment, using a randomized field experiment. Our findings offer a potential explanation for why appointments work in some settings but not others. Scheduling an appointment is an effective way for a person with self-control problems to voluntarily commit to care. For people who actively avoid care, stronger incentives or policies may be required.

Increasing HIV testing is a major public health challenge in sub-Saharan Africa, as delays in treatment lead to unnecessary deaths and new infections.[3] Our data comes from a randomized experiment among 1,232 high-risk men recruited at urban bars in Zomba, Malawi. Participants were randomized to four study arms: a control group, appointments, hard commitment devices, or appointments plus hard commitment devices. Men in either appointments arm could commit to an HIV test at a clinic, date, and time of their choosing over the next two months. They received a phone call reminder two days before their appointment with the option to reschedule. The hard commitment device allowed men to stake USD $1.38 (approximately 50 percent of the median daily income per capita in our sample), taken out of their study compensation, on showing up at any testing site in Zomba City (where tests are available for free).

Offering men appointments sharply increases the HIV testing rate, raising it by 16 percentage points.[4] This is a 140 percent increase relative to the control-group testing rate of 11 percent. Using the randomized assignment as an instrument, we find that the appointments increase HIV testing among the two-thirds of men who sign up for them by 23 percentage points. We also find evidence suggesting that appointments increase the detection of actual HIV cases, and help guide HIV-positive men into treatment for the disease.

In contrast, hard commitment devices are less effective in our setting, and backfire for a

[3] Approximately 680,000 people died of AIDS in 2020, and 1.5 million became newly infected with HIV (UNAIDS 2020). Antiretroviral therapy prevents both death and HIV transmission, but many people living with HIV remain undiagnosed.

[4] All our inferences are robust to variations in the choice of controls and outcome variable, and to the Benjamini, Krieger, and Yekutieli (2006) false discovery rate correction for multiple testing.

substantial fraction of the men who sign up for them. About 50 percent of study participants demand a hard commitment, suggesting that they know that self-control issues are a barrier to HIV testing. But the effect on HIV testing is just 8 percentage points, half that of the appointments, and we can reject that the two effects are equal at the 0.05 level. Moreover, most of the men who signed up for a hard commitment device were made worse off: 64 percent of those who received only a commitment device lost the money they staked on getting tested. This failure of commitment parallels findings by John (2020), Bai et al. (2021), and Buehren et al. (2022), and is consistent with the average rate of commitment failure in health studies (Table 1).

We find that health appointments substitute nearly perfectly for hard commitment devices, and are particularly effective for people with known self-control problems. The cross-randomized experimental design allows us to test whether the interventions act as complements or substitutes, by comparing the combined arm to each intervention alone. Combining both appointments and hard commitment devices leads to an 18 percentage-point increase in HIV testing. This effect is only slightly higher than, and not significantly different from, the impact of appointments alone, suggesting that for our subjects appointments are a superior substitute for hard commitment devices. Looking at the study arm that received only appointments, the appointments are far more effective for men who wanted a hard commitment device but did not receive one—and for the subset of men who want a hard commitment device, getting only an appointment is almost as good as getting an appointment plus the commitment device.

Appointments are also a more cost-effective way of increasing HIV testing. We find that the cost per additional HIV test induced by an appointment was USD $2.69, as compared with USD $3.01 for a hard commitment device.[5] Appointments thus compare favorably with cash payments, which cost $11 per additional person who learns their test result (Thornton 2008). The cost-effectiveness of appointments for promoting HIV testing is particularly policy relevant given the importance of early diagnosis and treatment in preventing both AIDS deaths (INSIGHT START Study Group 2015) and the spread of HIV (Cohen et al.

[5] This calculation considers only the incremental cost of each intervention. We also consider alternative choices about how to compute costs.

2011). That these effects are for men is important: men in Malawi are less likely than women to seek treatment for HIV and more likely to die of AIDS (Dovel et al. 2015).

Standard healthcare appointments address several different behavioral biases simultaneously. Appointments are soft commitments that create social pressure to follow through, as men who fail to show up for their appointments waste the time of clinic staff. Indeed, we observe stronger effects among men who demand commitment. Appointments can also make a person feel expected, or welcomed, as in Nyondo et al. (2015). These social elements may explain why appointments appear to be more effective than making a private plan (e.g. Macis et al. 2021). Another related mechanism is that appointments might increase HIV testing through social pressure at the time of sign-up, due to experimenter demand effects (de Quidt, Haushofer, and Roth 2018). Experimenter demand effects are unlikely to fully explain our results for three reasons. First, all participants received information about HIV testing as well as an HIV testing voucher, which likely would have signaled the value of HIV testing more strongly than the appointment offer. Second, both demand and follow-through are higher for appointments than for hard commitments, despite similar levels of experimenter demand. Finally, the appointments had larger effects on HIV testing for men who demanded commitment devices, providing support for self control as a mechanism. Even if our results were fully explained by social pressure at sign-up, they would remain important from a policy perspective. Appointments are a cheap and scalable way to increase the take-up of healthcare services.

Appointments also come with reminders, which help overcome limited memory issues (Ericson 2011; Haushofer 2015; Ericson 2017). Reminders are a well-established public health intervention, with proven benefits for many health behaviors (Pop-Eleches et al. 2011; Gurol-Urganci et al. 2013; Altmann and Traxler 2014; Jacobson Vann et al. 2018; Banerjee et al. 2021). Indeed, the time pattern of HIV testing suggests reminders were important. Men in the appointments arm tested later in the study period, with many coming in even later than their appointment date. But reminders are not the only mechanism at play. In our study, conditional on receiving a reminder, 67 percent of participants who get tested do so on their exact appointment date. If the impact of appointments were due to reminders alone, we would expect visits to occur throughout the days or weeks after the reminder.[6] The spike in testing on the appointment date could reflect an increase in the salience of the date or a desire to avoid wait time at the clinic. This latter possibility is unlikely in our study context: supplementary data indicates that the HIV testing clinics in our sample (and across Malawi) operate below capacity and there is typically no wait. Indeed, we might expect larger effect sizes in other settings where wait times are longer.

Our findings provide insight on why appointments work in some contexts but not others: appointments appear to work best for those who would like to seek healthcare, but have problems following through. Appointments improve efficiency in clinics (Steenland et al. 2019), and have also been shown to increase HIV testing rates in a lower-prevalence setting (Salvadori et al. 2020). By focusing on high-risk men in a high-prevalence region, we show that appointments can be effective even when stakes are high. Moreover, by directly comparing appointments to hard commitment devices, and capturing each participant's demand for commitment, we are able to clarify *for whom* appointments are likely to make a difference. Scheduling an appointment is voluntary, and acts as a substitute for a harder commitment. Appointments appear to be effective specifically among those who deliberately demand such a commitment. We therefore expect to see large effects in settings where there is an underlying desire for care, but follow-through is limited by self-control problems. We would not expect to see large effects among those who make a conscious choice to avoid care. Indeed, Chang et al. (2021) find that appointments do not increase COVID vaccinations among a highly hesitant population.

Our findings also build on the literature on using commitment devices to promote health behaviors, by showing that appointments can overcome self-control problems more effectively than hard commitments, with fewer drawbacks. We compare the two types of commitments directly, using random assignment of participants from a single sample, in a setting where healthcare is both critical and subject to self-control problems. Other studies of health commitment devices evaluate a single intervention, or compare different types of hard commitments; Table 1 summarizes the existing literature on hard commitment devices to encourage

[6] This is consistent with Salvadori et al. (2020), who find that reminders explain only a third of the increase in HIV testing induced by appointments.

health behaviors.[7] The effects we estimate for hard commitment devices are within the range of those found in previous studies. In contrast, the effects we estimate for appointments compare favorably to hard commitments in other health contexts. Their take-up rate ranks in the top quintile, and unlike hard commitments, appointments do not cause people to lose money due to a failure to follow through. The treatment effect is four times as large as that of the average hard commitment device, and 54 percent larger than the most successful intervention from the existing literature (the combination of a personalized hard commitment and a discount from Bai et al. 2021).

More broadly, our paper relates to an extensive literature in behavioral finance comparing savings accounts with various levels of commitment and flexibility (Amador, Werning, and Angeletos 2006; Karlan and Linden 2023; Beshears et al. 2020; John 2020). Many effective soft commitment devices involve social contracts or opportunities for signalling (Ashraf, Karlan, and Yin 2006*a*; Kast, Meier, and Pomeranz 2018; Dupas and Robinson 2013). In the context of a health savings account, Dupas and Robinson (2013) show that while earmarking funds is largely ineffective, applying social pressure through savings groups is highly effective. This is consistent with the fact that healthcare appointments appear to be more effective than simply prompting health seekers to make a plan. The effectiveness of a soft commitment might also depend on how difficult the goal is to achieve. Social pressure from deposit collectors appears to have a significant, but moderate effect on savings (Ashraf, Karlan, and Yin 2006*a*). Visiting a clinic is likely easier than saving money; this might explain why healthcare appointments produce larger effects.[8]

This paper also contributes to the literature on behavioral nudges for health. Our study evaluates appointments as they are typically implemented, including all of their behavioral features and reminders, as well as any perceived reduction in wait time. Indeed, these features may be more effective as a bundle than as a sum of parts. The separate impacts of many of these behavioral nudges have been studied extensively. These include studies that involve nudges to increase appointment attendance, including defaults and reminders (Chapman et al. 2016; Milkman et al. 2021), as well as interventions that invite participants

[7] Schilbach (2019) presents a similar table focused only on take-up, including non-health studies as well.

[8] Saving money can be very difficult, but once it is saved, a soft commitment can promote expenditure for educational purposes (Karlan and Linden 2023).

to make a plan, or a private commitment, without offering a formal appointment (Milkman et al. 2011; Kavanagh et al. 2020; John and Orkin 2022). Planning prompts appear to affect behavior in some health contexts, but in the absence of any social commitment or reminder their measured effects are often small (Rogers et al., 2015; Macis et al., 2021). Reminders alone, without a planned appointment also appear to be less effective (Salvadori et al. 2020).

Our results suggest that wider use of medical appointments in developing countries could reduce healthcare under-utilization, which is common in such settings (Glasziou et al. 2017). The use of appointments in healthcare appears to be quite rare in Malawi: qualitative data from clinics in Zomba shows that only a few highly specialized services use appointments at all; those that do tend to only specify a day rather than a specific time, and do not provide reminders. Because many health services are underutilized, and are performed by low-skilled health workers, health providers may view appointments as logistically unnecessary and even burdensome. Indeed, introducing and integrating appointments into an existing healthcare system is complex and requires upfront investment. Some of these investments are already in place: mobile phones and smartphones are quite widespread in developing countries, making healthcare appointment scheduling and reminders technologically feasible. Moreover, health providers and policymakers may underestimate the demand for appointments among healthcare users, who benefit from the soft commitment they entail. Appointments represent a promising tool for addressing behavioral barriers to healthcare around the world.

## 2 Experiment and Data

We conducted a randomized controlled trial to assess and compare the impact of a soft commitment to seek an HIV test, in the form of a scheduled appointment, to a hard commitment device. The experiment took place in the city of Zomba in southern Malawi, where the prevalence of HIV is approximately 13 percent (DHS 2016). In Malawi, HIV testing, as well as most other services, are offered primarily on a walk-in basis, as opposed to by appointment. This fact is validated by supplemental data we collected by interviewing clinic staff (Appendix A.1 and Appendix A.2). HIV testing services are provided for free, and ART is provided for free to anyone who tests positive. There are 11 clinics that offer HIV testing in the city of Zomba. The study lasted for 3 months, from June 21$^{\text{st}}$ to September 30$^{\text{th}}$ of 2019.

Our sample consists of men from Zomba who are at high risk of HIV infection. We focus on men because they are less likely than women to be tested and treated for HIV, and more likely to die of AIDS (Dovel et al. 2015). One reason women in Malawi are more likely to be in HIV care is because they receive routine tests and treatment initiation as a part of antenatal care (National Statistical Office/Malawi and ICF 2017). HIV testing is supposed to be a voluntary part of antenatal care visits, but is perceived as compulsory (Angotti, Dionne, and Gaydosh 2011). Despite the fact that most Malawian men report a fairly recent clinic visit, more than half have not been tested for HIV in the past year (Dovel et al. 2020*a*).

We worked with local stakeholders to design a recruitment strategy that would target men at high risk of HIV infection, and to design an intervention that could be scaled to target a population that is otherwise difficult to reach. We decided to recruit participants at bars and nightclubs in Zomba. We selected these locations because they are commonly used by sex workers to find clients. Men who agreed to take part in the study and satisfied the screening criteria completed a baseline survey.[9] Every participant was offered a MK2,000 (approximately USD $2.76 at market exchange rates) gift of mobile phone credit at the end of the survey, and a MK500 mobile phone credit voucher that they could redeem by visiting any HIV testing clinic in Zomba within the three month study period. To redeem the voucher, the participant had to show the testing staff a text message on their mobile phone that contained a unique identifier code. This enables us to accurately link HIV tests to study participants, and capture all HIV tests that took place in the study area within the three-month study period. The voucher also helped to defray time or transportation costs of going to the clinic, and might have mitigated the stigma associated with HIV testing (Angelucci and Bennett 2021; Derksen, Muula, and van Oosterhout 2022). A financial reward for getting tested gives men a different "excuse" to go to the clinic (Thornton 2008; Ngatia 2016).

The list of participating clinics included all 11 HIV testing clinics within and around the city of Zomba, and this list was shared with all participants. The men in the study were

[9] We screened potential participants for mobile phone ownership and other criteria. Screening criteria and other implementation details can be found in Appendix B.1.

told that the vouchers had a two-month deadline for use, but this was not emphasized in the recruitment script. It also was not strictly enforced: HDAs were instructed to accept any vouchers presented to them while the study was ongoing, and had no way to check the date validity. We did ensure that the clinics continued to be staffed and accept men for testing for two months after the last participant was recruited into the study. Participants did not have to agree to an HIV test to redeem the MK500 voucher; both voucher redemption and actual HIV testing were recorded. Our main outcome variable is HIV testing, rather than simply voucher redemption; we show that our main results are robust to using any voucher redemption in Appendix Table D10.

We used a factorial design that randomly assigned men to one of four study arms: (1) a control group, (2) a group that was offered hard commitment devices only, (3) a group that was offered appointments only, and (4) a group that was offered both hard commitment devices and appointments. This experimental design has several key advantages. First, it allows for a head-to-head comparison of appointments and hard commitment devices in a single context. This is particularly important because there is substantial heterogeneity in the estimated effects of commitment devices across contexts—Table 1 shows treatment effects that range from 3 percent to 91 percent—so drawing conclusions from cross-study comparisons is difficult. Second, the factorial design lets us test whether appointments and hard commitments are substitutes or complements by comparing the combined study arm to the individual treatments. Specifically, it allows us to test the impact of adding a hard commitment device on top of an appointment, and vice versa. If appointments are superior substitutes for hard commitments, then we would expect the marginal effect of adding a hard commitment on top of an appointment to be small, and for the marginal effect of an appointment on top of a hard commitment to be large. Third, it allows us to better measure demand for a hard commitment device at the individual level, because we are able to observe credibly incentivized choices for the entire sample. We can therefore study the effect of appointments specifically for the subset of men who demand commitment.

Both interventions were offered to participants at the end of the baseline survey. Before offering either intervention, we elicited demand for the hard commitment device from all participants. For men who were assigned to the combined treatment, we offered the hard commitment device first, followed by the appointment. Figure 1 depicts the implementation process in more detail.

The hard commitment device allowed participants to stake half (MK1,000) of their study compensation on visiting an HIV testing clinic. That is, rather than receiving MK2,000 immediately and MK500 upon visiting the clinic, they could choose to receive MK1,000 immediately and MK1,500 at the clinic. Since this payment was conditional on visiting a clinic rather than on getting tested, it was possible for participants to collect the commitment payment without actually getting an HIV test.[10]

Before offering the commitment device, we elicited incentivized demand for the commitment device from each participant. All participants, including those not randomized to the hard commitment arm, were given an explanation of the purpose of the hard commitment device and how it worked. The device was framed as a way of helping people overcome self-control problems, and described as a voluntary form of collateral people could use to commit to an HIV test. We explained this with reference to *chikole* (Chichewa for "collateral"), which is commonly used for loans among our study population.[11]

Participants were then asked if they were interested in receiving such a commitment device. They were told that their choice would only be implemented with 50 percent probability, and that they would otherwise receive MK2,000 immediately regardless of their choice. Participants were informed that their decision would be final; those who expressed demand for the commitment device could not opt out of it afterwards. After recording demand for the commitment device, we revealed the result of the random assignment. Participants who were randomized into one of the commitment device arms and who had requested the commitment device received MK1,000, as well as a MK1,500 voucher to be redeemed at any clinic. Collecting this voucher was not conditioned on agreeing to an actual HIV test; men simply had to appear at the clinic. Everyone else received MK2,000 immediately, as well as the standard MK500 clinic voucher (see Figure 2).

After receiving the vouchers, participants who were randomized into one of the two

[10] We deducted the price of the commitment device from the respondent gift rather than accepting cash payments in order to reduce the effects of liquidity constraints. Demand for hard commitments would likely be lower if they had to be paid for out of pocket. It was made clear that the commitment device was voluntary.

[11] For the full script that we used to explain the commitment devices, see Appendix B.2.

appointments arms were given the opportunity to schedule an appointment at any of the 11 study clinics during regular clinic hours.[12] These clinics do not offer appointments for HIV tests themselves, so men in the non-appointment study arms could not make a testing appointment independently of our study. Participants could choose the clinic, as well as the date and time of the appointment, starting from the day after their baseline survey and ending two and a half months after the start of the study. We scheduled appointments over a relatively short time horizon, as in many contexts health appointments are scheduled a few weeks or months in advance. They could also decline the appointment. Those who chose to schedule an appointment received a phone call from testing staff two days before their scheduled appointment, reminding them of the time and place.

## 2.1 Data

Our study uses four sources of data: a baseline survey, administrative data on HIV testing collected in all clinics offering HIV testing within and around the city of Zomba, records of appointment reminder calls, and qualitative interviews with clinic staff.

After recruiting participants, our team of interviewers conducted a baseline survey, which included questions on demographic characteristics (gender, ethnicity, age, place of birth, educational achievement, marital status), socio-economic characteristics (employment, income, assets, expenditures), sexual behaviors (number of partners, risk, perceptions), past HIV tests, and intentions to get tested for HIV in the future. The baseline data was used for three purposes: (1) as a source of control variables to increase statistical power, (2) to explore mechanisms via heterogeneous treatment effect analyses, and (3) to prevent impostors from claiming the vouchers on behalf of our participants, and to remove those who did from the data. Using the baseline survey, we exclude from our study sample all respondents who listed a preferred HIV testing clinic outside of the study area (24 exclusions). This exclusion was specified in our pre-analysis plan, and was done to ensure that we observe all HIV tests that take place post-intervention.

To collect HIV testing data, we collaborated with the 11 clinics that provide HIV testing

[12] We did not mention the possibility of appointments to other participants, to avoid inducing John Henry effects (Saretsky 1972).

services within and around the city of Zomba. We hired qualified HIV Diagnostic Assistants (HDAs) and integrated them into each clinic.[13] This was done to ensure that our participants would not face wait times for their appointments, as well as to avoid disruptions to the clinic's usual operations due to our study. In any case, demand for HIV testing is low at all clinics in Zomba, and at the time of the study HIV test seekers did not typically have to wait. See Appendix A.3 for information on wait times from supplementary interviews we conducted with clinic staff.

To ensure participants' privacy, and to make them more comfortable with revealing personal information about their sexual behavior and HIV testing decisions, we did not collect any names as part of the study. Instead, we verified participants' identities at the time of voucher redemption by requiring them to show testing staff the text messages they received on their mobile phones during the baseline survey, which included unique voucher codes, and by asking them to answer specific security questions we asked during the baseline survey (see Appendix B.3 for details).

Our project's HDAs were in charge of collecting vouchers, recording data, and performing HIV tests for study participants. In line with local protocols, those who tested negative were encouraged to seek a second confirmatory test after 3 months, and the HDAs worked with the clinics to integrate newly-diagnosed individuals into ART initiation and care. The HDAs recorded data in handwritten notebooks. This included participants' voucher codes and phone numbers, in order to link them to our baseline data, as well as whether they agreed to be tested for HIV. HIV test results and ART initiation were recorded only by study arm, and are not linkable to individual participants. The information in the notebooks was digitized only at the end of the experiment.

Our sample consists of 1,232 men. Of those, 301 participants were assigned to the control group, 329 were offered only appointments, 295 were offered only hard commitment devices, and 307 were offered both appointments and hard commitment devices. Descriptive statistics and balance tests are provided in Appendix Table D3. The average age of the men in our sample is just under 32 years old. Roughly 98 percent of our sample can read and write,

[13] In some cases we placed new HDAs in the clinic, and in others we directly hired existing clinic staff to work for our team for the duration of the project.

and the average participant had finished just under 11 years of schooling. In our sample, 66 percent of participants are married and 43 percent report having a girlfriend (26 percent of married men also report having a girlfriend). The most common ethnic group is Lomwe (30 percent) followed by Yao (21 percent) and then Chewa and Ngoni (13 percent each). Over 90 percent of participants report having been tested for HIV at least once, and the average man in the sample reports having been tested for HIV five times. 87 percent of the sample report being willing to get an HIV test on the spot and 94 percent are willing to get tested in the future. However, 73 percent did not get tested in the three months preceding the survey.

The two cross-randomized treatments are well-balanced on baseline variables. Omnibus $F$-tests of the joint significance from regressions of the treatment indicators on all the baseline covariates yield $p$-values that are above usual significance thresholds. As an alternative to significance tests for balance, the table also shows pairwise normalized differences across study arms (Imbens and Rubin 2015, p. 310). These are small: out of 68 differences, 66 are below 0.1 and all are below 0.15, which indicates good balance. Appendix Table D3 follows our pre-analysis plan in analyzing balance for the two orthogonal treatments separately. Another way of testing balance is to compare the baseline covariates of the four individual study arms (control, appointments only, hard commitments only, and appointments plus hard commitments). We present this alternative balance test in Appendix Table D4. These results also indicate good balance: out of 204 normalized differences, 195 (96 percent) are below 0.15, and only 14 (7 percent) of the differences are statistically significant at the 10 percent level.

To better understand the study context, especially the use of health appointments and healthcare wait times in Malawi, we also conducted supplementary qualitative interviews with staff at clinics in the country (see Appendix A). We also conducted a follow-up survey in 2020, but high attrition prevents us from drawing meaningful conclusions from this data (see Appendix B.3).

# 3 Empirical Strategy

Our empirical strategy is based on a pre-analysis plan that we filed prior to receiving the outcome data (Derksen et al. 2019).[14] We discuss several minor deviations from the pre-analysis plan in Appendix E; these deviations increase the rigor of the analysis and do not alter our substantive conclusions.

Our primary outcome of interest is the decision to get an HIV test. This is captured by an indicator $T_i$ that is equal to one for men who redeemed their voucher at one of the study clinics and followed through with an HIV test, and zero otherwise.

## 3.1 Intention-to-Treat Estimates

To estimate the intention-to-treat effects of the two treatments on HIV testing, we use linear regressions of the following form:

$$T_i = \alpha + \beta_1 A_i + \beta_2 HC_i + \beta_3 Both_i + X_i' \gamma + \varepsilon_i \quad (1)$$

where $A_i$ indicates a participant was assigned to the appointments-only treatment, $HC_i$ indicates assignment to the hard commitment-only treatment, and $Both_i$ indicates being assigned to be offered both an appointment and a hard commitment.[15] $X_i$ is a list of baseline characteristics, which are included in the regression to increase precision.

We consider three different sets of control variables $X_i$. Our first specification includes no controls at all. In the second specification, we control for the variables and fixed effects that we specified in our pre-analysis plan. We include 10 variables that were significant predictors of past testing behavior as recorded in the baseline survey (these variables are described in Table D1). We also include fixed effects for the date of the baseline survey, the baseline survey interviewer, and the participant's preferred testing clinic as reported at baseline.[16]

[14] The analysis plan can be accessed at `https://www.socialscienceregistry.org/versions/57507/docs/version/document`.

[15] In Appendix Table D5 and Appendix F, we also show the "short" specification that omits this interaction term. The treatment effect estimates using the short specification are biased downward because this specification ignores the important negative interaction between the two treatments (Muralidharan, Romero, and Wüthrich 2019).

[16] Results are not statistically different when we also include bar fixed effects (not pre-specified in our

We focus on the results from this second specification, and show that our results are robust to varying the controls we use. The third specification includes all the pre-specified fixed effects, but selects other controls using the double LASSO method of Chernozhukov et al. (2017). Specifically, we use the `pdslasso` command in Stata (Ahrens, Hansen, and Schaffer 2019), and ask it to select variables from Table D1 and Table D2.[17]

The random assignment of participants to study arms ensures that $\hat{\beta}_1$ and $\hat{\beta}_2$ are consistent estimates of the intention-to-treat effects of the two treatments (when offered on their own), provided that the stable unit treatment value assumption holds—that is, that there are no spillovers between the men in the study. This is a plausible assumption in our case: the men were interviewed individually and privately, and there was no direct way for them to share the interventions with one another nor any incentive for them to do so.

## 3.2 Treatment-on-the-Treated Effects

We next estimate the treatment-on-the-treated effect of each intervention—the average effect of the intervention on the people who actually opted into it.[18] For the appointments intervention, we estimate equation (1) via 2SLS, using $A_i$ as an instrument for an appointment-only take-up indicator $A_i^{TOT}$:

$$\text{First stage:} \qquad A_i^{TOT} = \mu_0 + \mu_1 A_i + \mu_2 HC_i + \mu_3 Both_i + X_i'\lambda + \tau_i \tag{2}$$

$$\text{Second stage:} \qquad T_i = \alpha + \beta_1 \widehat{A_i^{TOT}} + \beta_2 HC_i + \beta_3 Both_i + X_i'\gamma + \varepsilon_i \tag{3}$$

pre-analysis plan).

[17] In addition to linear terms for each variable, we follow Knaus, Lechner, and Strittmatter (2020) in feeding the algorithm fourth-degree polynomials and logarithmic terms in each variable, along with all first-order interactions between variables. For variables with zeros or negative values, instead of using the logarithm, we use both the inverse hyperbolic sine transformation (Burbidge, Magee, and Robb 1988) and the Ravallion (2017) concave log-like transformation. Prior to variable selection, we partial out fixed effects for the date of the baseline survey, the baseline survey interviewer, and the participant's preferred testing clinic as reported at baseline. The exact variables chosen by the double LASSO vary by specification but, for our main treatment effects analysis (Table 2, Column 3), it selects three variables: 1) the Ravallion transformation of the index of willingness to get an HIV test, 2) the interaction between ever having been tested and demand for the hard commitment, and 3) the interaction between the number of sexual partners in the past 12 months and the demand for the hard commitment.

