Economics as an Experimental Science: Using Field Experiments to Test Models of Economic Behavior

by

Jason Theodore Kerwin

**A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in the University of Michigan
2015**

Doctoral Committee:

**Assistant Professor Rebecca L. Thornton, Co-Chair
Professor Jeffrey A. Smith, Co-Chair
Professor John E. DiNardo
Professor David A. Lam**

For Brad, Lynda, and Audrey.

ACKNOWLEDGEMENTS

I am indebted to my dissertation committee chairs, Rebecca Thornton and Jeff Smith, for their support, guidance, and counsel during my doctoral studies, and to my dissertation committee members David Lam and John DiNardo for their feedback on my research. I also thank Achyuta Adhvaryu, Manuela Angelucci, Raj Arunachalam, Martha Bailey, John Bound, Matias Cattaneo, Mel Stephens, Robert Willis, Dean Yang, Susan Watkins, and Jobiba Chinkhumba for their invaluable comments on my research. This dissertation also benefited from the suggestions of seminar and conference participants at the University of Michigan, PAA, the World Bank, the University of Malawi College of Medicine, PACDEV, MWIEDC, NEUDC, IFPRI and the University of Minnesota. In both my research and the rest of my life, I have benefited from many conversations with Lasse Brune, Eric Chyn, Anne Fitzpatrick, Enda Hargaden, Chenyu Yang, Justin Ladner, Joe Golden, Evan Herrnstadt, Ophira Vishkin, and Olga Malkova.

I am grateful for the many friends who have shared my time in graduate school with me, both in Ann Arbor and in Africa: Anna Antoniou, Tyler Burns, Eric Burnstein, Deanna Chyn, Alan Griffith, Jonathan Hershaff, Dan Hirschman, Isaac Sorkin, Daniel Marcin, Florence Pache, Alena Perez, Doug Piper, Daniel Schaffa, Desmond Toohey, James Wang, Pierre Pratley, James Aidini, and everyone on the Devil Bears. And most of all, I thank the people who have been my bedrock of support during graduate school: my parents, Lynda and Brad Kerwin, my brother, Adrian Kerwin, my lifelong friend, Sam King, and my amazing partner, Audrey Dorélien. I could not have done it without you.

This dissertation is the result of extensive empirical research conducted in Malawi and Uganda, which would not have been possible without the hard work and help of a number of people. I thank Christopher Nyirenda, Anderson Moyo, and Synab Njerenga for their excellent management of the fieldwork for the first chapter of this dissertation, and Ndema Longwe for his exemplary fieldwork management for the second chapter. Moffat Kayembe and Carl Bruessow from the Mulanje Mountain Conservation Trust played a crucial role in making the research for my second chapter possible, and Esperanza Martinez Maldonado provided excellent research assistance on that project. The randomized evaluation of the Northern Uganda Literacy Project would not have been possible without the collaboration of

Victoria Brown, Bernadette Jerome, Benson Ocan, and everyone at Mango Tree Educational Enterprises Uganda; I deeply appreciate their role in making it happen. My research also benefited from the efforts of support staff at the University of Michigan: I thank Mary Braun, Heather MacFarland, Mary Mangum, and Lauren Pulay for all their work in making my dissertation a reality.

My doctoral studies, and the research contained in this dissertation, was funded by the generous contributions of many institutions. I am grateful for internal University of Michigan research funding from the Department of Economics, the Michigan Institute for Teaching and Research in Economics, the Center for Global Health, the Population Studies Center, the Rackham Graduate School, and the Center for the Education of Women. I am also thankful for external research grant support from the Russell Sage Foundation's Small Grants in Behavioral Economics program, the IPA/Yale Savings and Payments Research Fund (funded by the Bill and Melinda Gates Foundation), and the William and Flora Hewlett Foundation's Quality Education in Developing Countries initiative. This research was supported in part by an NIA training grant to the Population Studies Center at the University of Michigan (T32 AG000221), as well as by fellowship funding from the Rackham Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Economics as an Experimental Science:
Using Field Experiments to Test
Models of Economic Behavior
by
Jason Theodore Kerwin

Co-Chairs: Rebecca L. Thornton and Jeffrey A. Smith

Economics is famously limited by the fact that it is not an experimental science: our ability to test core models of economic behavior is limited by the fact that we cannot run experiments, for logistical and ethical reasons. Over the last two decades, however, the credibility revolution in econometrics has reshaped that view, and economists now increasingly demand credible identification of causal effects in empirical work. More recently, a set of prominent randomized trials in developing countries has shown that experiments are indeed feasible for testing a wide range of economic models. A major critique of that movement, as well as of the instrumental variables approach, is that the results do not generalize and also that they are "looking for one's key's under the lamppost", simply studying convenient topics rather than important ones.

It is possible to overcome this critique by building on the work done by the pioneers of IV methods and RCTs to design experiments that are motivated by testing economic models, running so-called "mechanism experiments." In addition, economic theory can help researchers to isolate specific results and understand their source, thus showing which results will generalize to other settings and why. Using this approach, one can learn about topics that are not amenable to experiments, such as how people's sexual behavior might change if HIV became more prevalent, by running experiments that capture the same theoretical object. This dissertation applies this approach to three open, policy-relevant, first-order questions in health and labor economics, described in the chapter abstracts below.

Chapter 1: The effect of HIV risk beliefs on risky sexual behavior: Scared straight or scared to death?

Economists typically assume that risk compensation is uniformly self-protective – that people become more careful as the health risks of their actions increase. However, risk-seeking, or fatalistic, responses can also be rational: increased risks can lead people to take *fewer* precautions. I extend the typical model of risk compensation to show that fatalism is a rational response to sufficiently high risks if people do not have perfect control over all possible exposures, and if the condition in question is irreversible. This result holds even for people who do not understand how to add up probabilities. I test this model's implications by randomizing the provision of information on HIV transmission risks to people in Malawi, a country with a severe HIV epidemic where there is qualitative evidence of fatalistic responses to the virus. Average risk responses are self-protective and statistically significant, but small in magnitude: the mean risk elasticity of sexual behavior is roughly -0.6. To test the model of rational fatalism, I develop a method of decomposing 2SLS estimates of the risk elasticity of sexual behavior by baseline risk beliefs. Consistent with the predictions of my theoretical framework, I find that this elasticity varies sharply by baseline risk beliefs: the risk elasticity varies from -2.3 for the lowest initial beliefs to 2.9 for the highest initial beliefs. 13.8% of the population has a positive elasticity, suggesting they are fatalistic.

Chapter 2: Income Timing, Temptation and Expenditures: Field Experimental Evidence from Malawi (with Lasse Brune)

The canonical model of savings and spending predicts that expenditure should be independent of the precise timing of individuals' income streams. Given market and psychological constraints, however, income timing may matter. We report results from a randomized field experiment in Malawi that varied the timing of workers' income receipt in two ways. First, payments were made either in weekly installments or as a deferred lump sum. Second, payments at a local market were made either on the weekend market day (Saturday) or the day before the market day (Friday), in order to vary the degree of temptation workers faced when receiving payments. We provide novel evidence that the frequency of payments matters for workers' ability to benefit from high-return investment opportunities. When workers are aware of the investment opportunity ahead of time, workers in the monthly group have more cash on hand than the weekly payment group and are then more likely to invest in a risk-free short-term "bond" that required a large payment and that was offered by the project in the week after the lump sum payday. We argue that this result is driven by the

lump sum group's decreased savings constraints. In contrast, despite anecdotal evidence and suggestive survey data to the contrary, being paid at the site of the local market on Saturday compared to Friday did not strongly matter for expenditure levels or temptation spending.

Chapter 3: Making the Grade: Understanding what works for teaching literacy in rural Uganda (with Rebecca Thornton)

This paper evaluates an early primary literacy program in Northern Uganda. Through a randomized experiment, we measure the effects of the literacy program as implemented by the organization that developed it. We compare those results to a second treatment group, which received a reduced-cost version of the program that was implemented through the government and designed to simulate how the program could be implemented at scale. The full version of the program has extremely large impacts on student learning: it improves student recognition of letter names by 1.0 SD, which is among the largest impacts ever measured in a randomized trial of an education program. The reduced-cost version improves letter-name knowledge by 0.4 SDs making it slightly more cost-effective than the full version. However, its effects on overall literacy are statistically-insignificant and it generates large negative effects on certain aspects of writing. This suggests that cost-effectiveness in improving the "headline" outcome measures emphasized by programs can come at the cost of lower performance in other areas.

CHAPTER I

The Effect of HIV Infection Risk Beliefs on Risky Sexual Behaviors: Scared Straight or Scared to Death?

1.1 Introduction

Risk compensation is central to our understanding of how people make decisions about protecting their health. Beginning with the seminal Peltzman (1975) paper on automobile safety regulation, economists have realized that a decline in the risk associated with a particular behavior is often offset by a rational increase in risk-taking. Empirical research on risk compensation typically assumes that people are uniformly risk-avoiding. When the per-act risk of an activity goes up, people are presumed to take fewer chances, a pattern that can be described as "self-protective." This paper considers the possibility that for some people rational responses to health risks are instead risk-seeking, or "fatalistic" – that the optimal choice may be to *increase* one's risk-taking when the per-act risk rises.

Risk-avoiding behavior is always rational if the expected cost of risk-taking can be approximated as a linear function of the per-act risk.[1] In this paper I show that fatalistic behavior is rational if a) the linear approximation is replaced with any reasonable function that is bounded above by a 100% chance of the negative outcome occurring and b) the per-act risk is sufficiently high.[2] This happens because an increase in the per-act risk (in my example, the risk of contracting HIV) affects not only the marginal cost of the acts the agent is deciding over, but also a stock of previously-chosen acts over which one no longer has any control. If the per-sex-act risk of contracting HIV rises, this raises the marginal

[1] This model is explicit in Oster (2012), but is used implicitly in many empirical analyses which restrict the relationship between risks and behavior to be linear and therefore monotonic.

[2] This result does not depend on agents using the *true* expected cost, which is based on the binomial CDF. A number of theoretical papers (Kremer 1996; O'Donoghue and Rabin 2001; Sterck 2014) have used the assumption that agents can compute the true expected cost to make the point that rational responses to increased risks will be fatalistic rather than self-protective for certain individuals.

cost of additional sex acts, by increasing the chance that they will lead to HIV infection. But it also increases the probability that the agent *already* has HIV, which *decreases* the marginal cost of more risky sex. When the second effect dominates, increases in perceived risks will lead to more risk-taking rather than less. Furthermore, if people cannot perfectly avoid all future exposures to HIV – for example, because condoms sometimes break – then unpreventable future exposures can also drive fatalistic behavior, and HIV testing will not prevent people from becoming fatalistic.[3] This mechanism is conceptually similar to the one that drives models of rational habit formation. In Becker and Murphy (1988), for example, past consumption has a large effect on current choices by changing the marginal utility of consumption. In my model, the linkage across periods operates through the marginal cost of consumption instead.

This theory of rationally fatalistic behavior suggests that people with sufficiently high beliefs about the risks of their behaviors, and imperfect control of their entire risk-taking history, will tend to become fatalistic, because they feel that they are doomed irrespective of what they do. I test this implication using data from field experiment that I conducted in southern Malawi, an area with a severe HIV epidemic. It is also an area where qualitative evidence suggests that some people are responding fatalistically to the virus[4], and where HIV prevention education emphasizes that the risk of contracting HIV is extremely high (see Figure 1.1). The experiment recruited 1292 respondents from 70 villages, and randomly assigned the respondents from 35 of the villages to be taught medically-accurate information about HIV transmission risks. A baseline survey was conducted prior to the information treatment, followed by an endline survey four to twelve weeks later.

The randomized information treatment substantially decreased people's beliefs about the risks of unprotected sex: at the endline survey, the average person in the treatment group believed the risk of HIV transmission from unprotected sex with an infected partner was 33% per sex act, as opposed to 74% in the control group.[5] Using the experimental treatment as an instrumental variable, I estimate that the risk belief elasticity of sexual activity is small but statistically significant at about -0.6. This elasticity estimate is larger than those found in studies that measure the response of sexual behavior to the actual prevalence of HIV in sub-Saharan Africa, and comparable to estimates for the United States. However, because people do not accurately know the true prevalence of the virus, the implied prevalence elasticity from my results would be smaller, and could be consistent with previous estimates for Africa.

[3] This basic idea relies on the condition in question being irreversible: if the disease in question can easily be cured, risks will not continue to aggregate because the probability of infection will reset to zero after it reaches one.

[4] E.g. Kaler (2003), Kaler and Watkins (2010)

[5] People in Malawi greatly overestimate how easily HIV is transmitted: the actual rate is just 0.1%.

This estimated mean elasticity follows the literature in assuming that the risk term enters linearly into the regression function.[6] This assumption is consistent with monotonically self-protective responses to risks. A model of rationally fatalistic behavior, however, implies that risk responses are non-monotonic, and so a linear regression is misspecified. I therefore examine whether responses to the information treatment are heterogeneous by people's baseline (pre-treatment) beliefs. I find that people with initially low risk beliefs respond self-protectively to the new information (which lowers their risk beliefs), while people with initially high risk beliefs respond fatalistically. This is the same non-monotonic pattern of responses predicted by a model of rational fatalism; the results reject the typical model of monotonically self-protective responses to risks. I can rule out that this heterogeneity is due to correlations between beliefs and other respondent attributes, and find no other factors that cause statistically-significant heterogeneity in responses.

Having demonstrated that the estimated mean elasticity from simple 2SLS is misspecified, I develop a method for decomposing instrumental-variables estimates by exogenous covariates and show that this method gives consistent estimates of the underlying conditional parameter. This approach reveals that the risk elasticity of sexual behavior varies substantially across the population, from -2.3 for the lowest initial risk beliefs to 2.9 for the highest initial beliefs; 13.8% of the population has a positive elasticity. The fatalistic group has higher-than-average risk factors for HIV, such as years of sexual experience and perceived HIV-positive status. This suggests that they may be more important in driving the overall prevalence of HIV. Therefore, the effect of the status quo policy – in which health educators encourage people to greatly overestimate HIV transmission risks, for their own good – is ambiguous from both an ethical and an epidemiological standpoint. More generally, these results militate against programs that attempt to "scare people straight" via messages that emphasize that risks are extremely high – especially when they actually are not.

This paper contributes to four bodies of research in economics. First, it builds on our understanding of risk compensation by providing what I believe to be the first experimental evidence on the elasticity of risk-taking behavior with respect to perceived risks. Moreover, it shows that that elasticity cannot be meaningfully summarized by a population average, because the subgroup of the population with the highest baseline risk beliefs may respond positively (fatalistically) to risks. This implies that future empirical work on risk compensation should take into account the possibility of non-monotonicity.

Second, it contributes to a growing empirical literature that studies how people's subjec-

[6] An extreme example of this approach is the Viscusi (1990) study of cancer risk perceptions and smoking behavior, which employs one-sided rather than two-sided t-tests. This eliminates any possibility of fatalistic responses, although Vicusi's estimated standard errors are small enough that this assumption does not affect inference.

tive expectations affect their behavior. Expectations have long played an important role in economic models, but recent research has shown that it is possible to collect meaningful information on people's subjective expectations both in the developed world (e.g. Lillard and Willis 2001, Manski 2004) as well as in developing countries (e.g. Attanasio 2009; Delavande, Giné and McKenzie 2011; Delavande 2014). I provide the first experimental evidence that subjective expectations about risks have a measurable, causal effect on people's behavior, lending credence to the broader idea that we should be asking people about their subjective beliefs rather than assuming they know the true probabilities of events.

Third, it helps reconcile the substantial responses to HIV risks found in America (Ahituv, Hotz and Philipson 1996) with very small ones in Africa (Oster 2012). Self-protective responses by the majority of people may be offset by opposite-signed, fatalistic responses by a subset of the population, yielding an average response that is self-protective but low in magnitude. This is particularly plausible because gay men in the US perceive the prevalence of HIV to be much lower than Africans do (White and Stephenson 2014). The same reasoning may also help explain why recent field experiments in Africa have found large responses to relative HIV risk information for specific population groups despite the fact that overall risk responses appear to be small in magnitude.[7]

Fourth, it also helps explain the small measured responses of sexual behavior to HIV testing. As Philipson and Posner (1993) point out, the effect of learning one's HIV status is theoretically ambiguous, because learning that you are HIV-positive can have two opposite-signed effects on your behavior. Purely self-interested people should see little or no marginal cost from further risky sex if they are already infected, while altruistic people would want to take measures to protect their prospective partners. A parallel logic applies to those who learn they are HIV-negative. Experimental research on HIV testing has found fairly small responses: Thornton (2008) finds zero average effects for HIV-negatives and very small average reductions in risk-taking for HIV-positives in Malawi. One possible explanation for the small responses she finds is that people's high perceived risk of contracting HIV means that testing has a limited effect on their perceived lifetime risk of becoming HIV positive: even if a person tests negative today, they may continue to think that contracting HIV is highly likely in the future. Likewise, a current positive test may not be a substantial surprise. Consistent with this argument, Gong (2015), studying people in urban Kenya, finds that responses to HIV testing vary by people's priors about their HIV status. People who are surprised by a test result respond in a selfishly rational manner, with large increases

[7] Godlonton, Munthali and Thornton (2015) find that uncircumcised men in Malawi take fewer sexual risks when they are told that circumcised men face a lower risk of HIV infection. In a study in Kenya, Dupas (2011) finds that girls in secondary school choose younger partners when they are told that older partners are riskier.

in risk-taking when people are surprised by positive test results and large declines in risk-taking in response to surprise negative test results.

The remainder of this paper is organized as follows: I begin in Section 1.2 by laying out a model of responses to risks that extends the typical approach to allow for the possibility of rational fatalism, showing that under very general conditions people may rationally respond to high perceived risks fatalistically (as opposed to self-protectively). In Section 1.3, I describe a randomized field experiment that I conducted in southern Malawi to test the implications of this model, as well as the data on risk beliefs and sexual risk-taking that I rely on. Section 1.4 lays out my empirical strategy and results, and in Section 1.5 I address the mechanisms behind my results, address some potential limitations of this paper, and discuss the implications of my findings for the design of HIV prevention policies. Section 1.6 concludes.

1.2 Theoretical Framework

This section outlines a model of behavioral responses to HIV risks that relaxes a key assumption made by the previous empirical literature. Most empirical work on responses to risks, HIV or otherwise, relies on the assumption that the stochastic cost of risk-taking is linear in the riskiness of each individual act. My model allows that cost to follow a concave shape that asymptotes to a probability of 1, which is consistent with the risks of individual sex acts adding up into a sensible total probability of HIV infection. The core result is that the comparative static in question – the derivative of risk-taking with respect to per-act risks – is not always negative, or self-protective. In general, the sign of the comparative static will flip from negative to positive if an agent's risk beliefs and stock of unavoidable risks are sufficiently high. This happens because the marginal cost of risk-taking will approach zero as the total chance of HIV infection gets close to 100%.

1.2.1 Model Basics

In this model, I assume that agents weigh the benefits of choosing a level of risky sex, y, against its costs. These costs include both a fixed per-act cost (which could be a pecuniary cost but also time cost) q, and a stochastic component due to the risk of HIV infection. An agent's *perceived* risk per sex act is x. The expected cost of HIV infection is the agent's subjective belief about the total probability of it occurring, P, times its perceived cost, c. The subjective probability can be written as a continuously differentiable function $P = P(x, n)$, where $n = y + m_0 + m_1$ is the total number of sex acts, including both the current choice y, and an immutable stock of acts $m_0 + m_1$. This stock includes all previous sex acts since

one's most recent HIV test, m_0, and also all future risky acts that are unavoidable, m_1. The latter captures accidental exposures through things like condom breakage, situations where an agent may lack the bargaining power to turn down some future sex acts, imperfect self control, and so forth.

Throughout the model I will treat HIV infection as irreversible, so that all risky acts aggregate into a single probability P. This is true of HIV if we consider fatalism to be driven only by inevitable future exposures, or if testing is unavailable. It will only hold for certain other risks, and depends on perceived rather than actual irreversibility of the condition. For example, if people perceive lung cancer to be a binary and irreversible condition, the model results will go through, but if a condition is widely known to be curable, such as Chlamydia, then they will not. It is possible to compute the actual value of P using the binomial distribution, but my results will be robust to agents potentially not understanding how to correctly compute probabilities. The benefit of y sex acts is described by a continuously differentiable benefit function, $B(y)$, with positive and diminishing marginal benefits.

To focus the exposition on the mechanism that drives fatalistic risk responses, rather than on mathematical derivations, I model the agent's choice as a one-shot, static decision. This collapses the future into the expected cost of HIV infection $P(x, y + m_0 + m_1)c$. The results in this section can be generalized to a multi-period setting – see Appendix A.4 for details. The single-period optimization problem is:

$$\max_{y \geq 0} \{U(y; x, m_0, m_1, q, c)\} = \max_{y \geq 0} \{B(y) - qy - P(x, y + m_0 + m_1)c\} \tag{1.1}$$

By the assumption that y is continuous, the maximand $U(n; m_0, m_1, p, c, x)$ is a sum of continuously differentiable functions and therefore continuously differentiable itself.

I do not assume that agents can correctly convert levels of risk-taking and per-act risks into an aggregate probability of HIV infection. Instead, I simply assume that $P(x, y + m_0 + m_1)$ corresponds to sensible probabilities: it must lie between 0 and 1, and be equal to zero if either sex is risk-free ($x = 0$) or an agent engages in no risky sex ($y + m_0 + m_1 = 0$). I also assume that higher riskiness x is in fact interpreted as leading to a higher subjective probability of HIV infection, and more risk-taking $y + m_0 + m_1$ also increases the chance of contracting HIV.[8] The subjective probability also approaches 1 as riskiness rises toward 1 or as total risk-taking goes to infinity.[9]

The model formulated above is similar in spirit to those used in the literature on rational habit formation and addiction. In Becker and Murphy (1988), consumption choices are linked

[8] Formally, $P_1 \geq 0$, with $P_1(0, y + m_0 + m_1) > 0$ if $y + m > 0$ and $P_1(x, 0) = 0$; $P_2 \geq 0$, with $P_2(x, 0) > 0$ if $x > 0$ and $P_2(0, y + m_0 + m_1) = 0$.

[9] $P \to 1$ as $y + m \to \infty$ as long as $x > 0$, and $P = 1$ if $x = 1$ and $y + m_0 + m_1 \neq 0$.

across periods by the effect of past consumption on the marginal benefit of current-period consumption. In this model, both past and future consumption of the risky act have a large impact on the marginal cost of current risk-taking. This analogy is made even more clear by the multi-period formulation of the model in Appendix A.4.

For most possible functional forms of $B(\cdot)$ and $P(\cdot,\cdot)$ this optimization problem will have no closed-form solutions for the optimal number of sex acts y^*. However, there must be *some* interior solution as long as the marginal benefit of risky sex outweighs the costs for at least one act, and approaches zero as $y \to \infty$. A sufficient condition for the existence of interior optima is that $q > 0$, so there is some fixed price or time cost to risky sex.[10]

1.2.2 Comparative Statics

Given the existence of an interior solution, we are interested in a specific comparative static: how does risk-taking y^* respond to a change in the per-act risk x? I derive the properties of $\partial y^*/\partial x$ using the implicit function theorem. For an interior solution, the optimal number of sex acts y^* must satisfy the following first- and second-order conditions:

$$B'(y^*) - q - P_2(x, y^* + m_0 + m_1)c = 0 \tag{1.2}$$

$$B''(y^*) - P_{22}(x, y^* + m_0 + m_1)c \leq 0 \tag{1.3}$$

The first-order condition in equation 1.2 is a function $G(y^*, x, m_0, m_1, q, c) = 0$. Therefore the implicit function theorem allows us to compute the comparative static for changes in y^* in response to changes in x:

$$\frac{\partial y^*}{\partial x} = -\frac{\frac{\partial G}{\partial x}}{\frac{\partial G}{\partial y^*}} = \frac{P_{21}(x, y^* + m_0 + m_1)c}{B''(y^*) - P_{22}(x, y^* + m_0 + m_1)c} \tag{1.4}$$

The denominator is just the left-hand side of the second-order condition, and is thus negative.[11] Since $c > 0$, $\text{sign}(\partial y^*/\partial x) = -\text{sign}(P_{21}(x, y^* + m_0 + m_1))$. It is typical in the literature to approximate P by a linear function, $P(x, y + m_0 + m_1) \approx x(y + m_0 + m_1)$. This is done explicitly in Oster (2012) and implicitly by Viscusi (1990), for example. In that case $P_{21} = 1 > 0$ always, so $\partial y^*/\partial x < 0$. This is analogous to the Oster (2012) result that sexual activity should fall as the prevalence of HIV rises. More broadly, it says that behavior is uniformly self-protective: people always choose fewer risky acts as the per-act risk of each

[10] See Appendix A.1 for a proof.