[18] There are no always-takers: control-group participants could not gain access to the commitment devices nor to an HIV testing appointment, as clinics do not offer these services under the *status quo*.

For the hard commitment, we know who the compliers are because we used incentivized choices to elicit demand for hard commitment devices for all the men in the study, and did not allow participants to deviate from their stated preferences post-randomization. We can therefore define an indicator $D_i$ for men who demand the hard commitment, and estimate the treatment-on-the-treated effect:

$$\begin{aligned} T_i = & D_i \times (\beta_0 + \beta_1 A_i + \beta_2 HC_i + \beta_3 Both_i) \\ & + (1 - D_i) \times (\beta_4 + \beta_5 A_i) + X_i' \gamma + \varepsilon_i \end{aligned} \tag{4}$$

The first part of Equation (4) specifies treatment effects on men who demand the hard commitment device. For these men, $\beta_0$ captures the average testing rate in the control group, $\beta_1$ is the intention-to-treat effect of the appointments alone, and $\beta_2$ is the treatment-on-the-treated effect of the hard commitment treatment alone. $\beta_3$ captures the combined effect of (i) being offered an appointment and (ii) receiving a hard commitment, for men who demand the hard commitment device. The second part of Equation (4) specifies the control-arm testing rate ($\beta_4$) and intention-to-treat effect of appointments ($\beta_5$) among men who do not demand the hard commitment device.[19]

## 3.3 Exploratory Analyses

We conduct three additional exploratory analyses. First, we estimate the impacts on HIV diagnosis and antiretroviral treatment initiation. These outcomes are anonymized and linked only to the study arm the participant was in, so we cannot include covariates in these regressions. Our study was not powered to detect effects on these outcomes, as HIV diagnosis is a rare outcome (less than 1 percent of the control group in our study tested positive for HIV).

Second, we analyze treatment effect heterogeneity by interacting treatment indicators

[19] Here, $A_i$ indicates men who were assigned to any arm that included appointments. For these men, the variables $HC_i$ and $Both_i$ are irrelevant as none of them received the hard commitment device (and they were not even told whether they would have been assigned to the hard commitment arm or not).

with de-meaned baseline covariates:[20]

$$
\begin{aligned}
T_i = \alpha &+ \beta_1 A_i + \beta_2 HC_i + \beta_3 A \times HC_i \\
&+ \beta_4 A_i \times W_i + \beta_5 HC_i \times W_i \\
&+ \beta_6 A_i \times HC_i \times W_i + \beta_7 W_i + X_i'\gamma + \varepsilon_i.
\end{aligned} \tag{5}
$$

Here, $W_i$ is the baseline covariate of interest. We also conduct a pooled analysis in which we include all interactions between the treatment indicators and the variables in Table D1, while controlling for main effects.

Finally, we investigate heterogeneous treatment effects by propensity to seek an HIV test, using the repeat split-sample (RSS) endogenous stratification procedure of Abadie, Chingos, and West (2018). [21]

## 3.4 Inference

We report heteroskedasticity-robust standard errors, with no adjustment for clustering, as our treatment was randomized at the individual level (Abadie et al. 2017). We also show that our main results are robust to randomization inference.[22] To address multiple testing concerns we compute sharpened $q$-values that control the false discovery rate (FDR) using the Benjamini, Krieger, and Yekutieli (2006) two-step procedure.[23] We conduct our FDR

[20] We de-mean all variables so that the main effects can be interpreted as average treatment effects (Imbens and Rubin 2015, p. 247).

[21] We conduct 100 random splits of our sample, and use half of the data to predict the outcome variable (the first stage) and the other half for treatment effect heterogeneity analysis (the second stage). We use the `estrat` Stata command (Ferwerda 2014). Point estimates and standard errors are the mean and standard deviation of the estimates. In the first stage, we use the two different sets of control variables from Equation 1 as predictors, not including the fixed effects. For the double LASSO controls, to avoid selecting different controls for each sample split, we use the predictors selected for estimating Equation 1. In the second stage, we estimate treatment effects for terciles of the predicted outcome calculated in the first stage, including the fixed effects. We run the procedure only on the control group plus the appointments-only arm, and then separately on the control group plus the hard commitment-only arm.

[22] The conventional sampling-based standard errors that we report will be conservative on average, i.e., larger than the correct standard errors that capture design-based uncertainty (Abadie et al. 2020).

[23] We use Anderson (2008)'s approach and report the lowest value for which the Benjamini, Krieger, and Yekutieli approach rejects the null hypothesis, so our $q$-values are comparable to $p$-values.

corrections across all the $p$-values for treatment effects reported in the paper and appendix.[24] The tables show significance stars based on $p$-values, and sharpened $q$-values in brackets. All discussions of statistical significance in the text are based on the $q$-values.

# 4 Results

Demand for both types of commitment devices is quite high in our setting. Among the men randomly offered an appointment, 65 percent signed up for one. We elicited demand for hard commitment for the entire sample: 51 percent of all the men in the study wanted a hard commitment device, including 49 percent of men assigned to one of the hard commitment study arms. Since we measure demand using incentivized choices, and participants could not back out post-randomization, this means that the take-up of the commitment devices was 49 percent as well. The high take-up of the hard commitment device suggests that many of the men in the study were aware of their self-control issues when it comes to HIV testing, and believed the intervention could help them overcome these issues. Yet these rates of take-up may understate the true extent of self-control problems, if some men are too naive to recognize their need for a commitment device. Moreover, for many study participants this sophistication about self-control problems was only partial: just 41 percent of men who signed up for a hard commitment device actually followed through and visited a clinic to redeem their voucher. As a result, 59 percent of men who enrolled in the hard commitment—and 64 percent in the hard-commitment-only arm—simply lost their investment.[25] Not only did these participants fail to seek an HIV test, they did not even visit the clinic to redeem their vouchers. This parallels the results in Bai et al. (2021) and John (2020), who also find substantial failures to follow through on commitment devices—and thus find that people are made worse off by the offer of a hard commitment device. In contrast, none of the men in

[24] This includes Tables 2 through 4 and Appendix Tables D5 through D11, as well as the tests associated with Figures 5 and 6. It does not include any of the hypothesis tests in the balance tables (Appendix Table D3 and Appendix Table D4) or supplementary tests that are reported only in the text and not in tables or figures.

[25] Regression analyses (controlling for the pre-specified list of baseline covariates) suggest that there is a small negative correlation between the time of the baseline survey and the take up of hard commitment (p-value = 0.051), but no significant correlation with the demand for appointments. We find no evidence of significant heterogeneous treatment effects with respect to the time of the baseline survey (this part of the analysis was not pre-specified).

the appointments-only arm lost money, as no money was at stake.

Both treatments significantly increase HIV testing. Table 2 shows the intention-to-treat effects of the two interventions on HIV testing. The bar chart in Figure 3 illustrates the results, showing testing rates by study arm based on our preferred specification (Column 2 of Table 2). Appointments, offered on their own, cause a 16 percentage-point increase in HIV testing—a 141 percent increase relative to the control group mean. This effect is statistically significant at the 0.01 level. Appointments are much more effective than hard commitment devices at increasing HIV testing. Hard commitment devices, offered on their own, increase HIV testing by eight percentage points—about half of the effect of the appointments treatment. This effect is also statistically significant at the 0.01 level. We can reject the equality of the appointment and hard commitment effects at the 0.05 level.

Appointments are strong substitutes for hard commitment devices, while hard commitment devices are imperfect substitutes for appointments. Figure 3 shows that the combination of both types of commitments has roughly the same effect on HIV testing as the appointments-only treatment ($q$-value = 0.35). In contrast, the marginal effect of the appointments treatment on top of the hard commitment treatment is positive—an increase of nine percentage points—and statistically significant at the 0.05 level. The coefficient on the interaction term between the two treatments is large and negative, showing that appointments and hard commitments are substitutes. Its $q$-value is slightly above the conventional significance threshold of 0.1 in our preferred specification (Column 2 of Table 2). Conversely, we cannot reject that the effect of an appointments alone is equal to the effect of an appointment combined with a hard commitment device ($q$-value = 0.295). In other words, getting the commitment device on top of an appointment generates next to zero additional effect on testing rates, while getting an appointment on top of the hard commitment device more than doubles the effect on testing rates.

Indeed, for the subset of men who benefit from the hard commitment, a soft commitment, in the form of an appointment, is just as effective. To see this, consider a scenario in which a subset of men get tested if offered only a hard commitment device, but do not get tested if offered only an appointment. Then, we would expect these men to get tested when offered a hard commitment on top of an appointment. The fact that we do not observe an increase in the treatment effect suggests that either such a subset does not exist, or that there is a negative interaction effect—that is, adding a second intervention on top of the first actually decreases demand for HIV testing for some men. While we cannot rule out this sort of negative interaction effect *a priori*, this possibility seems fairly unlikely.

We estimate the treatment-on-the-treated effects of appointments in Table 3. The average treatment effect of an appointment on men who actually signed up for one is a 24 percentage-point increase in HIV testing. The treatment-on-the-treated effect of the hard commitment treatment is estimated in Table 4.[26] The results are also shown in Figure 4. Recall from equation (4) that this specification does not include a constant, so the first and the fifth rows of Table 4 represent absolute testing levels in the control arm. The effect of a hard commitment device on men who demand one is between 11 and 13 percentage points.

Table 4 also shows that appointments work best for men who have self-control problems and therefore demand commitment devices. The table shows that the appointments treatment increases HIV testing by nearly 20 percentage points among the men who wanted a hard commitment but did *not* receive one. This is more than twice the effect of the appointments treatment on men who did not want a hard commitment—even though both groups received the exact same intervention (appointments alone, without hard commitment). We can reject the equality of the two effects at the 0.1 level. We conclude that appointments are more effective for men who have self-control issues that they are (at least partially) aware of.

Among men who wanted a hard commitment, appointments are again superior substitutes. For this subset of men, appointments are nearly twice as effective as hard commitment devices. The effect of the combined treatment (appointments plus hard commitment devices) on men who wanted a hard commitment device is around 23 percentage points. This is only slightly larger than the effect of the appointments-only treatment for men who wanted a hard commitment device, and we cannot reject the null of equal treatment effects ($q$-value $\approx 0.3$). Indeed, men who wanted a hard commitment device are almost as well-served by just getting an appointment as they are if a hard commitment device is layered on top of an

[26] Table 4 focuses on the long specification. Results for the equivalent short specification are shown in Appendix Table D6.

appointment—and much better off than if they received only a hard commitment device.

Another piece of evidence that appointments substitute for hard commitment devices is that the demand for the two is positively correlated ($\rho = 0.21$). Of the 308 men in the appointments arms who wanted a hard commitment device, 76 percent signed up for an appointment; for the 328 men who did not want a hard commitment device, just 56 percent signed up for an appointment. The combined treatment arm also allows us to examine how offers of the hard commitment device affected uptake of appointments. If hard commitments substitute perfectly for appointments then we would expect men who receive hard commitments to have lower demand for appointments. We see no such pattern: out of the men who wanted a hard commitment device, 78 percent of those who received one signed up for an appointment, as compared with 73 percent of those who did not ($p = 0.31$). This is consistent with the fact that hard commitments do not act as perfect substitutes for appointments. Note that since participants made decisions about the hard commitment devices prior to finding out about the appointments, being offered an appointment could not affect uptake of the hard commitment.

## 4.1 Cost Effectiveness

Appointments are highly cost effective. At a cost of \$0.43 per person they increase testing by 16 percentage points (Table 5). The cost per additional person tested is \$2.69, as compared with \$3.01 for the hard commitment devices. Our preferred cost-effectiveness calculations (Columns 1 and 5) show only the incremental cost of each intervention. This means we exclude the cost of finding and contacting men for the study, since that was done in the control group as well. These costs do not include any of the costs of testing itself, since testing is the outcome of interest rather than an input. For details of the cost-effectiveness calculations, see Appendix C. We also assume that the increase in testing does not run up against any capacity constraints at the clinics. This is based on our qualitative work (see Appendix A), which indicates that there is substantial excess capacity in terms of HIV testing and treatment. In other contexts with tighter capacity constraints, this would be an additional cost we would need to consider. This also puts our results on even footing with other interventions that attempt to increase HIV testing, which would run up against similar constraints.

A simple basis for comparison for these results is directly paying people to get tested. Thornton (2008) did something quite similar to this, testing her entire sample and paying people to pick up their results at a clinic. (Our participants got their test results immediately, whereas when the Thornton study was conducted blood samples had to be sent out to a lab for analysis, creating a delay between testing and learning the results.) Cash incentives increase the rate of picking up one's results by 9.1 percent per dollar of incentive, so each additional person tested cost $10.99.[27] Our increase is also measured off of a lower base rate of HIV testing (11 percent in our study vs. 34 percent in Thornton's study). In percentage terms, appointments increase testing by 141 percent; to achieve the same relative increase by paying people to learn their HIV status would cost $5.27 per person, as compared with $0.43 for appointments.

The other columns of Table 5 present alternative ways of computing the cost effectiveness of the two interventions. Columns 2 and 3 calculate the cost effectiveness of appointments with two key changes. Column 2 includes the cost of the testing voucher (even though this is not an incremental cost, as it was provided to the control group as well). Column 3 provides an upper bound on the cost of contacting people for their appointment reminders, by adding the cost of an additional appointment reminder call for each participant. This allows for one call attempt that the respondent does not pick up, and one successful one; our preferred estimates from Column 1 assume a single call per person. These changes do not substantially alter our results: adding the cost of the voucher raises the cost per additional person tested to $3.81, and raising the number of call attempts to two changes the cost only slightly, to $2.93. Even under these alternative assumptions, appointments remain more cost-effective than paying people to learn their HIV test results. Column 4 presents a calculation that also includes the cost of recruiting the men into the study and conducting the baseline survey, which is also incurred for the control group. The total cost comes to $4.50 per participant. Since these are not incremental costs of the intervention, they are not normally included in cost-effectiveness calculations, but they may be relevant for assessing the cost of actually

[27] Another point of comparison is Macis et al. (2021), who find that cash incentives greatly increase HIV testing—but have much smaller effects on actually learning one's results. Our results are more comparable to Thornton's because the test results were revealed immediately.

implementing appointments as a policy. With those costs included, the cost per additional person tested rises to \$29.04. While this is higher than the cost from Thornton (2008), her results also rely on people already being surveyed, so they would have to be adjusted as well.

We also try several variations in how we calculate the costs of the hard commitment devices. In Column 6, we add the cost of the MK1,000 of the survey gift that was used as the hard commitment. Our preferred estimates from Column 4 exclude this amount because this gift was provided to all participants. If this amount is included in the cost of the hard commitment devices then they cost \$19.61 per additional person tested. This larger number may better reflect the true cost of implementing a policy where participants are offered gifts to use as hard commitment devices (as opposed to using their money). Column 7 instead adds the cost of the testing voucher, which raises the cost per additional person tested to \$4.56. Column 8 adds both the commitment device cost and the voucher cost, raising the cost per person tested to \$21.16. It is worth noting that we include the commitment device cost for all participants (since that is the amount they need to buy the hard commitment), but the testing voucher costs only for those who get tested. Thus including the cost of the commitment device has a much larger effect on costs than including the cost of the vouchers. Finally, Column 9 adds in the cost of recruiting men into the study, similar to Column 4. Including these costs (which also apply to the control group) raises the cost per additional person tested to \$67.79. Again, these variations do not affect our qualitative conclusion that appointments are more cost-effective than hard commitments.

## 4.2 Exploratory Analyses

One key exploratory analysis is to examine treatment effects on positive HIV tests and on ART initiation. Because both interventions have large effects on HIV testing, they may also increase the detection of actual HIV cases and guide HIV-positive men into treatment for the disease. Appointments lead to diagnoses at the same rate as control-group tests (6 percent of tests are positive), and at a higher rate than the hard commitment device alone (2 percent positive). This suggests that appointments are not disproportionately selecting for low risk men, and that the hard commitment device may be.[28] We examine treatment effects on positive HIV tests and ART initiation in Table 6. The hard commitment devices had no significant effect on either outcome, but the appointments increase both the rate of HIV detection and the rate of ART initiation by 1 percentage point—more than doubling the control-group rates. Using the short specification, these effects are statistically significant at the 0.1 level and the effect on positive tests is essentially robust to FDR correction (the $q$-value is 0.104 for ART initiation). However, we are underpowered to study these very rare outcomes, and these results are not significant in the long specification, which has lower statistical power (Muralidharan, Romero, and Wüthrich 2019). We therefore interpret this as suggestive—but not conclusive—evidence that appointments not only increase HIV testing rates, but also help to locate HIV-positive men and bring them into the treatment pipeline.

We also explore variation in the treatment effects of the two interventions in two different ways. First, we present conventional heterogeneous treatment effect analyses in Appendix Table D7. We see no evidence of statistically-significant heterogeneity in treatment effects by men's perceived probability of having HIV, their self-reported tendency to postpone HIV testing, the extent to which they say they live for today, or their perceived benefits from taking ART.[29]

Second, in Appendix Table D8, we use the Abadie, Chingos, and West (2018) method to study how the treatment effects vary by propensity to get an HIV test.[30] Both appointments and hard commitment devices work better for men who are more likely to get tested for HIV in the absence of either intervention. In many settings, interventions are viewed as more impactful if they target demographics for whom outcomes are otherwise low. But HIV testing is a somewhat unusual outcome in this regard; those who are unlikely to test without the intervention may be at lower risk infection, and have little to gain from an HIV test.

[28] At scale, a diagnosis rate of 6 percent would be policy relevant. HIV prevalence among urban Malawian men is 11 percent (DHS 2016); if we assume that approximately half are already diagnosed, then our appointments treatment is selecting on men with average risk. However, our study is not well-powered for this analysis and we therefore cannot draw strong conclusions.

[29] Theoretically, these variables are expected to affect risky behaviors, including related to HIV testing (Sterck 2013, 2014; Kerwin 2020).

[30] The `estrat` command returns individual standard errors for each tercile, but not for the differences between them. We thus compute the standard errors of the differences as $\sqrt{Var(X) + Var(Y) - 2Cov(X,Y)}$. We treat the covariance term as zero, as it is also not reported by the command, and thus our $p$-values are likely to be conservative.

Data from Zomba suggests this is the case in our setting: HIV test-positivity rates are much higher than the local prevalence of the virus (Derksen, Muula, and van Oosterhout 2022), so people who are likely to get tested for HIV under the status quo are at higher risk of HIV infection. This implies that the intervention may be successfully targeting high-risk men.

## 4.3 Robustness Checks

Our primary robustness check is presented in each of the tables in the paper: all our results are basically unchanged by our choice of controls, or by using no controls at all. Columns 1 and 3 of Table 2 through Table 4 show results that are nearly identical to our preferred specification in Column 2, with occasional minor changes in the coefficients and $q$-values. Our main results from Table 2 are also qualitatively similar when we use the "short" specification that omits the interaction term (Appendix Table D5), but smaller in magnitude because omitting the interaction term biases the coefficients downward (Muralidharan, Romero, and Wüthrich 2019).

We also show our findings are qualitatively robust to switching all our analyses to the exact approach we pre-specified in our analysis plan. Appendix F shows all the results of the original specifications from our analysis plan, which are primarily focused on the short regression that omits the interaction between the two treatments. All our findings are qualitatively identical, and our inferences—based on $q$-values that adjust for multiple testing across all tests in Appendix F—are substantively unchanged.

Our inferences are robust to the use of randomization inference instead of $p$-values based on Eicker-Huber-White standard errors. Appendix Table D9 shows randomization inference $p$-values for our main results from Columns 1 to 3 of Table 2. Some of the $p$-values are somewhat larger, and others are slightly smaller, but we continue to reject a zero treatment effect for each intervention separately, and the equality of the two interventions, at the 0.05 level.

Our results are robust to different ways of defining our outcome variable. One alternative definition is to use any voucher redemption, without conditioning on an HIV test. A meaningful proportion of participants visited a clinic to redeem their voucher without getting an HIV test: about 5 percent of the control group, 6 percent of men in the appointments-only arm, 9 percent of men in the hard commitment-only arm, and 9 percent of men in the combined arm. This suggests another margin of partial sophistication by the men who enrolled in the hard commitment device. While this subset of men did not lose their investments, they wasted their time and effort in coming to the clinic but not actually following through with a test, likely due to fear of learning their HIV status. Columns 1 to 3 of Appendix Table D10 show treatment effects on any voucher redemption. The estimated effect of the appointments treatment is nearly unchanged, while the effect of the hard commitment treatment is somewhat larger; we can no longer reject the equality of the two effects in this table. Columns 4 to 6 show that appointments had no effect on voucher redemption alone, with no HIV test, while hard commitment devices increased it by four percentage points.

Finally, there are potential sources of measurement error in the outcome data. First, it is possible that some participants got tested without using a voucher. This concern is mitigated by the fact that our staff were present at all local clinics during opening hours, and the fact that the value of the voucher was fairly high, at 25% of median daily earnings in our sample. Our records suggest that this type of error, as well as other potential sources of measurement error are uncommon (see Appendix B.3). This type of unobserved attrition is likely to be differential: more testing due to the appointments will lead to a higher number of attritors in that arm. In contrast, the added incentive to redeem the voucher in the hard commitment arm would lead to systematically lower attrition rates in that arm. However, the study arm with both appointments and hard commitments has a high testing rate relative to the hard-commitments-only arm; since participants in both of those arms have a strong incentive to redeem their vouchers, this mitigates concerns about attrition.

# 5 Potential Mechanisms

Why are appointments effective at increasing HIV testing? Appointments can be viewed as a bundle of several behavioral interventions that help overcome self-control problems. Many of these interventions, or nudges, have previously been studied in isolation, including in the context of HIV testing.[31] By combining them, we obtain a large effect that likely depends

[31] See for example, Tenthani et al. (2014), Nyondo et al. (2015), Rana et al. (2015), Mugo et al. (2016), Mayer and Fontelo (2017), Taylor et al. (2019), Salvadori et al. (2020), Friedman and Wilson (2021), and

on several overlapping mechanisms. This section discusses the potential mechanisms at play, and provides evidence that two particular mechanisms play an important role in their success. First, appointments appear to help address self-control problems, as they work best for people who demand commitment. We discuss potential reasons why participants follow through on the soft commitment, including both personal reasons and social pressure. We also discuss the possibility that social pressure applies not only to follow-through, but also to demand for appointments at the time of sign-up. Next, appointments help overcome limited memory problems, making people less likely to forget to follow through with an HIV test. At the end of this section, we discuss a number of other potential mechanisms. Most notably, we explain why wait times are unlikely to be important in our setting.

## 5.1 Appointments Address Self-Control Problems

Committing to a health appointment could help people overcome self-control problems. Visiting a clinic for an HIV test is immediately costly in terms of time, effort and anxiety, while the benefits of treatment accrue over the longer term. This combination of short term costs with long term benefits can lead to self-control problems (O'Donoghue and Rabin 1999). Responses from the baseline survey indicate that self-control problems may be an important barrier to HIV testing. When asked about reasons for avoiding an HIV test, men most commonly answer that testing is not needed, either due to low risk, a recent test, or a lack of symptoms (34 percent of participants). The next most common answer is that the participant says he is too busy, too lazy, or too forgetful to seek a test (33 percent), with 8 percent of participants mentioning laziness specifically.

Consistent with this, in Section 4, we saw that appointments are more effective for men who demand hard commitment devices. We interpret demand for a hard commitment device as an indication of self-control problems. Moreover, appointments strongly substitute for hard commitment devices, both overall and for men who demand commitment in particular.

Somewhat surprisingly, treatment effects do not appear to vary by some other measures such as participants' self-reported tendency to delay HIV testing or whether they live for today (Appendix Table D7). One possible explanation for this difference in results is that the

Macis et al. (2021).

self-reported variables are measured with greater error; our commitment demand variable was measured using incentivized choices, which should reduce measurement error.[32] Moreover, we do not expect every person with a self-control problem to demand the commitment device. Many of those who know that they delay testing may also know that the commitment will not be strong enough to make them follow through. For this type of sophisticated but severe procrastinator, a commitment device will not suffice.

Another possibility is that our measure of the demand for hard commitment could capture something other than self-control problems. One potential issue with our metric is experimenter demand effects: since the commitment device was offered as part of an experiment, participants might have signed up for it in order to please the interviewer (de Quidt, Haushofer, and Roth 2018). This is also a challenge for the other hard commitment devices summarized in Table 1, and indeed for any measure of demand that uses revealed preferences within a survey or experiment. This concern is mitigated by the fact that these were costly, real choices rather than just cheap talk—signing up for the commitment device meant you were actually committed to visiting the clinic. A second potential issue is default effects, or the tendency to go with whatever the default option is on a survey (Jachimowicz et al. 2019). This is unlikely in our case because the commitment device was not a default; participants only received it if they explicitly opted in.

A third potential issue with our hard commitment demand measure is that participants might have misunderstood the commitment device as an additional gift, and thus missed that they could choose to receive the MK1,000 immediately. However, the fact that they could get the money right away was completely explicit in the script:

> If you say yes, we will have a lottery, and 50/50 chance you receive the 1000K gift or the 1000K goes to the 'chikole'. If you say no, you will just get the 1000K gift now. Are you sure about your answer? (Appendix B.2)

Moreover, the concept of collateral ("chikole") is widely understood in Malawi; 92 percent of our participants stated that they were familiar with it. We also have an empirical measure of people misunderstanding the setup: we recorded both people's initial responses and also

[32] For this reason, we pre-specified demand for hard commitment device as the relevant source of heterogeneity for our main analysis, with the self-reported variables pre-specified as exploratory.

their revised responses after the followup question. 62.8 percent of people said yes to the initial question, and 15.2 percent of respondents changed their mind about their answers. 13.6 percent of people switched their answer from yes to no, and 1.5 percent changed their mind from no to yes. This indicates that the bulk of respondents understood the question initially. Reassuringly, revisions to choices were not correlated with education levels, which are a proxy for how well the respondents understood the questions. These revisions were, however, exceedingly rare among respondents who were already familiar with the concept of collateral. The main results are similar if we restrict to this subsample.

Participants might also have wanted the commitment device because they had a preference for receiving their money later. Savings constraints are common in developing countries, and have specifically been documented in Malawi; Brune, Chyn, and Kerwin (2021) find high demand for a zero-interest deferred-payment savings product. People might have trouble saving because of temptation spending, security concerns (particularly because the study recruitment took place in a bar), or kin taxes. Relatedly, it is possible that some of these men did not want an HIV test, but did want to collect the MK500 from their voucher for appearing at the clinic, and used the hard commitment to encourage themselves to do so.

All of these concerns about alternative potential drivers of demand for the hard commitment devices are mitigated by the empirical evidence we have on the take-up and impacts of the devices. First, demand for hard commitments is positively correlated with the respondent saying that they live for today, their willingness to get tested, and their tendency to postpone HIV testing (Appendix Table D12). It is uncorrelated with our measure of time preferences, but self-control problems are characterized by differences in time preferences between different time periods, not the levels of time preferences. Second, hard commitments do induce an increase in actual HIV testing, not just appearing at the clinic (Appendix Table D10). This means they changed behavior in exactly the way we would expect them to for people with self-control problems.

### 5.2 Social Pressure

Signing up for an appointment is a soft commitment to seek care—while they create a social expectation to follow through, there is no financial penalty for failing to do so. Why do so many people honor the commitment? One possibility is that once an appointment is made, there is a social cost associated with failing to appear. While health appointments are rare in some parts of the world, other types of appointments, i.e. plans to meet at a particular time and place, are common, and there are typically social costs of failing to appear. Missing a health appointment means wasting the time of the health provider who was expecting you. Health appointments typically involve personal interaction as they are often scheduled in-person or over the telephone. This might make the social commitment more powerful. Indeed, automated appointment nudges sent by text message appear to be less effective (Chang et al. 2021).

Many participants do honor the social commitment to attend the clinic on the day of their appointment. We observe a large spike in HIV testing on the exact appointment date (Figure 7). Over 50 percent of men in the two appointments arms who got an HIV test did so on the date of their appointment. On the other hand, almost half of the men with a scheduled appointment who got tested did so on *other* days, so the value of coming in on the day of the appointment is not sufficiently large to cause everyone to do so. The overall rate of missed appointments is large, at 80.8%: 63.7% of men who had an appointment did not get tested at all, while 17.1% got tested on a different date. This suggests that other mechanisms do also play a role.

Social commitments work because individuals care about how they are perceived and are willing to modify their behavior to signal socially-desirable traits or to comply with social standards, such as honoring a previous commitment. The degree to which an individual's actions will be affected by social pressure can depend on how socially desirable an action is and on how much the individual cares about how they are perceived (Bursztyn and Jensen 2017). This has been found to be true for behaviors ranging from vaccination (e.g., Rao, Möbius, and Rosenblat 2007, Karing 2018, Brewer et al. 2017) and HIV testing (e.g., Godlonton and Thornton 2012) to productivity at work (Mas and Moretti 2009) and savings (e.g., Ashraf et al. 2003, Breza and Chandrasekhar 2019, Kast, Meier, and Pomeranz 2018, Kast, Meier, and Pomeranz 2012). The fact that social costs are more effective than financial costs is consistent with Gneezy and Rustichini (2000), who find that social pressure is more effective than incentives at inducing people to pick up their children from daycare on time. Socially-enforced commitments have been used as an explicit intervention by Karlan and Zinman (2012) to address indebtedness. They also play a role in the traditional design of microcredit products (de Aghion and Morduch 2005).

A different but related possibility is that the *demand* for appointments is the result of social pressure, as opposed to a genuine desire for commitment. Giving participants an appointment for HIV testing might reveal that the researchers think HIV testing is important, and lead to experimenter demand effects (de Quidt, Haushofer, and Roth 2018). In particular, participants are asked whether they want an appointment for an HIV test and might feel bad saying no, especially because they have just answered a number of questions about HIV testing. Then the social pressure of disappointing clinic staff could cause them to both schedule an appointment and subsequently follow through with an HIV test. This is related to the use of appointments as a social commitment device, but somewhat separate: social pressure might induce more testing even among men who do not actually want an HIV test. On the other hand, Chang et al. (2021) found appointments to be ineffective for people who do not want to seek care, suggesting that social pressure to accept a health appointment may play a limited role.

Similarly, our evidence suggests that while social pressure at the time of sign-up may well play a role, it cannot account for the full effect of appointments on HIV testing. All of our participants, including those in the control arm, took part in a baseline survey with many questions pertaining to HIV testing, and received a voucher to visit an HIV testing site. This was likely a larger signal from the research team than the marginal effect of additionally being offered an HIV testing appointment. Despite this, the HIV testing rate was more than twice as high in the appointments-only arm as in the control arm. Moreover, as discussed in Section 5.1, if experimenter demand effects play a role in our setting, this should also affect demand for the commitment device. At the moment of sign-up there should have been similar social pressure to take an appointment as a hard commitment device: in each case the intervention was offered by the interviewer as part of the baseline survey. But we see higher take-up of appointments (65 percent) than hard commitments (49 percent). We also see significantly higher rates of HIV testing among those in the appointments-only arm (15.9 percent) compared to the hard-commitment-only arm (8.2 percent).

The fact that those in the appointments-only arm are more likely to actually follow through with an HIV test suggests that social pressure plays an important role in encouraging people to honor their previous commitment. Moreover, appointments are more than twice as effective for men who wanted a hard commitment device. Even among men who were not randomly assigned to actually *receive* a hard commitment, the impact of an appointment is 10.3 percentage points higher for those who wanted one. This pattern could indicate that appointments work better for men who are more susceptible to social pressure, who are also more likely to say yes to a hard commitment device. But, given the asymmetry in follow-through between appointments and commitment devices, it is unlikely that social pressure to simply sign up for an appointment fully explains the effect on HIV testing.

Even if the effectiveness of appointments were driven entirely by social pressure to accept the appointment, and unrelated to self-control or limited memory problems, our results would remain important from a policy perspective. Appointments are a cheap and scalable intervention that can address the important public health challenge of promoting HIV testing among high-risk groups. These impacts also have the potential to translate to other health problems. HIV testing is a particularly fraught and stressful decision; social pressure may work even better for increasing the uptake of other health services.

## 5.3 Appointment Reminders and Limited Memory

Our data suggests that another mechanism drives the success of appointments as well: appointment reminders help address problems of limited memory. People often forget to do things they want to do, and moreover they are overconfident about remembering their plans (Ericson 2011). Models of limited memory predict large effects of reminders, particularly for people who are present-biased (Ericson 2017).[33] An important literature shows that

[33] Haushofer (2015) shows that a model of limited memory can also generate several departures from standard neoclassical utility maximization, including status quo bias and loss aversion.

reminders are very effective at nudging people into health care (e.g., Vervloet et al. 2012, Gurol-Urganci et al. 2013, Altmann and Traxler 2014, Jacobson Vann et al. 2018, Banerjee et al. 2021). Reminders can also help people make better financial decisions (Karlan et al. 2016), raise charitable giving rates (Rogers and Milkman 2016), and increase the uptake of tutoring (Pugatch and Wilson 2018). More broadly, forgetting also leads to important failures to optimize in life insurance (Gottlieb and Smetters 2021) and retirement savings (Goodman, Mukherjee, and Ramnath 2021).

To examine this mechanism in our data, we plot histograms of the delay between recruitment and getting an HIV test (Figure 5). Relative to both the control group (Panel A) and the hard commitment arm (Panel B), men in the appointments arm got tested for HIV substantially later than those in the other two arms, and these differences are significant at the 0.01 level.[34] The hollow red bars in the histograms show that in the non-appointments arms, testing trails off to nearly zero within about 15 days of enrollment into the study. This pattern is consistent with a limited memory mechanism, and suggests that reminders do play a role.

One alternative explanation for the patterns in Figure 5 is simply that men signed up for appointments on later dates, and followed through with them. To rule out this possibility, Figure 6 breaks down the appointments vs. control comparison by whether the men in the appointments arm came in on the day of their appointment or on a different day.[35] The solid gray bars are substantially shifted to the right relative to the hollow red bars in both panels, indicating that men in the appointments arm got tested later even when they did not come in on the day of their appointment. $t$-tests show that these differences are statistically significant at the 0.01 level, both for men who got tested on their appointment date, and

[34] We find no statistically significant difference between the appointment arm and the combined treatment arm.

[35] Men were allowed to reschedule their appointments when they received the reminder phone call; in these cases, we use the date of the rescheduled appointment. In order to do this, we drop 62 men from the appointments arm for whom appointment rescheduling information is not available, 20 of whom took an HIV test. No men are dropped from the control group because appointment rescheduling does not apply to them. Overall testing rates are similar for men who kept their original appointment and those who rescheduled their appointments. For men who kept their original appointments, 19 percent got tested on the day of their appointment, 19 percent got tested on another day, and 62 percent did not get tested at all. For those who rescheduled, 22 percent came in on the new (rescheduled) appointment date, 18 percent came in on another date, and 61 percent did not get tested at all.

men who did not.

The simplest explanation for the fact that men in the appointments arm test later in the study period is that the appointment reminders helped them overcome limited memory problems. The pattern in Panel B of Figure 6 implies that a number of men in the appointments arm forgot about their plan to get tested for HIV, and then were reminded about it by the phone call about their appointment. They then followed through, but *not* on their appointment date—driven to get an HIV test by the reminder, and not the social cost of a missed appointment.[36] Beyond the literature on forgetting *per se*, this pattern is also consistent with models of limited attention (Gabaix 2019).

However, appointment reminders, on their own, do not appear to fully explain our results. First, the baseline survey and HIV testing voucher already serve as a fairly significant initial reminder to get tested for HIV. But appointments more than double the HIV testing rate relative to the control group, who also received this initial "reminder". It is possible that a simple phone call reminder has a much larger marginal effect on testing than the in-person recruitment, but this seems unlikely. Indeed, Salvadori et al. (2020), find that reminders explain only one third of the increase in HIV testing caused by an appointment. Moreover, we observe a large spike in testing on the exact appointment date (Figure 7). Conditional on testing after the reminder, 71 percent of men in the appointments-only arm visit the clinic exactly two days later, on the appointment date. This spike in testing is inconsistent with a pure reminder effect. If the appointments increased testing only by reminding people to go to the clinic at some point, then we would expect a smoother increase in visits across a wide range of days. The spike in testing on the date of the appointment implies that that specific date is special, consistent with the social commitment and salience mechanisms.

[36] An alternative explanation of this pattern is that men who miss their appointments feel rising shame from their missed appointments (Butera et al. 2021), inducing them to continue to come in for tests. We cannot completely rule out this potential mechanism, but it would likely lead to increased testing over time in the appointments arm, which we do not observe. It also cannot explain why the men miss their appointments in the first place.

## 5.4 Other Mechanisms

Several other mechanisms might play a role in driving the effects of appointments. First, participants might honor the commitment for personal, rather than social reasons. Appointments may help people to make a mental plan to get an HIV test on the appointment date. Prompting people to make a plan is effective at improving follow-through for many behaviors (Rogers et al. 2015), particularly when the plan includes specific details such as time and place. This type of mechanism might well play a role in our setting, although the magnitude of the effect we observe is substantially larger than that typically seen for interventions that simply encourage participants to make a plan. Indeed, simple planning prompts do not appear to impact HIV testing (Macis et al. 2021).

Alternatively, the appointment may be viewed as a social *invitation* or *encouragement* to visit the clinic, or might make a particular date salient. Clinics in Malawi are predominantly female spaces: they primarily target women through a combination of policy, practice and gender norms (Dovel et al. 2020*b*). Offering appointments might reassure men that they will be welcomed at the clinic and provided quality care. Nyondo et al. (2015) find that men are more likely to attend antenatal care visits with their partners if they receive a formal invitation, and appointments may play a similar role. Finally, participants who schedule an appointment for reasons other than commitment may nevertheless follow through because the date has been made salient, in a sense similar to Bordalo, Gennaioli, and Shleifer (2013).

Appointments may reduce the expected time cost of getting tested for HIV on the appointment date. This is one reason that appointments are common in developed-country healthcare systems: they solve coordination problems, leading to less wasted time. We can shed light on this possible mechanism using the supplementary interview data we collected on typical wait times and absenteeism in 2019 for clinics in Malawi (see Appendix A.3). We find that at the time of the study, there was typically no wait for an HIV test at any of the clinics in Zomba, though at busy times the wait could be as long as 60 minutes. At urban clinics, passersby can often see whether there is a line of people waiting, and decide to wait or return later. Moreover, unlike evidence from other contexts (e.g. Banerjee and Duflo 2006), provider absence is rare: fewer than two percent of HDAs were absent on a typical day. These patterns are similar to those for Malawi as a whole. The fact that there is no shortage of HIV testing staff may be explained by the fact that this occupation requires fewer qualifications than other health occupations; HDAs only need a high school diploma and two months of training. Given these facts about the clinics under the *status quo*, it is highly unlikely that expected wait times and provider availability are major drivers of the effects of appointments. Indeed, in the baseline survey fewer than 0.1 percent of respondents named wait time as a reason they had avoided HIV testing. This does not rule out the possibility that reductions in perceived wait time play some role. Indeed, in a setting where wait times are long, appointments may be even more effective, and the logistical challenges might be quite different (Hakimov et al. 2021).

Finally, because healthcare appointments are rare in Malawi, novelty may play a role in their effectiveness. On one hand, where appointments are novel, offering one may act as a salient social signal about the importance of a particular healthcare service. On the other hand, novelty might make appointments less effective, due to misunderstanding or mistrust. The concept of an appointment more broadly is not novel in the broader Malawian context. Outside of the healthcare sector, appointments are commonly scheduled in Malawi for social and business purposes, and social costs for failing to appear are typically incurred. In fact, a hard commitment device for healthcare is arguably a more novel intervention than an appointment, and also signals the importance of HIV testing, yet appears to generate significantly lower follow-through.

Nevertheless, our estimated effect size represents a *marginal* effect in a specific context in which appointments are rare. If appointments were used for many different health services, we may expect to find different effects. Yet for many Malawians, HIV testing is one of the most important interactions with the healthcare system. HIV diagnosis is the first step towards effective treatment for AIDS, a prevalent and fatal disease. Because treatment also prevents transmission, promoting HIV testing is a priority for many countries, and therefore represents a natural starting point for the use of healthcare appointments at scale. In fact, for many young men in Malawi, HIV testing may be their only interaction with the healthcare system.[37] Whether appointments would work in a context where frequent healthcare visits

[37] According to the 2015-16 Malawi DHS, 42 percent of men in Malawi say they have been tested for HIV

are required, for example, to access HIV treatment, is another question. Encouragingly, Pop-Eleches et al. (2011) show that reminders alone are effective for HIV treatment adherence, even over a one-year time horizon. Yet, they find that daily reminders are not effective, suggesting that at very high frequency, appointment setting may also lose its power as a behavioral nudge.

While not a mechanism *per se*, another explanation for our results is that they are driven purely by displacement: men who would have gotten tested in the near future simply move their tests forward in time, and the intervention causes no meaningful increase in the HIV testing rate. This explanation seems unlikely for three reasons. First, the time pattern of testing in the control group shows almost no testing after the first couple of weeks. If testing were simply displaced slightly earlier in time in the appointments arm, we would expect testing to continue in the other arms, rather than stopping. Second, the time window for the study was intentionally vague, with no sharp cutoff. Most men who wanted to get tested more than two months after their recruitment into the study could still do so, and could still redeem their voucher, so we would pick up some fraction of any later HIV tests rather than missing them. Third, the annual rate of HIV testing is fairly low in Zomba. Derksen, Muula, and van Oosterhout (2022) find that on average just 5 percent of the population per year voluntarily seek out an HIV test; 26 percent of our appointments arm got an HIV test in less than three months.

## 6 Conclusion

We show that appointments can increase HIV testing more effectively than hard commitment devices, and that their effects are concentrated primarily among those who would like to seek care but are held back by self-control problems. Using a randomized controlled trial in Malawi, we find that offering men appointments for HIV testing more than doubles testing rates. There is also high demand for hard commitment devices, and these devices significantly increase HIV testing rates. However, appointments dominate hard commitment devices, in terms of demand, treatment effects, and cost effectiveness. Appointments also avoid

in the past year, while fewer than 2 percent visited a clinic to seek care for a sexually transmitted infection.

an important downside of hard commitment devices, which is that they can be welfare-diminishing due to failed commitments (John and Orkin 2022; Bai et al. 2021). None of the men in our study lost money due to missed appointments because no money was at stake. It is possible that men who missed appointments suffered mild psychological costs such as shame, however. In our context, appointments act as superior substitutes for the hard commitment devices, and the treatment effects are strongest among those who express *ex ante* demand for commitment.

Appointments can be viewed as a coherent bundle of behavioral interventions. These interventions, many of which have been shown effective on their own, address different and overlapping biases simultaneously. Our results suggest that appointments help people overcome self-control and limited-memory problems. People are compelled to follow through on the soft commitment, perhaps due to social pressure, or for personal reasons. Appointments address limited memory problems by providing men with reminders to come to their appointments: testing quickly trails off to zero in the control group, but continues in the appointments arm—even for men who do not actually come in on their appointment date. While these two mechanisms appear important, the large effects we observe suggest that the whole may be greater than the sum of the parts.

Health appointments are a natural and relatively inexpensive policy tool. While the introduction of a healthcare appointment system can be initially costly and complex, in the long run, appointments are more straightforward to implement than traditional hard commitment devices. Appointments are attractive to many health providers and policymakers because they allocate time and resources efficiently. Even in poor countries, mobile phone ownership is becoming increasingly common, so it is possible to schedule appointments and issue reminders. Using a hard commitment device for healthcare requires collecting money, or providing and withholding incentives at scale. Adding financial transactions to a healthcare system is likely to introduce managerial and logistical challenges, with no added benefit relative to simply using appointments.

This raises the question of why appointments are not already widespread in low-income countries. In our qualitative data (Appendix A), we do not see evidence for social or cultural barriers to appointments on the part of healthcare users. On the contrary, clinic staff report that some health-seekers do express a desire for fixed appointments, both for HIV testing and for other health services. Yet clinics view appointments as largely unnecessary. The vast majority of health services in Malawi, as in many low-income settings, are provided by low-skilled health workers. This includes HIV testing and treatment, family planning services, and outpatient services for e.g. sexually transmitted infections or malaria. These services typically have staff and supplies on hand, and clinics do not need to manage patient flows with an appointment system. Instead, they appear to prefer the flexibility of walk-in services, which facilitate triage. Notably, clinic staff do not report technological limitations as a barrier to appointment scheduling. As mobile phone use spreads, the use of scheduled appointments will likely become more commonplace in low-resource settings.

How generalizable are these results likely to be? Appointments are not likely to increase demand for healthcare in all settings. Our findings suggest that appointments are most effective for people with known self-control problems as opposed to those who purposefully avoid care (Chang et al. 2021). Indeed, appointments appear to work best for those who want to voluntarily commit to a particular health behavior. This suggests that appointment effectiveness depends both on features of the healthcare service and on the personal beliefs of the user. Many health services offer long-run benefits but involve immediate physical or mental costs, such as a painful cancer screening or a diagnostic test with life-changing implications. Self-control problems are a likely barrier to these types of care, and appointments may be effective. On the other hand, some health services, such as vaccination, are not particularly costly, even in the short term, but are prone to hesitancy. If potential users do not believe in the benefits of care, self control is not the primary barrier, and offering appointments is unlikely to work.

Yet the fact that offering a simple appointment can impact HIV testing is a promising sign that appointments may be effective in other settings. Seeking an HIV test is an important yet fraught decision. The stakes are high, and access to treatment is lifesaving. We might expect even larger impacts in settings where appointments also reduce wait times. Healthcare appointments are also novel in the specific context of HIV tests in Malawi, although appointments are broadly used for other purposes in the country. This could either increase or decrease their effectiveness. On the one hand, novelty helps to send a stronger signal that HIV testing is important; on the other, novelty also decreases trust that the appointment will be honored by health provider.

Another factor that could affect the generalizability of our results is the in-person recruitment of participants from a population with limited healthcare engagement, and the provision of the testing voucher to all participants in the study. Specifically, our findings speak to the effect of the appointments *conditional* on an in-person recruitment effort and a small cash payment to cover the time and transportation costs of going to the clinic. However, this aspect of our intervention could be scaled up: these vouchers are relatively cheap, and including their cost does not materially affect our cost-effectiveness results. In-person recruitment is likely to be more effective than recruitment online or by phone, especially for target groups that are otherwise difficult to reach. It can be done at relatively low cost by low-skilled healthcare support workers. In our study, the total cost of the appointment intervention, including the costs of recruiting the study participants and conducting the baseline survey, was $4.50 per person. This is likely to be comparable to the cost of scaling the intervention more broadly. Participants were recruited at local bars in one city in Malawi, and this could be done nationwide and also expanded to other recruitment locations.

Studying the role of appointments in other developing-country contexts and demographic groups is an important direction for future research. The use of appointments in developed countries—and their near absence in the developing world—may help explain the severe underutilization of healthcare in developing countries. Developing countries that make the necessary investments to integrate appointments into their healthcare systems may see an increase in demand among those who face behavioral barriers to health behaviors.

# References

**Abadie, Alberto, Matthew M Chingos, and Martin R West.** 2018. "Endogenous Stratification in Randomized Experiments." *Review of Economics and Statistics*, 100(4): 567–580.

**Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge.** 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econo-*

*metrica*, 88(1): 265–296.

**Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge.** 2017. "When Should You Adjust Standard Errors for Clustering?" NBER Working Paper 24003.

**Ahrens, Achim, Christian B. Hansen, and Mark E. Schaffer.** 2019. "PDSLASSO: Stata Module for Post-Selection and Post-Regularization OLS or IV Estimation and Inference." Statistical Software Components, Boston College Department of Economics.

**Alan, Sule, and Seda Ertac.** 2015. "Patience, Self-Control and the Demand for Commitment: Evidence from a Large-Scale Field Experiment." *Journal of Economic Behavior & Organization*, 115: 111–122.

**Altmann, Steffen, and Christian Traxler.** 2014. "Nudges at the Dentist." *European Economic Review*, 72: 19–38.

**Amador, Manuel, Iván Werning, and George-Marios Angeletos.** 2006. "Commitment vs. Flexibility." *Econometrica*, 74(2): 365–396.

**Anderson, Michael L.** 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association*, 103(484): 1481–1495.

**Angelucci, Manuela, and Daniel Bennett.** 2021. "Adverse Selection in the Marriage Market: HIV Testing and Marriage in Rural Malawi." *The Review of Economic Studies*, 88(5): 2119–2148.

**Angotti, Nicole, Kim Yi Dionne, and Lauren Gaydosh.** 2011. "An Offer You Can't Refuse? Provider-Initiated HIV Testing in Antenatal Clinics in Rural Malawi." *Health Policy and Planning*, 26(4): 307–315.

**Ashraf, Nava, Dean Karlan, and Wesley Yin.** 2006*a*. "Deposit collectors." *The BE Journal of Economic Analysis & Policy*, 6(2).

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

**Ashraf, Nava, Nathalie Gons, Dean S Karlan, and Wesley Yin.** 2003. "A Review of Commitment Savings Products in Developing Countries." ERD Working Paper.

**Bai, Liang, Benjamin Handel, Edward Miguel, and Gautam Rao.** 2021. "Self-Control and Demand for Preventive Health: Evidence from Hypertension in India." *The Review of Economics and Statistics*, 103(5): 835–856.

**Banerjee, Abhijit, and Esther Duflo.** 2006. "Addressing Absence." *Journal of Economic Perspectives*, 20(1): 117–132.

**Banerjee, Abhijit, Arun G. Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew O. Jackson, Harini Kannan, Francine N. Loza, Anirudh Sankar, Anna Schrimpf, and Maheshwor Shrestha.** 2021. "Selecting the Most Effective Nudge: Evidence from a Large-Scale Experiment on Immunization." NBER Working Paper 28726.

**Benjamini, Yoav, Abba M Krieger, and Daniel Yekutieli.** 2006. "Adaptive Linear Step-up Procedures That Control the False Discovery Rate." *Biometrika*, 93(3): 491–507.

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

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

**Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer.** 2013. "Salience and Consumer Choice." *Journal of Political Economy*, 121(5): 803–843.

**Brewer, Noel T., Gretchen B. Chapman, Alexander J. Rothman, Julie Leask, and Allison Kempe.** 2017. "Increasing Vaccination: Putting Psychological Science Into Action." *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, 18(3): 149–207.

**Breza, Emily, and Arun G. Chandrasekhar.** 2019. "Social Networks, Reputation, and Commitment: Evidence From a Savings Monitors Experiment." *Econometrica*, 87(1): 175–216.

**Brune, Lasse, Eric Chyn, and Jason Kerwin.** 2021. "Pay Me Later: Savings Constraints and the Demand for Deferred Payments." *American Economic Review*, 111(7): 2179–2212.

**Bryan, Gharad, Dean Karlan, and Scott Nelson.** 2010. "Commitment Devices." *Annual Review of Economics*, 2(1): 671–698.

**Buehren, Niklas, Markus Goldstein, Leora Klapper, Tricia Koroknay-Palicz, and Simone Schaner.** 2022. "The Limits of Commitment: Who Benefits from Illiquid Savings Products?" Working Paper.

**Burbidge, John B., Lonnie Magee, and A. Leslie Robb.** 1988. "Alternative Transformations to Handle Extreme Values of the Dependent Variable." *Journal of the American Statistical Association*, 83(401): 123–127.

**Burke, Jeremy, Jill Luoto, and Francisco Perez-Arce.** 2018. "Soft versus hard commitments: a test on savings behaviors." *Journal of Consumer Affairs*, 52(3): 733–745.

**Bursztyn, Leonardo, and Robert Jensen.** 2017. "Social Image and Economic Behavior in the Field: Identifying, Understanding, and Shaping Social Pressure." *Annual Review of Economics*, 9(1): 131–153.

**Butera, Luigi, Robert Metcalfe, William Morrison, and Dmitry Taubinsky.** 2021. "Measuring the Welfare Effects of Shame and Pride." NBER Working Paper 25637.

**Chang, Tom, Mireille Jacobson, Manisha Shah, Rajiv Pramanik, and Samir B. Shah.** 2021. "Financial Incentives and Other Nudges Do Not Increase COVID-19 Vaccinations among the Vaccine Hesitant." National Bureau of Economic Research Working Paper 29403.

**Chapman, Gretchen B, Meng Li, Howard Leventhal, and Elaine A Leventhal.** 2016. "Default Clinic Appointments Promote Influenza Vaccination Uptake without a Displacement Effect." *Behavioral Science & Policy*, 2(2): 40–50.

**Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey.** 2017. "Double/Debiased/Neyman Machine Learning of Treatment Effects." *American Economic Review*, 107(5): 261–265.