[11] Technically it is only weakly negative since the second-order condition is a weak inequality. The discussion that follows assumes strict negativity, since otherwise $\partial y^*/\partial x$ is undefined. However, all the results in this section will hold as the second-order condition approaches 0 from above.

act rises.

However, the linear approximation does not satisfy the requirements for being a sensible probability laid out in Section 1.2.1. For low values of x and $y + m_0 + m_1$ this is not an issue, since P will lie between 0 and 1. In the context of HIV risk beliefs, however, x is often quite high, since perceived risks are typically large overestimates, and $m_0 + m_1$ will reflect a potentially long sexual history and an extensive future of possible condom failures and so forth. These can easily push the linear approximation above 1, which is obviously wrong. One way of imposing sensible probabilities on p is to use the true probability function $P = \pi(x, y + m_0 + m_1) = 1 - (1 - x)^{y+m_0+m_1}$. O'Donoghue and Rabin (2001) point out that for this function, $\pi_{12} = (1 - x)^{y+m_0+m_1-1}[1 + y + m_0 + m_1 \ln(1 - x)]$, and hence $\pi_{12} > 0$ if $y + m_0 + m_1 < 1/[-\ln(1 - x)]$ and $\pi_{12} < 0$ if $y + m_0 + m_1 > 1/[-\ln(1 - x)]$. In words, P_{12} is not constant in sign, but shifts from positive to negative if x rises above a point defined by the total number of risky acts $y + m_0 + m_1$. This then implies that the sign of $\partial y^*/\partial x$ will shift from negative to positive when it crosses that tipping point.

This result is not specific to relying on the true function $\pi(x, y + m_0 + m_1)$ but is true for any function $P(x, y + m_0 + m_1)$ that satisfies the basic conditions laid out in Section 1.2.1. I prove this fact formally in Appendix A.2, but can readily be understood from the conceptual illustration in Figure 1.2. The horizontal axis shows the number of risky acts chosen, while the vertical axis shows the total subjective probability of contracting HIV. The dashed blue line shows the relationship between P and $y + m_0 + m_1$ for a low perceived per-act risk x, and the solid red line shows the relationship for a higher value of x. Consistent with the basic rules of sensible probabilities, and also with the linear approximation used in most empirical research on risk responses, the slope of the red line is initially higher. When sex is riskier, the total probability of contracting HIV initially rises faster for the same number of sex acts. But the total probability is capped at one, so there must be some point above which the slope of the red line is *lower* than that of the blue line.[12] Formally, this can be written as follows:

Proposition 1.2.1 (Tipping point in P_{21})**.**

$$\exists \tilde{x} = \tilde{x}(y+m_0+m_1) \text{ s.t. } P_{21}(x, y+m_0+m_1) > 0 \text{ if } x < \tilde{x} \text{ and } P_{21}(x, y+m_0+m_1) < 0 \text{ if } x > \tilde{x}$$

Recall that part of the total level of risk-taking is tied up in $m_0 + m_1$, which is out of the agent's control. It is useful to think about this as including the agent's sexual history (in a context where HIV testing is unavailable, for example), but it also contains all future

[12] The results here technically rely on $P(x, y + m_0 + m_1)$ being continuous, but as discussed in Appendix A.3 it is possible to reach similar conclusions even if people use heuristic methods for aggregating risks into total probabilities that are not continuous.

risks that the agent cannot avoid. To fix concepts, suppose that everyone thinks that they will experience at least one condom break some time in the future, so $m_1 \geq 1$. For $m_1 = 1$, and using the true function $\pi(x, y + m_0 + m_1)$, the tipping point occurs at $x = 0.63$. This is extremely high compared with the actual per-unprotected-act risk of contracting HIV from a randomly-selected partner, but it is not particularly high compared with the subjective beliefs expressed by people in Malawi. At baseline, 28% of my sample believed the risk was at least that high.

If we maintain the assumption that sexually active adults cannot eliminate all possible exposures to HIV (so $m_0 + m_1 \geq 1$ in general), this eliminates the possibility of a corner solution where $y + m_0 + m_1 = 0$, and guarantees that the tipping point value $\tilde{x}$ that changes the sign of P_{12} from positive to negative will be somewhere below 1. Proposition 1.2.1 then implies that $\partial y^*/\partial x$ will itself have a tipping point:

Proposition 1.2.2 (Tipping point in comparative static $\partial y^*/\partial x$)**.**

$$\exists \tilde{x} = \tilde{x}(y + m_0 + m_1) \ s.t. \ \frac{\partial y^*}{\partial x} < 0 \ \textit{if} \ x < \tilde{x} \ \textit{and} \ \frac{\partial y^*}{\partial x} > 0 \ \textit{if} \ x > \tilde{x}$$

Below the threshold value of the per-act HIV infection risk $\tilde{x}$, rational agents will behave self-protectively (reducing their risk-taking in response to increased risks); above $\tilde{x}$ they will behave fatalistically (increasing their risk-taking in response to increased risks).

This result is somewhat counterintuitive, but it captures a fairly simple logical conclusion: if the risks are sufficiently high and I can't totally avoid exposure, there is no value to limiting how much sex I have; I am doomed no matter what. It is a purely rational alternative to the psychologically-driven fatalism derived by Caplin (2003). In his model, agents do not compensate away from extremely high risks because not responding lets them ignore the problem and thereby avoid the stress and fear associated with it. In my model, agents do not compensate away from extremely high risks because the perceived marginal benefit of abatement is nearly zero.

This sort of rationally fatalistic response is a potential issue for a wide range of decisions. Anti-smoking campaigns, to take one example, often feature "Benefit Timelines" that emphasize the health benefits that accrue to ex-smokers 20 minutes after quitting, 24 hours, 3 months, and so forth (e.g. National Health Service 2013). These timelines can be understood as a way to combat the possibility that smokers will think they are doomed to eventual cancer, no matter what they now decide. Similar to the benefit timelines in logic, HIV prevention messaging targeted at HIV-positive people emphasizes the risk of "reinfection" with a different strain of HIV (e.g. Cichocki 2014). Actual cases of reinfection are rare enough

that the medical importance of this possibility is unclear (Smith, Richman and Little 2005), but one goal of this kind of messaging is to avoid a rise in risky sex by selfishly rational people who believe they have nothing to lose. Indeed, there is suggestive evidence that fatalistic reasoning about HIV infection is important in sub-Saharan Africa's HIV epidemic (Barnett and Blaikie 1992; Kaler 2003; Kaler and Watkins 2010; Wilson, Xiong and Mattson 2014).

It is possible to extend Proposition 1.2.2 to account for altruistic behavior on the part of people who know they are HIV-positive, and may choose to be careful to protect their sex partners. In this case, there is no stochastic *personal* cost of risky sex, and $P(x, y+m_0+m_1)$ can instead be interpreted as the total subjective probability of infecting one's *partner* given a perceived risk x and total risk-taking $y + m_0 + m_1$. c is then the extent to which agents care about their partners avoiding HIV. All the same results then go through: for relatively low values of perceived risks and low levels of risk-taking, agents will respond to rises in the per-act risk by reducing how much sex they have, but when the risks are sufficiently high they give up, assuming their partner is either already infected or doomed to infection in the future.

One consequence of Proposition 2 is that the linear relationship between x and y^* typically estimated in empirical analyses of risk responses may be misspecified, since y^* is in general a non-monotonic function of x. Estimated average partial effects of x on y^* will in general include both positive and negative ranges of $\partial y^*/\partial x$, which will tend to push the average toward zero. They will also ignore potentially-crucial heterogeneity in the effect of risk beliefs on risk-taking behavior. In my empirical analysis in Section 1.4, I explicitly examine risk responses for heterogeneity by initial beliefs.

1.3 Data and Experimental Design

This section outlines the data and experiment that I use to test the model laid out in Section 1.2. I begin by describing the randomized field experiment that I conducted in southern Malawi to collect data on how individuals' sexual behavior responds to changes in their beliefs about HIV infection risks. I then describe my preferred measures of sexual risk-taking, which come from retrospective sexual diaries collected as part of the survey. Throughout this paper, I use the word "sex" to refer to heterosexual vaginal intercourse. Other forms of sexual activity are extremely uncommon in Malawi and are potentially sensitive topics (cf. Kerwin, Thornton and Foley 2014), so they were not included in the survey. I conclude the section with a discussion of my measures of beliefs about HIV infection risks.

1.3.1 Experimental Design

This paper uses data from a field randomized controlled trial I conducted from August to December 2012. The experiment took place in Traditional Authority (TA) Mwambo, in the Zomba District of Malawi's Southern Region. I sampled roughly 30 sexually active adults aged 18-49 from each of 70 villages. Each participant was interviewed twice: once for a baseline survey, and again for an endline survey conducted 1-3 months later. At the end of the baseline survey, all participants were provided with basic information about the sexual transmission of HIV and the benefits of condoms.[13] Participants from half of the villages, chosen at random, were assigned to the treatment group. They were read an information script that presented the actual annual risk of HIV transmission in serodiscordant[14] couples that have unprotected sex, based on estimates from Wawer et al. (2005) and also figures from the Malawi National AIDS Commission.

The village sample for the study was constructed from the Malawi National Statistics Office GIS files for the 2008 Census. I began by removing all duplicate village entries from the dataset.[15] Because existing evidence indicates that fatalistic responses to HIV risks and risky sexual activity may be concentrated around major trading centers (Kaler 2003), I then constructed sampling strata based on the distance to the closest major trading center.[16] 24 of the sampled villages (34%) were within 2 km of a trading center[17]; another 24 (34%) were within 2 and 5 km from a trading center; and 22 (31%) were more than 5 km away from the closest center. This compares with overall proportions of 10%, 40% and 50% of all villages in TA Mwambo. Within each sampling stratum, I randomly assigned half of the villages to the treatment group and half to the control group.

In each village, a team of enumerators first conducted a comprehensive household census. Using this census, 15 men and 15 women aged 18-49 were then sampled from each village, with only one respondent allowed per household. The sample was thus stratified by both gender and distance to the nearest trading center, so the effective sampling strata are formed

[13] Knowledge of the basics of HIV transmission and prevention is already high in this population. In the 2010 DHS, nearly 100% of individuals said that HIV was sexually transmitted and over four fifths knew that condoms were effective prevention (Malawi National Statistical Office and ORC-MACRO 2010).

[14] Relationships with one HIV-positive and one HIV-negative partner.

[15] The Population and Housing Census uses Enumeration Areas as its basic sampling unit, rather than villages. The boundaries of these enumeration areas commonly cross through villages, leading to duplicate entries in the GIS datasets.

[16] Trading centers were identified based on their designation by the 2008 Malawi Population and Housing Census. Since TA Mwambo adjoins the city of Zomba, I also included the main markets in that city as trading center equivalents. In addition, based on conversations with key informants, I included several more trading centers in the local area that were not designated as such by the census.

[17] In discussions with key informants in TA Mwambo, 2 km was generally agreed to be the maximal distance people will walk for nightlife. These strata thus roughly proxy for how easily people could access the trading centers in order to drink and search for sex partners.

by combinations of gender and distance indicators. Some villages had too few households for 30 eligible-age adults to be selected, and hence the maximum feasible number was chosen instead.[18] This yielded a total of 2024 sampled individuals. The survey team then attempted to contact all sampled people for a baseline survey. Although refusals were rare ($< 1\%$ of respondents refused the baseline survey), 23% of sampled people could not be found at baseline, typically because they were temporarily away from the household.[19] A total of 1543 respondents had a successful baseline survey. Because the survey contained sensitive questions about sexual behavior, and the model of fatalism applies mainly to sexually active adults, the survey used an early screening question to eliminate people who had never had sex from the sample. This removed 2.6% of the respondents, leaving 1503 sexually-active adults in the baseline survey.

After a minimum delay of 30 days, the enumerator team attempted to recontact all 1503 sexually-active respondents from the baseline survey, successfully finding 1292.[20] There is no evidence of differential attrition: an indicator for inclusion in the final sample is not significantly correlated with treatment status, irrespective of whether I control for other baseline covariates. [21] There is also no evidence of differential attrition by baseline covariates, which I examine by interacting the treatment indicator with different baseline variables.[22]

Baseline summary statistics for the overall sample, as well as a comparison of the treatment and control groups, are presented in Table 1.1.[23] The sample is 43% male and 82% married, with a mean age just below 30. Respondents are fairly poor on average: household cash expenditures average just under $2 (at purchasing-power parity) per person per day. The sample is well-balanced across the treatment and control groups with the exception of household cash income, which is approximately $64/month higher in the control group. However, this discrepancy can be attributed to seasonal variation in income combined with the differential timing of the baseline surveys: for reasons discussed below in Section 1.3.2, the control group baseline surveys were done first and the treatment group baseline surveys were done second. A comparison of incomes at the endline survey is valid if we make the

[18] My respondents therefore form a weighted probability sample of TA Mwambo, with oversampling of villages closer to trading centers as well as oversampling of people from smaller villages. I do not adjust any of the results in the paper using sampling weights, but all of my main findings are robust to using such weights.

[19] It is common for people in this area of Malawi to travel during the agricultural off-season to look for casual wage labor.

[20] See Appendix Table B.1 for detailed figures on the number of people in each study arm and sampling stratum.

[21] See Appendix Table B.2.

[22] See Appendix Table B.3.

[23] In this table, and in all the other balance tests in this paper, the p-values are adjusted to account for the clustered design of the study, following Donner and Klar (2000).

plausible assumption that the information treatment had no impact on earnings. Monthly household income at the endline survey is still $23 higher in the control group, but this difference is not statistically significant. The summary statistics are consistent with the randomization having successfully generated balanced treatment and control groups.

1.3.2 Information Treatment

At the end of the baseline survey, all respondents from the treatment villages were read and shown information about the true risk of HIV infection between serodiscordant partners who have unprotected sex, as measured by the Wawer et al. (2005) study of serodiscordant couples in Rakai, Uganda. I used the annual risk for the information treatment because it is simpler to explain than the per-act risk, which is very small, and also because it is the figure available on the Malawi National AIDS Commission's website. For a discussion of the ethical dimensions of teaching people the true risk of HIV transmission, see Appendix C.

The information treatment was administered by the survey enumerators in a one-on-one setting. It involved both an oral component and an interactive visual component. In the oral component, the basic details of the original Rakai study were explained, with certain aspects simplified for clarity. Respondents were told that the study occurred in Uganda, and that 100 serodiscordant couples were followed for a single year.[24] They were told that all the couples had regular sex without using condoms, about once every three days on average, and asked how many people they thought would contract HIV. They were then informed that in fact only ten of the initially HIV-negative people became HIV-positive.[25] Respondents were asked if they believed the results of the study; enumerators were trained in how to respond to a number of common questions, such as whether the testing equipment was faulty.[26] The script listed the reasons that HIV transmission sometimes does not happen even when serodiscordant couples have unprotected sex, for example the fact that HIV sometimes cannot penetrate the genitalia. The script then emphasized that HIV transmission is something

[24] The actual figure for the Wawer et al. study is 235 couples, 188 of which never used condoms when they had sex (results are not broken out by condom use, but condom use was very inconsistent and had no impact on the estimated transmission rate). The time period was actually 10 months, with some couples being observed for multiple time windows. This was reduced to 100 couples over the course of 1 year for the sake of clarity and simplicity, and to match the 10% per year figure cited by Malawi's National AIDS Commission.

[25] This is the annual transmission rate cited by the Malawi National AIDS Commission. The exact annual rate implied by the Wawer results is 12%. The Hollingsworth, Anderson and Fraser (2008) reanalysis of the Wawer et al. data finds an annual transmission rate of 10.6% from asymptomatic partners (HIV-positive sex partners who have not just recently contracted the virus and do not yet have AIDS), which are the majority of cases, but does not provide an overall average.

[26] The questions respondents asked were recorded on the baseline survey. All my results are robust to excluding respondents who asked any follow-up questions.

that happens by chance, comparing it to popular games of chance used by local cell phone companies as marketing tools.

The interactive visual component complemented the oral component and occurred at the same time. It involved showing respondents a diagram with 100 pairs of stick figures representing serodiscordant couples, with a black stick figure indicating an HIV-negative partner and white stick figure indicating an HIV-positive partner. The respondent was asked to guess as to the number of people who would contract HIV after a year of regular unprotected sex with an infected partner, and this guess was indicated by circling an appropriate number of these stick figure couples. When the true rate was presented, the enumerator showed a second diagram in which ten of the initially HIV-negative individuals had turned from black to white. Enumerators then counted and circled these transmissions.

To minimize the risk of contaminating the control villages, all the baseline treatment surveys were done after the baseline control surveys were completed. This approach parallels that taken by Godlonton, Munthali and Thornton (2015). The survey enumerators were only taught to administer the information intervention after all the control surveys were completed.

1.3.3 Measures of sexual behavior

My primary outcome measure is self-reported sexual behavior as recorded using a detailed retrospective sexual diary. The diary walks respondents through the previous seven days beginning with yesterday. On each day, respondents were asked what time they woke up, how much alcohol they had, whether they were menstruating (or for men, whether their sex partner was menstruating), the value of gifts they received from their partner (or for men, gifts they gave to their partner),[27] how many times they had sex, and the time they went to sleep. Then, for each reported sex act, they were asked detailed questions such as the time of day, the length of the act, condom use, and whether the sex act was with their primary sex partner or a different partner. The surveys also contained single-question recall measures of sexual behavior, for example: "In the past 30 days, how many total times did you have sex, including serious and non-serious partners?"

The diary-based approach to measuring sexual behavior was initially developed and refined in previous research on sexual behavior in southern Malawi (Kerwin et al. 2011). It builds on research that shows that calendar-based methods reduce recall bias compared with single-question recall methods (Belli, Shay and Stafford 2001). Luke, Clark and Zulu (2011) have found that relationship history calendars improve the quality of responses to questions

[27] The culture of gift-giving in sexual relationships in Malawi is strongly gender-driven: with very few exceptions men give gifts to women and not the other way around.

on sexual behavior, showing that apparent biases due to social desirability effects are smaller. The sex diary approach adapts these insights to a much shorter time frame to assist respondents in the recall of all sex acts over the past 7 days. The improved accuracy of the sex diary over other methods is reflected in the data captured by the surveys. Column 1 of Table 1.2 shows that the two variables record fairly similar levels of sexual activity. The distributions of the two variables are also very different, with substantially more heaping at multiples of 5 in the single-question recall variable.[28] Given the lower quality of the single-question recall variables, and because I used total sex acts as recorded on the diary as my primary outcome in an earlier working paper that I wrote prior to the experiment (Kerwin 2012), I will rely primarily on the sex diary variables for my analysis.

Table 1.2 presents summary statistics for all the available measures of sexual activity in the data. Columns 3 and 4 show the means of my measures of sexual activity for the control and treatment groups respectively, while Column 5 shows the difference between the two. These are generally balanced across the two study arms, with the only statistically significant differences being a lower number of lifetime sex partners ($p < 0.05$). All the differences are fairly small in magnitude, but none of the variables has exactly equal means across the treatment and control groups at baseline. This is one reason my analyses will control for respondents' baseline values of self-reported sex, as described in Section 1.4.2.

An additional measure of the demand for safer sex comes from the sale of subsidized condoms to respondents that occurred immediately after the endline survey. All participants were given six coins worth five Malawi Kwacha each (30 Kwacha total, or about ten cents). They were then offered the chance to purchase 3-packs of condoms for five Kwacha apiece, or individual condoms for two kwacha. While this price represents a sizeable subsidy relative to the sale of condoms at local stores, the vast majority of respondents who had acquired condoms in the period leading up to the endline survey got them for free. When asked about the nearest place to acquire condoms, respondents commonly named heath centers and health extension workers, both of which offer condoms free of charge. The condom sales measure was only collected at the endline survey.

It is common in the literature to present results using a constructed combined outcome index, both to reduce concerns about multiple comparisons and to improve the precision of estimates (e.g. Kling, Liebman and Katz 2007). However, the value of such an index is unclear in situations where some outcomes are measured with greater error or where baseline data is not available for particular outcomes (for example, condom sales were only done at endline). I therefore present two versions of the sexual risk index. One uses all outcomes that can be constructed from the retrospective sexual diary, which I argue provides more

[28] See Appendix D for histograms and a discussion of the implications of heaping for regression estimates.

accurately-measured outcomes than the single-question recall variables. An alternative index includes both the sex diary outcomes as well as all other outcomes that can be constructed from the survey, including the condom sales.

Each index is constructed separately for the baseline and endline waves by normalizing all component variables (subtracting the control-group mean and then dividing by the control-group standard deviation). The normalization is reversed in sign for condom use, condom acquisition, and condoms purchased, for which positive numbers imply less risk-taking. These normalized values are then averaged for each respondent, weighted by the factor loadings for the first principal component of the matrix of the data for the control group. This follows Black and Smith (2006) in assuming that there is a single underlying sexual activity factor, and that the different outcomes measured in the data are noisy signals of that factor; the procedure selects the linear combination of the data that gives the best estimate of the underlying sexual activity factor.[29]

1.3.4 Measures of risk beliefs

The central prediction of the model I outline in Section 1.2 is that individuals' responses to risk information will depend on their initial perceptions of those risks. A key input for my analysis, therefore, is a quantitative measure of risk perceptions. Due to data limitations, one common strategy for this is to utilize some measure of the true risk.[30] However, an emerging literature has shown that it is feasible to collect meaningful data on subjective beliefs about probabilities using surveys, not just in the United States (e.g. Lillard and Willis 2001; Manski 2004) but also in the developing world (e.g. Attanasio 2009, Delavande, Giné and McKenzie 2011, Delavande 2014). Delavande and Kohler (2009) have developed a method of eliciting subjective expectations using visual aids that they show performs very well in Malawi.

Rather than following Delavande and Kohler, I rely on measures of subjective risk beliefs collected using concrete questions about proportions out of a fixed number of people. These are questions of the form "If 100 men, who do not have HIV, each sleep with a woman who is HIV-positive tonight and do not use a condom, how many of them do you think will have HIV after the night?" I then divide the reported number by the denominator used to construct a subjective probability. Question E1a in Figure 1.3 is an example of one of these questions. All the questions were gender-specific: for instance, when men were asked about HIV transmission they were asked about 100 men having sex with an HIV-positive woman,

[29] I also explored unweighted averages; these produce similar results with slightly smaller magnitudes.