**Cohen, Myron S., Ying Q. Chen, Marybeth McCauley, Theresa Gamble, Mina C. Hosseinipour, Nagalingeswaran Kumarasamy, James G. Hakim, Johnstone Kumwenda, Beatriz Grinsztejn, Jose H.S. Pilotto, Sheela V. Godbole, Sanjay Mehendale, Suwat Chariyalertsak, Breno R. Santos, Kenneth H. Mayer, Irving F. Hoffman, Susan H. Eshleman, Estelle Piwowar-Manning, Lei Wang,**

**Joseph Makhema, Lisa A. Mills, Guy de Bruyn, Ian Sanne, Joseph Eron, Joel Gallant, Diane Havlir, Susan Swindells, Heather Ribaudo, Vanessa Elharrar, David Burns, Taha E. Taha, Karin Nielsen-Saines, David Celentano, Max Essex, and Thomas R. Fleming.** 2011. "Prevention of HIV-1 Infection with Early Antiretroviral Therapy." *New England Journal of Medicine*, 365(6): 493–505.

**de Aghion, Beatriz Armendariz, and Jonathan Morduch.** 2005. *The Economics of Microfinance.* Cambridge, MA:MIT Press.

**DellaVigna, Stefano, and Ulrike Malmendier.** 2006. "Paying Not to Go to the Gym." *American Economic Review*, 96(3): 694–719.

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

**Derksen, Laura, Adamson Muula, and Joep van Oosterhout.** 2022. "Love in the Time of HIV: How Beliefs About Externalities Impact Health Behavior." *Journal of Development Economics*, 159: 102993.

**Derksen, Laura C., Jason T. Kerwin, Natalia Ordaz Reynoso, and Olivier Sterck.** 2019. "Soft and Hard Commitment Devices to Increase HIV Testing." AEA RCT Registry. November 22. https://doi.org/10.1257/rct.4295-2.2.

**DHS.** 2016. *Malawi Demographic and Health Survey 2015–2016.* National Statistical Office (NSO) and ICF Macro: Zomba, Malawi, and Rockville, MD.

**Dovel, Kathryn, Kelvin Balakasi, Sundeep Gupta, Misheck Mphande, Isabella Robson, Pericles Kalande, Eric Lungu, Alemeyehu Amberbir, Joep J. van Oosterhout, Shaukat Khan, Naoko Doi, and Brooke E. Nichols.** 2020*a*. "Missing Men or Missed Opportunity? Men's Frequent Use of Health Services in Malawi."

**Dovel, Kathryn, Sara Yeatman, Susan Watkins, and Michelle Poulin.** 2015. "Men's Heightened Risk of AIDS-Related Death: The Legacy of Gendered HIV Testing and Treatment Strategies." *AIDS*, 29(10): 1123–1125.

**Dovel, Kathryn, Shari L Dworkin, Morna Cornell, Thomas J. Coates, and Sara Yeatman.** 2020*b*. "Gendered Health Institutions: Examining the Organization of Health Services and Men's Use of HIV Testing in Malawi." *Journal of the International AIDS Society*, 23(Suppl 2).

**Dupas, Pascaline, and Jonathan Robinson.** 2013. "Why Don't the Poor Save More? Evidence from Health Savings Experiments." *American Economic Review*, 103(4): 1138–1171.

**Ericson, Keith Marzilli.** 2017. "On the Interaction of Memory and Procrastination: Implications for Reminders, Deadlines, and Empirical Estimation." *Journal of the European Economic Association*, 15(3): 692–719.

**Ericson, Keith M. Marzilli.** 2011. "Forgetting We Forget: Overconfidence and Memory." *Journal of the European Economic Association*, 9(1): 43–60.

**Ferwerda, Jeremy.** 2014. "ESTRAT: Stata module to perform Endogenous Stratification for Randomized Experiments." Statistical Software Components, Boston College Department of Economics.

**Fitzpatrick, Anne, Sabrin Beg, Laura Derksen, Anne Karing, Jason Kerwin, Adrienne M. Lucas, Natalia Ordaz Reynoso, and Munir Squires.** 2021. "Health Knowledge and Non-Pharmaceutical Interventions During the Covid-19 Pandemic in Africa." *Journal of Economic Behavior & Organization*, 190: 33–53.

**Friedman, Willa, and Nicholas Wilson.** 2021. "Can Nudging Overcome Procrastinating on Preventive Health Investments?" *Economics & Human Biology*, 101040.

**Gabaix, Xavier.** 2019. "Chapter 4 - Behavioral inattention." In *Handbook of Behavioral Economics: Applications and Foundations 1*. Vol. 2 of *Handbook of Behavioral Economics - Foundations and Applications 2*, , ed. B. Douglas Bernheim, Stefano DellaVigna and David Laibson, 261–343. North-Holland.

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

**Glasziou, Paul, Sharon Straus, Shannon Brownlee, Lyndal Trevena, Leonila Dans, Gordon Guyatt, Adam G. Elshaug, Robert Janett, and Vikas Saini.** 2017. "Evidence for Underuse of Effective Medical Services Around the World." *The Lancet*, 390(10090): 169–177.

**Gneezy, Uri, and Aldo Rustichini.** 2000. "A Fine Is a Price." *The Journal of Legal Studies*, 29(1): 1–17.

**Godlonton, Susan, and Rebecca Thornton.** 2012. "Peer Effects in Learning HIV Results." *Journal of Development Economics*, 97(1): 118–129.

**Goodman, Lucas, Anita Mukherjee, and Shanthi Ramnath.** 2021. "Abandoned Retirement Savings." Working Paper.

**Gottlieb, Daniel, and Kent Smetters.** 2021. "Lapse-Based Insurance." *American Economic Review*, forthcoming.

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

**Gurol-Urganci, Ipek, Thyra de Jongh, Vlasta Vodopivec-Jamsek, Rifat Atun, and Josip Car.** 2013. "Mobile Phone Messaging Reminders for Attendance at Healthcare Appointments." *Cochrane Database of Systematic Reviews*, 2013(12).

**Hakimov, Rustamdjan, C.-Philipp Heller, Dorothea Kübler, and Morimitsu Kurino.** 2021. "How to Avoid Black Markets for Appointments with Online Booking Systems." *American Economic Review*, 111(7): 2127–51.

**Halpern, Scott D., Benjamin French, Dylan S. Small, Kathryn Saulsgiver, Michael O. Harhay, Janet Audrain-McGovern, George Loewenstein, Troyen A. Brennan, David A. Asch, and Kevin G. Volpp.** 2015. "Randomized Trial of Four Financial-Incentive Programs for Smoking Cessation." *New England Journal of Medicine*, 372(22): 2108–2117.

**Halpern, Scott D., David A. Asch, and Kevin G. Volpp.** 2012. "Commitment Contracts as a Way to Health." *BMJ*, 344: e522.

**Haushofer, Johannes.** 2015. "The Cost of Keeping Track." Working Paper.

**Imbens, Guido W., and Donald B. Rubin.** 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences.* Cambridge University Press.

**INSIGHT START Study Group, The.** 2015. "Initiation of Antiretroviral Therapy in Early Asymptomatic HIV Infection." *New England Journal of Medicine*, 373(9): 795–807.

**Islam, MD, Susumu Wakai, Nobukatsu Ishikawa, AMR Chowdhury, J Patrick Vaughan, et al.** 2002. "Cost-effectiveness of community health workers in tuberculosis control in Bangladesh." *Bulletin of the World Health Organization*, 80: 445–450.

**Jachimowicz, Jon M., Shannon Duncan, Elke U. Weber, and Eric J. Johnson.** 2019. "When and Why Defaults Influence Decisions: A Meta-Analysis of Default Effects." *Behavioural Public Policy*, 3(2): 159–186. Publisher: Cambridge University Press.

**Jacobson Vann, Julie C., Robert M. Jacobson, Tamera Coyne-Beasley, Josephine K. Asafu-Adjei, and Peter G. Szilagyi.** 2018. "Patient Reminder and Recall Interventions to Improve Immunization Rates." *Cochrane Database of Systematic Reviews*, 2018(1): CD003941.

**John, Anett.** 2020. "When Commitment Fails: Evidence from a Field Experiment." *Management Science*, 66(2): 503–529.

**John, Anett, and Kate Orkin.** 2022. "Can Simple Psychological Interventions Increase Preventive Health Investment?" *Journal of the European Economic Association*, 20(3): 1001–1047.

**Karing, Anne.** 2018. "Social Signaling and Childhood Immunization: A Field Experiment in Sierra Leone." Working Paper.

**Karlan, Dean, and Leigh Linden.** 2023. "Loose Knots: Strong versus Weak Commitments to Save for Education in Uganda." Working Paper.

**Karlan, Dean, Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinman.** 2016. "Getting to the Top of Mind: How Reminders Increase Saving." *Management Science*, 62(12): 3393–3411.

**Karlan, Dean S., and Jonathan Zinman.** 2012. "Borrow Less Tomorrow: Behavioral Approaches to Debt Reduction." SSRN Scholarly Paper ID 2060548, Rochester, NY.

**Kast, Felipe, Stephan Meier, and Dina Pomeranz.** 2012. "Under-Savers Anonymous: Evidence on Self-Help Groups and Peer Pressure as a Savings Commitment Device." NBER Working Paper 18417.

**Kast, Felipe, Stephan Meier, and Dina Pomeranz.** 2018. "Saving More in Groups: Field Experimental Evidence from Chile." *Journal of Development Economics*, 133(C): 275–294.

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

**Kavanagh, Nolan M, Elisabeth M Schaffer, Alex Ndyabakira, Kara Marson, Diane V Havlir, Moses R Kamya, Dalsone Kwarisiima, Gabriel Chamie, and Harsha Thirumurthy.** 2020. "Planning Prompts to Promote Uptake of HIV Services Among Men: A Randomised Trial in Rural Uganda." *BMJ Global Health*, 5(11): e003390.

**Kerwin, Jason T.** 2020. "Scared straight or scared to death? Fatalism in response to disease risks." *mimeo.*

**Kessler, Judd B., and C. Yiwei Zhang.** 2015. "Behavioural Economics and Health." In *Oxford Textbook of Global Public Health.* Vol. 6. 6 ed., , ed. Roger Detels, Martin Gulliford, Quarraisha Abdool Karim and Chorh Chuan Tan. Oxford, UK:Oxford University Press.

**Kim, Jaejeung, Hayoung Jung, Minsam Ko, and Uichin Lee.** 2019. "GoalKeeper: Exploring Interaction Lockout Mechanisms for Regulating Smartphone Use." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1): 16:1–16:29.

**Knaus, Michael C., Michael Lechner, and Anthony Strittmatter.** 2020. "Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach." *Journal of Human Resources*, 0718–9615R1.

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

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

**Macis, Mario, Michelle Grunauer, Erika Gutierrez, Ricardo Izurieta, Phillip Phan, Miguel Reina Ortiz, Carlos Rosas, and Enrique Teran.** 2021. "Using Incentives and Nudging to Improve Non-Targeted HIV Testing in Ecuador: A Randomized Trial." *AIDS and Behavior.*

**Mas, Alexandre, and Enrico Moretti.** 2009. "Peers at Work." *American Economic Review*, 99(1): 112–145.

**Mayer, Jonathan E., and Paul Fontelo.** 2017. "Meta-Analysis on the Effect of Text Message Reminders for HIV-Related Compliance." *AIDS Care*, 29(4): 409–417.

**Milkman, Katherine L., John Beshears, James J. Choi, David Laibson, and Brigitte C. Madrian.** 2011. "Using Implementation Intentions Prompts to Enhance Influenza Vaccination Rates." *Proceedings of the National Academy of Sciences*, 108(26): 10415.

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

**Milkman, Katherine L., Mitesh S. Patel, Linnea Gandhi, Heather Graci, Dena Gromet, Hung Ho, Joseph Kay, Timothy Lee, Modupe Akinola, John Beshears, Jon Bogard, Alison Buttenheim, Christopher Chabris, Gretchen B. Chapman, James J. Choi, Hengchen Dai, Craig R. Fox, Amir Goren, Matthew Hilchey, Jillian Hmurovic, Leslie John, Dean Karlan, Melanie Kim, David Laibson, Cait Lamberton, Brigitte C. Madrian, Michelle N. Meyer, Maria Modanu, Jimin Nam, Todd Rogers, Renante Rondina, Silvia Saccardo, Maheen Shermohammed, Dilip Soman, Jehan Sparks, Caleb Warren, Megan Weber, Ron Berman, Chalanda Evans, Christopher Snider, Eli Tsukayama, Christophe Van den Bulte, Kevin Volpp, and Angela Duckworth.** 2021. "A Mega-Study of Text-Based Nudges Encouraging Patients to Get Vaccinated at an Upcoming Doctor's Appointment." SSRN Scholarly Paper ID 3780267, Rochester, NY.

**Montiel Olea, José Luis, and Carolin Pflueger.** 2013. "A Robust Test for Weak Instruments." *Journal of Business & Economic Statistics*, 31(3): 358–369.

**Mugo, Peter M., Elizabeth W. Wahome, Evanson N. Gichuru, Grace M. Mwashigadi, Alexander N. Thiong'o, Henrieke A. B. Prins, Tobias F. Rinke de Wit, Susan M. Graham, and Eduard J. Sanders.** 2016. "Effect of Text Message, Phone Call, and In-Person Appointment Reminders on Uptake of Repeat HIV Testing among Outpatients Screened for Acute HIV Infection in Kenya: A Randomized Controlled Trial." *PLOS ONE*, 11(4): e0153612.

**Muralidharan, Karthik, Mauricio Romero, and Kaspar Wüthrich.** 2019. "Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments." NBER Working Paper 26562.

**National Statistical Office/Malawi, and ICF.** 2017. "Malawi Demographic and Health Survey 2015-16." National Statistical Office and ICF, Zomba, Malawi.

**Ngatia, Mũthoni.** 2016. “Social Interactions, Stigma, and HIV Testing.” Working Paper.

**Nyondo, Alinane Linda, Augustine Talumba Choko, Angela Faith Chimwaza, and Adamson Sinjani Muula.** 2015. “Invitation Cards during Pregnancy Enhance Male Partner Involvement in Prevention of Mother to Child Transmission (PMTCT) of Human Immunodeficiency Virus (HIV) in Blantyre, Malawi: A Randomized Controlled Open Label Trial.” *PLOS ONE*, 10(3): e0119273.

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

**Pop-Eleches, Cristian, Harsha Thirumurthy, James P Habyarimana, Joshua G Zivin, Markus P Goldstein, Damien De Walque, Leslie Mackeen, Jessica Haberer, Sylvester Kimaiyo, John Sidle, et al.** 2011. “Mobile phone technologies improve adherence to antiretroviral treatment in a resource-limited setting: a randomized controlled trial of text message reminders.” *AIDS (London, England)*, 25(6): 825.

**Pugatch, Todd, and Nicholas Wilson.** 2018. “Nudging Study Habits: A Field Experiment on Peer Tutoring in Higher Education.” *Economics of Education Review*, 62: 151–161.

**Rana, Yashodhara, Jessica Haberer, Haijing Huang, Andrew Kambugu, Barbara Mukasa, Harsha Thirumurthy, Peter Wabukala, Glenn J. Wagner, and Sebastian Linnemayr.** 2015. “Short Message Service (SMS)-Based Intervention to Improve Treatment Adherence among HIV-Positive Youth in Uganda: Focus Group Findings.” *PLOS ONE*, 10(4): e0125187.

**Rao, Neel, Markus M. Möbius, and Tanya Rosenblat.** 2007. “Social Networks and Vaccination Decisions.” Federal Reserve Bank of Boston Working Paper 07-12.

**Ravallion, Martin.** 2017. “A Concave Log-Like Transformation Allowing Non-Positive Values.” *Economics Letters*, 161: 130–132.

**Rogers, Todd, and Katherine L. Milkman.** 2016. “Reminders Through Association.” *Psychological Science*, 27(7): 973–986.

**Rogers, Todd, Katherine L. Milkman, Leslie K. John, and Michael I. Norton.** 2015. “Beyond Good Intentions: Prompting People to Make Plans Improves Follow-Through on Important Tasks.” *Behavioral Science & Policy*, 1(2): 33–41.

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

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

**Sadoff, Sally, Anya Samek, and Charles Sprenger.** 2020. "Dynamic Inconsistency in Food Choice: Experimental Evidence from Two Food Deserts." *The Review of Economic Studies*, 87(4): 1954–1988.

**Salvadori, Nicolas, Pierrick Adam, Jean-Yves Mary, Luc Decker, Lucie Sabin, Sylvie Chevret, Surachet Arunothong, Woottichai Khamduang, Prapan Luangsook, Visitsak Suksa-ardphasu, et al.** 2020. "Appointment Reminders to Increase Uptake of Hiv Retesting by at-Risk Individuals: A Randomized Controlled Study in Thailand." *Journal of the International AIDS Society*, 23(4): e25478.

**Saretsky, Gary.** 1972. "The OEO P.C. Experiment and the John Henry Effect." *The Phi Delta Kappan*, 53(9): 579–581.

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

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

**Steenland, Maria, Janeth Dula, Amanda de Albuquerque, Quinhas Fernandes, Rosa Marlene Cuco, Sergio Chicumbe, Eduardo Samo Gudo, Sandra Sequeira, and Margaret McConnell.** 2019. "Effects of Appointment Scheduling on Waiting Time and Utilisation of Antenatal Care in Mozambique." *BMJ Global Health*, 4(6): e001788.

**Sterck, Olivier.** 2013. "Why are testing rates so low in Sub-Saharan Africa? Misconceptions and strategic behaviors." *Forum for Health Economics and Policy*, 16(1): 219–257.

**Sterck, Olivier.** 2014. "HIV/AIDS and fatalism: should prevention campaigns disclose the transmission rate of HIV?" *Journal of African Economies*, 23(1): 53–104.

**Taylor, Darlene, Carole Lunny, Petra Lolić, Orion Warje, Jasmina Geldman, Tom Wong, Mark Gilbert, Richard Lester, and Gina Ogilvie.** 2019. "Effectiveness of Text Messaging Interventions on Prevention, Detection, Treatment, and Knowledge Outcomes for Sexually Transmitted Infections (STIs)/HIV: A Systematic Review and Meta-Analysis." *Systematic Reviews*, 8(1): 12.

**Tenthani, Lyson, Andreas D. Haas, Hannock Tweya, Andreas Jahn, Joep J. van Oosterhout, Frank Chimbwandira, Zengani Chirwa, Wingston Ng'ambi, Alan Bakali, Sam Phiri, Landon Myer, Fabio Valeri, Marcel Zwahlen, Gilles Wandeler, and Olivia Keiser.** 2014. "Retention in Care under Universal Antiretroviral Therapy for HIV Infected Pregnant and Breastfeeding Women ("Option B+") in Malawi." *AIDS*, 28(4): 589–598.

**Thornton, Rebecca L.** 2008. "The Demand for, and Impact of, Learning HIV Status." *American Economic Review*, 98(5): 1829–1863.

**Toussaert, Severine.** 2019. "Revealing Temptation Through Menu Choice: Field Evidence." Working Paper.

**UNAIDS.** 2020. "UNAIDS Data 2020." *Programme on HIV/AIDS*.

**Vervloet, Marcia, Annemiek J Linn, Julia C M van Weert, Dinny H de Bakker, Marcel L Bouvy, and Liset van Dijk.** 2012. "The Effectiveness of Interventions Using Electronic Reminders to Improve Adherence to Chronic Medication: A Systematic Review of the Literature." *Journal of the American Medical Informatics Association*, 19(5): 696–704.

**Figure 1**
Experimental Design
(HC = Commitment Device)

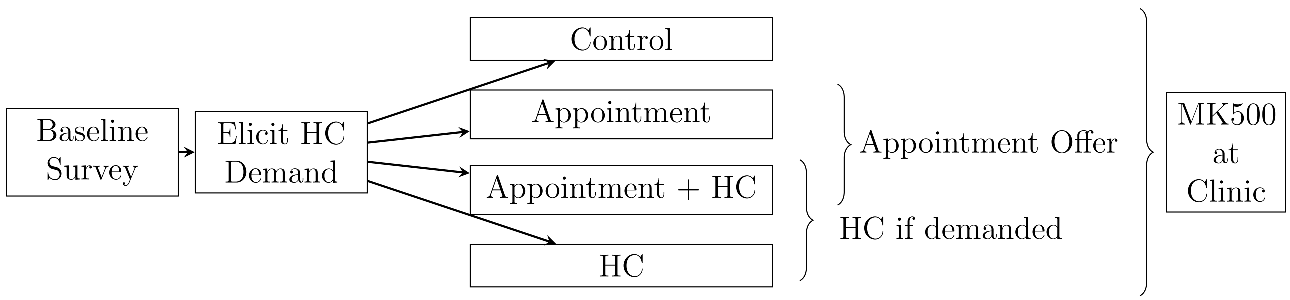



**Figure 2**
Participant Compensation with and without Commitment Device

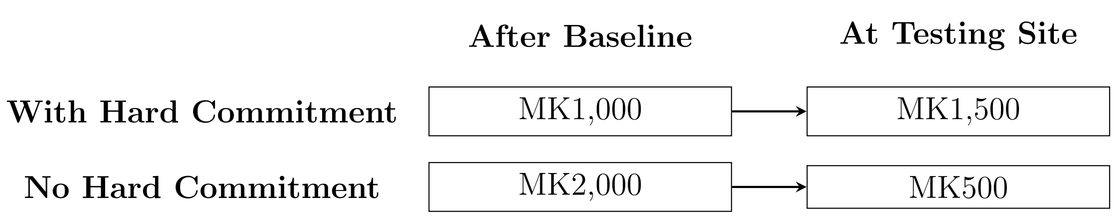
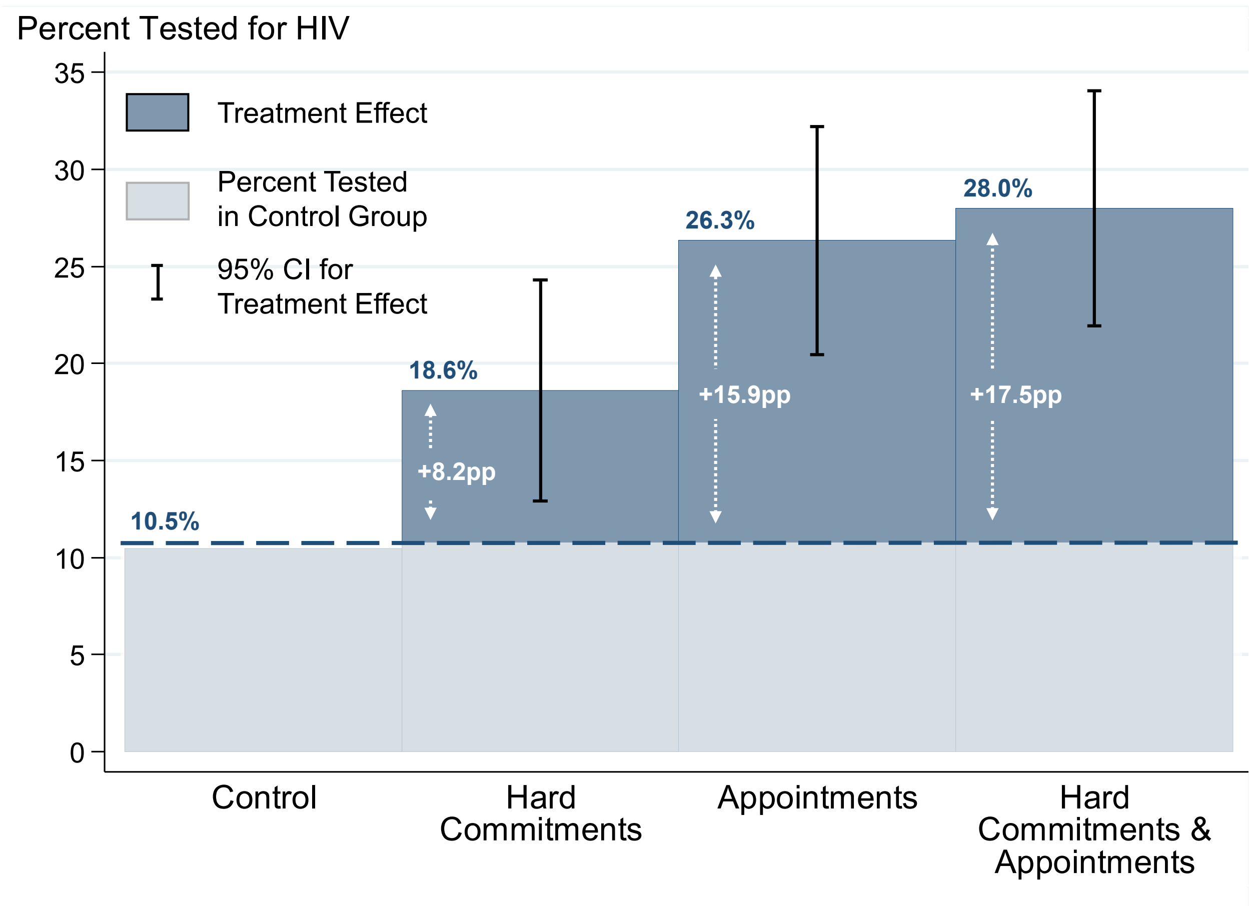


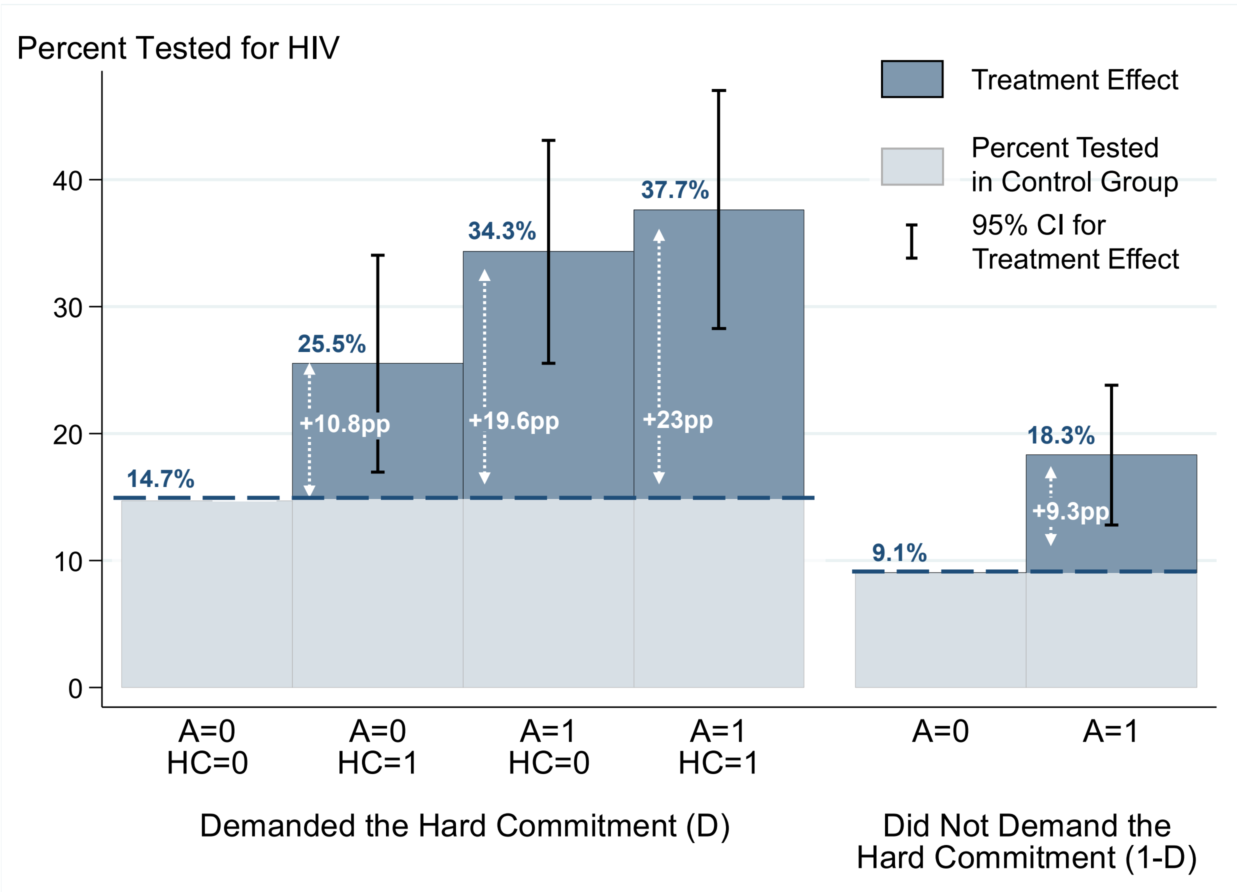
**Figure 3**
HIV Testing Rates by Study Arm




*Notes:* Bars represent predicted values for each study arm based on Column 2 of Table 2, which uses Equation 1 with our pre-specified control variables and the full set of fixed effects. Whiskers show 95 percent confidence intervals for the differences between the control group and each treatment arm. Confidence intervals for the differences between treatment arms are of similar width, and thus we can reject A = HC at the 5 percent level (and likewise for HC + A = HC), but cannot reject the equality of A and HC + A.

**Figure 4**

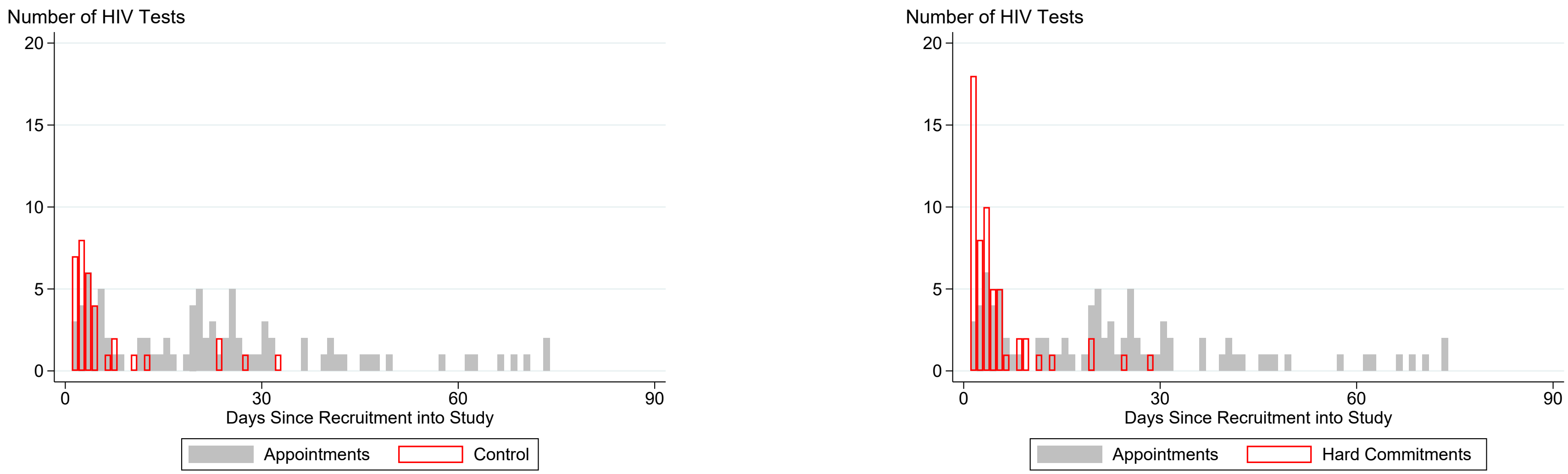
Variation in Treatment Effects by Demand for,
and Random Assignment to, the Hard Commitment Device




*Notes:* Predicted values of the outcome variable based on Column 2 of Table 4. $HC = 1$ for participants who were randomly assigned to one of the hard commitment arms and $HC = 0$ otherwise. $A = 1$ for participants who were randomly assigned to one of the appointment arms and $A = 0$ otherwise. Arrows show the magnitudes of the treatment effects. Whiskers show 95 percent confidence intervals for the differences between the control group and each treatment arm within each panel. Confidence intervals for the differences between treatment arms are of similar width, and thus (for example) we can reject the equality of A + HC and HC at the 5 percent level for those who demanded the hard commitment (left panel, second and fourth bars).

**Figure 5**

Histograms of Delays before HIV Test



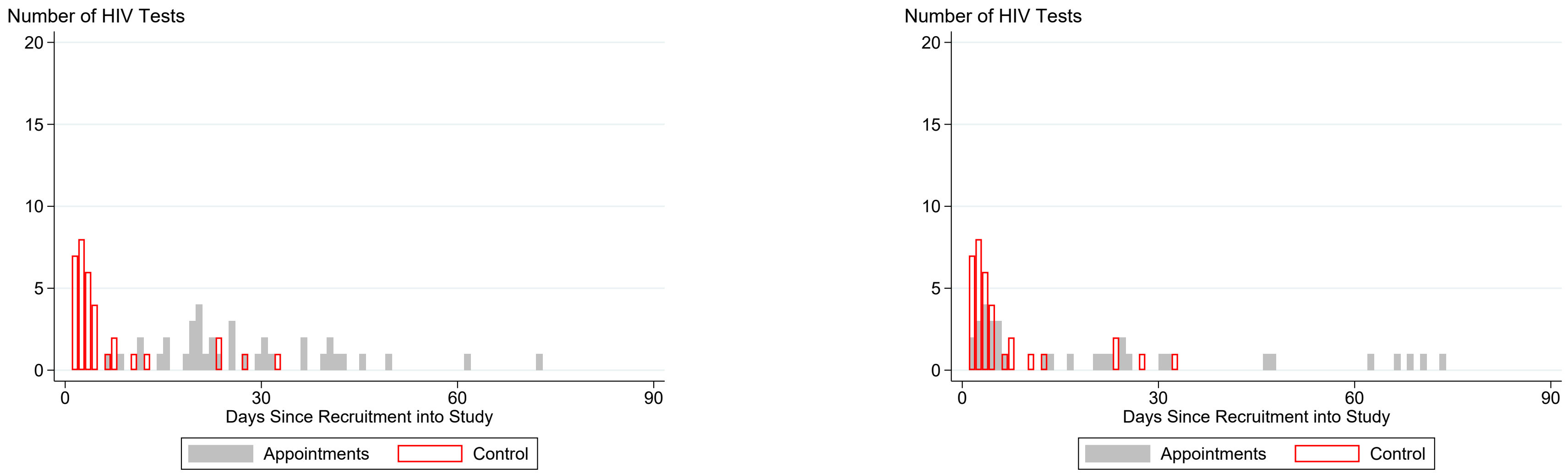

**Panel A:** Appointments vs. Control Group

**Panel B:** Appointments vs. Hard Commitment Devices

*Notes:* Histogram of HIV test timing relative to the date the subject was recruited, i.e. the baseline survey date. Sample includes only men who got an HIV test, from the control group ($N = 34$) the pure appointments arm ($N = 87$), and the pure hard commitment arm ($N = 57$). $t$-tests for the equality of the average delay before an HIV test: Panel A, $p < 0.001$, $q = 0.001$; Panel B, $p < 0.001$, $q = 0.001$.

**Figure 6**

Histograms of Delays before HIV Test by Whether Test was on Appointment Date



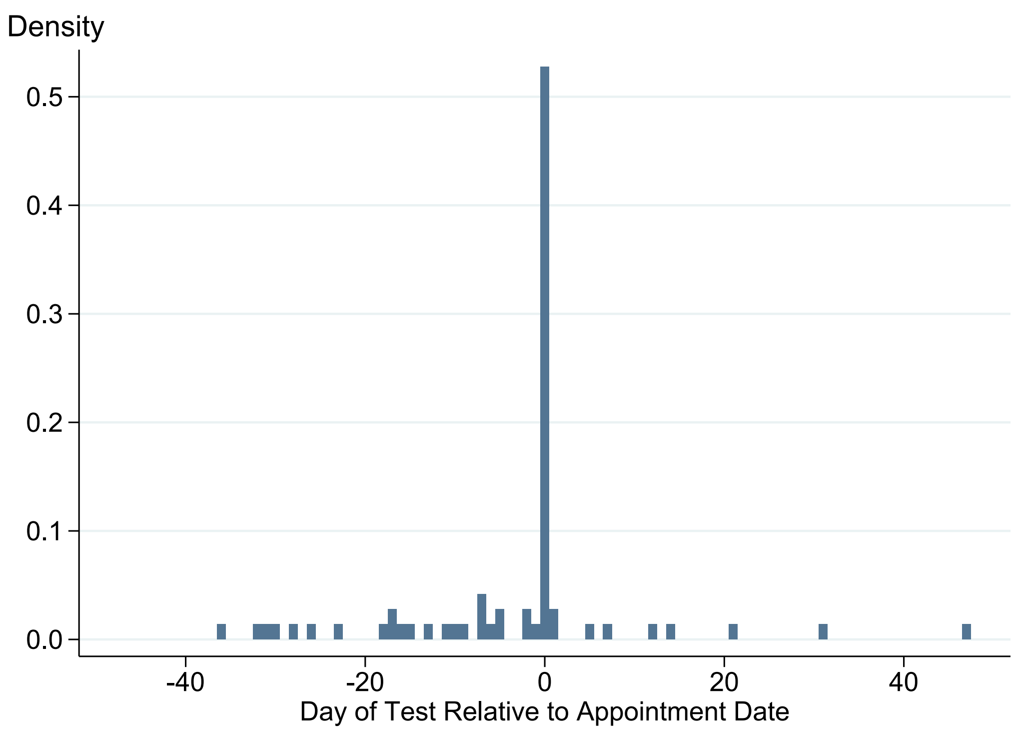

**Panel A:** Tested on Appointment Date

**Panel B:** Tested on a Another Date

*Notes:* Histogram of HIV test timing relative to the date the subject was recruited, i.e. the baseline survey date. Sample includes all control-group men who got an HIV test ($N = 34$), and men from the pure appointments arm who got an HIV test and for whom appointment rescheduling information is available ($N = 72$). $t$-tests for the equality of the average delay before an HIV test: Panel A, $p < 0.001$, $q = 0.001$; Panel B, $p < 0.001$, $q = 0.002$.

**Figure 7**

Timing of HIV Tests Relative to Appointment Date




*Notes:* Histogram of HIV test timing relative to appointment date (using the new date for men who rescheduled their appointments). Sample includes the 72 men of the appointments-only arm who scheduled an appointment, got an HIV test and for whom appointment rescheduling information is available.

## Table 1

Take-up and Effects of Commitment Devices for Health Behaviors

| Study | Outcome (1) | Treatment (2) | Take-up Rate (%) (3) | Rate of Forfeiting Money (%) (4) | Treatment Effect (% Increase) (5) |
|---|---|---|---|---|---|
| Sadoff et al. (2020) | Fruit and vegetable purchases | Commitment to previous food choices | 78 | 0 | 6 |
| Alan and Ertac (2015) | Chocolate bars allocated to day 2 | Lockbox for chocolates (not randomized) | 69 | 0 | — |
| Toussaert (2019) | Weight loss | Financial commitment (not randomized) | 65 | 56 | — |
| Milkman et al. (2014) | Attended gym $\geq$ 1x per week | Gym-only audiobook access | 61 | 0 | 33 |
| | | Encouragement to limit audiobooks to gym | — | 0 | 19 |
| Schilbach (2019) | Sobriety | Financial commitment | 48 | 47 | 36 |
| Schwartz et al. (2014) | Healthy grocery purchases | Financial commitment | 36 | 67 | 3 |
| Bai et al. (2020) | Doctor Visits | Fixed financial commitment | 14 | 62 | 17 |
| | | Personalized financial commitment | 14 | 70 | 43 |
| | | Fixed financial commitment + discount | 26 | 68 | 33 |
| | | Personalized financial commitment + discount | 39 | 77 | 91 |
| Bhattacharya et al. (2015) | Weeks of exercise completed | Nudge to 8-week financial commitment (control) | 23 | — | — |
| | | Nudge to 12-week financial commitment | 23 | — | 6 |
| | | Nudge to 20-week financial commitment | 22 | — | 28 |
| Sadoff and Samek (2019) | Proportion of healthy food choices | Menu with only healthy food choices (not randomized) | 20 | 0 | — |
| Halpern et al. (2015) | Smoking cessation | Individual financial commitment | 13 | 84 | 57 |
| | | Competitive financial commitment | 15 | 89 | 85 |
| Gine et al. (2010) | Smoking cessation | Financial commitment | 11 | 66 | 39 |
| Royer et al. (2015) | Attended gym $\geq$ 1x per week | Financial commitment + incentives | 12 | 27 | 17 |
| *Average* | | | *38* | *37* | *29* |
| *Average for financial commitments* | | | *26* | *65* | *38* |
| **Derksen et al. (2021)** | **HIV Testing** | **Financial commitment** | **49** | **64** | **73** |
| | | **Appointment** | **65** | **0** | **141** |

*Notes:* The table is sorted by the study-average value of Column 3 (the share of people offered the device who take it up). Column 4 presents the fraction of people who forfeit the money they put at stake, conditional on taking up the commitment device (if relevant). Column 5 shows the treatment effect of the commitment device on the outcome of interest, as a percent of the control-group mean. Fields marked with — are not reported by the study. Sadoff, Samek, and Sprenger (2020) results are for the Los Angeles site, because that is where the treatment was randomized; we show the treatment effect on fruit & vegetable purchases, which is the only one of the seven outcomes where increases were desirable. Milkman, Minson, and Volpp (2014) take-up rate is the share of the entire sample with positive WTP in a Becker–DeGroot–Marschak mechanism. Bai et al. (2021) arm-specific forfeiture results are from personal correspondence with the authors. Sadoff and Samek (2019) use randomized treatments to increase demand for commitment but do not look at effects on the outcome; we present the take-up rate from their control group. Royer, Stehr, and Sydnor (2015) treatment effect is the marginal effect of the hard commitment relative to the pure incentives arm.

**Table 2**

Intention-to-Treat Effects of Appointments and Hard Commitment Devices on HIV Testing

| | (1) | (2) | (3) |
|---|---|---|---|
| Appointment Only (A) | 0.151*** | 0.159*** | 0.154*** |
| | (0.030) | (0.030) | (0.029) |
| | [0.001] | [0.001] | [0.001] |
| Hard Commitment Only (HC) | 0.080*** | 0.082*** | 0.085*** |
| | (0.029) | (0.029) | (0.029) |
| | [0.010] | [0.008] | [0.006] |
| Both A and HC | 0.151*** | 0.175*** | 0.175*** |
| | (0.031) | (0.031) | (0.030) |
| | [0.001] | [0.001] | [0.001] |
| **Comparisons of Effects** | | | |
| A − HC | 0.071** | 0.077** | 0.069** |
| SE for $t$-test | (0.034) | (0.032) | (0.031) |
| $q$-value $t$-test | [0.034] | [0.021] | [0.029] |
| (Both A and HC) − HC | 0.071** | 0.094*** | 0.090*** |
| SE for $t$-test | (0.034) | (0.033) | (0.032) |
| $q$-value $t$-test | [0.038] | [0.008] | [0.009] |
| (Both A and HC) − A | -0.001 | 0.017 | 0.021 |
| SE for $t$-test | (0.035) | (0.034) | (0.033) |
| $q$-value $t$-test | [0.428] | [0.295] | [0.264] |
| Controls | None | Pre-Specified | Double Lasso |
| Observations | 1232 | 1232 | 1232 |
| Adjusted R-squared | 0.021 | 0.123 | 0.120 |
| Control-group Mean | 0.113 | 0.113 | 0.113 |

*Notes:* Sample is 1,232 men who completed a baseline survey. Pre-specified controls include all the variables in Table D1. Double Lasso controls uses the Chernozhukov et al. (2017) method for variable selection and inference, as described in Section 3.1. Columns 2 and 3 also control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

**Table 3**

2SLS Estimates of the Treatment-on-the-Treated Effect of Appointments

| **Panel A**: Second Stage | *Outcome: 1(Tested for HIV)* | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Signed up for an Appointment | 0.228*** | 0.240*** | 0.225*** |
| | (0.044) | (0.043) | (0.043) |
| | [0.001] | [0.001] | [0.001] |
| Controls | None | Pre-Specified | Double Lasso |
| Observations | 1232 | 1232 | 1232 |
| Adjusted R-squared | 0.052 | 0.142 | 0.142 |
| Control-group Mean | 0.113 | 0.113 | 0.113 |
| **Panel B**: First Stage | *Outcome: 1(Signed up for an Appointment)* | | |
| Appointment Only (A) | 0.666*** | 0.660*** | 0.660*** |
| | (0.026) | (0.026) | (0.026) |
| | [0.001] | [0.001] | [0.001] |
| Effective $F$-Statistic | 652.9 | 680.5 | 667.0 |
| Controls | None | Pre-Specified | Double Lasso |
| Observations | 1232 | 1232 | 1232 |
| Adjusted R-squared | 0.477 | 0.545 | 0.543 |
| Control-group Mean | 0.000 | 0.000 | 0.000 |

*Notes:* Sample is 1,232 men who completed a baseline survey. The first- and second-stage regressions are estimated using equations 2 and 3; we control for $HC$ and $Both$ but show only the coefficients for variables of interest in the table. Pre-specified controls include all the variables in Table D1. Double Lasso controls uses the Chernozhukov et al. (2017) method for variable selection and inference, as described in Section 3.1. Columns 2 and 3 both control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. Panel B shows the effective $F$-statistic of Montiel Olea and Pflueger (2013). Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

**Table 4**

Treatment Effect Heterogeneity by Demand for Hard Commitment

| | *Outcome: 1(Tested for HIV)* | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| **Demanded the Hard Commitment (D)** | | | |
| D | 0.141*** | 0.147*** | 0.146*** |
| | (0.027) | (0.027) | (0.027) |
| | [0.001] | [0.001] | [0.001] |
| D × Appointment Only (A) | 0.196*** | 0.196*** | 0.188*** |
| | (0.046) | (0.045) | (0.044) |
| | [0.001] | [0.001] | [0.001] |
| D × Hard Commitment Only (HC) | 0.125*** | 0.108** | 0.110** |
| | (0.045) | (0.044) | (0.043) |
| | [0.009] | [0.017] | [0.014] |
| D × Both A and HC | 0.233*** | 0.230*** | 0.228*** |
| | (0.049) | (0.048) | (0.047) |
| | [0.001] | [0.001] | [0.001] |
| **Did Not Demand the Hard Commitment (1–D)** | | | |
| (1–D) | 0.097*** | 0.091*** | 0.097*** |
| | (0.018) | (0.020) | (0.019) |
| | [0.001] | [0.001] | [0.001] |
| (1–D) × Appointment Only (A) | 0.083*** | 0.093*** | 0.088*** |
| | (0.028) | (0.028) | (0.027) |
| | [0.005] | [0.002] | [0.003] |
| **Comparisons of Effects** | | | |
| D × A − (1–D) × A | 0.113** | 0.104** | 0.100* |
| SE for $t$-test | (0.053) | (0.053) | (0.052) |
| $q$-value $t$-test | [0.035] | [0.044] | [0.046] |
| D × (Both A and HC) − D × HC | 0.108** | 0.121** | 0.118** |
| SE for $t$-test | (0.054) | (0.052) | (0.051) |
| $q$-value $t$-test | [0.043] | [0.023] | [0.023] |
| D × (Both A and HC) − D × A | 0.037 | 0.033 | 0.040 |
| SE for $t$-test | (0.055) | (0.053) | (0.052) |
| $q$-value $t$-test | [0.256] | [0.269] | [0.235] |
| Controls | None | Pre-Specified | Double Lasso |
| Study Arms Included | All | All | All |
| Observations | 1232 | 1232 | 1232 |
| Adjusted $R$-squared | 0.251 | 0.308 | 0.304 |
| Control-group Mean | 0.113 | 0.113 | 0.113 |

*Notes:* Sample is 1,232 men who completed a baseline survey. These regressions do not include a constant; (HC) x (1-D) is omitted and assumed to be zero. Pre-specified controls include all the variables in Table D1. Double Lasso controls uses the Chernozhukov et al. (2017) method for variable selection and inference, as described in Section 3.1. Columns 2 and 3 also control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. All controls are standardized prior to running the regressions. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

## Table 5

Cost Effectiveness

| | Appointments | | | | Hard Commitments | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Including Testing Voucher | Including Two Call attempts | Including Testing Voucher & Recruiment Costs | Excluding Cost of HC | Including Cost of HC | Excluding Cost of HC, Including Testing Voucher | Including Cost of HC & Testing Voucher | Including Cost of HC, Testing Voucher & Recruitment Costs |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Cost per Participant | | | | | | | | | |
| MK | 314.07 | 445.57 | 342.82 | 3,307.17 | 181.46 | 1,181.46 | 274.46 | 1,274.46 | 4,083.60 |
| USD | $0.43 | $0.61 | $0.47 | $4.50 | $0.25 | $1.61 | $0.37 | $1.73 | $5.56 |
| Effect Size (percentage points) | 0.159 | 0.159 | 0.159 | 0.155 | 0.082 | 0.082 | 0.082 | 0.082 | 0.082 |
| Cost per Additional Person | $2.69 | $3.81 | $2.93 | $29.04 | $3.01 | $19.61 | $4.56 | $21.16 | $67.79 |
| Effect Size (percent) | 140.7% | 140.7% | 140.7% | 137.2% | 72.6% | 72.6% | 72.6% | 72.6% | 72.6% |
| Cost per 100% Increase | $0.30 | $0.43 | $0.33 | $3.28 | $0.34 | $2.22 | $0.51 | $2.39 | $7.66 |

*Notes:* Sample is 1,232 men who completed a baseline survey. This table focuses on the marginal cost of running each intervention relative to the control group; unless otherwise specified we do not include the costs of the HIV testing vouchers. We compute costs based on the additional time needed to recruit participants in each of the two pure treatment arms, priced at the wage we paid to the survey enumerators. For the appointments, we also include the time and airtime costs of the appointment reminders, and the rescheduling calls (for the respondents who needed to reschedule). Column 3 includes the MK1,000 that participants were allowed to stake on the hard commitment device, since the commitment devices cannot be offered without incurring this cost. The costs of the tests themselves are not included, as the goal is to encourage the uptake of HIV tests. The effect sizes come from the main results, in Column 2 of Table 2.

**Table 6**
Effects on Positive HIV Tests and ART Initiation

| | *Outcome:* | | | |
|---|---|---|---|---|
| | *HIV-Positive* | *Initiated ART* | *HIV-Positive* | *Initiated ART* |
| | (1) | (2) | (3) | (4) |
| Appointment Treatment (A) | 0.011* | 0.009* | | |
| | (0.006) | (0.006) | | |
| | [0.053] | [0.073] | | |
| Hard Commitment Treatment (HC) | 0.002 | 0.004 | | |
| | (0.006) | (0.006) | | |
| | [0.320] | [0.256] | | |
| Appointment Only (A) | | | 0.006 | 0.002 |
| | | | (0.008) | (0.007) |
| | | | [0.243] | [0.331] |
| Hard Commitment Only (HC) | | | -0.003 | -0.003 |
| | | | (0.006) | (0.006) |
| | | | [0.280] | [0.280] |
| Both A and HC | | | 0.013 | 0.013 |
| | | | (0.009) | (0.009) |
| | | | [0.108] | [0.108] |
| **Comparisons of Effects** | | | | |
| A − HC | 0.008 | 0.005 | 0.009 | 0.006 |
| SE for $t$-test | (0.007) | (0.006) | (0.007) | (0.006) |
| $q$-value $t$-test | [0.138] | [0.216] | [0.128] | [0.202] |
| Controls | None | None | None | None |
| Observations | 1232 | 1232 | 1232 | 1232 |
| Adjusted R-squared | 0.001 | 0.001 | 0.001 | 0.001 |
| Control-group Mean | 0.007 | 0.007 | 0.007 | 0.007 |

*Notes:* Sample is 1,232 men who completed a baseline survey. The regressions in this table include no controls or fixed effects because HIV status and ART initiation data was anonymized and linked only to participants' study arms. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

# Supplemental Online Appendix,
# Not Intended for Publication

May 27, 2024

## Appendix A Qualitative Data

In this section we summarize qualitative data collected from staff at clinics in Zomba and across Malawi. We also include a selection of direct quotes that illustrate our key findings. The full, deidentified qualitative dataset will be made available upon request.

These interviews included staff at all clinics in the study area, as well as four other large clinics in Zomba District, and seven hospitals from different regions of Malawi. We interviewed both clinic coordinators and HDAs. This supplementary data collection was approved separately from the main study, by IRBs at the Malawi College of Medicine Research Ethics Committee (2572/2019) and at the University of Toronto (36913/2020).

### Appendix A.1 Appointments for HIV Testing

In general, appointments are not used for HIV testing. There are some exceptions, in some clinics, for people that are referred for HIV testing from the outpatient department (OPD) and patients that were treated for a sexually-transmitted infection (STI). Clinics also sometimes contact the sexual partners of people who have tested positive and invite them to get tested on a particular day. More generally, in the rare cases where "appointments" are scheduled for HIV testing, they simply involve inviting a person to come in for a test within a certain number of weeks, or sometimes on a particular day, without scheduling a specific time or using formal reminders.