[30] E.g. Ahituv, Hotz and Philipson (1996) and Auld (2006) in the US and Oster (2012) and Juhn, Kalemli-Ozcan and Turan (2009) in Africa.

and likewise women were asked about 100 women having sex with an HIV-positive man.[31]

I use these concrete expectations questions for two reasons. First, the Delavande and Kohler approach adds considerably to the logistical complexity of surveys, as well as the time needed to conduct them. Second, this concrete style of expectation question has been validated through extensive use in previous research across a variety of contexts in Malawi, including in urban areas[32] as well as in areas of rural southern Malawi near my study site.[33] They also appear to be fairly scale-invariant: switching the denominator from 100 to 1000 or 10,000 yields nearly the same average subjective probabilities, and individual respondents give the exact same answer roughly 60% of the time.[34] The questions also perform well in terms of respecting nested probabilities: if the chance of event B occurring includes all possible instances of event A, then respondents should ideally report a weakly higher probability for B than for A. Delavande and Kohler emphasize this as one of the major strengths of their approach.

My data do not afford many direct comparisons with Delavande and Kohler's on HIV transmission and HIV prevalence, because their survey instrument did not ask many HIV-related questions that are necessarily nested within one another. One comparison, however, is the per-unprotected-sex-act risk of contracting HIV from an infected partner, compared with the annual risk. In my data, the latter probability was weakly higher 92.2% of the time, whereas this was the case 91.9% of the time in the Delavande and Kohler data.[35] In addition to performing comparably to the Delavande and Kohler approach in terms of nesting probabilities, the concrete probability method also produces similar results in terms of the mean expectation of the risk of HIV transmission: this is 82.8% per act for the control group at baseline using concrete probabilities, and 85.9% per act using Delavande and Kohler's method.

One potential concern with eliciting subjective expectations is the tendency for probabilities to heap at the "focal" probability of 50%. The typical interpretation, cited by Delavande and Kohler, is that this heaping reflects a misunderstanding of the question, or

[31] Six HIV risk belief variables were collected: the unprotected transmission rate (both per-act and annual), the condom-protected transmission rate (both per-act and with a condom), and two questions about the prevalence of the virus: the share of all members of the opposite sex that respondents thought were HIV-positive, and the share of members of the opposite sex that they find attractive.

[32] Chinkhumba, Godlonton and Thornton (2014)

[33] Godlonton, Munthali and Thornton (2015), Kerwin et al. (2011)

[34] Author's calculations based on Chinkhumba, Godlonton and Thornton (2014)

[35] The annual question for Delavande and Kohler actually asks about someone who is married to an HIV-positive person, and does not explicitly specify unprotected sex. However, social norms in Malawi strongly proscribe the use of condoms within marriages (Tavory and Swidler 2009) and married couples use condoms just 11.2% of the time in my sample. Repeating this analysis just for people in the Delavande and Kohler sample who say there is no chance they would use condoms with their own spouse yields a similar nesting rate of 94.1%.

simple uncertainty, rather than a true belief. People commonly use 50% (or in my case, report half of the total denominator), when they are simply unsure about the answer. To address this issue, respondents who reported beliefs of 50% were prompted with a followup question about whether they really believed the chance was 50%, or if they were just not sure, which is an approach taken on the Health and Retirement Study's subjective expectations questions (Hudomiet, Kézdi and Willis 2011). Building on that approach, respondents who said they were just not sure were then prompted for their best guess. Question E1b in Figure 1.3 illustrates these followup questions. In my measure of risk beliefs I use the response to the followup question for people who change their answer.

1.3.5 Enumerator-knowledge contamination of measured beliefs

As described in Section 1.3.2, the enumerators were only trained to provide the information intervention after the baseline interviews for the control group were finished. This was done to minimize any chance of the information intervention contaminating the control group. However, it also meant that this was the first time the enumerators themselves were taught the true risk of HIV transmission. As a result, enumerators brought different beliefs with them into the baseline treatment and control surveys. This had a relatively small but statistically-significant effect on the measured beliefs of treatment-group respondents at baseline.

Figure 1.4 shows the daily average recorded risk belief, separately for treatment and control surveys, and including both baseline and endline surveys. The lines show linear time trends fit to the data. One thing that is immediately clear is that the measured difference at baseline is much smaller than the impact of the information intervention. This can be confirmed numerically by comparing Panel A of Appendix Table B.4, which shows the enumerator effects on measured baseline beliefs, to Table 1.3, which presents the effects of the information intervention people's beliefs about the transmission rate of the virus. The treatment effects are at least four times as large as the enumerator effects, no matter what specification is used.

There are two potential explanations for this pattern. One is that different knowledge may have led enumerators to prime subjects differently, possibly even subconsciously. Enumerators were trained to follow up with probing questions when respondents answered a question by just saying that they did not know. The phrasing of these probing questions could have been affected by the knowledge enumerators brought to the surveys. A second possibility is enumerator experience with the questions. While the sex diary questions that form my outcome measure use very simple statements that enumerators were already familiar with using, the phrasing used on the subjective expectations questions was fairly complex.

This may have lead to some temporal pattern in reported risk beliefs as the phrasing of the probing questions used was refined over time.

There is evidence for both explanations in Figure 1.4. A downward trend in measured risk beliefs is evident prior to the enumerators being taught the information about HIV transmission, and there is a large drop in beliefs after the first vertical line that marks the training session. Further confirmation of the importance of enumerator knowledge for measured risk beliefs can be seen based on the light blue dots that appear after the first vertical line. These are average beliefs for "cleanup" surveys – a handful of control-group interviews that were done after the treatment surveys had begun, because respondents were not home when surveys were attempted prior to the information treatment training. Excluding the large negative outlier (which is the average for a day when just a single control-group survey was done) these generally match the measured beliefs for the baseline treatment group surveys.

Another way of understanding the importance of enumerator knowledge is to compare the beliefs recorded at baseline for the treatment group to the endline beliefs for the control group; this can be done by comparing the hollow triangles to the solid circles in Figure 1.4. These are both surveys during which the enumerators' knowledge is identical (they know the information about HIV) and the respondents in the treatment and control groups have identical information sets (neither has been told the HIV risk information). This is reflected in the recorded values, which look the same in the two groups.[36]

To correct for the evident contamination of measured risk beliefs due to differential enumerator knowledge, I adjust reported beliefs based on time trends with a trend break. This involves estimating the following regression:

$$x_i = \rho_0 + \rho_1 Date_i + \rho_2 After_i + \rho_3 After_i * Date_i + v_i \tag{1.5}$$

$Date_i$ is the date of the baseline survey for respondent i and $After_i$ is an indicator for whether the baseline survey was done after the information treatment training session. I then construct $x_i^{Adj} = x_i^{resid} + \hat{\rho}_0$, and bound the resulting variable to lie within $[0, 1]$ by replacing values below 0 with 0 and those above 1 with 1. Panel C of Appendix Table B.4 presents the trend-adjusted risk beliefs for the control and treatment groups. They are unsurprisingly similar across groups. As robustness checks, I also replicate my analysis using the raw (unadjusted) risk belief measures, as well as two other kinds of trend adjustment: using a single trend across the whole baseline period, and using just a level shift in reported

[36] Panel B of Appendix Table B.4 does formal t-tests for this comparison. The only statistically-significant differences are in annual unprotected transmission risks and the prevalence of HIV among attractive people of the opposite sex.

beliefs. My results are not sensitive to any of these variations, but my preferred specifications use the adjustments described above. These have a simple interpretation: they are my best estimate of how a respondent's initial beliefs compare with the rest of the sample, given the known time trend and trend break evident in the data due to enumerator-knowledge contamination.

1.3.6 Composite belief measures

My analysis focuses on a composite measure of the perceived risk of contracting HIV from unprotected sex with a randomly-chosen potential sex partner. This is the product of the perceived per-act risk of HIV transmission from unprotected sex with an infected partner and the perceived prevalence of HIV among attractive people of the opposite gender. I use this composite variable for three reasons. First, it is the same risk belief variable I used in the working paper that laid out the key theoretical results that motivated this project (Kerwin 2012). Second, using the perceived HIV prevalence among attractive people of the opposite sex mitigates concerns that people's self-beliefs about risks may differ from their beliefs about the risks faced by the rest of the population. Recent research on subjective expectations has highlighted that people's self-beliefs can be very different from what they believe about people in general and that people are more responsive to self-beliefs (e.g. Wiswall and Zafar 2014). For risks, this commonly takes the form of unrealistic optimism about one's own risk relative to the rest of the population (Weinstein and Klein 1996). In my context, there is also the potential for unrealistic pessimism: people's stated perceptions of both HIV prevalence and transmission risks are much higher than the truth, and they may feel more at risk personally than they believe to be the case for the broader population. While I cannot totally eliminate the potential for differences between self-beliefs and general beliefs, focusing on the risk from unprotected sex with a random attractive member of the opposite sex (rather than all local people of the opposite sex) is likely to be a superior measure of the level of risk people feel they actually face.

Third, relying on variation in the other beliefs allows me to avoid one of the shortcomings of using perceived per-act HIV risks, which is that they are extremely concentrated in the right tail. At baseline, over four in ten respondents believe that the per-act risk of HIV transmission from unprotected sex is 100% (Figure 1.5, Panel A). If I use this variable to conduct the heterogeneous treatment effects analyses in Sections 1.4.5 and 1.4.6, I find the same basic pattern of heterogeneity as with my preferred risk measure. People with the highest risk beliefs have sharply lower treatment effects that people with the lowest beliefs, and I estimate a zero treatment effect for people with beliefs above 65% per act. However, by clustering 40% of people at the very top of the belief distribution, this approach hides

the fact that people in the highest category of per-act risk beliefs actually perceive sharply different risks. Interacting the per-act risk belief variable with the respondent's perceived prevalence breaks up the mass point of people who think the per-act risk is 100%, and does so according to their perception of how risky they think having unprotected sex actually is. The resulting product also has a natural interpretation: it is how risky people perceive any given sex act to be if they do not know the HIV status of their partner, given their perceptions about the prevalence of HIV among potential sex partners and the transmission rate of the virus. Panel B of Figure 1.5 shows the distribution of this combined variable, which has a much smaller mass point at 100%.

In Figure 1.6 I present the baseline CDFs of the combined risk measure I focus on in this paper, constructed two different ways. Panel A uses unadjusted values of the per-act risk and prevalence belief variables, while Panel B uses values that have been adjusted for a linear time trend with a trend break as described in Section 1.3.5. In each panel the solid line shows the control group's beliefs while the dashed line shows the treatment group's beliefs. The treatment and control group distributions are different using the raw values, and this is largely corrected by the regression adjustment.

1.4 Empirical Results

This section details the empirical results of the study. I begin by showing that the information treatment has large effects on people's risk beliefs. I then show that the average effect of the information treatment is to slightly (but statistically significantly) increase the amount of risky sex people have. This is consistent with a small negative risk elasticity of sexual behavior, which I estimate directly using two-stage least squares. I then construct semiparametric decompositions of the treatment effect by people's initial risk beliefs, and show that the overall average masks substantial heterogeneity by baseline beliefs. I extend this analysis to 2SLS estimates of partial effects as well: I use indirect least squares to develop an estimator of the local average treatment effect (LATE) that allows for heterogeneity by baseline covariates. Using this heterogeneous LATE estimator, I show that the elasticity of sexual behavior with respect to risk beliefs is negative for individuals with low risk beliefs, and becomes positive for individuals at the high end of the risk belief distribution.

1.4.1 Impact of the information treatment on risk beliefs

The information treatment has large effects on respondents' risk beliefs. Panel A of Table 1.3 shows the endline treatment-control differences for all the measures of people's beliefs about HIV transmission and prevalence. The treatment group believes the annual risk from

unprotected sex is 38 percentage points lower than the control group does. Their belief about the per-act risk decreases even further, by 41 percentage points.[37] Note that the respondents do not update their beliefs perfectly: the actual annual transmission rate is about 10%; just 2% of the treatment group reports beliefs that low. The alternative specifications in Panels B and C confirm that these results are robust to controlling for baseline values of the outcome variable and running a difference-in-differences respectively.

Respondents also update their beliefs about HIV risk variables other than the transmission rate from unprotected sex. For example, beliefs about the risk of condom-protected sex and about HIV prevalence are both reduced. This suggests that instead of simply memorizing the numbers they were told, respondents learned the information and updated their beliefs accordingly: if they understand that the current prevalence of HIV depends on infected people transmitting the virus to others, then a reduction in the transmission rate implies the a reduction in the prevalence of the virus. The information treatment contained no direct information about the prevalence of the virus nor about condom-protected sex, so the effects on these variables can be ascribed purely to this learning process.

1.4.2 Estimation Strategy

All my regressions control for baseline values of the outcome variable. Frison and Pocock (1992) and McKenzie (2012) show that this generates estimated treatment effects with a lower variance than either a) relying the endline values of the outcome alone or b) using changes in the outcome (i.e. a difference-in-differences). When there are baseline differences in outcomes across study arms, this approach also generates estimates with a lower bias than either alternative. (See Appendix E for a mathematical derivation). Controlling for the baseline value of the outcome will reduce the bias anytime the outcome variable is not exactly equal across study arms – even if the difference is not statistically significant. Since there are small but non-zero differences in the means of outcome variables across study arms, this is the preferred estimator for my sample. The specifications used in this paper also control for the stratification cells (combinations of distance categories and gender) used to draw the original sample, which improves statistical efficiency (Bruhn and McKenzie 2009). My regressions have the following form:

$$y_i^e = \alpha + \beta T_i + \gamma y_i^b + Z_i'\eta + e_i \tag{1.6}$$

[37] The larger impact on per-act risks is a consequence of the ceiling of 100% on transmission rates; 50% of treatment group respondents who think the annual transmission rate is 100% believe the per-act transmission rate is less than that.

where y_i^e is the endline value of the outcome variable, T_i is an indicator of whether the respondent was in the treatment group, y_i^b is the baseline value of the outcome variable, Z_i is a vector of categorical dummy variables for the sampling strata, and e_i is an error term.

1.4.3 Reduced form effects of the information treatment

The results of the reduced-form specifications are shown in Table 1.4; all continuous outcomes are presented in logs so the coefficient estimates can be interpreted as percentage-point changes.[38] The estimated impact is small in magnitude: it is possible to rule out magnitudes larger than 20 percentage points, or greater than 0.16 standard deviations for the indices. The number of sex acts in the past week rises by 10 percentage points. Focusing specifically on the margin of abstinence (whether people have any sex at all), this shifts by 5 percentage points, which is roughly 0.1 standard deviations. The risk indices confirm that these results are robust to multiple hypothesis testing: both the overall and sex diary risk indices rise by 6%, significant at the 10% and the 5% level respectively. The treatment has no effect on condom use, nor on condom purchases. This is consistent with the extremely high rates of unprotected sex: at baseline just 1 in 10 sex acts involved a condom, leaving limited room for increases in risk-taking at this margin.

1.4.4 The risk belief elasticity of sexual behavior

The effect of this specific information treatment on sexual behavior is less generalizable than the marginal effect of HIV risk beliefs on sexual risk-taking, which can be used to design other policy interventions involving responses to HIV infection risks.[39] Consider the OLS regression

$$y_i^e = \alpha + \delta x_i^e + \gamma y_i^b + Z_i'\eta + e_i \tag{1.7}$$

$\hat{\delta}$ is an estimate of $\partial y^*/\partial x$, the partial effect of risk beliefs on risky sex. The results of running this regression with various outcomes are shown in Panel A of Table 1.5, and discussed below. However, for these estimates to be consistent, x_i^e must be independent of the error term. This is unlikely to be true. One reason it may fail to be true is that individuals may form their risk beliefs based in part on sexual experience, and sexual experience is highly autocorrelated. Another reason, noted by Oster (2012), is that the subjective risk will probably have some

[38] Because many outcomes contain zeroes, I use the inverse hyperbolic sine transformation of Burbidge, Magee and Robb (1988) rather than logging the variable directly, constructing $log_{ihs}(y) = ln(y+sqrt(y^2+1))$.

[39] As noted above, I use the adjusted versions of the belief variables, which removing time-varying trends in beliefs. All the results are robust to using the original belief variables instead.

association with the actual prevalence of HIV – and that the prevalence is itself the outcome of local sexual behavior.

I therefore estimate $\hat{\delta}$ via two-stage least squares, using T_i as an instrument for x_i^e. T_i is plausibly excludable from the second-stage regression. Because the treatment was randomized, membership in the treatment group should have no association with sexual behavior other than through the information treatment. Furthermore, the information treatment is very unlikely to affect sexual behavior through any channel other than individuals' risk beliefs: it does not contain any guidance or information about sex. The instrument also easily satisfies the relevance condition. The F-statistic on T_i in the first-stage regressions is roughly 220 for all specifications.[40] This allows me to estimate two-stage regressions as follows:

$$x_i^e = \alpha + \beta T_i + \gamma y_i^b + \rho x_i^b + Z_i'\eta + e_i \tag{1.8}$$

$$y_i^e = \alpha + \delta \hat{x}_i^e + \gamma y_i^b + \rho x_i^b + Z_i'\eta + e_i \tag{1.9}$$

x_i^b is included as a control in the first stage in order to improve efficiency and reduce bias, for the same reason discussed in Section 1.4.2 above.

The 2SLS estimates are shown in Panel B of Table 1.5, with OLS results (estimated on the control group only) shown in Panel A for comparison. The OLS results have a uniform positive bias relative to 2SLS, confirming that OLS is not consistent in this context. This is consistent with Oster (2012), who finds that OLS estimates of the elasticity of sexual behavior with respect to the true prevalence of HIV are biased and wrong-signed. The fact that the omitted variable in the second-stage regression is positively correlated with risk beliefs can be explained in one of two ways. First, people may form their risk beliefs through a process in which sexual activity plays a part. For example, people who have more sex may be exposed to more gossip, which (if the tone is frightening) leads them to raise their risk beliefs. Second, people who have a latent desire for more sex may select into opportunities to learn about HIV risks; since HIV risk messaging tends to overstate transmission risks, this would lead them to have upward-biased beliefs.

The elasticity of sex acts in the past week with respect to HIV risk beliefs is approximately -0.6. The other elasticities are smaller in magnitude: they are mostly around -0.3, which is the estimate yielded by the sexual activity index method. These results are much larger than Oster (2012), which estimates prevalence elasticities of about -0.01 to -0.02 for binary outcomes (compared with -0.3 for my binary outcome in column 1). My estimates are closer

[40] It is not possible to conduct a formal test for weak instruments unless the number of excluded instruments is at least two more than the number of endogenous regressors (Stock and Yogo 2005). However, the informal "rule of thumb" generally used in applied econometrics is an F-statistic of at least 10; by this standard, my instrument easily passes.

to the Ahituv, Hotz and Philipson (1996) estimates for the US: they find elasticities of about -0.2 for binary outcomes. My estimates for continuous outcomes are also close to those found in US studies: focusing on gay men in San Francisco, Auld (2006) estimates a prevalence elasticity of sexual acticity of -0.5. However, my results are not directly comparable with this earlier work, which uses the true prevalence as the regressor of interest. People do not accurately know the true prevalence, so changes in the true prevalence are unlikely to show up 1-for-1 as changes in perceived prevalence. This means that the implied prevalence elasticities from my results are likely to be smaller than those for the US, and closer to the Oster (2012) findings.

The population-average reduced form and marginal effects both fit a model of self-protective risk-compensation, which is consistent with the existing literature. However, the specifications in Tables 1.4 and 1.5 impose common effects across all respondents, and hence across all levels of risk beliefs. To explore the importance of this restriction, I explore heterogeneity in ITT and marginal effects by baseline covariates, with a focus on baseline risk beliefs.

1.4.5 Heterogeneity in the reduced-form effect of the risk information treatment

The key prediction of the rational fatalism model is that responses to risks will be heterogeneous by individuals' baseline characteristics. Specifically, it predicts that the magnitude and sign of the comparative static will vary by baseline beliefs about risks. This implies that, provided the first-stage effect of the information treatment on risk beliefs is uniformly negative, the sign of the effect of the information treatment should vary by baseline risk beliefs as well, I test this prediction by estimating a modified version of the reduced-form regression:

$$y_i^e = \alpha + \beta T + \sum_{j=1}^{J} \left[\beta^{Tw^j} T_i w_i^j + \delta_j w_i^j\right] + \gamma y_i^b + Z_i'\eta + e_i \tag{1.10}$$

Here $w_i^1, ..., w_i^J$ are a set of J baseline covariates. My primary focus is on heterogeneity by baseline risk beliefs x_i^b. I also examine other potential sources of heterogeneity in responses, such as gender, baseline sexual activity, and previous HIV exposures.

The results of these heterogeneous treatment effects analyses for the total number of sex acts in the past week are presented in Table 1.6. Responses to the information treatment are strongly heterogeneous by baseline risk beliefs (Column 1). Using this linear specification, people with baseline risk beliefs of 0% respond to the information treatment by increasing

their sex acts per week by 32%. For people with baseline beliefs of 100%, the response is lower by 50%, meaning that weekly sexual activity *declines* by 18%. I can reject that responses for people with high risk beliefs are the same as for those with low beliefs at the 1% level; the negative response for people with the highest risk beliefs is statistically significant at the 10% level. The positive treatment effect for people who have baseline beliefs of 0% suggests that a linear specification for the treatment effect heterogeneity is misspecified, since their risk beliefs should increase rather than decrease in response to the treatment. This lends further support to the flexible analyses I conduct below.

In Columns 3 through 6 I look for heterogeneous responses by gender,[41] baseline sexual activity, perceived previous exposure to HIV,[42] and whether the respondent believes he or she may currently be HIV-positive.[43] There is also no statistically-significant heterogeneity by any of these factors. Moreover, the results for baseline risk beliefs are also robust to including three-way interactions with gender, as well as the other variables in Table 1.6.

The specification in Table 1.6 assumes that the heterogeneity in treatment effects is linear in form. While this is not a concern for binary w_j such as gender, it is a more substantive restriction for continuous variables like baseline beliefs. As an alternative, I estimate semiparametric regressions of dy/dT by baseline risk beliefs for the treatment and control groups:

$$y_i^e = \beta^T + f^T(w_i) + \gamma^T y_i^b + Z_i'\eta^T + \varepsilon_i \text{ if Treatment} = 1 \tag{1.11}$$

$$y_i^e = \beta^C + f^C(w_i) + \gamma^C y_i^b + Z_i'\eta^C + \nu_i \text{ if Treatment} = 0 \tag{1.12}$$

These regressions give me estimates of $\mathbb{E}[y|T=1]$ and $\mathbb{E}[y|T=0]$ for each value of w_i.[44] Thus taking the difference gives me estimates of the w_i-specific treatment effect $\hat{\tau}_y(w_i) = \hat{f}^T(w_i) - \hat{f}^C(w_i)$.[45]

[41] The effect of gender on responses to the information treatment is theoretically ambiguous. Malawian women commonly have less bargaining power in sexual relationships than men. However, most of my sample comprises matrilocal villages, which grant women more power to divorce their husbands and hence may increase bargaining power within relationships as well (Schatz 2005).

[42] Perceived previous exposure to HIV is an indicator that is coded to 1 if the respondent believes any of their past sex partners was HIV-positive and zero if they do not. This ignores the possibility that a condom was used for the sex acts with an HIV-positive partner, but given the low rates of condom use in this population that should not affect the results appreciably.

[43] The perceived HIV status variable is an indicator that collapses a Likert scale question in which respondents report how likely they think it is that they are HIV-positive now on a scale from "No Likelihood" up to "High Likelihood." "No Likelihood" is coded as a zero, while any other response is coded as a one. "Don't Know" is coded as a missing value.

[44] Technically these are $\mathbb{E}[y|T=1, y_i^b, Z_i]$, but the randomization of T_i means I can ignore the expectation over the control variables.

[45] A purely nonparametric version of this estimator is used in the Bennear et al. (2013) study of behavioral responses to information about arsenic in drinking water.