> "[When asked if they schedule appointments for HIV testing]: There were some who would do that but it was not common. These would be the people who are very busy like the business people they would call and tell us they want to come for testing on a particular day and time. We also have other clients who would tell us that they can

only come to the clinic on Saturday because of the nature of their work they cannot come during the week." – HDA at a large public clinic

> "We have appointments for the AIT [Active Index Testing][38] clients because we have already made arrangements that they will come to the clinic on a day that we have booked them. [. . . ] We also have the STI clients because we give them an appointment to come after four weeks." – HDA at a large public clinic

> "We don't specify the time when the client should come we just tell them come on Wednesday or Thursday [...] [On the appointment day] we first see the person who booked an appointment, but we inform the people on the queue that this client had an appointment and we have to see him first because we already started the process in the past and we would want to complete the process." – HDA at a large public clinic

## Appendix A.2 Appointments for Other Services

Appointments are also rare in clinics more generally. They are not typically used for OPD services. Patients receiving treatment for HIV are told to come back around a certain date or whenever they need a prescription refill. Clinics do report using informal appointments for various services as needed, but most patients are walk-in, and wait times are short. At large hospitals, highly specialized services, such as surgery and oncology visits, are scheduled for a particular date. But, even in this case patients are not given a particular time, nor formal reminders.

> "It is the nurses and the clinicians in the wards who assess the clients and for those that need medical attention are taken to OPD so there is no need for appointment. For the ART we know and we keep their medication and we give them at the time they are supposed to take the medication; this time we are giving them in the morning. And when the drugs are coming to an end we go and refill at the ART clinic so we don't have to make any appointments." – Staff at a small public clinic

[38] "Active Index Testing" refers to the practice of offering HIV tests to all sexual contacts identified by a person recently diagnosed with HIV.

"[When asked if they use appointments]: If we have the surgical clinic and other clinics; for the surgical clinic they are done on Mondays. [...] We write these clients in the appointment books; we have hard covers which we enter the names of the clients. [...] For this specific clinic we do not give them specific time because we have a room where the clinic is conducted and it is done on first come and first served and the clinician can also make consideration if there are clients who are very sick and needed to be seen first. The clinic runs from 7:30 am and ends at 4:30 pm and people are advised to come within this time." – Staff at a large public hospital

"[When making appointments]: We give them the specific date but we don't give them the specific time. [...] To be honest we do not give them that chance [to choose the appointment date] [...] we just do the booking from our perspectives not considering the client's choice." – Staff at a large public clinic

"For cough, flu and malaria do not require appointments and we have the chronic conditions these require appointments; like those which needs further investigations and treatment they require appointments. [Of those who visit the clinic] 10 percent of the clients [have] appointments." – Staff at a small public clinic

## Appendix A.3 HIV Testing Wait Times and Absenteeism

In July 2019, during the study period, every clinic in our sample consistently had excess capacity for HIV testing. In small clinics, there were multiple HDAs on duty, yet on the average day fewer than ten clients came for HIV testing. Larger clinics saw as many as 12 patients per HDA. HIV testing takes as little as 5 minutes (if the result is negative) or as long as 45 minutes (with counselling). So, it appears that even large clinics were operating below capacity. HDAs report that there was typically no wait for HIV testing, but at busy times, at large clinics, wait times could be as long as one hour. Clinics are well-staffed, and absenteeism among HDAs is extremely rare.

"It was 4 on average. [. . . ] 3 HDAs work in the clinic." – HDA at a small public clinic

“We would have 36 clients on a daily basis. We open every day except for the weekend. [...] We were two HDAs now we have a new one.” – HDA at a large public clinic

“The patients would wait between 15 to 20 minutes because the patient flow was very fast.” – HDA at a large public clinic

“The client will wait for his turn between 40 to 60 minutes. But if there is no client in the counselling room it means he will enter immediately.” – HDA at a large public hospital

“I don’t think there was anyone who was absent. People [HDAs] work up to the knocking off time.” – Senior Staff at a large public clinic

# Appendix B Experiment and Data Details

## Appendix B.1 Recruitment and Randomization

We screened potential participants for mobile phone ownership, as mobile phones were essential to our implementation strategy. Only one potential participant was screened out for this reason; mobile phone ownership is very common in this population. According to the 2015-16 DHS, 82.97 percent of urban men in southern Malawi owned a mobile phone; this fraction is probably even higher among men who frequent bars and engage in transactional sex. We also screened out men who were already diagnosed and on ART treatment (183 exclusions), as well as men who appeared to be intoxicated (1 exclusion), who knew the interviewers personally (1 exclusion), who did not live in the city of Zomba (1 exclusion), or who were under 18 years old (0 exclusions). These exclusion criteria were applied prior to the baseline survey taking place; we have no data for the men who were excluded for this reason.

The survey was administered by trained enumerators who worked in mixed-gender pairs and wore clothing appropriate for a professional setting. The survey took place near where the participant was recruited in an area that afforded privacy.

Randomization was done using pre-randomized lists of intervention assignments that were loaded onto the tablets used for the baseline survey. The random assignments were linked to respondents via sequential ID numbers for each survey interviewer and day. For example, interviewer 1's first respondent on day 1 was ID # 010101, their second respondent was 010102, etc.; each of these IDs had a pre-specified random study arm assignment linked to it.

## Appendix B.2 Commitment Device Explanation Script

In this section we reproduce the script that enumerators used to explain the commitment device to participants. Enumerators first explain the commitment device drawing a parallel to the concept of a *chikole*. After that, enumerators elicit demand for the commitment device. Then, randomization into the commitment device arm is revealed.

*Script:* Before sending the clinic voucher, there is one more thing I would like to discuss with you, it's about self control and HIV testing. Have you ever asked an employer to hold your pay until the end of the job because you don't trust yourself not to spend it? Have you ever taken a loan where you had to give collateral ("chikole") which you would get back when you repay the loan? Are you familiar with this idea of "chikole"? Sometimes you really want to do something, but you don't have the will power or self control. For example, a person might want to go for HIV testing, but he keeps postponing because he is afraid, or he is busy, or maybe he is just lazy. Are you someone that has been postponing HIV testing? It can help to make a commitment, for example, by using a collateral "chikole" to ensure that you go for testing. I want to offer you a way to give some collateral "chikole" for HIV testing. In addition to the study gift, we will add another 1000 Kwacha of airtime. For some people, we will give the choice of moving this airtime from the gift to the HIV testing voucher. This means, you would only receive that airtime if you visit an HIV testing facility. It would be a "chikole" for HIV testing. Are you interested in this "chikole"? If you say yes, we will have a lottery, and 50/50 chance you receive the 1000K gift or the 1000K goes to the "chikole". If you say no, you will just get the 1000K gift now. Are you sure about your answer?

*A: If participant wants the commitment device:* The program is going to randomly choose whether you get the "chikole" or not.

**Randomization revealed**

*A1: If randomized into Commitment Device arm:* The program selected you for chikole. You will receive 1,000 kwacha worth of airtime now. If you go to the HIV testing clinic, you will get a total of 1,500 kwacha: the normal voucher of 500 plus an additional 1,000 kwacha "chikole".

*A2: If not randomized into commitment device arm:* The program did not select you for chikole. You will receive 2,000 kwacha worth of airtime now. You will still receive the 500 kwacha airtime voucher if you go to the HIV testing clinic.

*B If participant does not want the commitment device:*

Ok, since you did not want chikole, I will give you your gift and the extra 1,000 in airtime now I will send you your voucher for 500 kwacha via text message now. Remember you can redeem it at any one of the participating HIV testing clinics any time during clinic hours for the next two months.

## Appendix B.3 HIV Testing and Appointments

HIV testing is free in Malawi, and was free for our study participants. The HIV tests used at the clinics are rapid antigen tests, and the results, as well as post-test counselling, were provided to participants immediately.

During HIV testing appointment reminder calls, for privacy purposes, HIV testing was not explicitly mentioned; the calls mentioned only the time and place of the appointment. To further maintain privacy, and because we did not collect participants' names, the calls did not attempt to verify that the person who answered was the original participant. The reminder call also allowed participants to reschedule their appointments if they wished.

Participants in the appointments arms still received MK500 clinic vouchers which could be redeemed at any clinic and at any time; they did not need to attend their appointment to redeem this voucher, and there was no financial penalty for missing the appointment.

We used three questions to confirm the identity of the participant at the time of HIV testing: the name of the participant's primary school, their year of birth, and their mother's district of birth. To reduce the likelihood of intentional impersonation, these security questions were inconspicuous to participants, and we did not specify how participants' identities would be verified at the clinic. The HDAs verified the answers to these security questions at the time of voucher redemption. Just ten of the participants who got an HIV test gave more than one incorrect answer to our security questions, which was our pre-specified criterion for labeling a test-taker as an impostor. We code the outcome variables for these impostors as zeroes rather than ones; our results are also robust to coding their outcomes as ones (see Appendix Table D11). Fourteen people sought an HIV test and claimed to have lost their vouchers, but were linked to a study identifier based on phone number, and were able to answer all security questions. These participants' HIV tests are included in our main analysis, though they did not receive a voucher payment. One person claimed to have lost his voucher but could not be linked to any study participant based on phone number.

There are potential sources of measurement error in the outcome data. First, it is possible that not all participants who got tested used a voucher: they might have forgotten about it or changed phones or SIM cards (which is common in Malawi). This can be thought of as (unobserved) attrition: we would not observe actual HIV testing outcomes for such participants. We impute their outcomes implicitly, setting the HIV testing indicator equal to zero, which results in a type of measurement error. While some participants visited the clinic claiming to have lost their voucher, the vast majority of these were successfully linked to a study identifier using phone number and security questions, and their HIV testing outcomes are therefore measured correctly. Our voucher redemption data records show just one instance of a man who claimed to be part of the study and got an HIV test, yet could not be linked to a study identifier. This person was likely an impostor. It is nevertheless possible that some HIV tests do not appear in our register data at all.

Another possible source of measurement error is in the process of hand-writing the voucher codes and phone numbers and digitizing them. Our match rate is very high: we are able to match 97 percent of observations from the voucher registers to observations in the baseline data. The remaining measurement error is likely to generate small attenuation bias.

A related consideration in defining the outcome variable is the handling of impostors—men who come to the clinic in order to collect the voucher payment, but do not appear to be the original participant who was recruited into the study. There are only ten of these individuals in our sample; our main analysis codes them as zeroes for the outcome variable. In Appendix Table D11 we code them as ones instead. This leaves the coefficient estimates nearly unchanged, and does not affect any of our inferences based on the $q$-values.

Because our main analysis relies on in-person voucher redemption to link individual HIV test seekers to study identifiers, we are limited in our ability to measure HIV testing behavior after the study ends. We conducted a follow-up survey in 2020, to study longer-run treatment effects and analyze beliefs and behaviors related to COVID-19 (Fitzpatrick et al. 2021). We did attempt to contact all 1,232 men from our original sample to conduct a phone survey. However, our effective attrition rate (after removing impostors) was 81 percent because people in Malawi frequently change SIM cards. We are therefore unable to draw meaningful conclusions from this data.

# Appendix C Cost Calculations

This section details how we calculated the costs of each intervention. We compute only the incremental cost of each pure treatment arm: the *additional* labor costs and other expenses needed to carry out the intervention, beyond the costs that apply to all study arms including the control group. Our cost calculations thus exclude the cost of finding and contacting men for the survey, since that was done in the control group as well. The labor costs are primarily calculated based on differences in the total survey time. To compute these, we only use survey times for respondents who were assigned either into the control group or into one of the two pure intervention arms (that is, we omit the combined intervention arm).

## Appendix C.1 Appointments Arm

The cost of offering an appointment has two main components: (i) the cost of explaining and scheduling an appointment during the baseline survey, and (ii) the cost of reminding the participant of his appointment and rescheduling if necessary.

Part (i) is the product of enumerator wages and the average time it takes to explain and schedule an appointment, taking into account the fact that some participants who are offered an appointment do not schedule one. This time difference has two elements: (a) the time explaining appointments and eliciting demand, and (b) the time spent adding the chosen time slot into the calendar for those who want an appointment. We can measure (a) by calculating the time the SurveyCTO application was active for each participant and taking the difference between the average time in the appointments-only group and the average time in the control group. Part (b) was done in a different application on the phone, so it is not included in the active on-screen time on the survey app. We estimate this time to be one minute per person, based on having our research team carry out the steps needed to schedule the appointment and timing themselves. We therefore add one minute to the times of those participants who actually scheduled an appointment. The total time that corresponds to part (i) equals 139.4 seconds. Enumerators were paid MK20,000 for 6 hour days, so MK3,333.33 per hour. The cost of part (i) is therefore MK129. The following equation shows the formula used to compute part (i) of the cost of the appointments:

$$\text{Part (i)}^A = \big(\text{Extra Survey Time} + \mathbb{1}[\text{Scheduled Appt.}] \times (60 \text{ seconds})\big) \times \frac{\text{Enumerator Hourly Wage}}{3,600 \text{ seconds per hour}}$$

Part (ii) has two components: wages paid to HDAs who make the calls and phone credit spent on successful calls. Times per call were not estimated during the experiment. We asked our research team to estimate these times by calling people and reading them the appointment reminder script and timing themselves. We found that it could take up to 40 seconds for a respondent to pick up, and we estimated the time it takes to read the reminder script to be 65 seconds. Wages must be paid for this entire time (105 seconds) but phone credit is only charged for active call time. Every person who schedules an appointment receives a reminder call. Out of those calls, 42 percent of men who initially scheduled an appointment rescheduled their appointment once, and thus received a second reminder for the rescheduled appointment. At that reminder, 16 percent of men who initially scheduled an appointment rescheduled for a second time, and thus received a third reminder. We account for rescheduling times differently than reminder times. Reminder calls took 65 seconds each, and some of those included rescheduling time, which is estimated to be an additional 40 seconds.

To price the time used on the calls we need to know the wage rate for the HDAs who made the calls, and the cost of phone credit. HDAs were paid MK20,000 for 8 hour days. There are different telecommunication companies in Malawi, and they charge different rates per minute of calls. In our study we used the two main companies that operate in the area. One of them charges MK65 per minute and the other charges MK72. In these calculations we take a conservative approach by using the higher of the two rates to calculate phone credit costs. This brings the cost of part (ii) to be MK185 The following equation depicts the formula used to compute part (ii) of the cost of the appointments.

$$
\begin{aligned}
\text{Part (ii)}^A =& \left( \frac{\text{Calls Made}}{\text{Appointments Offered}} \times \text{ Reminder Script Time} \right. \\
&+ \left. \frac{\text{Rescheduled Appointments}}{\text{Appointments Offered}} \times \text{ Rescheduling Time} \right) \\
&\times \left( \frac{\text{Credit cost/minute}}{\text{60 seconds/hour}} + \frac{\text{HDA Hourly Wage}}{3,600 \text{ seconds/hour}} \right) \\
&+ \frac{\text{Calls made}}{\text{Appointments Offered}} \times \text{ Waiting Time} \times \frac{\text{HDA Hourly Wage}}{3,600 \text{ seconds per hour}}
\end{aligned}
$$

The total cost of adding a person to the appointment arm is the sum of parts (i) and (ii), which comes to MK314.07

## Appendix C.2 Commitment Device Arm

The cost of offering a commitment device has two parts: (i) the extra survey time needed to explain and enroll men in the commitment device, and (ii) the cost of the voucher used for the commitment device. Since we elicited demand for the commitment device from all of our participants, average survey time differences between those assigned only a commitment device and those in the pure control group (16 seconds) do not measure part (i) completely. In order to estimate it, we asked our field team to read the survey script that explains the intervention and time themselves. We find that this part of the survey takes on average 180 seconds. Therefore we add 180 seconds to the average survey time differences between pure control and those only offered a commitment device. These time differences are then multiplied by the wages of enumerators, yielding a total cost for part (i) of MK181.46. The following equation shows the formula used to compute part (i) of the cost of the commitment devices.

$$
\text{Part (i)}^{HC} = (\text{Extra Survey Time} + \text{HC Explanation Script Time}) \times \frac{\text{Enumerator Hourly Wage}}{3,600 \text{ seconds per hour}}
$$

Computing the correct cost to include for part (ii) is less straightforward. If the commitment device is implemented with the participant's own money, the total cost of adding a participant to the commitment device arm would be simply part (i). If the program were to use its own funds, then the amount to be committed needs to be budgeted for. It may be necessary to budget for the MK1,000 for all participants, bringing the total cost to MK1,181.46. But not all participants who accept a commitment device will follow through and collect the money. In our study, 49 percent of men in the commitment device-only study arm enrolled in a commitment device, but just 41 percent of those men redeemed their voucher. Considering the fact that the project kept the MK1,000 for the 29 percent of men who did not redeem their voucher, the net cost of offering the vouchers is just MK710.9 per person. Thus the cost of enrolling one man in the commitment device arm would come to MK892.36.

The above describes in detail how Columns 1, 4, and 5 are calculated. Column 2 shows the cost effectiveness of appointments as described plus the inclusion of the testing voucher (MK500) for those who tested (26.3%)[39] while Column 3 includes an additional unsuccessful reminder call attempt per person (an additional 40 seconds of HDA time). Finally, Columns 6 and 7 correspond to Columns 2 and 3 plus the inclusion of the testing voucher cost (MK500) for those who claim it (18.6%). Looking at the additional comparable columns, appointments are still more cost-effective than hard commitments.

[39] Again, like in the case of the commitment device, it may be necessary to budget for all participants, which would add the full MK500 to the cost per participant from Column 1, bringing it to 814.07 for appointments and 681.07 or 1681.07 for hard commitments

# Appendix D Appendix Tables

**Appendix Table D1**

Pre-Specified Control Variables and Definitions

| **Variable** | **Definition** |
|---|---|
| Age | Age in years |
| Literate | Equal to 1 if the participant can read a letter in Chichewa and 0 otherwise |
| Married | Equal to 1 if the participant is married and 0 otherwise |
| Index of Willingness to Get an HIV Test | First principal component of two indicators: the first capturing willingness to get tested immediately (mean = 0.87), and the second capturing plans to get an HIV test in the future (mean = 0.94). |
| Any Previous HIV Test | Equal to 1 if the participant ever tested for HIV and 0 otherwise |
| Number of Previous HIV Tests | Number of times the participant tested since born |
| Tendency to Postpone HIV Testing | Indicator variable capturing any positive response to the open-ended question "Are you someone that has been postponing HIV testing?" (manually coded) |
| Perceived Life Expect. Gain from ARVs | Perceived difference between the life expectancy of an HIV+ individual with and without ART |
| Perceived Prob. of being HIV+ | Respondent's perceived likelihood of being HIV+ |
| Demanded Hard Commitment | Equal to 1 if the participant opted into the Hard Commitment and 0 otherwise |

*Notes:* Pre-specified list of control variables from our analysis plan. These controls were selected as statistically significant predictors of past HIV testing choices based on the baseline data from the study. We made this selection before the outcome data for the study was available.

## Appendix Table D2

Other Control Variables and Definitions

| **Variable** | **Definition** |
|---|---|
| Ethnicity | Categorical variable identifying the five largest ethnic groups in our sample (Nyanja, Lomwe, Yao, Chewa, Ngoni) plus "other" |
| Years of Education | Number of completed years of schooling. Ranges from 0 to 13; values of 14 or higher are top-coded at 13 |
| Number of Children | Respondent's total number of children. |
| Religion | Categorical variable identifying the respondent as being Muslim, Christian, or other. |
| Asset Index | First principal component of a set of 14 indicators for owning each of the following: landline, smartphone, sofa, table, chair, lantern, television, radio, bednet, bicycle, motorcycle, Ox-cart, mattress, bed |
| Has a Girlfriend | Equal to 1 if the respondent has a girlfriend and 0 otherwise |
| Any Sex in Past Week | Equal to 1 if the respondent had sex in the past 7 days and 0 otherwise |
| Number of Sex Partners in Past 12 Mo. | Number of sexual partners in the last 12 months. |
| Lives for Today | Measures the extent to which the participant agrees with the statement "I live for today and do not think about tomorrow" on a scale from 1 ("completely disagree") to 5 ("completely agree") |
| Occupation | Categorical variable including the following occupations: military/police/security, skilled activity, transport sector, manual activity, none/student, other (e.g. trade, agriculture) |
| Ln(Expenditures) in Past Month | Sum of the seven categories of household expenditures (food, health, schooling, business, transport, entertainment, accomodation, other) in the past month, transformed via the IHST |
| Ln(Earnings) in Past 7 Days | Respondent's total earnings in the past 7 days, transformed via the IHST |
| Zero Earnings in Past 7 Days | Equal to 1 if respondent has zero earnings in the past 7 days, 0 otherwise |
| HIV Knowledge Index | First principal component of five questions assessing the respondent's knowledge about HIV |
| Perceived Prob. of Serodiscordance | Answer to the question: "Think of 100 women who are HIV positive. How many of their husbands do you think are HIV negative?" |

*Notes:* Other variables used in our balance table and in the double LASSO procedure. This list includes all other baseline variables that were not included in the pre-specified list in Table D1

## Appendix Table D3

Balance

| | Panel A: Appointments | | | | | Panel B: Hard Commitment | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ctrl. Mean (SD) | Treat. Mean (SD) | Diff. ($p$-val.) | Norm. Diff. | Omnibus Test | Ctrl. Mean (SD) | Treat. Mean (SD) | Diff. ($p$-val.) | Norm. Diff. | Omnibus Test |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| Age | 32.250 | 31.638 | -0.612 | 0.065 | -0.004* | 32.100 | 31.761 | -0.339 | 0.036 | -0.000 |
| | (9.549) | (9.321) | (0.256) | | (0.003) | (9.683) | (9.169) | (0.528) | | (0.003) |
| Literate | 0.978 | 0.976 | -0.002 | 0.012 | -0.045 | 0.984 | 0.970 | -0.014 | 0.094 | -0.124 |
| | (0.146) | (0.152) | (0.835) | | (0.108) | (0.125) | (0.170) | (0.101) | | (0.105) |
| Married | 0.644 | 0.671 | 0.027 | -0.057 | 0.063 | 0.646 | 0.671 | 0.025 | -0.053 | 0.114** |
| | (0.479) | (0.470) | (0.317) | | (0.045) | (0.479) | (0.470) | (0.354) | | (0.046) |
| Index of Willingness | 0.049 | -0.046 | -0.095 | 0.086 | -0.013 | -0.044 | 0.046 | 0.090 | -0.081 | 0.018 |
| to Get an HIV Test | (1.047) | (1.161) | (0.132) | | (0.014) | (1.136) | (1.077) | (0.155) | | (0.014) |
| Any Previous HIV | 0.918 | 0.914 | -0.004 | 0.015 | -0.031 | 0.908 | 0.924 | 0.016 | -0.056 | 0.035 |
| Test | (0.275) | (0.281) | (0.788) | | (0.056) | (0.289) | (0.266) | (0.323) | | (0.059) |
| Number of Previous | 4.850 | 4.990 | 0.140 | -0.025 | 0.003 | 4.975 | 4.867 | -0.107 | 0.019 | -0.001 |
| HIV Tests | (5.439) | (5.606) | (0.657) | | (0.003) | (5.466) | (5.588) | (0.733) | | (0.003) |
| Tendency to Postpone | 0.362 | 0.404 | 0.042 | -0.086 | 0.063* | 0.394 | 0.374 | -0.020 | 0.041 | -0.028 |
| HIV Testing | (0.481) | (0.491) | (0.133) | | (0.034) | (0.489) | (0.484) | (0.473) | | (0.035) |
| Perceived Life Exp. | 23.069 | 24.690 | 1.621* | -0.095 | 0.001 | 24.010 | 23.797 | -0.212 | 0.012 | -0.000 |
| Gain from ARVs | (17.002) | (17.250) | (0.097) | | (0.001) | (17.192) | (17.104) | (0.828) | | (0.001) |
| Perceived Prob. of | 17.819 | 17.799 | -0.020 | 0.001 | 0.000 | 17.190 | 18.455 | 1.265 | -0.054 | 0.000 |
| being HIV+ | (22.868) | (23.880) | (0.988) | | (0.001) | (22.944) | (23.843) | (0.343) | | (0.001) |
| Demanded Hard | 0.532 | 0.484 | -0.048* | 0.095 | -0.067** | 0.527 | 0.487 | -0.040 | 0.081 | -0.060* |
| Commitment | (0.499) | (0.500) | (0.095) | | (0.032) | (0.500) | (0.500) | (0.158) | | (0.033) |
| Ethnicity: | | | | | | | | | | |
| Nyanja | 0.077 | 0.080 | 0.003 | -0.011 | -0.017 | 0.070 | 0.088 | 0.018 | -0.068 | 0.030 |
| | (0.267) | (0.272) | (0.845) | | (0.066) | (0.255) | (0.284) | (0.237) | | (0.066) |
| Lomwe | 0.300 | 0.302 | 0.002 | -0.003 | -0.011 | 0.302 | 0.301 | -0.001 | 0.002 | -0.022 |
| | (0.459) | (0.459) | (0.953) | | (0.047) | (0.459) | (0.459) | (0.972) | | (0.048) |
| Yao | 0.211 | 0.214 | 0.002 | -0.006 | -0.024 | 0.221 | 0.204 | -0.016 | 0.040 | -0.037 |
| | (0.409) | (0.410) | (0.917) | | (0.055) | (0.415) | (0.404) | (0.484) | | (0.056) |
| Chewa | 0.138 | 0.129 | -0.009 | 0.025 | -0.049 | 0.144 | 0.121 | -0.023 | 0.068 | -0.105* |
| | (0.345) | (0.335) | (0.656) | | (0.057) | (0.352) | (0.327) | (0.231) | | (0.058) |
| Ngoni | 0.138 | 0.115 | -0.023 | 0.069 | -0.046 | 0.125 | 0.126 | 0.001 | -0.003 | -0.039 |
| | (0.345) | (0.319) | (0.229) | | (0.057) | (0.331) | (0.332) | (0.964) | | (0.057) |
| Years of Education | 10.554 | 10.616 | 0.063 | -0.023 | 0.006 | 10.635 | 10.535 | -0.100 | 0.037 | -0.006 |
| | (2.784) | (2.638) | (0.686) | | (0.007) | (2.684) | (2.736) | (0.517) | | (0.007) |
| Number of Children | 1.911 | 1.918 | 0.007 | -0.004 | 0.010 | 1.933 | 1.895 | -0.038 | 0.019 | -0.004 |
| | (1.942) | (2.012) | (0.949) | | (0.012) | (2.068) | (1.879) | (0.736) | | (0.011) |

## Appendix Table D3

Balance (continued)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Religion: | | | | | | | | | | |
| Muslim | 0.161 | 0.165 | 0.004 | -0.011 | 0.010 | 0.175 | 0.151 | -0.023 | 0.063 | -0.056 |
| | (0.368) | (0.372) | (0.849) | | (0.049) | (0.380) | (0.359) | (0.265) | | (0.048) |
| Other | 0.013 | 0.008 | -0.006 | 0.054 | -0.146 | 0.011 | 0.010 | -0.001 | 0.011 | -0.062 |
| | (0.115) | (0.088) | (0.344) | | (0.142) | (0.105) | (0.099) | (0.844) | | (0.160) |
| Asset Index | 0.008 | -0.007 | -0.015 | 0.008 | -0.005 | -0.008 | 0.009 | 0.017 | -0.009 | 0.002 |
| | (1.846) | (1.915) | (0.891) | | (0.010) | (1.888) | (1.876) | (0.873) | | (0.010) |
| Has a Girlfriend | 0.430 | 0.425 | -0.005 | 0.010 | 0.002 | 0.411 | 0.444 | 0.032 | -0.065 | 0.072** |
| | (0.495) | (0.495) | (0.859) | | (0.035) | (0.492) | (0.497) | (0.251) | | (0.035) |
| Any Sex in Past Week | 0.547 | 0.582 | 0.035 | -0.070 | 0.030 | 0.568 | 0.561 | -0.007 | 0.014 | -0.044 |
| | (0.498) | (0.494) | (0.219) | | (0.035) | (0.496) | (0.497) | (0.810) | | (0.035) |
| Sex Partners in Past | 2.264 | 2.348 | 0.084 | -0.025 | 0.003 | 2.279 | 2.337 | 0.058 | -0.018 | -0.000 |
| 12 Months | (3.065) | (3.506) | (0.654) | | (0.005) | (3.478) | (3.103) | (0.758) | | (0.005) |
| Lives for Today | 2.710 | 2.895 | 0.185** | -0.113 | 0.026** | 2.722 | 2.892 | 0.170* | -0.104 | 0.027** |
| | (1.633) | (1.634) | (0.047) | | (0.012) | (1.633) | (1.636) | (0.069) | | (0.012) |
| Occupation: | | | | | | | | | | |
| Military/Police/ | 0.106 | 0.101 | -0.005 | 0.017 | 0.060 | 0.090 | 0.116 | 0.026 | -0.085 | 0.077 |
| Security | (0.308) | (0.301) | (0.770) | | (0.059) | (0.287) | (0.321) | (0.138) | | (0.059) |
| Skilled Activity | 0.128 | 0.143 | 0.016 | -0.045 | 0.077 | 0.141 | 0.130 | -0.012 | 0.034 | -0.036 |
| | (0.334) | (0.350) | (0.425) | | (0.050) | (0.349) | (0.336) | (0.549) | | (0.051) |
| Transportation | 0.107 | 0.116 | 0.009 | -0.028 | 0.027 | 0.114 | 0.110 | -0.005 | 0.015 | -0.042 |
| | (0.310) | (0.321) | (0.618) | | (0.053) | (0.318) | (0.313) | (0.796) | | (0.052) |
| Manual Labor | 0.128 | 0.138 | 0.011 | -0.032 | 0.047 | 0.149 | 0.116 | -0.033* | 0.097 | -0.074 |
| | (0.334) | (0.346) | (0.575) | | (0.049) | (0.357) | (0.321) | (0.088) | | (0.047) |
| None/Student | 0.104 | 0.113 | 0.009 | -0.029 | 0.045 | 0.098 | 0.120 | 0.021 | -0.068 | 0.046 |
| | (0.306) | (0.317) | (0.605) | | (0.057) | (0.298) | (0.325) | (0.234) | | (0.058) |
| Ln(Expenditures) in | 11.860 | 11.829 | -0.032 | 0.021 | -0.002 | 11.887 | 11.799 | -0.088 | 0.059 | -0.018 |
| Past Month | (1.528) | (1.437) | (0.708) | | (0.013) | (1.429) | (1.534) | (0.298) | | (0.013) |
| Ln(Earnings) in Past | 9.412 | 9.570 | 0.159 | -0.048 | -0.015 | 9.514 | 9.472 | -0.043 | 0.013 | -0.005 |
| 7 Days | (3.475) | (3.161) | (0.403) | | (0.013) | (3.340) | (3.293) | (0.822) | | (0.013) |
| Zero Earnings in Past | 0.104 | 0.080 | -0.024 | 0.083 | -0.235 | 0.092 | 0.091 | -0.001 | 0.002 | -0.047 |
| 7 Days | (0.306) | (0.272) | (0.149) | | (0.148) | (0.289) | (0.288) | (0.966) | | (0.146) |
| | -0.032 | 0.030 | 0.062 | -0.053 | 0.010 | -0.002 | 0.002 | 0.003 | -0.003 | -0.001 |
| HIV Knowledge Index | (1.197) | (1.115) | (0.350) | | (0.013) | (1.131) | (1.182) | (0.958) | | (0.013) |
| Perceived Prob. of | 41.322 | 42.695 | 1.373 | -0.044 | 0.000 | 41.502 | 42.585 | 1.083 | -0.035 | 0.000 |
| Serodiscordance | (31.132) | (31.630) | (0.443) | | (0.001) | (31.895) | (30.858) | (0.545) | | (0.001) |
| Omnibus Test: F-Statistic | | | | | 1.16 | | | | | 1.08 |
| ($p$-value) | | | | | (0.16) | | | | | (0.31) |

*Notes:* Sample is 1,232 men who completed a baseline survey. In Panel A, we compare the characteristics of all participants who were offered an appointment (either alone or in combination with a hard commitment) versus participants who were not (i.e., the control group and the hard commitment only arm). In Panel B, we compare the characteristics of all participants who were offered a hard commitment device (either alone or in combination with an appointment) versus participants who were not (i.e., the control group and the appointment-only arm). Columns 3 and 8 present treatment-control differences, with $p$-values from $t$-tests in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Columns 4 and 9 show differences in standard deviations. Columns 5 and 10 present OLS regressions of the treatment indicators on all the baseline covariates in the table; the bottom row of the table shows $F$-statistics (and associated $p$-values) for the joint significance of all the variables in the regression.

## Appendix Table D4

Balance by Individual Study Arm

| | Mean (SD) Ctrl. | HC Only | A Only | Both A & HC | Normalized Difference (*p*-value) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (1)-(2) | (1)-(3) | (1)-(4) | (2)-(3) | (2)-(4) | (3)-(4) |
| Age | 32.711 | 31.780 | 31.541 | 31.743 | 0.098 | 0.121 | 0.100 | 0.026 | 0.004 | -0.022 |
| | (10.060) | (8.990) | (9.304) | (9.354) | (0.234) | (0.13) | (0.219) | (0.745) | (0.961) | (0.785) |
| Literate | 0.983 | 0.973 | 0.985 | 0.967 | 0.072 | -0.011 | 0.103 | -0.083 | 0.032 | 0.114 |
| | (0.128) | (0.163) | (0.123) | (0.178) | (0.381) | (0.887) | (0.205) | (0.299) | (0.695) | (0.15) |
| Married | 0.628 | 0.661 | 0.663 | 0.681 | -0.069 | -0.072 | -0.111 | -0.003 | -0.042 | -0.039 |
| | (0.484) | (0.474) | (0.474) | (0.467) | (0.399) | (0.364) | (0.171) | (0.966) | (0.607) | (0.627) |
| Index of Willingness | 0.042 | 0.056 | -0.122 | 0.036 | -0.013 | 0.146* | 0.006 | 0.153* | 0.019 | -0.137* |
| to Get an HIV Test | (1.009) | (1.086) | (1.236) | (1.071) | (0.87) | (0.07) | (0.944) | (0.057) | (0.82) | (0.086) |
| Any Previous HIV | 0.910 | 0.925 | 0.906 | 0.922 | -0.055 | 0.016 | -0.042 | 0.071 | 0.014 | -0.057 |
| | (0.286) | (0.263) | (0.293) | (0.269) | (0.502) | (0.845) | (0.609) | (0.38) | (0.868) | (0.473) |
| Number of Previous | 4.791 | 4.910 | 5.143 | 4.826 | -0.022 | -0.065 | -0.007 | -0.041 | 0.015 | 0.057 |
| HIV Tests | (5.203) | (5.678) | (5.699) | (5.509) | (0.789) | (0.42) | (0.936) | (0.61) | (0.853) | (0.476) |
| Tendency to | 0.375 | 0.349 | 0.410 | 0.397 | 0.055 | -0.071 | -0.045 | -0.126 | -0.100 | 0.026 |
| Postpone HIV | (0.485) | (0.478) | (0.493) | (0.490) | (0.506) | (0.371) | (0.579) | (0.117) | (0.222) | (0.74) |
| Perceived Life Exp. | 23.266 | 22.868 | 24.690 | 24.691 | 0.023 | -0.083 | -0.082 | -0.108 | -0.107 | -0.000 |
| Gain from ARVs | (17.369) | (16.647) | (17.027) | (17.513) | (0.775) | (0.299) | (0.314) | (0.178) | (0.191) | (1) |
| Perceived Prob. of | 16.047 | 19.627 | 18.237 | 17.329 | -0.157* | -0.096 | -0.056 | 0.059 | 0.096 | 0.038 |
| being HIV+ | (21.984) | (23.637) | (23.773) | (24.024) | (0.056) | (0.232) | (0.493) | (0.465) | (0.237) | (0.632) |
| Demanded | 0.542 | 0.522 | 0.514 | 0.453 | 0.039 | 0.056 | 0.178** | 0.017 | 0.139* | 0.122 |
| | (0.499) | (0.500) | (0.501) | (0.499) | (0.634) | (0.485) | (0.029) | (0.835) | (0.089) | (0.125) |
| Ethnicity: | | | | | | | | | | |
| Nyanja | 0.060 | 0.095 | 0.079 | 0.081 | -0.132 | -0.076 | -0.084 | 0.056 | 0.047 | -0.009 |
| | (0.238) | (0.294) | (0.270) | (0.274) | (0.109) | (0.345) | (0.299) | (0.482) | (0.56) | (0.911) |
| Lomwe | 0.292 | 0.308 | 0.310 | 0.293 | -0.035 | -0.038 | -0.002 | -0.003 | 0.033 | 0.037 |
| | (0.456) | (0.463) | (0.463) | (0.456) | (0.668) | (0.63) | (0.983) | (0.967) | (0.683) | (0.644) |
| Yao | 0.226 | 0.197 | 0.216 | 0.212 | 0.072 | 0.024 | 0.034 | -0.047 | -0.037 | 0.010 |
| | (0.419) | (0.398) | (0.412) | (0.409) | (0.382) | (0.76) | (0.673) | (0.555) | (0.646) | (0.9) |
| Chewa | 0.159 | 0.115 | 0.131 | 0.127 | 0.128 | 0.082 | 0.093 | -0.047 | -0.036 | 0.011 |
| | (0.367) | (0.320) | (0.338) | (0.334) | (0.118) | (0.306) | (0.254) | (0.559) | (0.659) | (0.891) |
| Ngoni | 0.140 | 0.136 | 0.112 | 0.117 | 0.011 | 0.082 | 0.067 | 0.070 | 0.055 | -0.015 |
| | (0.347) | (0.343) | (0.316) | (0.322) | (0.889) | (0.306) | (0.412) | (0.381) | (0.499) | (0.85) |
| Years of Education | 10.585 | 10.522 | 10.681 | 10.547 | 0.022 | -0.036 | 0.014 | -0.058 | -0.009 | 0.051 |
| | (2.757) | (2.816) | (2.619) | (2.662) | (0.784) | (0.654) | (0.865) | (0.466) | (0.91) | (0.524) |
| Number of Children | 1.997 | 1.824 | 1.875 | 1.964 | 0.089 | 0.059 | 0.016 | -0.026 | -0.075 | -0.044 |
| | (2.031) | (1.847) | (2.104) | (1.910) | (0.277) | (0.463) | (0.839) | (0.746) | (0.36) | (0.578) |

## Appendix Table D4

Balance by Individual Study Arm (continued)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Religion: | | | | | | | | | | |
| Muslim | 0.163 | 0.159 | 0.185 | 0.143 | 0.009 | -0.060 | 0.054 | -0.069 | 0.045 | 0.114 |
| | (0.370) | (0.367) | (0.389) | (0.351) | (0.908) | (0.456) | (0.506) | (0.391) | (0.585) | (0.154) |
| Other | 0.013 | 0.014 | 0.009 | 0.007 | -0.002 | 0.040 | 0.068 | 0.042 | 0.071 | 0.030 |
| | (0.115) | (0.116) | (0.095) | (0.081) | (0.977) | (0.619) | (0.399) | (0.6) | (0.385) | (0.711) |
| Asset Index | 0.082 | -0.069 | -0.091 | 0.083 | 0.082 | 0.092 | -0.000 | 0.012 | -0.081 | -0.091 |
| | (1.816) | (1.877) | (1.951) | (1.875) | (0.318) | (0.249) | (0.995) | (0.883) | (0.321) | (0.251) |
| Has a Girlfriend | 0.439 | 0.420 | 0.386 | 0.466 | 0.037 | 0.107 | -0.055 | 0.070 | -0.091 | -0.162** |
| | (0.497) | (0.494) | (0.488) | (0.500) | (0.654) | (0.181) | (0.5) | (0.384) | (0.262) | (0.042) |
| Any Sex in Past Week | 0.548 | 0.546 | 0.587 | 0.577 | 0.005 | -0.078 | -0.057 | -0.082 | -0.062 | 0.020 |
| | (0.499) | (0.499) | (0.493) | (0.495) | (0.953) | (0.331) | (0.482) | (0.304) | (0.448) | (0.797) |
| Sex Partners in Past | 2.078 | 2.454 | 2.463 | 2.225 | -0.123 | -0.112 | -0.053 | -0.002 | 0.074 | 0.068 |
| 12 Months | (2.738) | (3.360) | (4.033) | (2.835) | (0.134) | (0.165) | (0.517) | (0.975) | (0.365) | (0.391) |
| Lives for Today | 2.535 | 2.888 | 2.894 | 2.896 | -0.217*** | -0.221*** | -0.221*** | -0.003 | -0.005 | -0.001 |
| | (1.624) | (1.626) | (1.624) | (1.648) | (0.008) | (0.006) | (0.007) | (0.966) | (0.954) | (0.987) |
| Occupation: | | | | | | | | | | |
| Military/Police/ | 0.090 | 0.122 | 0.091 | 0.111 | -0.105 | -0.005 | -0.070 | 0.100 | 0.035 | -0.065 |
| | (0.286) | (0.328) | (0.288) | (0.314) | (0.2) | (0.948) | (0.389) | (0.212) | (0.667) | (0.413) |
| Skilled Activity | 0.130 | 0.125 | 0.152 | 0.134 | 0.012 | -0.064 | -0.012 | -0.077 | -0.024 | 0.053 |
| | (0.336) | (0.332) | (0.360) | (0.341) | (0.88) | (0.421) | (0.885) | (0.34) | (0.767) | (0.508) |
| Transportation | 0.103 | 0.112 | 0.125 | 0.107 | -0.029 | -0.068 | -0.015 | -0.039 | 0.014 | 0.053 |
| | (0.304) | (0.316) | (0.331) | (0.310) | (0.727) | (0.395) | (0.857) | (0.623) | (0.864) | (0.502) |
| Manual Labor | 0.133 | 0.122 | 0.164 | 0.111 | 0.033 | -0.088 | 0.068 | -0.120 | 0.035 | 0.155* |
| | (0.340) | (0.328) | (0.371) | (0.314) | (0.692) | (0.272) | (0.405) | (0.135) | (0.667) | (0.051) |
| None/Student | 0.090 | 0.119 | 0.106 | 0.121 | -0.095 | -0.056 | -0.100 | 0.039 | -0.006 | -0.045 |
| | (0.286) | (0.324) | (0.309) | (0.326) | (0.248) | (0.483) | (0.216) | (0.629) | (0.944) | (0.575) |
| Ln(Expenditures) in | 11.842 | 11.879 | 11.928 | 11.722 | -0.024 | -0.060 | 0.076 | -0.035 | 0.102 | 0.143* |
| Past Month | (1.570) | (1.486) | (1.287) | (1.578) | (0.77) | (0.452) | (0.347) | (0.658) | (0.21) | (0.071) |
| Ln(Earnings) in Past | 9.517 | 9.303 | 9.511 | 9.633 | 0.062 | 0.002 | -0.035 | -0.062 | -0.100 | -0.039 |
| | (3.447) | (3.505) | (3.245) | (3.073) | (0.453) | (0.982) | (0.662) | (0.442) | (0.22) | (0.627) |
| Zero Earnings in Past | 0.100 | 0.108 | 0.085 | 0.075 | -0.029 | 0.050 | 0.088 | 0.079 | 0.116 | 0.037 |
| | (0.300) | (0.312) | (0.279) | (0.264) | (0.725) | (0.528) | (0.28) | (0.324) | (0.154) | (0.637) |
| HIV Knowledge Index | -0.075 | 0.012 | 0.066 | -0.008 | -0.073 | -0.124 | -0.056 | -0.047 | 0.017 | 0.066 |
| | (1.206) | (1.188) | (1.054) | (1.177) | (0.373) | (0.119) | (0.489) | (0.553) | (0.831) | (0.404) |
| Perceived Prob. of | 38.342 | 44.363 | 44.392 | 40.876 | -0.194** | -0.191** | -0.082 | -0.001 | 0.113 | 0.111 |
| Serodiscordance | (30.978) | (31.046) | (32.489) | (30.630) | (0.018) | (0.017) | (0.311) | (0.991) | (0.166) | (0.161) |
| N | 301 | 295 | 329 | 307 | 596 | 630 | 608 | 624 | 602 | 636 |

*Notes:* Sample is 1,232 men who completed a baseline survey. Columns 1-4 show the means and standard deviations of each variable within each study arm. Columns 5-8 present normalized treatment-control differences, with $p$-values from $t$-tests in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

**Appendix Table D5**

Intention-to-Treat Effects of Appointments and Commitment Devices—Short Specification

| | *Outcome: 1(Tested for HIV)* | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Appointment Treatment (A) | 0.112*** | 0.127*** | 0.124*** |
| | (0.023) | (0.022) | (0.022) |
| | [0.001] | [0.001] | [0.001] |
| Hard Commitment Treatment (HC) | 0.039* | 0.048** | 0.050** |
| | (0.023) | (0.022) | (0.022) |
| | [0.071] | [0.032] | [0.023] |
| **Comparisons of Effects** | | | |
| A − HC | 0.073** | 0.079** | 0.074** |
| SE for $t$-test | (0.034) | (0.032) | (0.032) |
| $q$-value $t$-test | [0.031] | [0.019] | [0.022] |
| Controls | None | Pre-Specified | Double Lasso |
| Observations | 1232 | 1232 | 1232 |
| Adjusted R-squared | 0.019 | 0.123 | 0.120 |
| Control-group Mean | 0.113 | 0.113 | 0.113 |

*Notes:* Sample is 1,232 men who completed a baseline survey. Pre-specified controls include all the variables in Table D1. Double Lasso controls uses the Chernozhukov et al. (2017) method for variable selection and inference, as described in Section 3.1. Columns 2 and 3 also control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

**Appendix Table D6**

Treatment Effect Heterogeneity by Demand for Commitment—Short Specification

| | *Outcome: 1(Tested for HIV)* | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| **Demanded the Hard Commitment (D)** | | | |
| D | 0.162*** | 0.165*** | 0.163*** |
| | (0.026) | (0.026) | (0.026) |
| | [0.001] | [0.001] | [0.001] |
| D × Appointment Treatment (A) | 0.155*** | 0.161*** | 0.156*** |
| | (0.035) | (0.034) | (0.033) |
| | [0.001] | [0.001] | [0.001] |
| D × Hard Commitment Treatment (HC) | 0.082** | 0.072** | 0.075** |
| | (0.035) | (0.034) | (0.034) |
| | [0.024] | [0.037] | [0.028] |
| **Did Not Demand the Hard Commitment (1−D)** | | | |
| (1−D) | 0.097*** | 0.091*** | 0.097*** |
| | (0.018) | (0.020) | (0.019) |
| | [0.001] | [0.001] | [0.001] |
| (1−D) x Appointment Treatment (A) | 0.083*** | 0.092*** | 0.088*** |
| | (0.028) | (0.028) | (0.027) |
| | [0.005] | [0.002] | [0.003] |
| **Comparisons of Effects** | | | |
| D × A − (1−D) × A | 0.072 | 0.069 | 0.068 |
| SE for $t$-test | (0.045) | (0.044) | (0.043) |
| $q$-value $t$-test | [0.081] | [0.084] | [0.083] |
| Controls | None | Pre-Specified | Double Lasso |
| Study Arms Included | All | All | All |
| Observations | 1232 | 1232 | 1232 |
| Adjusted $R$-squared | 0.251 | 0.307 | 0.303 |
| Control-group Mean | 0.113 | 0.113 | 0.113 |

*Notes:* Sample is 1,232 men who completed a baseline survey. These regressions do not include a constant; (HC) x (1-D) is omitted and assumed to be zero. Pre-specified controls include all the variables in Table D1. Double Lasso controls uses the Chernozhukov et al. (2017) method for variable selection and inference, as described in Section 3.1. Columns 2 and 3 also control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. All controls are standardized prior to running the regressions. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

**Appendix Table D7**

Treatment Effect Heterogeneity by Baseline Covariates

| | *Outcome: 1(Tested for HIV)* | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Appointment (A) | 0.159*** | 0.160*** | 0.160*** | 0.158*** | 0.155*** |
| | (0.030) | (0.030) | (0.030) | (0.030) | (0.030) |
| | [0.001] | [0.001] | [0.001] | [0.001] | [0.001] |
| Hard Commitment (HC) | 0.081*** | 0.082*** | 0.085*** | 0.082*** | 0.074*** |
| | (0.029) | (0.029) | (0.029) | (0.029) | (0.029) |
| | [0.009] | [0.008] | [0.007] | [0.008] | [0.013] |
| A × (Perceived Prob. of being HIV+) | -0.007 | | | | -0.010 |
| | (0.031) | | | | (0.030) |
| | [0.367] | | | | [0.332] |
| HC × (Perceived Prob. of being HIV+) | 0.020 | | | | 0.028 |
| | (0.030) | | | | (0.030) |
| | [0.261] | | | | [0.202] |
| A × (Tendency to Postpone HIV Testing) | | -0.008 | | | 0.000 |
| | | (0.029) | | | (0.030) |
| | | [0.359] | | | [0.429] |
| HC × (Tendency to Postpone HIV Testing) | | 0.004 | | | -0.007 |
| | | (0.029) | | | (0.030) |
| | | [0.400] | | | [0.367] |
| A × (Lives for Today) | | | 0.005 | | 0.002 |
| | | | (0.030) | | (0.032) |
| | | | [0.400] | | [0.420] |
| HC × (Lives for Today) | | | -0.022 | | -0.031 |
| | | | (0.029) | | (0.031) |
| | | | [0.234] | | [0.181] |
| A × (Perceived Life Expect. Gain from ARVs) | | | | 0.007 | 0.013 |
| | | | | (0.031) | (0.033) |
| | | | | [0.367] | [0.320] |
| HC × (Perceived Life Expect. Gain from ARVs) | | | | 0.007 | 0.015 |
| | | | | (0.030) | (0.031) |
| | | | | [0.367] | [0.295] |
| Controls | Pre-Specified | Pre-Specified | Pre-Specified | Pre-Specified | Pre-Specified |
| Treatments Interacted with Other BL Covariates | No | No | No | No | Yes |
| Observations | 1232 | 1232 | 1232 | 1232 | 1232 |
| Adjusted R-squared | 0.122 | 0.122 | 0.123 | 0.123 | 0.127 |
| Control-group Mean | 0.113 | 0.113 | 0.113 | 0.113 | 0.113 |

*Notes:* Sample is 1,232 men who completed a baseline survey. Pre-specified controls include all the variables in Table D1. Other baseline covariates interacted with treatments in Column 5 include all the other variables from Table D1. Main effects are included for all variables that are interacted with the treatment indicators. We also control for $A \times HC$ and its interactions with all the baseline variables in the table, but do not show these results. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

## Appendix Table D8

Treatment Effect Heterogeneity by Predicted Probability of Getting Tested

| | *Outcome: 1(Tested for HIV)* | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Appointment × | | | | |
| Low Likelihood of Testing (L) | 0.064 | 0.113* | | |
| | (0.042) | (0.060) | | |
| | [0.090] | [0.052] | | |
| Medium Likelihood of Testing (M) | 0.146*** | 0.087 | | |
| | (0.036) | (0.069) | | |
| | [0.001] | [0.129] | | |
| High Likelihood of Testing (H) | 0.236*** | 0.162** | | |
| | (0.051) | (0.065) | | |
| | [0.001] | [0.016] | | |
| Hard Commitment × | | | | |
| Low Likelihood of Testing | | | 0.012 | 0.005 |
| | | | (0.039) | (0.043) |
| | | | [0.343] | [0.404] |
| Medium Likelihood of Testing | | | 0.099** | 0.088 |
| | | | (0.042) | (0.070) |
| | | | [0.021] | [0.129] |
| High Likelihood of Testing | | | 0.102** | 0.199*** |
| | | | (0.051) | (0.073) |
| | | | [0.043] | [0.010] |
| **Comparisons of Effects** | | | | |
| H − L | 0.172*** | 0.049 | 0.090 | 0.194** |
| SE for $t$-test | (0.066) | (0.088) | (0.065) | (0.085) |
| $q$-value $t$-test | [0.013] | [0.281] | [0.111] | [0.025] |
| H − M | 0.091 | 0.075 | 0.003 | 0.111 |
| SE for $t$-test | (0.063) | (0.095) | (0.066) | (0.102) |
| $q$-value $t$-test | [0.101] | [0.228] | [0.420] | [0.162] |
| M − L | 0.081 | -0.026 | 0.087 | 0.083 |
| SE for $t$-test | (0.055) | (0.092) | (0.057) | (0.082) |
| $q$-value $t$-test | [0.098] | [0.354] | [0.091] | [0.180] |
| Predictors | Pre-Specified | Double Lasso | Pre-Specified | Double Lasso |
| Controls | All FEs | All FEs | All FEs | All FEs |
| Study Arms Included | C A | C A | C HC | C HC |
| Control-group Mean | 0.113 | 0.113 | 0.113 | 0.113 |

*Notes:* Sample is 1,232 men who completed a baseline survey; Columns 1 and 2 use only data from the control (C) and appointments-only (A) arms, while Columns 3 and 4 use only data from the control and commitment devices-only (HC) arms. This table uses the repeat split-sample method of Abadie, Chingos, and West (2018) to predict the outcome variable in the control group, then estimates treatment effects separately by tercile of the predicted outcome. Pre-specified predictors include all the variables in Table D1. Double Lasso predictors include the same variables chosen using the Chernozhukov et al. (2017) in Table 2. All columns control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets. $p$-values for the tests of equality between terciles ignore the covariance term and thus are likely to be conservative.