I implement the semiparametric regressions using the Robinson (1988) double residual estimator for partially linear regressions. The basic logic of the Robinson estimator is as follows: consider the regression function for the control group. If we take its conditional expectation given w_i, and subtract that from the original equation, the $f(w_i)$ component drops out and we have

$$y_i^e - \mathbb{E}[y_i^e|w_i] = \gamma^C(y_i^b - \mathbb{E}[y_i^b|w_i]) + (Z_i' - \mathbb{E}[Z_i'|w_i])\eta^C + \nu_i$$

The conditional expectations of y_i^e given w_i and of the controls given w_i are estimated by separate nonparametric regressions for each variable. These estimates are plugged in to the equation above, which is estimated by OLS. Finally, the parametric component of y_i^e is removed using the estimates of γ^C and η^C, allowing the function $f^C(w_i)$ to be estimated non-parametrically. I choose data-driven bandwidths to minimize the mean-squared prediction error using the generalized cross-validation (GCV) statistic of Loader (2004). My results are qualitatively robust to halving all the bandwidths as well (see Appendix Figures F.1 to F.3). The underlying semiparametric regressions do not have boundary bias problems because they are fit using local linear regressions. However, my estimates (which are the difference of two sets of local linear regressions) show a high degree of variability at the very edges of the distribution, so I truncate the display of my graphs to eliminate points outside (0.05, 0.95).

I apply this approach to heterogeneity in my first-stage regressions of endline risk beliefs x_i^e on the information treatment, and construct a function $\tau_x(x_i^b)$. I also apply it to my reduced-form regressions of treatment effects on sexual activity, estimating a function $\tau_y(x_i^b)$. I then construct confidence intervals via a clustered bootstrap with 1000 repetitions; for each bootstrap repetition, I repeat the procedure of adjusting belief variable to correct my estimated confidence intervals for the fact that it is a generated regressor. In each bootstrap sample, I trim observations with estimated densities lower than the minimum observed in the original dataset. The original sample has no estimated densities that are near zero, so my point estimates do not have trimming issues. Replicating the results while trimming at zero instead does not appreciably change the estimates, suggesting that very few observations have extremely small estimated densities.

Figure 1.7 shows the results of this semiparametric regression for the first stage, and Figure 1.8 shows the results for the reduced form. The first-stage results show that the change in risk beliefs is largest for people with the highest beliefs, and drops fairly steadily as baseline beliefs fall.[46] This pattern is reasonable, since people with the highest risk beliefs

[46] Near the low end of the scale the estimated dx/dT is larger in magnitude than the baseline beliefs x^b.

should update their priors by a larger amount than people with lower beliefs. These results are robust to an alternative semi-parametric approach, using brackets of the baseline belief distribution instead of the Robinson estimator. In that approach, I construct indicator variables for eight quantiles of the baseline risk belief variable, and interact those with the treatment indicator; I then regress the outcome on the full set of interactions plus my controls (see Appendix Figure F.5).

The semiparametric reduced-form estimates are consistent with those from the linear approximation in Table 1.6: the treatment effect is initially positive, and then becomes negative for people with extremely high baseline risk beliefs. For people with the highest baseline beliefs, I can reject the null that the treatment effect is ≥ 0 at the 1% level. The pattern of heterogeneity is also confirmed by the bracketed approach described above (see Appendix Figure F.6). In addition, I try a wide range of alternative specifications, several alternative methods of handling the baseline risk beliefs, and a number of different outcome measures (Appendix F, Figures F.7 to F.19). The results uniformly confirm the same pattern of heterogeneity: people with the highest baseline risk beliefs respond negatively, rather than positively, to the information treatment. By pooling the data for the middle 6 brackets in the bracketed approach, I can also confirm that the point estimates are positive in the middle range of the data. Even though the pointwise CIs include zero, the estimated treatment effects are all similar to one another, and so I can reject the null hypothesis of a zero effect in the middle range of the data.

1.4.6 Heterogeneity in the risk belief elasticity of sexual behavior

My theoretical framework predicts not just heterogeneity in treatment effects but also heterogeneity in the effect of risk beliefs x on sexual behavior y^*. In particular, it implies that the partial effect of x on y^* will be initially negative, and then positive for sufficiently high x. I therefore also examine heterogeneity in the instrumental-variables estimate of the effect of x on y^*.

To do this, I develop an estimation strategy that can be applied to any baseline covariate w_i. I begin by defining subgroup k of the sample as those individuals with $w_i = w^k$. It is possible to construct an estimator of the group k-specific marginal effect $\hat{\delta}^k_{IV} = \hat{\delta}^k_{IV}(w^k)$, which will in general be a function of w^k. Since T_i and w_i are independent, the treatment

This happens because dx/dT is estimated off of endline beliefs, which tend to revert toward the mean for the control group. For example, for people with baseline beliefs below 0.10 the average endline belief was 0.18 in the control group and 0.10 in the treatment group. My randomized treatment is orthogonal to this mean-reverting measurement error, so the consistency of my estimates should not be affected, but they may represent the wrong points on the baseline belief spectrum. If some of the respondents in the high tail at baseline were actually lower on the belief spectrum, my results will understand the extent of fatalism among the people whose initial beliefs were actually high.

remains a valid instrument for this subsample. Selection on right-hand side variables likewise does not affect the consistency of an estimator, so any valid instrumental variables estimator for the whole sample will be valid for this subsample (Heckman 1996). While I could rely on 2SLS estimation, in general I will want to estimate the relationships semiparametrically, so I instead use the indirect least squares (ILS) estimator. I estimate the following separate regressions:

$$x_i^e = \alpha^x + \beta^x T_i + \gamma^x y_i^b + Z_i'\delta^x + e_i \text{ for } w_i = w^k \tag{1.13}$$

$$y_i^e = \alpha^y + \beta^y T_i + \gamma^y y_i^b + Z_i'\delta^y + v_i \text{ for } w_i = w^k \tag{1.14}$$

with w_i being the baseline belief variable and w_k represents each of its values. I then construct

$$\hat{\delta}_{ILS,j}(w^k) = \frac{\hat{\beta}^y(w^k)}{\hat{\beta}^x(w^k)} \xrightarrow{p} \frac{\frac{dy}{dT}(w^k)}{\frac{dx}{dT}(w^k)} = \frac{dy}{dx}(w^k),$$

where convergence in probability comes from Slutsky's theorem.[47] I estimate the w_k-specific treatment effects $\hat{\beta}^x(w^k)$ and $\hat{\beta}^y(w^k)$ using $\tau^x(w^k)$ and $\tau^y(w^k)$ as described above. While it is possible to construct analytic standard errors for ILS, I rely instead on cluster-bootstrapped confidence intervals since my preferred underlying estimator is already semiparametric and has standard errors without a known analytical form.

The results of this procedure, using the log of sex acts in the past week as the outcome variable, are shown in Figure 1.9. These elasticities are consistent with the theoretical framework from Section 1.2, in which the relationship between risk beliefs and risky sex has an overall U-shape: the slope is initially negative and then becomes positive for people with sufficiently high risk beliefs. My confidence intervals are pointwise, rather than simultaneous; due to the nature of my estimation procedure, constructing simultaneous confidence intervals is difficult. However, using the bracketed version of the results I reject the null that marginal effects are less than zero for the highest risk belief category at beyond the 0.01 level; the Bonferroni-adjusted p-value is below 0.02. The bracketed approach thus suggests that the top octile, or highest 12.5%, of respondents are fatalistic. Looking instead to the results using the Robinson estimator, I find that 13.8% of people have elasticities greater than zero: the risk elasticity of risky sex varies from -2.3 for the lowest risk beliefs to 2.9 for the highest ones. Note that although this evidence suggests a U-shaped relationship, I am unable to recover the underlying function: I can estimate heterogeneity in the marginal effect of endline

[47] The overall LATE can be recovered from these w_i-specific LATEs by taking a weighted average of them, where the weights are the product of the share of the data that has a given value of w^k and the strength of the first stage for w^k. See Appendix G for a derivation.

risk beliefs on risky sex only by *baseline* risk beliefs, not by endline beliefs.

1.5 Discussion

In this section I discuss the implications and limitations of the results of this study. I begin by showing that the fatalistic responses I observe are consistent with the mechanisms of rationally fatalistic responses described in Section 1.2. Then I show that my results are not driven by baseline risk beliefs capturing other observed variation in the baseline data, such as education or sexual activity. I then discuss several potential limitations of this study. Finally, I consider what my results imply for HIV prevention policy.

1.5.1 Mechanisms for Fatalistic Responses

The theoretical framework in Section 1.2 predicts fatalistic responses to risks in two different situations. First, people may have an accumulated stock of past risks they have taken whose outcome has not yet been realized. Second, they may not have perfect control over their future risky behavior: condoms may break, they may be tempted into mistakes, and so forth. If the first mechanism alone is driving the fatalism measured in our sample, then people's responses to the information treatment should be fatalistic if (and only if) they believe they are currently HIV-positive. There is no evidence of this pattern in my sample: Column 6 of Table 1.6 shows that there is no statistically-significant difference in the treatment effect by people's baseline beliefs about their HIV status.[48] This result does not differ for people who are in the highest category of risk beliefs (not shown).

Another implication of the model is that the information treatment should shift people's beliefs about their current HIV status or about whether they will contract HIV in the future. To examine this, I use endline data about respondent's perceived likelihoods of current or future HIV infection. I run multinomial logits of the endline perceived likelihood variables on a treatment indicator, controlling for sampling strata and categorical indicators for the values of the baseline perceived likelihood variable.[49] These consider the different likelihood values, as well as "Don't Know," as discrete choices. I estimate these regressions separately for each quantile of risk beliefs. Figure 1.10 reports the mean marginal effects on people reporting there is "No Likelihood" that they have HIV from these regressions, multiplied by negative 1. These can be interpreted as the effect of the information treatment on people believing there is any chance that they have HIV now (Panel A) or will get it in the future

[48] See Section 1.4.5 for a description of how this variable is defined.

[49] No data for perceived likelihood of contracting HIV in the future was collected at baseline, so the baseline data for the respondent's perceived likelihood of having HIV currently was used as a proxy.

(Panel B).

I find evidence for both potential mechanisms for fatalism. The information treatment decreases the probability that people with high initial risk beliefs think there is any chance they currently have HIV by 18 percentage points compared to a control-group mean of 38%. The effect on perceiving there is any chance that you will contract HIV in the future is even stronger: it decreases by 19 percentage points. This suggests that the results presented in Figures 1.8 and 1.9 can indeed be explained by reductions in fatalism among the highest-risk group. It also implies that HIV testing may not on its own be able to eliminate fatalistic behavior: the response in terms of changes in qualitative beliefs is slightly stronger for contracting HIV in the future, rather than having it at present. Even if someone is tested for HIV, this may have a limited impact on their perceived life-cycle probability of HIV infection, because perceived transmission rates are so high. Someone who receives a positive HIV test result will only be expected to change their behavior if they were not already convinced that they have HIV (Gong 2015), and high priors about currently having HIV are encouraged by people's exaggerated risk beliefs. Furthermore, even someone who is surprised by a negative HIV test result may continue to believe they are extremely likely to contract HIV in the future, since they perceive the transmission rate to be extremely high.

The results on the perceived chance of getting HIV in the future are also robust to conditioning on respondents saying there is no likelihood that they currently have HIV. These findings demonstrate that the fatalistic responses I observe are consistent with my model of rational fatalism. They also rule out the possibility that beliefs about current HIV status alone are the sole source of fatalistic responses: perceptions about contracting HIV in the future are also important.

1.5.2 Is heterogeneity by beliefs driven by correlations with other variables?

The results shown in Figures 1.7, 1.8, and 1.9 show that responses to HIV risks vary by respondents' baseline beliefs. However, these beliefs are not assigned at random, and therefore may be correlated with the respondents' other characteristics. For example, people form their risk beliefs partly through experience with sexual partners, so their sexual behavior may affect their beliefs. Also, qualitative evidence suggests that Malawi's education system plays an important role in the formation of risk beliefs, hence it is likely that baseline risk beliefs are also capturing variation in education. As a result, it is possible that some of the heterogeneity in risk responses is coming from other factors correlated with risk beliefs, rather than from the beliefs themselves.

To explore this possibility, I run a regression of baseline HIV risk beliefs on an extensive list of demographic, socioeconomic, and sexual behavior variables measured at baseline that

could plausibly play a role in shaping respondents' beliefs.[50] Observable factors can explain only a tiny share of the variation in beliefs: this regression (omitted for space) has an R-squared of 0.067. I also repeat the analysis from Column 2 of Table 1.6, including interactions between the treatment indicator and the full set of baseline covariates. The results, shown in Column 7 of the table, show no significant heterogeneity by any other baseline factor, and leave the coefficient on the interaction between the information treatment and risk beliefs nearly unchanged. Thus the heterogeneity in risk responses by baseline risk beliefs is not due to those beliefs being correlated with other respondent attributes.

1.5.3 Potential limitations

The estimates in Section 1.4 are representative of the local population in the region where the experiment took place. Because my sample was chosen to mirror the overall population, where marriage is nearly universal among sexually-active adults, over 80% of my respondents are married. The effects I estimate, are therefore mostly for married people, and so represent changes in either marital sex or extra-marital activity. My experimental results confirm this: responses to the information treatment are not statistically different by marital status, but the magnitude of the response is much larger for married individuals (results not shown). This suggests that my results do mostly represent changes in sexual activity by married individuals.

Both changes in sexual activity within marriages and changes in infidelity are reasonable to expect in this setting, because southern Malawi has high rates of perceived and actual infidelity. 18% of married women and 10% of married men think their spouse is unfaithful (Conroy 2014). My survey did not ask whether reported sex partners were the respondent's spouse, in order to enhance respondents' comfort with revealing details of their sex lives, but did instruct enumerators to record this information if the respondent happened to mention it. Nearly a quarter of married respondents volunteered this information; of those, 5% of men and 19% of women said their primary sex partner was not their spouse. As a result, both the perceived and actual risk of contracting HIV from one's spouse is high. Longitudinal studies have estimated that up to 70% of all people newly-infected with HIV in Africa are married (Gray et al. 2011). My respondents are aware of this channel of infection: baseline, 36% of married people in my sample think there is some chance their primary sex partner

[50] The independent variables in this regression were three sexual behavior variables (lifetime number of sex partners, total sex acts in the past week, any sex in the past week), four measures of cognitive ability (immediate word recall, delayed word recall, numeracy quiz score, and Raven's Progressive Matrices score) and categorical indicators for gender, marital status, age bracket, ethnic group (collapsing small cells), education level, whether respondent read a newspaper in the past week, whether respondent listened to the radio in the past week, and whether respondent watched television in the past week.

has HIV.

The changes in behavior that I measure should be considered in light of the risk environment my respondents face. The majority of the population, having realized sex is less risky than they thought, is more open to sex with a spouse they might see as high-risk: perhaps a husband who is away a lot, or who is rumored to have another sex partner. Alternatively, they may be more open to sex with high-risk outside partners themselves. The fatalistic group that has the highest initial risk beliefs has realized that previous unprotected sex has not, in fact, doomed them to share the fate of their high-risk sex partner, and they reduce how much sex they are willing to have with that person.

A related issue is that both the theoretical model in Section 1.2 and the estimates in Section 1.4 assume that people can independently choose how much sex they have. In reality, sexual activity is a matching market, and people must find willing partners in order to have sex. I can close the model by assuming that people have a number of opportunities for sexual activity, and can choose how many to take advantage of, with their choices ranging from zero to some upper bound. My estimated effects can then be interpreted as the partial equilibrium effect of changing the risk beliefs of a single person, or a small number of people within the community. The general-equilibrium effect of changing everyone's beliefs would differ, and depend on how people sort into couples by their initial risk beliefs. In an additional set of analyses (not shown), I find no differential responses by village size, suggesting that the sexual markets are broader than individual villages. Thus my results are not affected by these general-equilibrium issues. I also find no differences in treatment effects by respondents' number of lifetime sex partners nor by the length of time they have been in their current relationship.

A separate potential limitation of this paper is that it relies almost exclusively on self-reported sexual behavior as a measure of sexual risk-taking. This could conceivably bias my results, but in my specific context there is no reason to believe that there would be differential social-desirability bias across study arms: the information treatment provided no direct modeling of "good" behavior nor encouragement to behave in a specific way. While Baird et al. (2012) find that self-reports do not yield accurate estimates of treatment effects, they study a specific treatment that may have led to differential self-report bias. Their intervention was focused on keeping girls in school, and one of the treatment arms conditioned cash transfers on school attendance. It is commonly believed in Malawi that girls who become sexually active automatically drop out of school (Grant 2012). Thus respondents who are being incentivized to stay in school may be reluctant to reveal that they are having sex. In contrast, treatments where there is no reason to expect differential self-report bias have fewer problems: de Walque, Dow and Gong (2014) find that STI incidence measures and

self-reports yield similar estimates of the effect of economic shocks on sexual activity. Beyond concerns about social-desirability bias being minimal, my approach also has the advantage of capturing changes in behavior among high-risk individuals. This cannot be done when using STIs as outcome measures unless treatable STIs are used and individuals are treated for existing STIs at baseline.

A final potential limitation is that my analyses of heterogeneous treatment effects are potentially subject to the Deaton (2009) critique that subgroup analyses can constitute ex post "fishing expeditions." However, that concern is mitigated due to the fact that my main theoretical results were laid out in earlier work done prior to the experiment (Kerwin 2012). I also use the same primary outcome variable as well as the same risk belief variable as I employed in the preliminary empirical analysis in that paper, limiting the number of researcher degrees of freedom involved in my analysis.

1.5.4 Implications for HIV Prevention Policy

The randomized treatment provided by my experiment – information about the true risk of HIV transmission – slightly increases sexual activity for most people, but sharply decreased it for people with the highest risk beliefs. The effect of the information treatment on overall HIV transmissions is therefore ambiguous: HIV transmission depends strongly on high-activity groups, who are responsible for keeping the epidemic alive and spreading it to the rest of the population (Koopman, Simon and Riolo 2005). Determining the overall effect my information treatment on the HIV epidemic would require detailed knowledge of the epidemiological model for the virus in my region, and is beyond the scope of this paper. However, it is informative to look at how risk factors for HIV transmission vary with the baseline beliefs that determine who responds fatalistically to the information treatment.

Figure 1.11 presents this analysis for four variables that are significant determinants of HIV prevalence and spread: age, total years of sexual activity, total lifetime sex partners, and perceiving that one may be HIV-positive. All four are positively correlated with risk beliefs, and the fatalistic group is significantly higher than the lowest risk belief category at the 0.10 level for all of them and at the 0.05 level for three of them. This suggests that people with extremely high risk beliefs may be crucial for the HIV epidemic, and that even if the information treatment increases the sexual activity of most people, it may decrease the overall spread of the virus by reducing risk-taking in this key group. A targeted information campaign, that restricted access to the information only to fatalistic people, could be even more beneficial; however, it may be difficult to prevent the information from spreading to other groups.

1.6 Conclusion

Empirical research on behavioral responses to health risks has traditionally assumed that responses are uniformly self-protective, and has therefore focused on mean elasticities as summaries of risk compensation across a population. I use a randomized field experiment in rural southern Malawi to explore the validity of this assumption in the context of behavioral responses to HIV infection risks. The experiment provided the treatment group with information on the true risk of HIV transmission from unprotected sex with an infected partner, which is much lower than most respondents thought. I find that the mean elasticity of sexual behavior with respect to HIV risk beliefs is small but statistically significant, with an elasticity of about -0.6. This is similar in magnitude to estimated responses to changes in HIV prevalence in the United States, and larger than previous estimates of prevalence elasticities in sub-Saharan Africa. However, because people do not accurately know the prevalence of HIV, I would expect the corresponding prevalence elasticity for my sample to be smaller, and possibly in line with the small measured responses for Africa. I develop a method to allow for heterogeneity in marginal effects (as opposed to just the reduced form effect of the treatment indicator) and find that the average marginal effect masks significant heterogeneity. The effect of risk beliefs on risky sex is negative (consistent with self-protective responses) for people who initially hold low risk beliefs, and becomes positive (consistent with fatalism) as initial risk beliefs become sufficiently high.

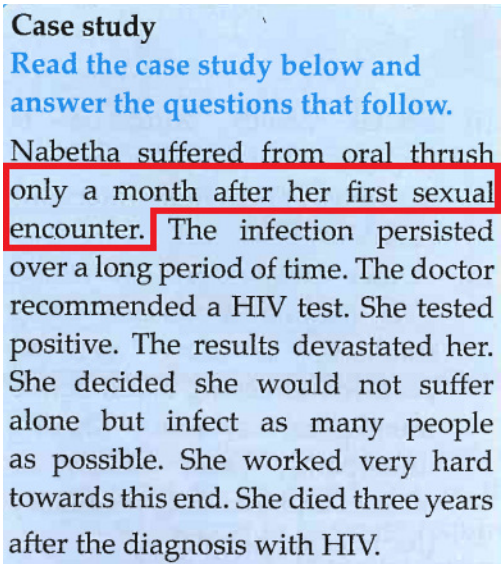
This heterogeneity is consistent with a model of rationally fatalistic behavior in which changes in perceived risks affect agent's choices not only via the risky sex acts being chosen at present, but also through a stock of previous – or unavoidable future – risky sex acts. A rise in the per-act risk increases the marginal cost of more risky sex due to the first channel, but also raises the chance that HIV is simply unavoidable, which lowers the marginal cost of additional risk-taking. I show that for this population, fatalistic responses appear to be driven not only by people who think they already have HIV, but also by those who believe that they are doomed to contract HIV in the future - for example, because of condom breaks. Moreover, even people who test negative now may maintain high priors about their chance of contracting HIV in the future, due to their exaggerated beliefs about HIV transmission rates. This suggests that HIV testing alone may not be sufficient to eliminate fatalism.

My results imply that the use of mean marginal effects as a way to summarize the response of health behaviors to health risks may be misleading. In the case of HIV in particular, epidemiologists have found that aggregate HIV transmission is dominated by high-sexual activity individuals. As a result, the effect of an increase in the perceived risk of HIV infection on the prevalence of the virus will depend predominantly on the response of

people with high sexual activity. If these individuals are fatalistic, the effect on prevalence may be the opposite of that implied by the mean marginal effect. My data suggests that this may in fact be true for HIV in Malawi: the 13.8% of people who respond fatalistically to the information treatment have an average of 4.4 lifetime sex partners, significantly higher than the rest of the population (p=0.07); they look worse in terms of other HIV risk factors as well. The extent to which mean marginal effects are a useful summary statistic for risk compensation for other health risks will depend on how many people hold extreme risk beliefs, whether the condition in question is incurable, and the dynamics of the broader economic or epidemiological system in which people are interacting.

Further research is needed on explicitly incorporating agents' perceived risk of HIV infection into rational epidemic models of HIV, rather than just assuming agents understand the true prevalence and transmission rate of the virus. Such models should also allow for responses to perceived risks to be heterogeneous by the level of the perceived risk, rather than imposing that they are the same across the whole population. The formation of people's risk beliefs is another important area for study. While anecdotal evidence suggests that people learn about HIV in school, the exact process by which many people arrive at gross overestimates of the prevalence and transmission rate of the virus is still unknown. Given that overestimating HIV risks seems to scare people to death, rather than scaring them straight, getting at the source of these overestimates may be crucial for understanding the continued spread of the African HIV epidemic.

Figure 1.1
Example of HIV Risk Messaging from a Malawian Life Skills Textbook



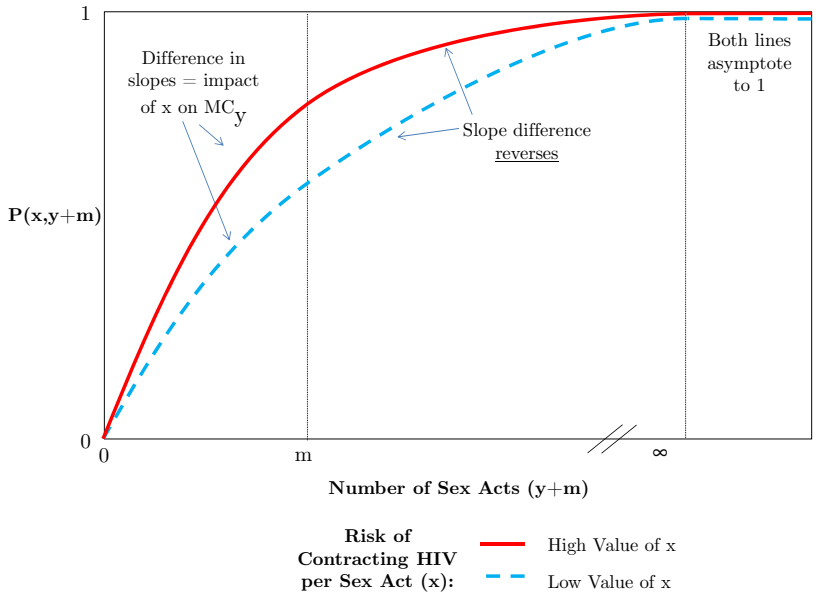
Case study

Read the case study below and answer the questions that follow.

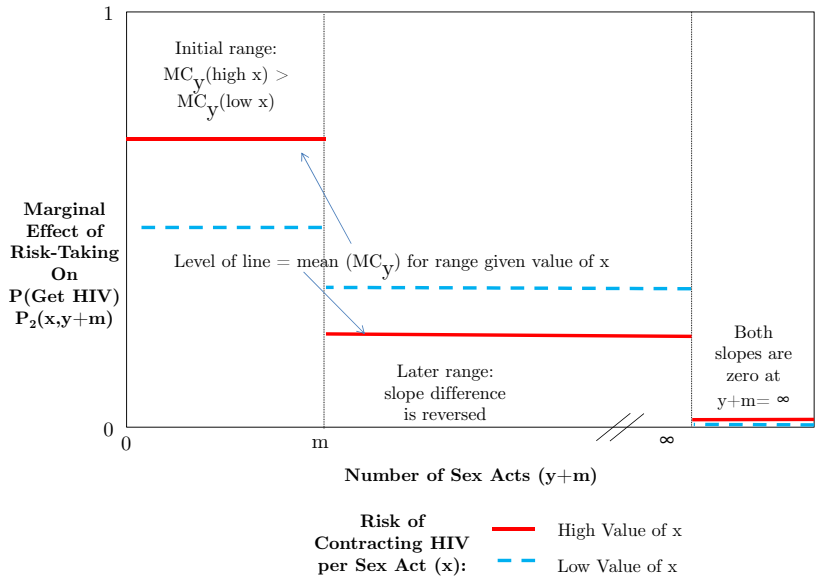
Nabetha suffered from oral thrush only a month after her first sexual encounter. The infection persisted over a long period of time. The doctor recommended a HIV test. She tested positive. The results devastated her. She decided she would not suffer alone but infect as many people as possible. She worked very hard towards this end. She died three years after the diagnosis with HIV.

Notes: Excerpted from the Form 4 Life Skills textbook used in Malawian secondary schools (highlighting added). The highlighted section suggests that the risk of contracting HIV from a single sex act is 100%. Author's conversations with Malawi Ministry of Education officials confirm that the Life Skills course taught from Form 1 to Form 4 (the equivalent of US high school) is the only course that covers HIV in the country's school system; this was the only explicit or implicit reference to HIV transmission rates found through an exhaustive review of the Life Skills text books and official curriculum.

Figure 1.2
Illustration of Tipping Point in Marginal Cost of Sexual Activity



Panel A: P[HIV Infection—Number of Sex Acts (y)] for Low and High Values of Per-Act Risk (x)



Panel B: MC[Sex Act (y)—(y)] for Low and High Values of Per-Act Risk (x)

Notes: Panel A illustrates the total probability of HIV infection, as a function of the number of sex acts chosen, y, for different levels of the per-act risk, x. The solid line is initially steeper because the chance of contracting HIV from each act is higher. Both lines asymptote to 1 as y goes to infinity; continuity and monotonicity therefore ensure that there exists a range (and hence at least one point) where the blue line is steeper. This leads to a tipping point combination of y and x: below the tipping point, the marginal cost of risky sex is higher when per-act risks are higher, while above the tipping point the marginal cost is *lower* when the per-act risk is higher.
Panel B directly illustrates the average marginal costs for different ranges of y given the two levels of the per-act risk; the mean marginal cost is higher for the lower per-act risk in the second portion of the graph, which is what generates the fatalistic range of responses.

Figure 1.3
Example Question about Subject's HIV Risk Beliefs

E1a. If 100 men, who do **not** have HIV, each sleep with a woman who is HIV positive tonight and do **not** use a condom, how many of them do you think will have HIV after the night?

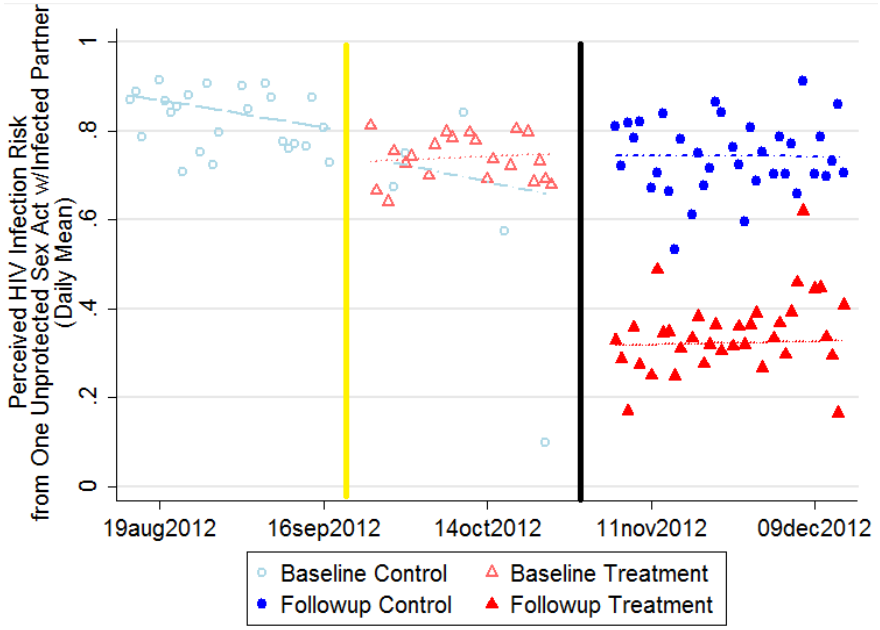
Number: [#] [#] [#]

E1b. **If answer to E1a is 50** Do you really think that 50 of the men would get HIV, or are you just not sure?

[ ] 1. I really think it's 50 [ ] 0. I'm just not sure → What is your best guess? [#] [#] [#]

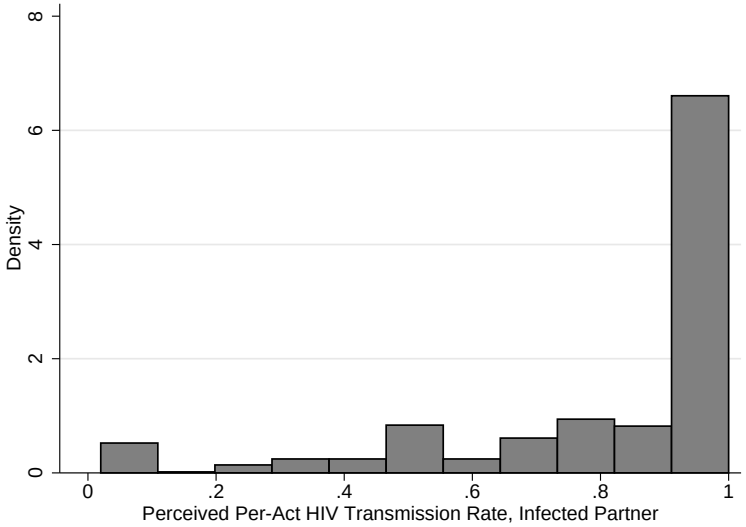
Notes: Example of one of the six different HIV-related expectations questions included on the survey. Enumerators were trained to ask a followup question along the lines of E1b if respondents answered 50% to any question; the data used in this paper replaces the initial response of 50 with the best guess if one was volunteered. The actual survey was conducted in Chichewa, the local language in southern Malawi; questions were translated by bilingual experts, tested extensively, and backtranslated to ensure accuracy.

Figure 1.4
Measured Risk Beliefs over Time, by Study Arm
(Per-act HIV transmission rate for unprotected sex w/infected partner)

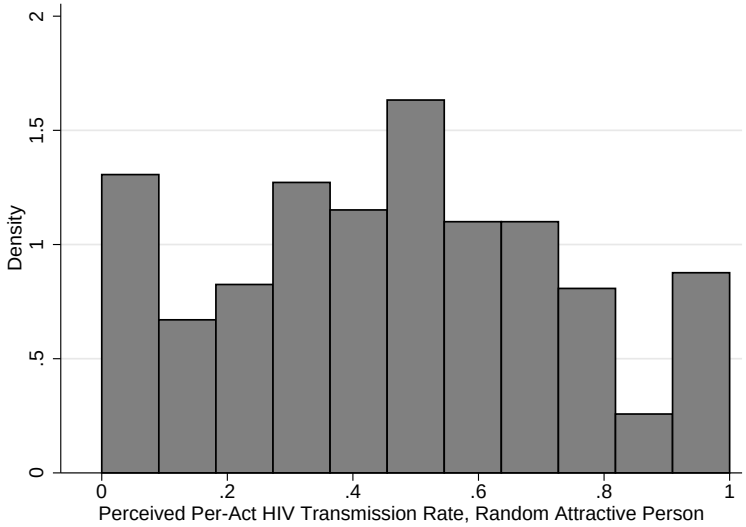


Notes: Each point represents the mean value of the risk beliefs for a given day; baseline control beliefs are hollow circles, endline control beliefs are solid circles, baseline treatment beliefs are hollow triangles, and endline treatment beliefs are solid triangles. The lines are linear fits of beliefs on date for a given date range and study arm. The light vertical line indicates the date of the training sessions when the survey enumerators were trained to provide the information treatment about HIV transmission risks. As shown on the plot, control-group baseline surveys were all conducted prior to this training session, with the exception of a handful of cleanup surveys. The pattern of Baseline beliefs suggests that the enumerators' knowledge about the information treatment affected the data they recorded in the surveys. This theory is supported by a comparison of the Baseline Treatment beliefs (hollow triangles) with the Endline Control beliefs (solid circles). This compares the groups when both the respondents and enumerators had identical information sets: it was after the enumerators were taught the HIV risk information, the baseline survey took place before treatment-group respondents were exposed to the information treatment, and the control-group respondents were never exposed to the information treatment. The post-training session cleanup surveys for the control group also lend support to this theory (the low outlier comes from a day with just a single cleanup survey). Sample is 1292 people from 70 villages for whom both baseline and endline surveys were successfully completed.

Figure 1.5
Histograms of Baseline HIV Infection Risk Beliefs, Control Group



Panel A: Per-Act Infection Risk from Unprotected Sex with an Infected Partner

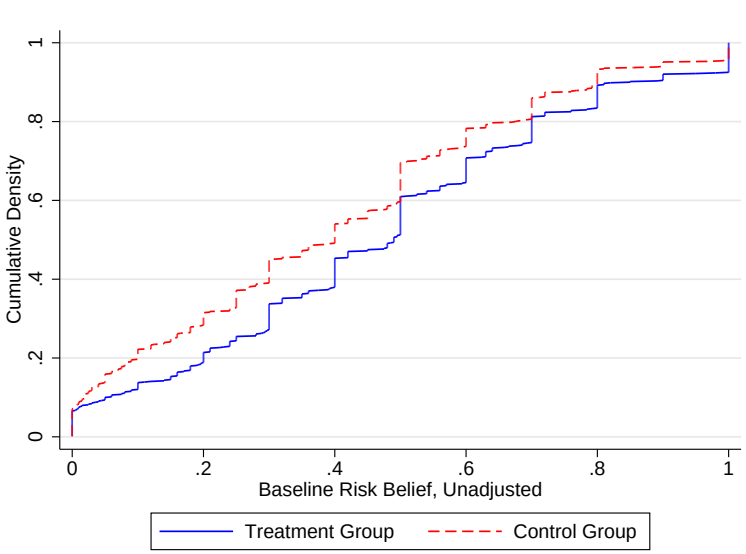


Panel B: Per-Act Infection Risk from Unprotected Sex with a Randomly-Selected Partner

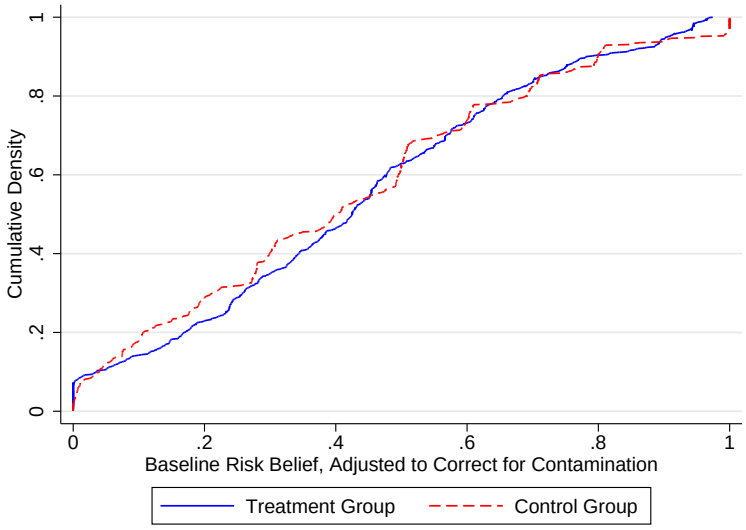
Notes: The two histograms plot the distribution of beliefs about the chance of contracting HIV from unprotected sex with either an infected partner (Panel A) or a randomly-selected person the respondent finds attractive (Panel B). Panel A has a large mass point at 100%. Panel B breaks up that mass point by accounting for the risk people perceive from unprotected sex with a randomly-selected partner, rather than conditioning on the partner being infected. Sample is 1292 people from 70 villages for whom both baseline and endline surveys were successfully completed.

Figure 1.6

CDFs of Baseline Beliefs about Per-Act HIV Infection Risk from a Random Attractive Sex Partner, by Study Arm



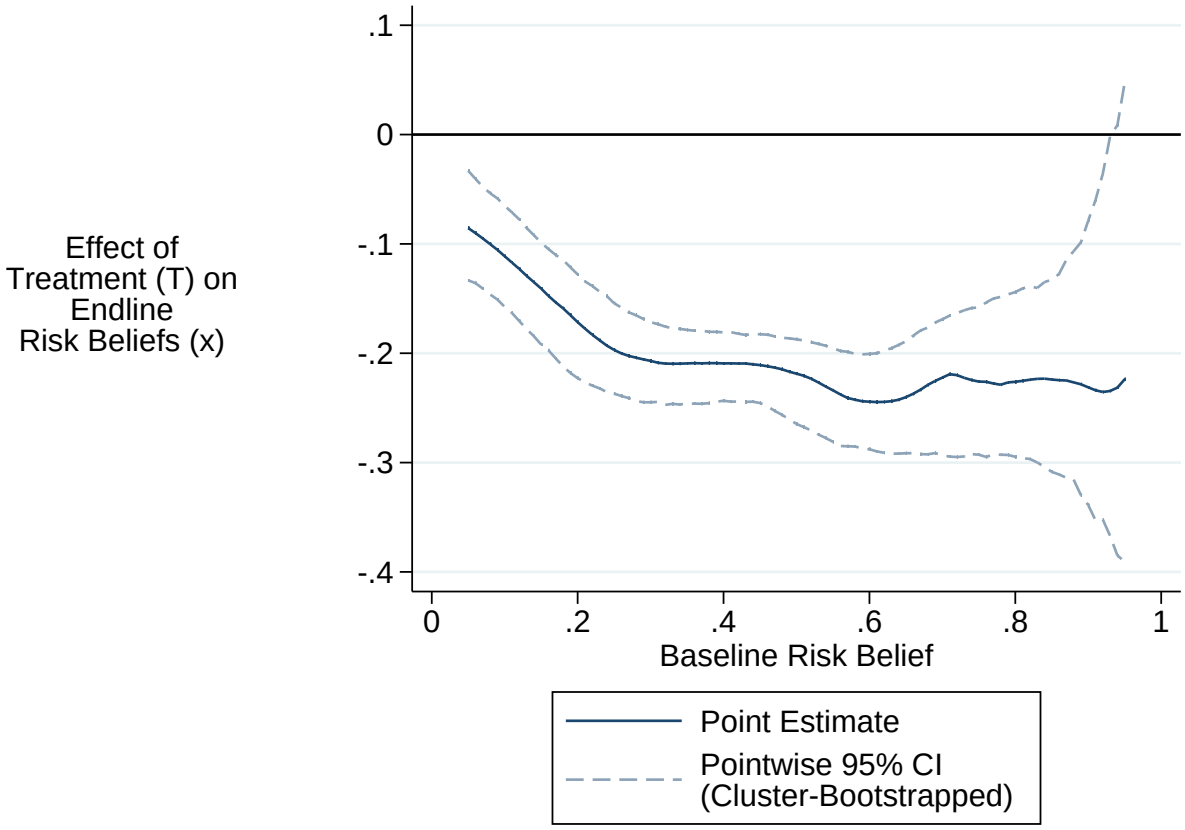
Panel A: Unadjusted



Panel B: Adjusted to Correct for Enumerator-Knowledge Contamination

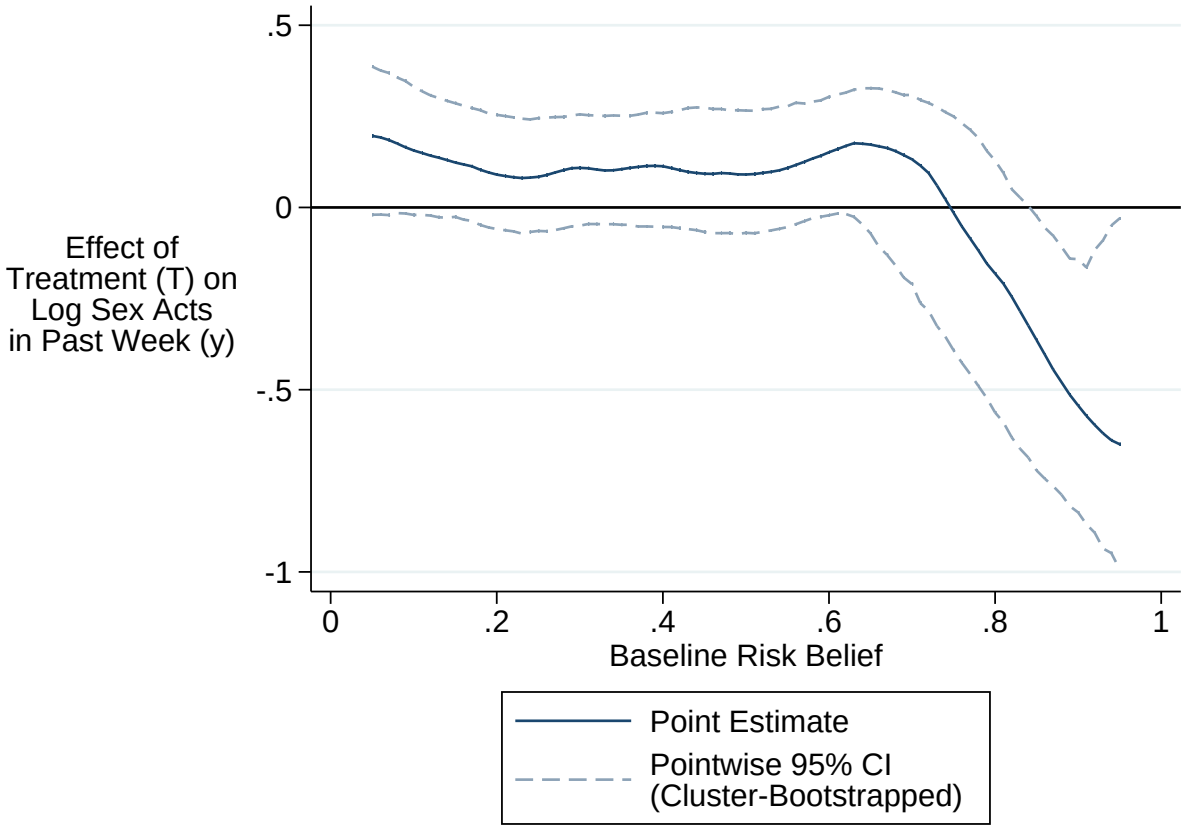
Notes: The two CDFs plot the distribution of beliefs about the chance of contracting HIV from a single unprotected sex act with a randomly-selected partner, separately for the treatment group (shaded red bars) and the control group (black outlined bars). This variable is constructed as $x_i = t_i * p_i$, where t_i is the perceived per-act HIV transmission rate from unprotected vaginal sex (for people of one's own gender) and p_i is the perceived prevalence of HIV among attractive members of the opposite sex from the local area. Panel A presents the raw data, while Panel B presents the data adjusted to correct for the contamination due to enumerator knowledge suggested by Figure 4. I run the regression $t_i = \beta_0 + \beta_1 AfterTraining_i + \beta_2 DaysAfterTraining_i + \beta_3 AfterTraining_i * DaysAfterTraining_i + \varepsilon_i$ and construct $t_i^{adj} = t_i^{resid} + \hat{\beta}_0$, bounding t_i to lie within $[0, 1]$. This preserves the scale on which the beliefs are measured. p_i^{adj} is constructed likewise, and x_i^{adj} is constructed as $x_i^{adj} = t_i^{adj} * p_i^{adj}$. A comparison of the two panels reveals that the adjustment mitigates the large excess of treatment-group respondents reporting beliefs in the lowest category, but does not perfectly harmonize the two distributions. Sample is 1292 people from 70 villages for whom both baseline and endline surveys were successfully completed.