**Appendix Table D9**

Randomization Inference

| | *Outcome: 1(Tested for HIV)* | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Appointment Treatment (A) | 0.151*** | 0.159*** | 0.154*** |
| EHW SE | (0.030) | (0.030) | (0.029) |
| EHW $p$-value | [0.000] | [0.000] | [0.000] |
| RI $p$-value | {0.000} | {0.000} | {0.000} |
| Hard Commitment Treatment (HC) | 0.080*** | 0.082*** | 0.085*** |
| EHW SE | (0.029) | (0.029) | (0.029) |
| EHW $p$-value | [0.006] | [0.005] | [0.003] |
| RI $p$-value | {0.020} | {0.020} | {0.020} |
| Both A and HC | 0.151*** | 0.175*** | 0.175*** |
| EHW SE | (0.031) | (0.031) | (0.030) |
| EHW $p$-value | [0.000] | [0.000] | [0.000] |
| RI $p$-value | {0.000} | {0.000} | {0.000} |
| *p-value A = HC* | | | |
| EHW $p$-value | [0.034] | [0.017] | [0.027] |
| RI $p$-value | {0.033} | {0.012} | {0.012} |
| Controls | None | Pre-Specified | Double Lasso |
| Observations | 1,232 | 1,232 | 1,232 |
| Adjusted R-squared | 0.021 | 0.123 | 0.120 |
| Control-group Mean | 0.113 | 0.113 | 0.113 |

*Notes:* Sample is 1,232 men who completed a baseline survey. This table presents the same results as Columns 4-6 of Table 2, but shows randomization inference (RI) $p$-values for comparison with the $p$-values that correspond to the Eicker–Huber–White (EHW) standard errors that we show in the rest of the tables in the paper. Pre-specified controls include all the variables in Table D1. Double Lasso controls uses the Chernozhukov et al. (2017) method for variable selection and inference, as described in Section 3.1. Columns 2 and 3 both control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

**Appendix Table D10**

Effects on Voucher Redemption

| | *Outcome: 1(Redeemed Voucher)* | | | *Outcome: 1(Redeemed Voucher, No Test)* | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Appointment Only (A) | 0.148*** | 0.168*** | 0.160*** | 0.011 | 0.022 | 0.016 |
| | (0.033) | (0.033) | (0.032) | (0.019) | (0.019) | (0.019) |
| | [0.001] | [0.001] | [0.001] | [0.280] | [0.150] | [0.216] |
| Hard Commitment Only (HC) | 0.125*** | 0.128*** | 0.129*** | 0.038* | 0.040* | 0.039* |
| | (0.034) | (0.033) | (0.032) | (0.021) | (0.021) | (0.021) |
| | [0.001] | [0.001] | [0.001] | [0.058] | [0.053] | [0.054] |
| Both A and HC | 0.179*** | 0.210*** | 0.207*** | 0.038* | 0.045** | 0.039* |
| | (0.034) | (0.033) | (0.033) | (0.021) | (0.021) | (0.021) |
| | [0.001] | [0.001] | [0.001] | [0.057] | [0.035] | [0.052] |
| **Comparisons of Effects** | | | | | | |
| A − HC | 0.022 | 0.040 | 0.031 | -0.028 | -0.017 | -0.023 |
| SE for $t$-test | (0.037) | (0.035) | (0.035) | (0.022) | (0.022) | (0.021) |
| $q$-value $t$-test | [0.273] | [0.158] | [0.208] | [0.125] | [0.224] | [0.167] |
| Controls | None | Pre-Specified | Double Lasso | None | Pre-Specified | Double Lasso |
| Observations | 1232 | 1232 | 1232 | 1232 | 1232 | 1232 |
| Adjusted R-squared | 0.021 | 0.129 | 0.118 | 0.002 | 0.040 | 0.031 |
| Control-group Mean | 0.159 | 0.159 | 0.159 | 0.053 | 0.053 | 0.053 |

*Notes:* Sample is 1,232 men who completed a baseline survey. Pre-specified controls include all the variables in Table D1. Double Lasso controls uses the Chernozhukov et al. (2017) method for variable selection and inference, as described in Section 3.1. Columns 2, 3, 5, and 6 also control both control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

**Appendix Table D11**

Robustness to Including Impostors

| | *Outcome:* | | |
|---|---|---|---|
| | *1(Tested for HIV, set to 1 for impostors)* | | |
| | (1) | (2) | (3) |
| Appointment Only (A) | 0.148*** | 0.157*** | 0.151*** |
| | (0.031) | (0.030) | (0.029) |
| | [0.001] | [0.001] | [0.001] |
| Hard Commitment Only (HC) | 0.080*** | 0.082*** | 0.084*** |
| | (0.030) | (0.029) | (0.029) |
| | [0.010] | [0.008] | [0.006] |
| Both A and HC | 0.154*** | 0.181*** | 0.179*** |
| | (0.031) | (0.031) | (0.030) |
| | [0.001] | [0.001] | [0.001] |
| **Comparisons of Effects** | | | |
| A − HC | 0.068** | 0.075** | 0.067** |
| SE for $t$-test | (0.034) | (0.032) | (0.031) |
| $q$-value $t$-test | [0.042] | [0.023] | [0.034] |
| Controls | None | Pre-Specified | Double Lasso |
| Observations | 1232 | 1232 | 1232 |
| Adjusted R-squared | 0.021 | 0.127 | 0.122 |
| Control-group Mean | 0.116 | 0.116 | 0.116 |

*Notes:* Sample is 1,232 men who completed a baseline survey. This table presents the same results as Table 2, but codes impostors as having a one for the outcome variable instead of a zero. Pre-specified controls include all the variables in Table D1. Double Lasso controls uses the Chernozhukov et al. (2017) method for variable selection and inference, as described in Section 3.1. Columns 2 and 3 both control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

**Appendix Table D12**

Regressions of Revealed Preference for Hard Commitment on Survey Measures of Self-Control Problems

| | *Outcome:* 1(Signed up for hard commitment device) | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Lives for Today | 0.018** | | | 0.020** |
| | (0.009) | | | (0.009) |
| Index of Willingness to Get an HIV Test | | 0.067*** | | 0.066*** |
| | | (0.011) | | (0.011) |
| Tendency to Postpone HIV Testing | | | 0.045 | 0.052* |
| | | | (0.029) | (0.029) |
| Observations | 1232 | 1232 | 1232 | 1232 |
| Adjusted R-squared | 0.003 | 0.021 | 0.001 | 0.025 |
| Control-group Mean | 0.542 | 0.542 | 0.542 | 0.542 |

*Notes:* Sample is 1,232 men who completed a baseline survey. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

# Appendix E  Deviations from the Pre-Analysis Plan

Our analyses deviate from the pre-analysis plan in several ways. First, while we did pre-specify a plan to examine the effect of the interaction term $A_i \times HC_i$ as a secondary analysis, our pre-specified analyses mainly did not include this interaction term. This "short" model has higher power when the interaction term does not matter, but leads to incorrect estimates when it does (Muralidharan, Romero, and Wüthrich 2019). Since our results show that the interaction between the two treatments is important, we focus on the fully-interacted specifications. However, none of our inferences or qualitative conclusions depend on this choice. We present the complete tables using the exact specifications from the pre-analysis plan in Appendix F.

Second, we handle the control variables slightly differently than specified in the pre-analysis plan. Our preferred specification, with our pre-specified list of controls, matches the analysis plan exactly. For the double LASSO, the analysis plan called for the fixed effects to be included in the selection procedure; we instead partial them out in advance. The analysis plan also did not specify the construction of the higher-order and logged terms and interactions. In addition to the pre-specified and double LASSO controls, we also show a specification with no controls, which was not in the analysis plan.

Third, the analysis plan specified that we would conduct the FDR adjustments only when testing related hypotheses, and not across estimands, and that the exploratory analyses would not be included in the adjustments at all. We take a broader approach, conducting the FDR procedure across all $p$-values included in the paper (including in the appendices).

Fourth, we conduct additional exploratory analyses that were not listed in the analysis plan. In particular, our analysis of the timing of test dates in Section 5 was not part of our analysis plan; rather, it is an exploration of a potential mechanism that we first conceived of after we had already seen the main results.

Note that the follow-up survey mentioned at the end of Section 2.1 was not part of our pre-analysis plan. We also did not file a separate analysis plan for this data, intending to treat any analyses of the data as exploratory. As noted in that section, we do not use any of that data in this paper.

# Appendix F Results Using Specification from Analysis Plan

This section presents the results of our analyses using the regression specification that we pre-specified in our analysis plan (Derksen et al. 2019, https://www.socialscienceregistry.org/versions/57507/docs/version/document). This differs from the version we present in the body of the paper in that it imposes the "short" model, which omits the interaction term. We present all the tables that include treatment effect estimates because the $q$-values are computed across all the tables, and thus even tables which show the exact same specification are affected by the change.

**Appendix Table F1**

Effects of Appointments and Commitment Devices on HIV Testing

| | *Outcome: 1(Tested for HIV)* | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Appointment Treatment (A) | 0.112*** | 0.127*** | 0.124*** | 0.151*** | 0.159*** | 0.155*** |
| | (0.023) | (0.022) | (0.022) | (0.030) | (0.030) | (0.029) |
| | [0.001] | [0.001] | [0.001] | [0.001] | [0.001] | [0.001] |
| Hard Commitment Treatment (HC) | 0.039* | 0.048** | 0.050** | 0.080*** | 0.082*** | 0.082*** |
| | (0.023) | (0.022) | (0.022) | (0.029) | (0.029) | (0.028) |
| | [0.066] | [0.035] | [0.027] | [0.012] | [0.010] | [0.009] |
| A $\times$ HC | | | | -0.081* | -0.065 | -0.063 |
| | | | | (0.046) | (0.044) | (0.044) |
| | | | | [0.060] | [0.087] | [0.090] |
| p-value A = HC | 0.029 | 0.015 | 0.019 | 0.034 | 0.017 | 0.021 |
| q-value A = HC | [0.034] | [0.023] | [0.027] | [0.037] | [0.025] | [0.027] |
| p-value A + A $\times$ HC = 0 | | | | 0.039 | 0.005 | 0.004 |
| q-value A + A $\times$ HC = 0 | | | | [0.040] | [0.009] | [0.009] |
| p-value HC + A $\times$ HC = 0 | | | | 0.986 | 0.620 | 0.554 |
| q-value HC + A $\times$ HC = 0 | | | | [0.399] | [0.263] | [0.244] |
| Controls | None | Pre-Specified | Double Lasso | None | Pre-Specified | Double Lasso |
| Observations | 1,232 | 1,232 | 1,232 | 1,232 | 1,232 | 1,232 |
| Adjusted R-squared | 0.019 | 0.123 | 0.120 | 0.021 | 0.123 | 0.120 |
| Control-group Mean | 0.113 | 0.113 | 0.113 | 0.113 | 0.113 | 0.113 |

*Notes:* Sample is 1,232 men who completed a baseline survey. Pre-specified controls include all the variables in Table D1. Double Lasso controls uses the Chernozhukov et al. (2017) method for variable selection and inference, as described in Section 3.1. Columns 2 and 3 both control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

**Appendix Table F2**

2SLS Estimates of the Treatment-on-the-Treated Effect of Appointments

| **Panel A**: Second Stage | *Outcome: 1(Tested for HIV)* | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Signed up for an Appointment | 0.171*** | 0.191*** | 0.188*** |
| | (0.034) | (0.032) | (0.032) |
| | [0.001] | [0.001] | [0.001] |
| Controls | None | Pre-Specified | Double Lasso |
| Observations | 1232 | 1232 | 1232 |
| Adjusted R-squared | 0.068 | 0.156 | 0.152 |
| Control-group Mean | 0.113 | 0.113 | 0.113 |
| **Panel B**: First Stage | *Outcome: 1(Signed up for an Appointment)* | | |
| Appointment Treatment (A) | 0.654*** | 0.663*** | 0.662*** |
| | (0.019) | (0.018) | (0.018) |
| | [0.001] | [0.001] | [0.001] |
| Effective $F$-Statistic | 1198.8 | 1344.9 | 1377.8 |
| Controls | None | Pre-Specified | Double Lasso |
| Observations | 1,232 | 1,232 | 1,232 |
| Adjusted R-squared | 0.477 | 0.545 | 0.550 |
| Control-group Mean | 0.000 | 0.000 | 0.000 |

*Notes:* Sample is 1,232 men who completed a baseline survey. Pre-specified controls include all the variables in Table D1. Double Lasso controls uses the Chernozhukov et al. (2017) method for variable selection and inference, as described in Section 3.1. Columns 2 and 3 both control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. In lieu of a conventional $F$-statistics, Panel B shows the effective $F$-statistic of Montiel Olea and Pflueger (2013). Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

**Appendix Table F3**

Effects on Voucher Redemption

| | *Outcome: 1(Redeemed Voucher)* | | | *Outcome: 1(Redeemed Voucher, No Test)* | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Appointment Treatment (A) | 0.102*** | 0.126*** | 0.121*** | 0.005 | 0.014 | 0.008 |
| | (0.025) | (0.024) | (0.024) | (0.015) | (0.015) | (0.015) |
| | [0.001] | [0.001] | [0.001] | [0.297] | [0.182] | [0.258] |
| Hard Commitment Treatment (HC) | 0.077*** | 0.084*** | 0.087*** | 0.033** | 0.031** | 0.032** |
| | (0.025) | (0.024) | (0.024) | (0.015) | (0.015) | (0.015) |
| | [0.006] | [0.002] | [0.001] | [0.035] | [0.041] | [0.034] |
| p-value A = HC | 0.499 | 0.239 | 0.325 | 0.202 | 0.429 | 0.256 |
| q-value A = HC | [0.226] | [0.134] | [0.170] | [0.117] | [0.200] | [0.141] |
| Controls | None | Pre-Specified | Double Lasso | None | Pre-Specified | Double Lasso |
| Observations | 1,232 | 1,232 | 1,232 | 1,232 | 1,232 | 1,232 |
| Adjusted R-squared | 0.019 | 0.128 | 0.118 | 0.002 | 0.040 | 0.033 |
| Control-group Mean | 0.159 | 0.159 | 0.159 | 0.053 | 0.053 | 0.053 |

*Notes:* Sample is 1,232 men who completed a baseline survey. Pre-specified controls include all the variables in Table D1. Double Lasso controls uses the Chernozhukov et al. (2017) method for variable selection and inference, as described in Section 3.1. Columns 2 and 3 both control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

**Appendix Table F4**

Robustness to Including Impostors

| | *Outcome: 1(Tested for HIV, set to 1 for impostors )* | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Appointment Treatment (A) | 0.112*** | 0.128*** | 0.125*** | 0.148*** | 0.157*** | 0.152*** |
| | (0.023) | (0.022) | (0.022) | (0.031) | (0.030) | (0.029) |
| | [0.001] | [0.001] | [0.001] | [0.001] | [0.001] | [0.001] |
| Hard Commitment Treatment (HC) | 0.042* | 0.052** | 0.054** | 0.080*** | 0.082*** | 0.082*** |
| | (0.023) | (0.022) | (0.022) | (0.030) | (0.029) | (0.029) |
| | [0.057] | [0.027] | [0.022] | [0.013] | [0.010] | [0.009] |
| A × HC | | | | -0.074 | -0.059 | -0.056 |
| | | | | (0.046) | (0.045) | (0.044) |
| | | | | [0.071] | [0.110] | [0.118] |
| p-value A = HC | 0.038 | 0.018 | 0.024 | 0.044 | 0.020 | 0.026 |
| q-value A = HC | [0.040] | [0.027] | [0.030] | [0.044] | [0.027] | [0.032] |
| Controls | None | Pre-Specified | Double Lasso | None | Pre-Specified | Double Lasso |
| Observations | 1,232 | 1,232 | 1,232 | 1,232 | 1,232 | 1,232 |
| Adjusted R-squared | 0.019 | 0.126 | 0.122 | 0.021 | 0.127 | 0.121 |
| Control-group Mean | 0.116 | 0.116 | 0.116 | 0.116 | 0.113 | 0.113 |

*Notes:* Sample is 1,232 men who completed a baseline survey. This table presents the same results as Table 2, but codes impostors as having a one for the outcome variable instead of a zero. Pre-specified controls include all the variables in Table D1. Double Lasso controls uses the Chernozhukov et al. (2017) method for variable selection and inference, as described in Section 3.1. Columns 2 and 3 both control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

**Appendix Table F5**

Effects on Positive HIV Tests and ART Initiation

| | *Outcome:* | |
|---|---|---|
| | *HIV-Positive* | *Initiated ART* |
| | (1) | (2) |
| Appointment Treatment (A) | 0.011* | 0.009* |
| | (0.006) | (0.006) |
| | [0.052] | [0.068] |
| Hard Commitment Treatment (HC) | 0.002 | 0.004 |
| | (0.006) | (0.006) |
| | [0.288] | [0.226] |
| p-value A = HC | 0.229 | 0.397 |
| q-value A = HC | [0.128] | [0.194] |
| Controls | None | None |
| Observations | 1,232 | 1,232 |
| Adjusted R-squared | 0.001 | 0.001 |
| Control-group Mean | 0.007 | 0.007 |

*Notes:* Sample is 1,232 men who completed a baseline survey. The regressions in this table include no controls or fixed effects because HIV status and ART initiation data was anonymized and linked only to participants' study arms. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

**Appendix Table F6**

Treatment Effect Heterogeneity by Baseline Covariates

| | *Outcome: 1(Tested for HIV)* | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Appointment Treatment (A) | 0.128*** | 0.127*** | 0.128*** | 0.127*** | 0.129*** |
| | (0.022) | (0.022) | (0.022) | (0.022) | (0.022) |
| | [0.001] | [0.001] | [0.001] | [0.001] | [0.001] |
| Hard Commitment Treatment (HC) | 0.048** | 0.048** | 0.051** | 0.048** | 0.048** |
| | (0.022) | (0.022) | (0.022) | (0.022) | (0.022) |
| | [0.036] | [0.035] | [0.030] | [0.035] | [0.036] |
| A × (Perceived Prob. of being HIV+) | -0.013 | | | | -0.015 |
| | (0.023) | | | | (0.023) |
| | [0.247] | | | | [0.234] |
| C × (Perceived Prob. of being HIV+) | 0.014 | | | | 0.025 |
| | (0.023) | | | | (0.023) |
| | [0.238] | | | | [0.152] |
| A × (Tendency to Postpone HIV Testing) | | 0.009 | | | 0.013 |
| | | (0.022) | | | (0.022) |
| | | [0.277] | | | [0.247] |
| C × (Tendency to Postpone HIV Testing) | | 0.021 | | | 0.007 |
| | | (0.022) | | | (0.023) |
| | | [0.175] | | | [0.313] |
| A × (Lives for Today) | | | 0.001 | | 0.012 |
| | | | (0.022) | | (0.023) |
| | | | [0.390] | | [0.260] |
| C × (Lives for Today) | | | -0.025 | | -0.021 |
| | | | (0.022) | | (0.023) |
| | | | [0.144] | | [0.184] |
| A × (Perceived Life Expect. Gain from ARVs) | | | | -0.015 | -0.014 |
| | | | | (0.024) | (0.024) |
| | | | | [0.234] | [0.247] |
| C × (Perceived Life Expect. Gain from ARVs) | | | | -0.019 | -0.014 |
| | | | | (0.023) | (0.024) |
| | | | | [0.200] | [0.247] |
| Controls | Pre-Specified | Pre-Specified | Pre-Specified | Pre-Specified | Pre-Specified |
| Treatments Interacted with Other BL Covariates | No | No | No | No | Yes |
| Observations | 1,232 | 1,232 | 1,232 | 1,232 | 1,232 |
| Adjusted R-squared | 0.122 | 0.122 | 0.123 | 0.122 | 0.129 |
| Control-group Mean | 0.113 | 0.113 | 0.113 | 0.113 | 0.113 |

*Notes:* Sample is 1,232 men who completed a baseline survey. Pre-specified controls include all the variables in Table D1. Other baseline covariates interacted with treatments in Column 5 include all the other variables from Table D1. Main effects are included for all variables that are interacted with the treatment indicators. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

## Appendix Table F7

Treatment Effect Heterogeneity by Demand for Commitment

| | Outcome: 1(Tested for HIV) | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Demanded the Hard Commitment (D) | 0.162*** | 0.165*** | 0.163*** | 0.141*** | 0.147*** | 0.146*** |
| | (0.026) | (0.026) | (0.026) | (0.027) | (0.027) | (0.027) |
| | [0.001] | [0.001] | [0.001] | [0.001] | [0.001] | [0.001] |
| Did not Demand the Hard Commitment (N) | 0.097*** | 0.091*** | 0.097*** | 0.097*** | 0.091*** | 0.096*** |
| | (0.018) | (0.020) | (0.019) | (0.018) | (0.020) | (0.019) |
| | [0.001] | [0.001] | [0.003] | [0.001] | [0.001] | [0.001] |
| Appointment Treatment (A) × D | 0.155*** | 0.161*** | 0.156*** | | | |
| | (0.035) | (0.034) | (0.033) | | | |
| | [0.001] | [0.001] | [0.001] | | | |
| Appointment Treatment (A) × N | 0.083*** | 0.092*** | 0.088*** | 0.083*** | 0.093*** | 0.088*** |
| | (0.028) | (0.028) | (0.027) | (0.028) | (0.028) | (0.027) |
| | [0.006] | [0.003] | [0.032] | [0.006] | [0.003] | [0.003] |
| Hard Commitment Treatment (HC) × D | 0.082** | 0.072** | 0.075** | 0.125*** | 0.108** | 0.110** |
| | (0.035) | (0.034) | (0.034) | (0.045) | (0.044) | (0.043) |
| | [0.027] | [0.039] | [0.001] | [0.011] | [0.021] | [0.018] |
| A × D × HC | | | | 0.108** | 0.121** | 0.118** |
| | | | | (0.054) | (0.052) | (0.051) |
| | | | | [0.045] | [0.027] | [0.027] |
| A × D × No Hard Commitment Treatment (NHC) | | | | 0.196*** | 0.196*** | 0.189*** |
| | | | | (0.046) | (0.045) | (0.044) |
| | | | | [0.001] | [0.001] | [0.001] |
| $p$-value A × D = A × N | 0.11 | 0.116 | 0.113 | | | |
| $q$-value A × D = A × N | [0.071] | [0.074] | [0.073] | | | |
| p-value A × D × NHC = A × N | | | | .034 | .049 | .05 |
| q-value A × D × NHC = A × N | | | | [0.037] | [0.045] | [0.045] |
| p-value A × D × NHC = A × D × HC + HC × D | | | | .503 | .533 | .457 |
| q-value A × D × NHC = A × D × HC + HC × D | | | | [0.226] | [0.237] | [0.206] |
| Controls | None | Pre-Specified | Double Lasso | None | Pre-Specified | Double Lasso |
| Observations | 1,232 | 1,232 | 1,232 | 1,232 | 1,232 | 1,232 |
| Adjusted R-squared | 0.251 | 0.307 | 0.303 | 0.251 | 0.308 | 0.303 |
| Control-group Mean | 0.113 | 0.113 | 0.113 | 0.113 | 0.113 | 0.113 |

*Notes:* Sample is 1,232 men who completed a baseline survey. Pre-specified controls include all the variables in Table D1. Double Lasso controls uses the Chernozhukov et al. (2017) method for variable selection and inference, as described in Section 3.1. Columns 2 and 3 both control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets.

**Appendix Table F8**

Treatment Effect Heterogeneity by Predicted Probability of Getting Tested

| | *Outcome: 1(Tested for HIV)* | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Appointment Treatment × | | | | |
| Low Likelihood of Testing (L) | 0.081*** | 0.071*** | | |
| | (0.029) | (0.027) | | |
| | [0.011] | [0.015] | | |
| Medium Likelihood of Testing (M) | 0.121*** | 0.143** | | |
| | (0.032) | (0.061) | | |
| | [0.001] | [0.027] | | |
| High Likelihood of Testing (H) | 0.147*** | 0.118** | | |
| | (0.044) | (0.050) | | |
| | [0.003] | [0.027] | | |
| Hard Commitment Treatment × | | | | |
| Low Likelihood of Testing | | | 0.015 | -0.029 |
| | | | (0.032) | (0.027) |
| | | | [0.266] | [0.150] |
| Medium Likelihood of Testing | | | 0.072** | 0.085* |
| | | | (0.030) | (0.049) |
| | | | [0.025] | [0.060] |
| High Likelihood of Testing | | | 0.032 | 0.092* |
| | | | (0.040) | (0.048) |
| | | | [0.200] | [0.048] |
| $p$-value H = L | 0.214 | 0.415 | 0.745 | 0.028 |
| $q$-value H = L | [0.122] | [0.199] | [0.306] | [0.034] |
| $p$-value H = M | 0.640 | 0.749 | 0.419 | 0.910 |
| $q$-value H = M | [0.266] | [0.306] | [0.200] | [0.378] |
| $p$-value M = L | 0.358 | 0.279 | 0.193 | 0.041 |
| $q$-value M = L | [0.181] | [0.150] | [0.112] | [0.042] |
| Predictors | Pre-Specified | Double Lasso | Pre-Specified | Double Lasso |
| Controls | All FEs | All FEs | All FEs | All FEs |
| Control-group Mean | 0.153 | 0.153 | 0.192 | 0.192 |

*Notes:* Sample is 1,232 men who completed a baseline survey. This table uses the repeat split-sample method of Abadie, Chingos, and West (2018) to predict the outcome variable in the control group, then estimates treatment effects separately by tercile of the predicted outcome. Pre-specified predictors include all the variables in Table D1. Double Lasso predictors include the same variables chosen using the Chernozhukov et al. (2017) in Table 2. All columns control for date-of-survey fixed effects, enumerator fixed effects, and preferred clinic fixed effects. Heteroskedasticity-robust standard errors in parentheses: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; Anderson (2008) sharpened $q$-values in brackets. $p$-values for the tests of equality between terciles ignore the covariance term and thus are likely to be conservative.