Figure 1.7
First-Stage Effect of Treatment (T) on Endline Risk Beliefs (x),
by Baseline Risk Belief



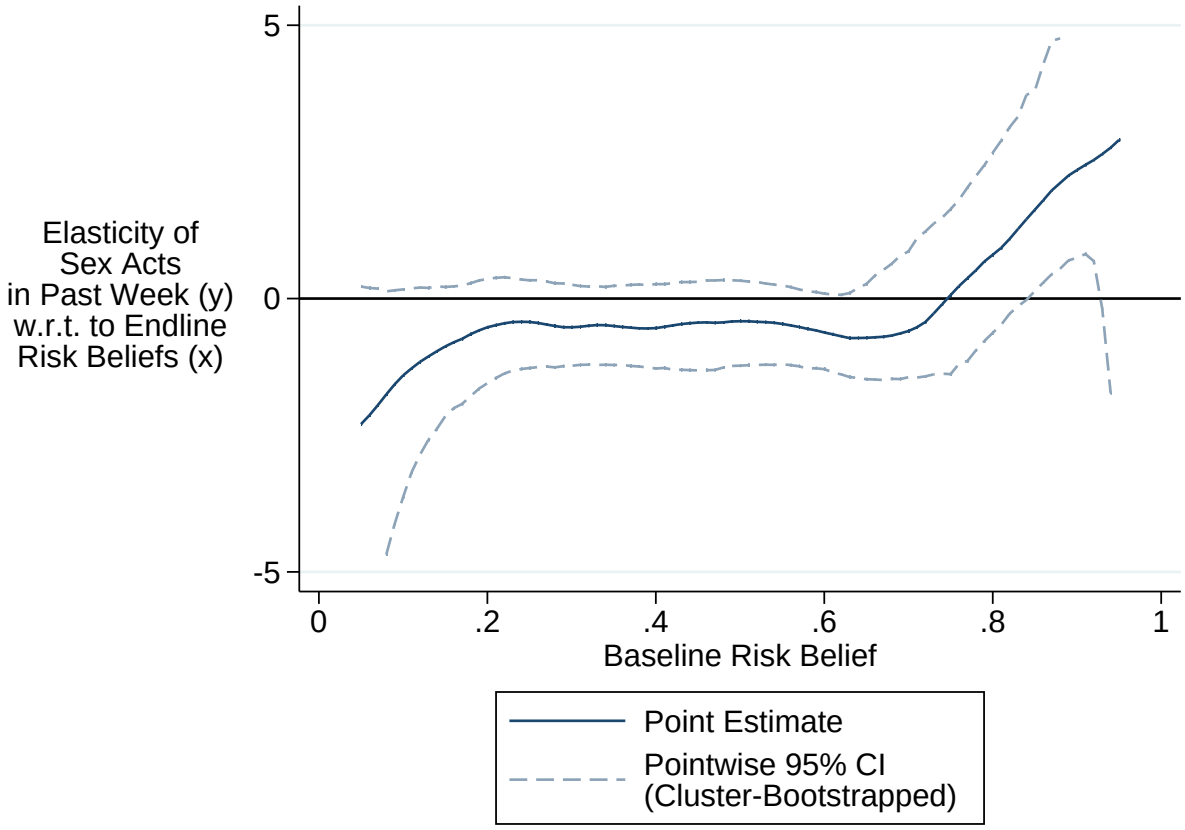
Notes: The graph illustrates the first-stage estimate of the effect of the information treatment on endline (post-treatment) risk beliefs, decomposed by individuals' baseline (pre-treatment) beliefs about HIV infection risks. The estimated effects on risk beliefs are negative for all levels of baseline beliefs because the true risk lies below the priors of virtually all respondents; the first stage is always negative, consistent with the monotonicity assumption. I estimate the underlying semiparametric regressions using the Robinson (1988) double-residual estimator to control for baseline values of the outcome and sampling strata; bandwidths are chosen to minimize the mean-squared error of the fitted values via the generalized cross-validation statistic of Loader (2004). See Section 1.4 for details on the estimation technique. The graph is restricted to Baseline Risk Belief values between 0.05 and 0.95 to mitigate boundary bias. Confidence intervals constructed via village-clustered bootstrap, with the Baseline Risk Belief variable re-generated for each resample to correct the confidence intervals for generated regressors. For each bootstrap sample, I trim observations with estimated densities below the minimum observed in the original sample. Baseline Risk Belief is the composite belief variable from Figure 6: the perceived chance of contracting HIV from a single unprotected sex act with a randomly-chosen attractive person of the opposite sex from the local area. Baseline Risk Belief is adjusted for non-constant time trends as in Panel B of Figure 6; omitting the adjustment does not change the qualitative results. Sample is 1292 people from 70 villages for whom both baseline and endline surveys were successfully completed.

Figure 1.8
Reduced-Form Effect of Treatment (T) on Log Sex Acts in Past Week ($\ln(y)$), by Baseline Risk Belief



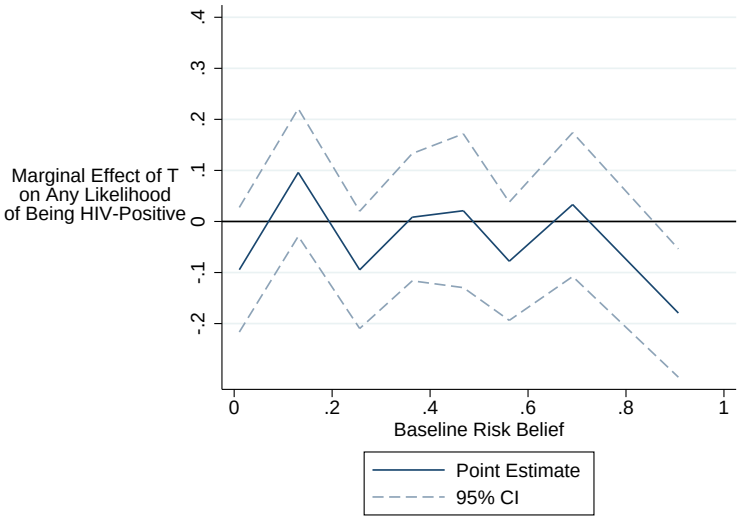
Notes: The graph illustrates the reduced form estimate of the effect of the information treatment on sexual behavior, decomposed by individuals' baseline (pre-treatment) beliefs about HIV infection risks. The treatment effect is positive for most respondents but negative for people with the highest initial beliefs, suggesting rationally fatalistic behavior. I estimate the underlying semiparametric regressions using the Robinson (1988) double-residual estimator to control for baseline values of the outcome and sampling strata; bandwidths are chosen to minimize the mean-squared error of the fitted values via the generalized cross-validation statistic of Loader (2004). See Section 1.4 for details on the estimation technique. The graph is restricted to Baseline Risk Belief values between 0.05 and 0.95 to mitigate boundary bias. Confidence intervals constructed via village-clustered bootstrap, with the Baseline Risk Belief variable re-generated for each resample to correct the confidence intervals for generated regressors. For each bootstrap sample, I trim observations with estimated densities below the minimum observed in the original sample. Log sex in past week constructed as $y' = \ln(y + \sqrt{(1+y^2)})$ to account for zeroes. Baseline Risk Belief is the composite belief variable from Figure 6: the perceived chance of contracting HIV from a single unprotected sex act with a randomly-chosen attractive person of the opposite sex from the local area. Baseline Risk Belief is adjusted for non-constant time trends as in Panel B of Figure 6; omitting the adjustment does not change the qualitative results. Sample is 1292 people from 70 villages for whom both baseline and endline surveys were successfully completed.

Figure 1.9
IV Estimates of the Elasticity of Sex Acts in Past Week (y) w.r.t. Endline Risk Beliefs (x), by Baseline Risk Belief

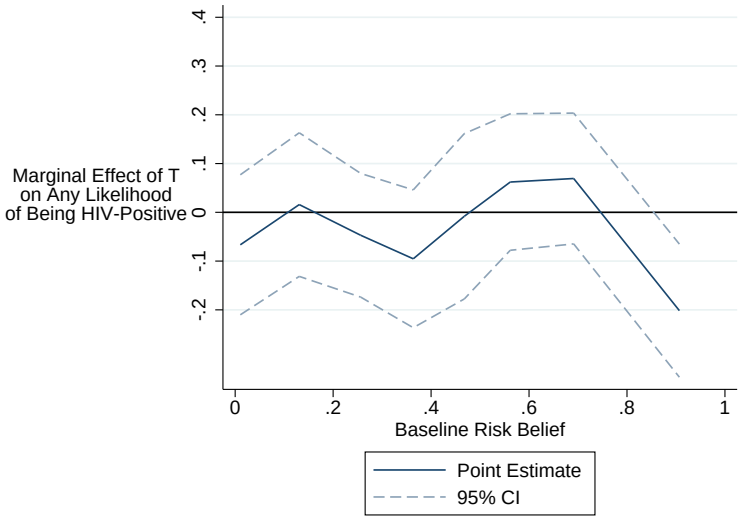


Notes: The graph illustrates the 2SLS estimate of the elasticity of sexual behavior with respect to endline (post-treatment) risk beliefs, decomposed by individuals' baseline (pre-treatment) beliefs about HIV infection risks. The estimated elasticity is negative for most people but positive for the highest baseline risk beliefs, consistent with rationally fatalistic behavior. I estimate the underlying semiparametric regressions using the Robinson (1988) double-residual estimator to control for baseline values of the outcome and sampling strata; bandwidths are chosen to minimize the mean-squared error of the fitted values via the generalized cross-validation statistic of Loader (2004). See Section 1.4 for details on the estimation technique. The graph is restricted to Baseline Risk Belief values between 0.05 and 0.95 to mitigate boundary bias. Confidence intervals constructed via village-clustered bootstrap, with the Baseline Risk Belief variable re-generated for each resample to correct the confidence intervals for generated regressors. For each bootstrap sample, I trim observations with estimated densities below the minimum observed in the original sample. Log sex in past week constructed as $y' = \ln(y + \sqrt{(1 + y^2)})$ to account for zeroes. Baseline Risk Belief is the composite belief variable from Figure 6: the perceived chance of contracting HIV from a single unprotected sex act with a randomly-chosen attractive person of the opposite sex from the local area. Baseline Risk Belief is adjusted for non-constant time trends as in Panel B of Figure 6; omitting the adjustment does not change the qualitative results. Sample is 1292 people from 70 villages for whom both baseline and endline surveys were successfully completed.

Figure 1.10
Multinomial Logit Estimates of Effect of Treatment on Perceived Likelihood of Having HIV Now (Panel A) in the Future (Panel B), by Baseline HIV Transmission Risk Belief



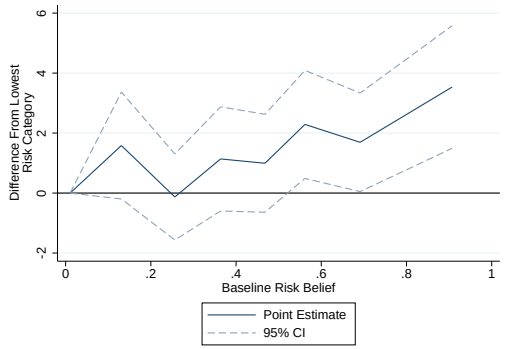
Panel A: Perceived Likelihood of Having HIV Now



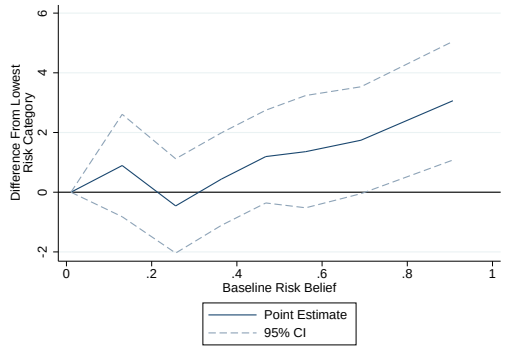
Panel B: Perceived Likelihood of Getting HIV in the Future

Notes: The graphs display the opposite of the mean marginal effects on the "No Likelihood" option from a multinomial logit of the categorical HIV status belief variable on a treatment indicator as well as controls for sampling strata and indicators for each category of the baseline value of the outcome; in Panel B no baseline data exists and so baseline data for "What is the likelihood that you have HIV now" are used as a proxy. Most changes are between some higher likelihood and "No Likelihood", thus the marginal effects for the latter summarize the effect of the information treatment, in terms of changes in believing there is any chance that one currently has HIV or will get it in the future. The treatment significantly increases the rate at which people reported any likelihood of having HIV now or getting it in the future for the highest category of risk beliefs, but had no effect for the rest of the population. This suggests that the mechanism of the risk-seeking responses observed in the sample is consistent with the model of rationally fatalistic responses laid out in Section 1.2. The results are not changed qualitatively if the "Don't know" category is excluded. Baseline Risk Belief is the composite belief variable from Figure 6: the perceived chance of contracting HIV from a single unprotected sex act with a randomly-chosen attractive person of the opposite sex from the local area. This is adjusted for non-constant time trends as in Panel B of Figure 6. Sample includes 1292 respondents who completed both baseline and endline surveys.

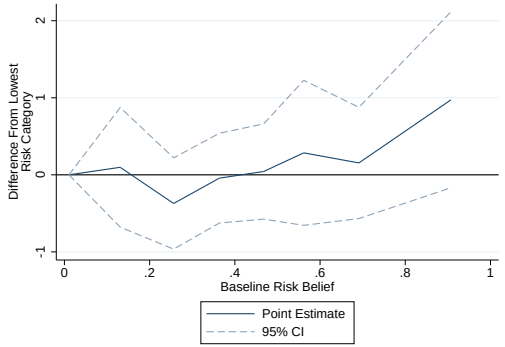
Figure 1.11
Differences in HIV Risk Factors by Baseline HIV Transmission Risk Belief



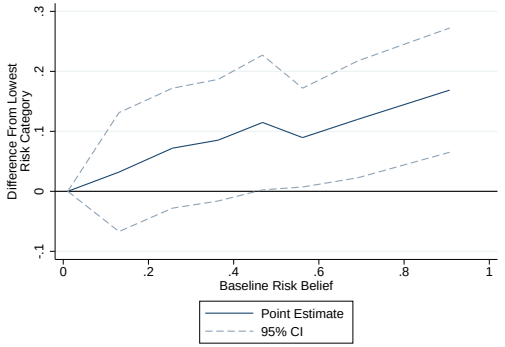
Panel A: Age



Panel B: Years Sexually Active



Panel C: Lifetime Sex Partners



Panel D: Perceives Any Likelihood of Being HIV-Positive

Notes: The graphs display the differences in baseline HIV risk factors between each risk category and the lowest one. People with the highest risk beliefs have consistently higher values for each risk factor; for all four graphs, the highest category is significantly different from the lowest category at the 0.10 level, and for three of the four the difference is significant at the 0.05 level. Baseline Risk Belief is the composite belief variable from Figure 6: the perceived chance of contracting HIV from a single unprotected sex act with a randomly-chosen attractive person of the opposite sex from the local area. This is adjusted for non-constant time trends as in Panel B of Figure 6. Sample includes 1292 respondents who completed both baseline and endline surveys.

Table 1.1
Demographic Covariate Baseline Balance

	N	Overall	Control	Treatment	C-T
	(1)	(2)	(3)	(4)	(5)
Demographics					
Male	1292	0.43	0.42	0.44	-0.01
Married	1290	0.82	0.83	0.80	0.03
Age	1292	29.36	29.13	29.59	-0.46
Grew up in village where currently residing	1289	0.62	0.65	0.60	0.05
Years of education	1292	5.81	5.76	5.86	-0.10
Number of people in household	1292	4.95	5.04	4.87	0.17
Total children still living	1292	2.99	2.94	3.05	-0.11
Desired future children	1289	1.36	1.31	1.41	-0.09
# media sources† used at least monthly	1292	1.18	1.16	1.20	-0.04
# common assets owned by household	1291	4.40	4.54	4.26	0.28
Household cash income past 30 days (PPP USD)					
Baseline (C and T observed at different times of year)	1292	250.29	282.46	218.23	64.23**‡
Endline (C and T observed simultaneously)	1292	190.28	201.94	178.66	23.29
Household expenditure past 30 days (PPP USD)	1292	292.70	292.39	293.01	-0.62
Religion					
Muslim	1292	0.07	0.09	0.06	0.02
Christian	1292	0.89	0.89	0.89	-0.01
Other	1292	0.04	0.03	0.05	-0.02
Ethnic Group					
Nyanja	1292	0.47	0.46	0.48	-0.02
Lomwe	1292	0.37	0.34	0.39	-0.05
Yao	1292	0.09	0.11	0.07	0.04
Chewa	1292	0.04	0.05	0.03	0.02
Other	1292	0.03	0.04	0.02	0.02

Notes: The t-tests shown in this table demonstrate that the sample is balanced on all observable demographics. The exception is income receipt at baseline due to seasonality; see (‡) below.
† Media sources are newspapers, radio, and television.
‡ Baseline income differs between treatment and control respondents due to seasonal patterns in income receipt. Endline income is not significantly different for the two groups; baseline expenditure is also almost equal as a result of consumption smoothing.
Sample is 1292 people from 70 villages for whom both baseline and endline surveys were successfully completed. Cluster-adjusted significance tests: * p< 0.1; ** p< 0.05; *** p<0.01.

Table 1.2
Sexual Activity Baseline Balance

	N	Overall	Control	Treatment	C-T
	(1)	(2)	(3)	(4)	(5)
Panel A - Single-Question Recall					
Years since sexual debut	1275	13.15	13.10	13.20	-0.10
Total lifetime sex partners	1288	3.34	3.12	3.56	-0.44**
Months since last sex act	1252	4.98	4.73	5.23	-0.50
Any sex in the past 30 days	1281	0.73	0.74	0.73	0.01
Sex partners during past 30 days	1290	0.81	0.82	0.80	0.02
Total sex acts during past 30 days	1281	7.37	7.48	7.27	0.21
Any unpro. sex acts in the past 30 days	1281	0.67	0.67	0.66	0.00
Total unpro. sex acts in the past 30 days	1281	6.66	6.75	6.57	0.18
Panel B - Retrospective Sex Diary - Sex Acts in Past 7 Days					
Any sex acts	1292	0.52	0.54	0.51	0.03
Total sex acts	1292	1.71	1.80	1.62	0.18
Any unpro. sex acts	1292	0.47	0.47	0.47	0.01
Total unpro. sex acts	1292	1.52	1.57	1.47	0.10
Sex with more than one partner	1292	0.01	0.02	0.01	0.01
Total sex acts with non-primary partners	1292	0.02	0.03	0.01	0.02
Any unpro. sex acts with non-primary partners	1292	0.01	0.01	0.00	0.00
Total unpro. sex with non-primary partners	1292	0.01	0.01	0.01	0.00

Notes: The t-tests presented in Column 5 suggest that the treatment and control group are well-balanced on observed sexual behavior. Because there are small differences between the two groups, however, controlling for baseline values of the outcome will reduce in less-biased regression estimates of treatmetn effects. Panel A shows data collected by the standard single-question recall method. Panel B shows data collected by a retrospective sex "diary" that walks respondents through the previous 7 days and asks them questions about a range of activities, both sexual and non-sexual, and collects details for each sex act. Sample is 1292 people from 70 villages for whom both baseline and endline surveys were successfully completed. Cluster-adjusted significance tests: * p< 0.1; ** p< 0.05; *** p<0.01.

Table 1.3

Regression Estimates of Effect of HIV Transmission Rate Information on HIV Risk Beliefs

	Perceived HIV Transmission Rate, if Partner Infected				Perceived HIV Prevalence		Composite Beliefs: P(Contract HIV from Unpro. Sex w/Random Attractive Person‡)	
	One Act		One Year†		All Local People‡	Attractive Local People‡	One Act	One Year†
	Unprotected	W/Condom	Unprotected	W/Condom				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A – Differences at Endline, without Controlling for Baseline Beliefs								
Treatment Group	-0.408***	-0.048***	-0.381***	-0.089***	-0.163***	-0.055***	-0.197***	-0.193***
	(0.019)	(0.006)	(0.016)	(0.013)	(0.018)	(0.018)	(0.015)	(0.016)
Observations	1,284	1,284	1,284	1,284	1,269	1,269	1,268	1,269
Adjusted R-squared	0.273	0.036	0.300	0.061	0.074	0.012	0.144	0.134
Panel B – Differences at Endline, Controlling for Baseline Beliefs								
Treatment Group	-0.384***	-0.045***	-0.371***	-0.071***	-0.162***	-0.047***	-0.182***	-0.185***
	(0.019)	(0.006)	(0.016)	(0.012)	(0.016)	(0.015)	(0.014)	(0.015)
Observations	1,281	1,283	1,276	1,276	1,257	1,254	1,252	1,251
Adjusted R-squared	0.315	0.066	0.328	0.142	0.157	0.081	0.200	0.182
Panel C – Difference-in-Differences								
Treatment Group	-0.316***	-0.022*	-0.336***	-0.010	-0.154***	-0.028	-0.127***	-0.148***
	(0.023)	(0.011)	(0.018)	(0.016)	(0.018)	(0.019)	(0.020)	(0.020)
Observations	1,281	1,283	1,276	1,276	1,257	1,254	1,252	1,251
Adjusted R-squared	0.149	0.002	0.225	0.008	0.046	0.006	0.049	0.066
Control Mean(Dep. Var)	0.742	0.082	0.905	0.176	0.485	0.463	0.351	0.424
Control SD(Dep. Var)	0.318	0.162	0.198	0.264	0.290	0.265	0.268	0.263

Notes: This table shows the information treatment has a strong negative effect on HIV risk beliefs that is robust different regression specifications. The treatment group received this information while the control group did not. Respondents update all their HIV-related beliefs, not just the one covered by the information treatment (the annual risk of infection from unprotected sex with an infected partner). This suggests that people learned and processed the information, and updated their other beliefs based on their new knowledge. All regressions include controls for sampling strata (distance category X gender). Panel A uses a simple regression of the endline value of the belief variable; Panel B adds controls for raw baseline values of the belief variable (not adjusted for enumerator contamination); Panel C uses the change in the belief variable from baseline to endline as the outcome.

† The question asked respondents to imagine couples having typical sexual behavior over the course of one year.

‡ Prevalence belief variables are questions specifically about members of the opposite sex.

Sample includes 1292 respondents who completed both baseline and endline surveys. Heteroskedasticity-robust standard errors, clustered by village, in parentheses. * p< 0.1; ** p< 0.05; *** p< 0.01

Table 1.4

Regression Estimates of the Effect of Information about HIV Transmission Risks on Sexual Behavior

	Any Sex in Past Week	Log Sex Acts in Past Week	Log Unprotected Sex Acts in Past Week	Log Sex Partners in Past 30 Days	Log Condoms acquired in past 30 days	Log Condoms Purchased	Log Overall Sexual Activity Index†	Log Diary Sexual Activity Index†
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment Group	0.050**	0.101**	0.071	0.012	0.080	0.054	0.063*	0.057**
	(0.024)	(0.047)	(0.045)	(0.019)	(0.075)	(0.105)	(0.032)	(0.024)
Observations	1,292	1,292	1,292	1,290	1,283	1,286	1,261	1,292
Adjusted R-squared	0.238	0.277	0.260	0.288	0.140	0.047	0.378	0.225
Ctrl Mean(Dep. Var)	0.490	1.67	1.48	0.77	2.52	5.08	-0.03	-0.02
Ctrl SD(Dep. Var)	0.500	2.39	2.29	0.58	9.65	6.59	0.99	1.03

Notes: Results illustrate that the information treatment had a small but statistically-significant effect on sexual behavior, increasing risky activities by 5 to 10 percentage points for most outcomes. I can reject effects above 20 percentage points in magnitude.

† The Sexual Activity Index variables are weighted averages of normalized values of all available outcome measures (Column 7) or just the outcomes measured on the Sex Diary, which are measured with less noise (Column 8). The weights used are factor loadings for the first principal component of the outcomes for the control group. Alternative indices using equal weights yield comparable, but slightly smaller, magnitudes.

Logged variables are constructed as $y'=\ln(y+\sqrt{1+y^2})$ to account for zeroes. All regressions include controls for sampling strata (distance category X gender). All regressions also control for baseline values of the outcome variable; the exception is Log Condoms Purchased (Column 6), where baseline Log Condoms Acquired in Past 30 Days was used as a proxy because condoms were not sold at baseline. Sample includes 1292 respondents who completed both baseline and endline surveys. Heteroskedasticity-robust standard errors, clustered by village, in parentheses. * $p< 0.1$; ** $p< 0.05$; *** $p<0.01$.

Table 1.5

OLS and 2SLS Estimates of the Partial Effect of Endline Risk Beliefs on Sexual Activity

	Any Sex in Past Week	Log Sex Acts in Past Week	Log Unprotected Sex Acts in Past Week	Log Sex Partners in Past 30 Days	Log Condoms acquired in past 30 days	Log Condoms Purchased	Log Overall Sexual Activity Index†	Log Diary Sexual Activity Index†
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: OLS Estimates (Control Group Only)								
Endline Risk Belief	0.155***	0.175*	0.106	0.196***	0.118	-0.337	0.318***	0.180**
	(0.054)	(0.102)	(0.103)	(0.058)	(0.172)	(0.224)	(0.100)	(0.078)
Observations	627	627	627	626	626	626	617	627
R-squared	0.210	0.277	0.240	0.258	0.165	0.049	0.340	0.219
Panel B: 2SLS Estimates								
Endline Risk Belief	-0.260**	-0.562**	-0.412*	-0.043	-0.375	-0.256	-0.327**	-0.317**
	(0.121)	(0.241)	(0.232)	(0.102)	(0.402)	(0.535)	(0.159)	(0.122)
Observations	1,252	1,252	1,252	1,250	1,243	1,246	1,222	1,252
R-squared	0.208	0.256	0.253	0.277	0.129	0.046	0.361	0.196
1^{st}-Stage F-Statistic	222.0	220.7	221.3	222.7	221.3	218.1	226.5	221.6

Notes: 2SLS estimates use the randomized treatment group assignment as an instrumental variable for endline beliefs. The results indicate that the elasticity of sexual activity with respect to HIV risk beliefs is between -0.3 and -0.6. OLS estimates use the endline data for the control group only, to estimate the relationship that would be observed in the absence of any exogenous variation in risk beliefs.

† The Sexual Activity Index variables are weighted averages of normalized values of all available outcome measures (Column 7) or just the outcomes measured on the Sex Diary, which are measured with less noise (Column 8). The weights used are factor loadings for the first principal component of the outcomes for the control group. Alternative indices using equal weights yield comparable, but slightly smaller, magnitudes.

Logged variables are constructed as $y'=\ln(y+\sqrt{1+y^2})$ to account for zeroes. Endline Risk Belief is the composite risk belief: the perceived chance of contracting HIV from a single unprotected sex act with a randomly-chosen attractive person of the opposite sex from the local area. All regressions include controls for sampling strata (distance category X gender) and baseline values of risk beliefs. All regressions also control for baseline values of the outcome variable; the exception is Log Condoms Purchased (Column 6), where baseline Log Condoms Acquired in Past 30 Days was used as a proxy because condoms were not sold at baseline. Sample includes 1292 respondents who completed both baseline and endline surveys. Heteroskedasticity-robust standard errors, clustered by village, in parentheses. * $p< 0.1$; ** $p< 0.05$; *** $p<0.01$.

Table 1.6

Non-Monotonic Responses to Information Treatment Effects by Baseline Risk Beliefs

	Outcome: Log Sex Acts in Past Week						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment (T)	0.101**	0.320***	0.123*	0.070	0.136**	0.095	0.412
	(0.047)	(0.083)	(0.072)	(0.057)	(0.060)	(0.062)	(0.309)
T*(Baseline Risk Belief [0-1])†		-0.499***					-.477***
		(0.162)					(0.168)
T*(Male)			-0.049				-0.039
			(0.131)				(0.155)
T*(Baseline Log Sex Acts in Past Wk.)				0.035			-0.075
				(0.051)			(0.112)
T*(Ever Exposed to HIV)					-0.151		-0.105
					(0.113)		(0.119)
T*(Any Chance I am HIV-positive)						0.009	0.075
						(0.115)	(0.122)
T Interacted with Other Baseline Covariates‡	No	No	No	No	No	No	Yes
Observations	1,292	1,275	1,292	1,292	1,275	1,277	1,245
R-squared	0.277	0.284	0.277	0.277	0.277	0.276	0.345

Notes: Results illustrate that there is substantial heterogeneity in responses to the information treatment by baseline HIV risk beliefs, but not by any other baseline covariate. This heterogeneity is robust to including interactions between the treatment indicator and a wide range of other baseline covariates.

†Baseline Risk Belief is the composite belief variable from Column 7 of Table 6: the perceived chance of contracting HIV from a single unprotected sex act with a randomly-chosen attractive person of the opposite sex from the local area. This is adjusted for non-constant time trends; omitting the adjustment does not change the qualitative results.

‡Other baseline covariates include immediate and delayed word recall [each 0-10], numeracy score [0-3], score on Raven's progressive matrices [0-3], lifetime sex partners, whether respondent had any sex in the past week, and indicators for marital status, age category, ethnic group, education level, frequency of listening to the radio, frequency of watching television, frequency of reading the newspaper.

All regressions include controls for baseline values of the outcome, and sampling strata (distance category X gender). In each specification, the factor being interacted with the treatment dummy also enters into the regression in levels. Logged variables are constructed as $y'=\ln(y+\sqrt{1+y^2})$ to account for zeroes. Sample includes 1292 respondents who completed both baseline and endline surveys. Heteroskedasticity-robust standard errors, clustered by village, in parentheses. * $p< 0.1$; ** $p< 0.05$; *** $p<0.01$. Standard errors in Columns 2 and 7 are cluster-bootstrapped to correct for generated regressors

CHAPTER II

Income Timing, Temptation and Expenditures: Field Experimental Evidence from Malawi

From a work with Lasse Brune.

2.1 Introduction

Savings rates in developing countries appear to be very low. People save little, whether in cash or other liquid assets. Moreover, despite evidently high returns to investment in domains ranging from health (Jones et al., 2003) to agriculture and small business (de Mel, McKenzie and Woodruff, 2008, 2012), people do not seem to be making those investments. In theory, even in the face of borrowing constraints, if returns are high enough households should be able to save up and invest. However, households appear to have trouble saving: households in developing countries act as if they are "savings constrained", meaning that shifting liquid wealth across time periods is costly.

Households in developing countries face a range of explicit and implicit "external" costs to savings, e.g. risk of theft, high transaction costs, lack of access to formal savings, or social pressure to share earned income or wealth.[1] In addition, savings constraints can be "internal" – people might be present-biased, causing them to save less than they would like. Present-biased preferences have been documented extensively in laboratory studies, and recent field research has confirmed that some people do exhibit present-biased preferences in the context of real-life choices (Giné et al., 2012). A number of papers have studied the potential of commitment savings accounts to manage this kind of internal savings constraint (Dupas and Robinson, 2013; Ashraf, Karlan and Yin, 2006; Brune et al., 2015).[2] However, the cause of

[1] e.g. Jakiela and Ozier (2012) and Goldberg (2011)

[2] In the developed world, research on self-control (e.g. Thaler and Shefrin (1981)) has identified Christmas clubs – savings accounts that pay no interest and lock up one's money until December 1st – as a form of commitment savings used to overcome internal savings constraints

present-biased preferences, and the best way to mitigate their impact on the poor's ability to save, remains unclear: in their review of the constraints that hinder savings among the poor, Karlan, Ratan and Zinman (2014) conclude that "remarkably little is known about which behavioral biases actually drive savings behavior." The canonical model of present bias is the Laibson (1997) model of quasi-hyperbolic discounting, but this sheds little light on why some people are present-biased and others are not. A possible explanation for variations in present bias comes from Banerjee and Mullainathan (2013, henceforth BM), who point out that one potential cause of variation in present bias is temptation: people may be biased toward present consumption because they are tempted to spend on goods and services that they later regret spending on, such as alcohol, tobacco, or fatty foods. Savings constraints could prevent people from saving up for large, discrete purchases (such as certain investments or durable goods), and could prevent people from having access to savings in the case of emergencies.

In light of this documented inability or unwillingness to save, the time structure of income streams is likely to be important. People in developing countries invest considerable effort and expenditure into aggregating streams of small installments of income into lump sums, in order to make purchases that cannot be broken up into small pieces (Collins et al., 2009). As a result, larger income installments may lead to more saving by easing this process. Lump-sum payments could also help savings under a BM-style temptation-based model of time inconsistency: BM show that having a larger sum of money on hand can help people overcome the fear that, if they do save, their future self will simply "waste" all of the money on temptation goods. This line of reasoning is consistent with previous research on self-control problems and lum-sum payments: Thaler and Shefrin (1981) argue that a worker who receives part of his salary in a lump-sum bonus (rather than always in equal monthly installments) would be able to save more, since typical rules-of-thumb used to constrain consumption would lead people to spend roughly the same amount as they make in each month.

However, an alternative possibility is that converting smooth income streams into larger, deferred sums will instead lead to increased temptation and potentially poor choices. Fudenberg and Levine (2006) note that ATMs are frequently placed in locations where lottery tickets are sold, or in nightclubs, in order to induce impulse purchases by myopic consumers. This proverbial effect of "money burning a hole in your pocket" is a potential concern in the microfinance industry, where recent research has studied whether access to microcredit can induce temptation spending due to the generation of large lump sums (Angelucci, Karlan and Zinman, 2013). In addition, this phenomenon is consistent with both theoretical and empirical work in developed countries (Ozdenoren, Salant and Silverman, 2012; Stephens Jr., 2003;

Shapiro, 2005) as well as with anecdotal reports of behavior around payday in developing countries.

In this paper we report results from a field experiment in Malawi designed to examine the role of timing of income for spending and savings decisions and its interaction with issues of self-control. We vary the time structure of wage payments for 363 casual laborers, with workers paid either in four weekly installments or a single lump sum at the end of the month. Our survey data demonstrates that this is a salient and potentially-important variation in how income is received. Before the start of the experiment we described to workers a non-incentivized, purely hypothetical situation in which they have two choices of wage payments: weekly payments or a lump sum payment in the end. Workers were informed that they would be required to come to the same location the same number of times (just as in the experiment we conducted; the hypothetical wage amounts were also nearly identical to the actual ones). 72% of workers said they preferred a lump sum payment. This preference appears to be related to savings constraints: of those 72%, a great majority (83%) stated, in an open ended question with at most one answer, that the reason for this preference is that enables people to "make a better plan" for the money, and an additional 13% openly stated that their reason was to avoid wasteful spending. These answers imply either a commitment problem as the reason for the lump sum preference, or at the least an expected inability to save – either due to internal constraints such as self-control problems or external constraints such as fear of theft.

This first treatment is cross-randomized against a second intervention in which we vary the day of the week on which workers are paid, with half of the sample being paid on Fridays, and half on Saturdays. All payments take place at the same location: the site of the local weekend market, which takes place on Saturdays and is reported to be an extremely tempting environment. Qualitative evidence from the study area found that people reported market days as tempting environments. This was confirmed by survey responses from our experimental sample. We also use respondents' own perceptions of regretted or mistaken expenditure, as reported on the surveys, as one of our measures of spending on temptation goods. While goods the respondents self-reported as regretted purchases included alcohol, tobacco, and sweets, the most common category was clothing. This is consistent with anecdotal reports from the local area: clothing is a major expenditure at the markets, with people making expensive purchases and then later regretting them. Workers who are paid on Saturdays are therefore exposed to a much more tempting environment at the time when they receive their pay (relative to members of the Friday group), with all other factors being held constant.[3]

[3] Friday was chosen as the control group, rather than Sunday or Monday, in order to eliminate the

Workers in all study arms receive the same total amount of money: about MK3000, or around 30% of their total cash income over the work period; they are employed in collaboration with a local NGO in two separate rounds of work that are followed by payments with re-randomization of experimental conditions after round 1. The travel and time costs of purchasing goods at the market are held constant across study arms by requiring attendance at the payday site by all participants on all potential paydays, even when they do not receive money.

The experiment has both a practical and a conceptual dimension: it was designed to evaluate the role of internal savings constraints in a practically relevant context – temptations to overspend on paydays and at weekend markets and local trading centers in particular – and to test conceptually the role of temptation in mediating the differential effects on spending of income stream frequency.

Research using randomized variation in the frequency of income streams is rare. To the best of our knowledge, the first experimental study of the effect of lump sum wage payments relative to smoother streams of labor income is Beegle, Galasso and Goldberg (2014), studying the Malawi Social Action Fund's Public Works Project. They compare outcomes for workers who receive their wages in a single lump sum against those of workers who are paid in 5 installments over the course of 15 days. The variation in the frequency of payments is cross-randomized with the season of employment. One important difference between the two seasons they study is that the marginal utility of immediate consumption is generally considered to be low in one season (agricultural investments are more important in the planting season) and high in the other (basic food consumption is more important in the lean season). Our study cross-randomizes the frequency of payments with whether payday is a market day, which anecdotally is considered to induce temptation for immediate spending, or whether payday is a non-market day. Hence we vary the immediate, short-run context in which pay is received rather than the larger seasonal context. In line with the differences in the exact type of variation in payment timing between the two studies the focus of data collection in our study is more short-term. While Beegle, Galasso and Goldberg (2014) collects consumption and expenditure data with a recall period of one week, within a period of up to a month after respondents receive their payments, our study documents short-run differences in spending and saving on the day of receipt of pay and immediately after. Along this dimension, therefore, our study can viewed as a short-run complement to Beegle, Galasso and Goldberg (2014). Another paper that randomizes whether income is received in lump sums is the Haushofer and Shapiro (2013) evaluation of the GiveDirectly

possibility that people in the Saturday group save less of their income simply because the time frame is longer.

program. The study randomizes the time structure of windfall income, rather than labor earnings, and looks at much longer-run changes in behavior, on the order of one year. They find a decrease in measured cortisol levels among people who receive annual lump-sum, as opposed to monthly installment, transfers, suggesting lower levels of stress.

This paper provides novel empirical evidence in three ways. First, we provide evidence that lump sum payments have an effect on purchases of an actual investment: a high-return, short-term "bond" offered by the project to all respondents. Second, we study the effect of the timing of payments within a week, which has not been examined in the previous literature. Third, we exploit the effect of the timing of payments within a week to explore the role of temptation in driving internal savings constraints.

The potential of temptation-driven waste due to market days, the frequency of payments and their interaction are not merely theoretical concerns. Many organizations in Malawi are presently moving to direct-deposit based payment schemes on an infrequent schedule that bring their employees to major cities on focal dates, potentially triggering the sorts of temptation issues discussed above. One example is Malawi's Ministry of Education; teachers now receive their pay via direct deposits into their bank accounts, as opposed to cash payments. This in turn induces a large fraction to travel to urban areas once a month to withdraw all their pay in a lump sum. A similar pattern holds for unconditional cash transfers like GiveDirectly: what makes that program logistically feasible is that the payments are sent through the M-Pesa mobile payments service. Haushofer and Shapiro (2013) state that GiveDirectly recipients "typically withdraw the entire balance of the transfer upon receipt." Since withdrawals must be done at a participating M-Pesa agent, this will tend to draw recipients to potentially-tempting trading centers at the same time as they receive their pay. This study evaluates how infrequent payments and payments on market days in particular influence spending decisions, for a highly-relevant category of income for people in rural Africa. Prior to the beginning of our study, 77% of our sample reported having done informal agricultural work; it is a more common source of cash income than any other activity except for selling one's own crops for cash. Our intervention also involves a smaller proportion of income these other contexts: GiveDirectly provided income worth more than two months of expenditures, and the Malawi Ministry of Education's direct deposit program covers all of a teacher's income. Our respondents received additional income worth approximately 50% of their existing cash income. This limits our ability to draw conclusions about the effect of changing the timing of larger proportions of income, but also means that our study more closely resembles realistic cash transfer programs for people in rural Africa, who are likely to have existing sources of cash income as well.

We present two sets of findings from the experiment: the effect of being paid during

the major local market, and the effect of being paid monthly rather than weekly. In our experimental context, being paid at the site of the local market during the market day, Saturday, does not strongly matter for expenditure decisions relative to being paid at the same location on a Friday – despite strong motivation from anecdotes and suggestive survey data. Drawing on a range of outcomes we document that neither the level nor composition of expenditures exhibits statistically-significant variation by the day of the week that people were paid, and that the frequency of payments does not affect this result. We focus on a set of outcomes related to spending at the market on each Friday and Saturday of the study, for which we can reject even moderate-sized effects of being paid on Saturdays relative to Fridays. However, some of our alternate outcome measures are noisy enough that we cannot conclude that the day of week of income receipt has moderate-sized effects. This result does not conclusively rule out important payday effects in settings other that of our specific experiment – we discuss external validity in the conclusion – and it does not necessarily imply that self-control more broadly is not a binding constraint for savings. The result should, however, lower our priors about the empirical relevance of the market payday effect, certainly in contexts that are similar to the ones of this study.

In contrast, we find strong effects on spending and savings patterns by payment frequency. While there is no evidence that the composition of expenditures (including in particular self-reported wasteful consumption) varies with payment frequency,[4] we do find strong evidence that the mode of payment frequency matters for workers' ability to benefit from high-return investment opportunities with a large minimum investment size. Workers in the monthly group have more cash left in the week after the last payday when the lump sum payment was made. Moreover, they are 9.5 percentage points more likely than the weekly payment group (a relative increase of 151% over the weekly mean of 6.3%) to invest in a risk-free short-term "bond" that required a large minimum installment size payment and that was offered by the project in the week after the last payday. The investment was returned to the respondent together with 33% interest after exactly two weeks. Workers knew about this opportunity before the beginning of round two of the experiment and had gained experience with the product in a pilot offer at the end of the first round.[5] In total, lump sum group workers spent about twice as much as weekly payment group workers on the investment opportunity. We

[4] We elaborate on the specific features of this experiment that maybe have mitigated potential effects in the discussion of the empirical results.

[5] We focus here on the effects for round 2 of the study, when all respondents knew about the possibility of purchasing the bond prior to receiving any payments or learning which study arm they were in. The results from round 1, in which the investment opportunity was anounced after three of the four weekly payments had been disbursed, are smaller and statistically insignificant. We discuss possible reasons for this difference in Section 5; the most likely explanation is that members of the monthly group had already committed their income to other purposes.

cannot entirely rule out borrowing constraints as an explanation for this result. However, based on other data, we argue that the result is driven by savings constraints.

These results, using a novel outcome measure for investments with a large minimum installment size, also make an important contribution to existing research on the relationship between savings constraints and high returns to investment. Previous research has found that the return to investment is high, but that people do not appear to make those investments – implying that people are constrained in their ability to save up for these investments. However, prior studies either have not measured objective returns (relying on e.g. purchases of health products), or have observed high average returns in a cross-section (e.g. cash drop experiments). Research that uses investments in health products as an outcome relies on the assumption that the return to health investment is actually high, and also that respondents understand these high returns. Cash drop experiments also do not necessarily show that people are failing to pursue high-return investments. Under heterogeneous returns and borrowing constraints it is possible to observe high average returns without a binding savings constraint. Those with access to high-return investments might be limited in how much they invest at any given time because they face either a) borrowing constraints or b) they prefer to not decrease present consumption too much. As a result, people do not take advantage of all their high-return investment opportunities, allowing high returns to persist over time. Our experiment resolves both of these concerns. First, we use an actual investment with high returns and zero risk as an outcome. Second, we ensure that returns are homogeneous. In our experiment everyone has access to the same high-return investment offer, but, compared to the lump sum group, the weekly group – who are otherwise identical due to randomization – need to save to be able to invest. We observe that they do invest, but to a much lesser extent. Thus this paper provides novel evidence for savings constraints being a relevant driver of the persistence of the observed high returns to capital in developing countries.

2.2 Study Design and Data

We designed a randomized experiment with informal agricultural workers from the Mulanje District of Southern Malawi. These workers took part in an expansion of an existing income-generation program that operates in Mulanje District. The subjects in the study received identical nominal[6] wages for their work, but were randomly assigned to receive the pay with different timing.

[6] The official inflation rate in Malawi was about 23% per annum during the study period (https://www.rbm.mw/inflation_rates_detailed.aspx), so prices would have risen just 1.7% per month. We therefore ignore the distinction between nominal and real wages for the purposes of our analysis.

We worked with the Mulanje District Executive Council to expand a previously-existing income-generation program to an additional 365 workers[7], who worked for a total of up to 15 days in two separate rounds of work and payments. This program was part of the Sustainable Livelihoods program run by Mulanje Mountain Conservation Trust (MMCT), an NGO based in Mulanje District that is focused on environmental protection and promoting sustainability in the Mulanje Mountain Forest Reserve and adjoining areas. MMCT provided detailed guidance on how to mirror their existing practices; as with the majority of MMCT's other projects, work oversight was conducted by officials from partnering government departments of Mulanje District.

The experiment was organized into two rounds that occurred over a period of three months from November 2013 to January 2014, with subjects randomized into treatment conditions separately by round. During each round, subjects worked for two weeks and then received their pay either a) in weekly installments beginning at the end of the second week of work; or b) in a single lump sum, about three weeks after the last day of work. Figure 2.1 shows the timing of the different components of the experiment: the two rounds of work and payments and the different rounds of data collection. In addition to variation in payment frequency, workers received their pay either c) on Fridays or d) on Saturdays. The two variations on the timing of pay – weekly vs. monthly and Friday vs. Saturday – were cross-randomized, creating four study arms in each round. The distribution of workers into experimental groups is shown in Table 2.1a (pooled) and Table 2.1b (separate by round); details of the randomization follow further below. The payments were made at the site of a major local market that occurs on Saturdays, with the intention of inducing variation in people's temptation to overspend. During the week after the last payday in each round, all workers were visited for a detailed survey about their expenditure and income.

2.2.1 Recruitment of Workers

We worked with MMCT to locate a set of villages that were potential targets for expanding their Sustainable Livelihoods program. The key criteria for a village to be eligible were:

1. *Location.* Villages had to lie within walking distance of the Forest Reserve, because the work activities supported by the program are centered around natural resource management and conservation.

[7] The original recruitment included 350 workers two of which dropped early (one never showed up for work; one never showed up to receive his wage); 15 workers were added for round 2 to replace workers who dropped out after the round 1.

2. *No previous Sustainable Livelihoods program participation.* Because this was an expansion of the program, we excluded areas that were already actively participating in the program, or which had been included in the past.

3. *Not included in any other recent income-generation programs.* The expansion was targeted toward underserved communities to maximize the benefits brought to the neediest people.

4. *Limited geographic range.* The villages for the study had to be physically close enough to each other to allow work and payroll to be organized across all of them together.

Given the criteria above, we settled on a region of Traditional Authority (TA) Nkanda near the Forest Reserve as the target location for the project; this area had not previously been included in the Sustainable Livelihoods program, nor recently participated in other major income-generating programs such as the Malawi government's Public Works Programme (PWP). Within that region, we picked seven villages that all lie within the catchment area of Mwanamulanje trading centre, one of the largest markets in TA Nkanda.

The selection of workers was handled by the standard operating procedure employed by the Sustainable Livelihoods program. The nature of the program, including the kind of work, the pay rate, and the expected length of employment, was explained at a meeting with the village head and the village development committee (VDC). Each VDC was then tasked with selecting a set of 50 participants and 15 substitutes. They were told to use the same criteria they generally use for deciding who should benefit from social programs. Discussions with MMCT and the VDCs revealed that the main criterion used was generally poverty, with some tendency to favor women as being more likely to be disadvantaged. The VDCs were asked to list the workers in order of preference from 1 to 65, and told we would replace workers who dropped out of the program by moving in order from position 51 to position 65 on the list of workers from their own village. This was done for a total of 15 workers at the end of the first round of the study.

This process generated an initial sample of 350 workers, all of whom were interviewed in a baseline survey. One person dropped out before the work started and one person never showed up at payday (only an additional nine people missed any day of work). After all payments of round 1 were done, 343 workers were successfully interviewed in the Midline 1 survey. Before the start of round 2 of the program, 13 workers left the study, and a total of 15 replacement workers were added.[8] A total of 352 workers participated in round

[8] The study protocol specified that only 13 new workers should have been added (to replace the dropouts); too many were mistakenly added, and the extra 2 workers were allowed to stay in the study in order to avoid disappointing them after they had already begun working.

2 of the study, of which all but 3 workers had full attendance and 346 were surveyed at Midline 2. The sample is similar to the broader population of the local region in most respects, differing chiefly in ways that are consistent with the selection criteria; for example, we recruited more women (69% compared to 55% in the district) and our sample is slightly worse off socio-economically than the rest of Mulanje District.[9] We consider the sample to be representative of the type of person likely to be involved in government- or non-government-provided income-generation programs in Mulanje district.

2.2.2 Random Variation in Income Timing

Our study exploits exogenous variation in the timing of individuals' pay. We designed this to vary in two ways. First, the payments are either in weekly installments for four weeks, or in a single lump sum at the end of the month. Second, the payments are made either on Fridays or Saturdays.

The effect of monthly lump sum payments, as opposed to weekly installments, is theoretically ambiguous. In a context where people have problems aggregating streams of income, receiving one's pay in a lump sum at the end of the payment period would increase take-up of profitable investments that are available after the end of the fourth week. However, if people's temptation to overspend is an increasing function of their potential immediate consumption, lump-sum payments could reduce savings instead. This would be the case if the lump sum were received concurrently with opportunities to purchase temptation goods, in which case the money could "burn a hole in people's pockets", causing them to spend money on things that *ex ante* they would prefer not to purchase. If these were the only two potential mechanisms, the variation in the frequency of pay would allow us to see which one dominates in our sample. However, the lump-sum payment could also increase savings through borrowing constraints, if people would prefer a smoother stream of income and would ideally prefer to borrow against the future lump sum payment.

The variation in the day of the week of the payment is designed to shed light on the mechanisms behind the savings constraints people face. If money is received in a tempting environment, like the local market day, then arguably costs to resisting that temptation increase and workers would decide to spend and consume more right at the market when receiving their pay.

We picked Saturdays at the local trading center – so that payroll for this group happened during the major market in the local area – as a tempting context for the receipt of income. This choice was based on extensive qualitative and descriptive work with people in the local area. Anecdotally, people in Mulanje District often describe market days as tempting

[9] See Appendix H.1 for detailed summary statistics on demographic characteristics.

situations, in which excitement can cause them to purchase things they would rather not. Our survey data confirms this: for a free-response question about situations that are tempting or in which respondents may "waste" money, 37% of all respondents volunteered Market Days as a tempting situation, by far the most common among those being ever tempted.[10,11] Multiple-choice questions confirmed this pattern: 69% of people said that market days are more tempting than the day before market days, and 65% of people said having a lot of cash on hand at the trading center was more tempting than having it on hand elsewhere. Based on these answers, payments during market days could exacerbate temptation-based psychological savings constraints, by inducing people to spend money on tempting goods that they would prefer to save. The alternate day – Friday – should not have the same effect on temptation spending, because the market does not take place on that day.

We chose Friday as the alternate day for several reasons. First, it was logistically simpler to manage payments on two consecutive days than on non-adjacent ones; Sunday was not an option because the vast majority of our sample goes to church on Sunday mornings. Second, using the day before the market ensured that all respondents had the liquid cash needed to make purchases at the market – if we had paid the control group on a later day, then for the first week they would not have had any money to spend at the market on Saturday. Third, and most important, if the control group was paid after the Saturday group, then any differences in savings could simply be a function of having to hang on to the money for a shorter period. By choosing Friday as the control group, we ensured that any such effects worked against the expected direction of the results.

There are also a number of reasons why the Saturday payday might not increase temptation, as well as mechanisms that might mute the effects. First, as noted above, many respondents report that having cash at the trading center is more tempting than having it elsewhere. While this is likely due to the market day itself, part of it could be independent of market days: people might just be more tempted to spend at the trading center even if the weekend market is not currently active; the selection of goods is always greater than at the village. Second, while Saturday is the major market day for the local region, there are other markets nearby that operate on Friday. Third, on an open-ended question about reasons they waste money (where the options were not read aloud), only 42% of people report being spending in response to temptation as one of the reasons they spend money they later

[10] Since 39% of respondents said they were never tempted, this constituted 58% of people who believe they ever waste money. The next-most frequent answer was "Going to the Trading Centre in general (not just market days)" with 4% mentioning it.

[11] The exact phrasing of the question in English was "In general, what are situations in which you waste money or are tempted to spend money that you would rather not spend?" The term used in the local language has a less-judgmental sense than "waste" does in American English.

regret spending. This is an appreciable fraction, but if it represents all the people who could possibly be affected by the Saturday treatment, any measured effects will tend to be muted.

We employed a within-person cross-randomized design in order to maximize statistical power. Individuals were randomly assigned to one study arm in the first round of the study and then to another study arm (potentially the same one) for the second round. The randomization for both rounds of the study was done prior to the baseline survey, but the group assignments were not revealed to the workers until the beginning of each round of work. For each round of the study, all workers were randomly assigned to one of four study arms: Weekly Installment payments on Fridays, Weekly Installment Payments on Saturdays, Single Monthly Payments on Fridays, or Single Monthly Payments on Saturdays. For the first round, the randomization assignment was stratified by village and gender. The randomization for round 2 was then stratified on the round 1 assignment and village.

2.2.3 Work Activities

Each subject worked for two weeks during each round of the project, for about four days per week, at a daily wage rate of MK400. There were 7 work days during the first round of the project and 8 days during round two. Workers were employed in conservation-oriented activities that promoted the sustainable use of natural resources. At the beginning of each round of work, representatives from the project met with the workers from each village to help them decide on the specific activities to pursue for that round, based on guidance from MMCT's Sustainable Livelihoods program. The two kinds of work done by the subjects during the study fell under the categories of *Tree Planting* and *Milambala*.

Tree Planting had two separate aspects. During the first round of the project, workers prepared pits for trees to be planted in, and nurseries to house the seedlings for later planting; the seedlings were provided by the Department of Forestry as part of a reforestation program in the area. During round two, which happened once the rainy season had begun, workers did the actual planting of trees. Milambala is a land conservation activity that focuses on building small bund walls to prevent the inundation of fields and limit environmentally harmful erosion of the topsoil. The principal tools needed for the work were hoes, which all the workers already owned. Milambala also required line levels and ropes, which were provided by the project.

Workers were trained in the tasks for each work activity by officials from Mulanje's District Forestry and District Agricultural Offices for Tree Planting and Milambala respectively. Progress on the work was also overseen by officials from the two departments, who set targets for the work to be done on each day and checked in to make sure it was accomplished.

2.2.4 Payroll

Payroll for the project was organized at Mwanamulanje Trading Centre, a major local market in TA Nkanda that was within 4 kilometers of all the villages included in the study. Subjects were informed about how they would be receiving their pay (weekly or monthly, Fridays or Saturdays) at the beginning of each round of work; the procedure was explained verbally, and they were also given a simple handout explaining their group assignment. Each round of work was followed by eight paydays: two per week for four weeks, starting on the Friday and Saturday immediately following the end of the work period.

To ensure that transit and time costs were held equal across the four study arms, all subjects were required to come to the payroll site on all eight paydays during each round – even when they were not being paid their wages. This also allowed us to collect high-frequency data on people's cash holdings and spending behavior, via questions that we asked during the payroll administration. In order to encourage attendance and defray some of people's time costs, all subjects received an MK100 show-up fee for each day, on top of any money they were slated to receive as part of their pay for the project. For example, a person who was paid monthly on Fridays was required to come to the market on all the preceding Fridays and Saturdays, and received MK100; on the day she received her pay, she received MK100 plus her entire wages for the project. The payment schedule in each round across the four payday weekends resulting from the show-up fees and payment of wages according to treatment group and number of work days are overviewed in Table 2.2. MMCT ordinarily manages payroll for its activities using experienced cashiers who work for the organization. For this project, the cashiers were instead employees from the Mulanje District council.

The location and timing of the payroll was specifically chosen to maximize the likelihood that people would be exposed to temptation goods. In pilot testing and qualitative work, people commonly reported market days as periods when they were tempted to spend against their ex ante plans, or tended to waste money. The market at Mwanamulanje happens only on Wednesdays and Saturdays (with Saturdays having the larger market out of the two days), and principally in the morning, which is when people were paid. Shops are still open on Fridays, and there are some mobile vendors, but the majority of market activity happens on Saturdays.

While the purpose of the show-up fee on non-payday days was to equalize transaction costs across treatment groups and make spending patterns comparable, the fact that some amount of money was paid each time may have reduced the potential to observe differences across groups: it is possible that workers satisfied most of their temptation consumption needs with the MK 100 they received each time they showed up at the market.

2.2.5 Data

Our data comes from three distinct sources. A detailed survey, focused on expenditures in the past week; several single-item recall questions administered during the payroll; and, as an objective measure of savings behaviors, respondents' choices about purchasing a short-term, high-return, zero-risk investment offered by the project at the end of the second round of the study.

The survey data was collected three times: once at baseline, and once after each round of the study. Subjects were interviewed at their homes, and answered questions about income, assets, savings, and financial transfers, as well as a detailed module about their expenditures since the previous Friday. This module went through a list of goods and asked respondents if they had bought the good since the previous Friday. If they said "yes" to a good, they were asked about how much they bought on each of Friday, Saturday, and Sunday up to now.

Also part of the survey data were a set of questions on wasting money and being tempted to buy things one should not. Respondents were asked about goods that they found particularly tempting, or that they thought they wasted money on, as well as situations in which they felt they wasted money. They were also asked for *ex post* judgments about whether they felt they had wasted money in the period since they received their pay; this question was only included on the survey after the second round.

Our second data source is a set of questions asked during the payroll process. On each of the eight paydays, all respondents were required to come to the payroll site as described above. Prior to receiving their pay or show-up fee, they were asked simple aggregate questions about the money they had on them at the time (not including their pay, which they had yet to receive) and the amount of money they spent at the market on the previous payday. Hence on Fridays, people were asked about the money they spent on the Friday of the previous week, and on Saturdays, they were asked about the money they spent yesterday. During the second round of the study, we also asked two additional questions as sensitivity checks: first, we asked people to recall their spending from the Friday of the previous week, to look at the influence of recall bias. Second, we asked people about money they spent outside of the market, in case there were differential patterns in non-market spending.

A third source of data comes from an investment opportunity offered to respondents at the end of each round of the study. Respondents were offered the chance to buy the investment good only once per round, immediately after we visited them for the midline survey for the round in question. The investment took the form of a "bond", with shares that cost MK1500 to purchase and that paid back the principal plus MK500 interest after exactly two weeks. Each respondent could buy a maximum of two shares, and no fractional shares were allowed. All respondents who purchased the bond were paid back on time according to the terms of

the investment.

The investment good was intentionally offered only once per round, in the week after the final payment was made. This allows us to use it to test for the existence of savings constraints, since members of the weekly group had to save their pay in order to use it for the investment good. An alternative design would have offered the investment opportunity each week. This would have lowered the amount of time that the weekly group needed to save in order to purchase it, thus relaxing the savings constraint somewhat. We chose this design in order maximize our statistical power to detect differences across the two groups.

Summary statistics from these data sources for all variables used in the regression analysis are presented in Table 2.3, separately for pre-experiment baseline and for outcome variables. At baseline, the households' total spending considering all expenditures from the last Friday prior to being interviewed up to the day of the survey averages MK2,257 (about US$5.6 or PPP$14). Respondents report having an average of MK670 (about US$1.7 or PPP$4.2) left out of the money they had received since the Friday prior to interviewing. Households spend about 69% of their total expenditures on food for preparation at home, another about 6% on immediate consumption away from home and about 28% on non-food items.[12] About a third of food expenditure was on maize, which is the principal staple crop in the region. Randomization led to a sample with no notable differences in pre-program characteristics across study arms (See discussion in Appendix H.1).

2.3 Empirical Specification

We study the effects of the experimentally-induced variation in payment timing on several sets of outcomes: expenditure at the market when payment was received; total expenditure levels and composition over the last weekend of each round, including self-reported wasteful expenditures; asset accumulation; and take-up of the large installment-size, risk-free, high-return investment opportunity.

We present two regression specifications reported as separate panels in the main results tables. The first tests the effect of being randomly assigned to a be paid in a single monthly lump sum as relative to four weekly installments. In Panel A of the subsequent tables (and in the only specification shown in Table 7), we run regressions of the form

$$Y_{ir} = \alpha SingleMonthlyPayment_{ir} + \boldsymbol{\beta}'\mathbf{X}_{ir} + \varepsilon_{ir} \tag{2.1}$$

[12] The shares do not add to 1 exactly due to Winsorizing.

Y_{ir} is the outcome of interest for worker i in round r. $SingleMonthlyPayment_{ir}$ is an indicator variable for individual-level assignment to receive one's wages in a single payment at the end of the month instead of in four weekly installments, during round r. The coefficient α measures the effect of receiving wages on in a single monthly lump-sum (on either Friday or Saturday). $\mathbf{X}_{ir}$ is a vector that includes stratification cell dummies; two household financial variables measured at baseline prior to the randomized assignment;[13] and a linear function of the weekday of the exogenously-assigned (first attempted) interview date. The available baseline controls are summarized in Table 3. ε_{ir} is a mean-zero error term.

Whenever data from both rounds are used (so r=1, 2 in the equation above) standard errors are clustered at the worker level to account for statistical dependence of outcome measures for the same individual across the two rounds. The stratification cells are defined separately by round, so these implicitly control for round fixed effects when multiple rounds are used.

Panel B is analogous to Panel A, except the included experimental group indicator compares the impact of being assigned to receive one's pay on Saturday as opposed to Friday. Regressions are of the form

$$Y_{ir} = \gamma Saturday_{ir} + \boldsymbol{\beta}'\mathbf{X}_{ir} + \varepsilon_{ir} \tag{2.2}$$

where Y_{ir} and $\mathbf{X}_{ir}$ are defined as above, and $Saturday_{ir}$ is an indicator for assignment to the Saturday payday group. The coefficient γ represents the effect of assignment to the Saturday payday group relative to the Friday payday group. Because the effect of being paid on during the tempting Saturday market may differ by the amount of pay received, these regressions are estimated separately for the workers in the monthly and the weekly study arms.

In general, workers in this project interact with each other and so in theory we cannot exclude that workers assigned to one experimental group had an impact on workers in another. Our design does not allows us to address potential spillovers of effects from one study arm to another. In the context of our design, any spillovers should bias our estimated effects toward zero: for example, if monthly payment group members gave loans to weekly payment group members, this should reduce any differences in expenditures between the two groups. Additionally, we find no empirical evidence of increased cash or in-kind transfers for any of the experimental groups (results not shown).

[13] Our baseline financial controls are an index of asset and livestock ownership (using principal component analysis) and the total amount of money the respondent spent out of their income received since the Friday prior to the baseline survey. Results are not sensitive to the specific choice of baseline financial controls.

2.4 Empirical Results

2.4.1 Lump Sum Payment vs. Weekly Payments

We begin by focusing on the effect of receiving a lump-sum payment relative to receiving weekly installments. Workers were randomized into one of the two payment frequency conditions; the lump sum group received wage payments on the last of four weekends at which the weekly payment condition received their wages. However, all workers were required to come to the site where payroll was administered every Friday and Saturday on all four payday weekends, even if no wages were received. Workers received a small "show-up fee" of MK 100 and were also asked the payday questions described in Section 2 above.

Table 2.4 columns 1 to 5 show the effect of the treatment on the amount people spent on specific days that they came to the market. Because people received income on different days, however, a better comparison is given by column 6, which presents the effect of the treatment on the amount of money people spend at the market on the day that they receive their wages. This variable measures expenditure on Fridays for the Friday condition and on Saturdays for the Saturday condition; it includes spending on all four paydays for the weekly condition, but only on the fourth week of paydays for the lump sum condition. Column 7 presents the same figure, but as a share of income received. Panel A of Table 2.4 shows the effect of lump sum payments vis-à-vis the weekly payment condition, all of which are strongly statistically significant. Focusing on column 6, we see that respondents in the monthly group spent MK 940 less of their total pay at the market on the same day that they received it. Column 7 shows that they reduced the share of their pay they spent on the day of receipt by 24 percentage points. In payday weekends 1 through 3, when the lump sum condition was not receiving any wages, market expenditure on paydays was lower in the lump sum condition: on Fridays 1, 2 and 3 in total workers only spend about 38% of the average in the weekly payment condition (column 1) and the same rate is about 42% for Saturdays (column 2). On the last payday weekend, when those in the lump sum group receive their wages, expenditures are higher by MK 318 and MK 495, respectively. The increase in the monthly group's expenditures during the fourth weekend is smaller than their decline in expenditures in weekends 1, 2 and 3.

Table 2.4 concerned expenditures at the market; Table 2.5 (Panel A) looks at survey measures of total expenditures during the fourth payday weekend. Table 2.5 columns 3 through 6 show effects on self-reported wasteful spending ("How much did you spend on items that you later thought you should not have spent money on?"), both in total for the last payday weekend as well as separately for Friday, Saturday and after. Consistent with the payday data about market expenditure, total expenditures over the weekend and into

the following week are higher for the lump sum group (by MK 1,451, column 1). Despite the higher spending, cash remaining on hand out of the money received since the Friday prior to the follow-up interview is marginally statistically significantly higher, with a point estimate of ca. MK 139. Wasteful spending, however, was not significantly different for the lump sum group (columns 3 through 6), suggesting that the higher receipt of cash in one chunk does not lead recipients to overspend on goods they later regret – at least in this context. While the standard errors are large enough that we cannot reject a doubling of wasteful expenditure, the results from Panel A of Table 2.6 are consistent with the idea that the composition of expenditure did not change in the monthly group. Table 2.6 columns 1 through 4 show expenditure shares in broad categories. These data are constructed from detailed, itemized listings. The shares of expenditure in different broad item categories were not significantly different between the monthly and weekly payment groups.

The wasteful spending variables in Table 2.5 are only available for round 2; we choose to show this set of outcomes as it most unambiguously reflects temptation spending and avoids constructing outcomes with researcher-imposed ideas of which expenditures are temptation purchases. There are multiple ways of constructing outcomes with the same intention. One variation that we have explored is based on reports of unplanned purchases of items: we have considered both items that are commonly unplanned purchases across the whole sample, as well as individual self-reports that a specific purchase was not planned. Neither of these variations affects the pattern of no significant treatment effects, and so we omit these alternative specifications for brevity.

Column 5 of the same table examines whether higher expenditures lead to differential asset purchases. The estimates show that net asset accumulation over the course of the all payday weekends does not appear to be different between lump sum and weekly payment conditions. However, the standard errors are large and so economically-significant effects cannot be ruled out by these estimates.

Lastly, in Table 2.7 we examine the effect of lump sum payments on take up of a large minimum-installment, high-return, risk-free "investment opportunity" that was offered to respondents right after the follow-up interview.[14] Workers were able to buy either 1 or 2 "shares" from the project that had a risk-free return of 33% and were repaid after exactly two weeks. This investment opportunity was offered to test whether the timing of payments affects respondents' ability to take up profitable investment opportunities that cannot be purchased in small parts. The main advantages of this novel outcome variable are that

[14] There is no effect of Saturday vs. Friday payments on these outcomes, consistent with the lack of difference in remaining cash after weekend 4. For clarity of presentation we omit the specifications of Panel B and focus only on the regressions analogous to Panel A in the preceding results tables.

it provides a controlled investment instrument with known features, and, moreover, that it makes a high-return investment opportunity, that requires a large minimum investment, homogeneously available to every respondent at the time of surveying. In real life respondents' opportunities vary widely cross-sectionally and, importantly, over time – e.g. farming investments are largely only available during a limited period of the year.

In round 1 the opportunity to invest was only announced in the week preceding the final payday. This limits the usefulness of the round 1 results, because workers already knew their treatment status but did not know about the investment opportunity until a week before it was made available to them. This could bias any estimated effects either upwards or downwards. An upward bias could occur because the weekly payment group members did not know about this opportunity until they had received three quarters of their wage. The wage amount remaining to paid in the last payday weekend was smaller than the minimum required amount for the investment opportunity (the remaining payment was MK 800 but one unit of the investment offer was priced at MK 1500); this would eliminate the subset of weekly workers who had less than MK700 in weekly income from being able to purchase the investment good. A downward bias could occur because lump sum payment group members may have already committed their pay to other expenditures. This would limit their ability to purchase the investment good, thus understating any measured effects.

In contrast, in round 2 the investment opportunity was announced before the start of the round, so all respondents across both groups knew they would have the opportunity prior to learning which payment group they were in. Workers therefore had advance notice of the prospect of this opportunity before any wage payments began, and before they could potentially commit any of their wages to other expenditures in a way that depended on their study arm assignment. Because of these differences in setup across rounds, we show results both from regressions on pooled data from both rounds and then specifically for round 2.

Table 2.7 columns 1 and 2 repeat outcome variables from Table 2.5 columns 1 and 2 (cf. Panel A) to be able to track differences due to changing sets of observations across the Round 2, Round 1, and Pooled specifications respectively. Columns 3 and 4 of Table 2.7 show effects on take-up of the investment opportunity. When we pool observations across the two rounds, lump sum payment group members had a 4.8 percentage-point higher probability of buying any share (significant at the 10% level) and the total amount spent on the investment opportunity was about MK 122 higher (significant at the 5% level). The comparison to the separate specifications for Round 1 and Round 2 show that this effect is concentrated in round 2 where the effect of lump sum payments on probability of taking up 9.5 percentage points, relative to a base of only 6.3% among the weekly payment group.[15] Total spending

[15] Takeup actually remains the same across rounds for the monthly group and declines from round 1 to

– number of shares times the price per share – was about MK 196 higher in the lump sum group, relative to a base of MK 172 in the weekly payment group. Both differences are statistically significant, at the 1 percent and 5 percent levels, respectively.

The results from Table 2.7 suggests that paying workers in a lump sum enabled them to hold enough cash to make use of a high-return large minimum installment size investment opportunity, while the weekly group did not have sufficient extra cash holdings at the time the opportunity was offered – despite experience with the product (from round 1) and sufficient advance notice.

In theory, the higher investment by the lump sum payment group could be driven by credit constraints alone, as opposed to savings constraints. Consider the case in which workers assigned to the lump sum payment group really wanted to smooth their consumption in the way the weekly payment group was able to, but could not due to a borrowing constraint. In that case, lump sum workers would "involuntarily" end up with more cash at the time the investment opportunity was offered and so they make use of it. While borrowing constraints are likely binding for many in the economic environment of this study, several arguments make this model an unlikely driver of our result: 72% of workers at baseline report preferring to be paid in a lump sum after four weeks as opposed to receiving four weekly installments (with the same twice-weekly attendance requirements in the hypothetical scenario that respondents were asked about as were imposed in this experiment). Of those 72%, a great majority (83%) state, in an open ended question with at most one answer, that the reason for this preference is that it enables them to "make a better plan" for the money. 13% outright list avoiding wasteful spending as the reason. These answers imply either a commitment problem as the reason for the lump sum preference, or at the least an expected inability to save – either due to internal constraints, such as self-control problems, or external constraints, such as fear of theft. Lastly, if lump sum payment group members truly preferred to smooth consumption in the way the weekly group was able to, then they should not prefer to invest in the shares offered in this project as it locks up half (if they bought one share) or all (if they bought two shares) of total received wage payments for two weeks without any opportunity to access it. While in theory workers could have potentially borrowed against the future income receipt to access the money in the investment, this would also have held for the receipt of their wages, implying borrowing constraints could not be driving the results.

If lump sum condition households were limited in their ability to smooth consumption in

round 2 for the weekly group. However, we cannot draw any strong conclusions from this pattern because of general seasonal variations in behavior - for example, spending levels are generally higher in round 1 before the start of the lean season in round 2.

the face of shocks then we would also expect that lump sum condition households would – relative to the weekly payment condition – receive more transfers from their social network over the course of the four payday weekends or request more loans – two of the most common risk coping mechanisms for workers of this study. However, we do not find statistically-significant effects on either of these outcomes; the point estimates are small, but the standard errors are large and so even sizeable effects along these dimensions cannot be ruled out (results not shown).

2.4.2 Saturday vs. Friday Paydays

Having demonstrated that receiving pay in a lump sum increases uptake of the investment good, and that this appears to be the result of savings constraints, we now look to whether receiving pay in a tempting environment alters this effect. To do this, we examine the effects of the experimentally-induced variation in whether workers are paid on Saturday compared to Friday on expenditures and saving. We consider estimates of equation (2.2) in Panel B, respectively, of Tables 4 through 6, with results shown separately for respondents in the monthly lump sum and weekly installment payment conditions.

We first examine how the specific day on which people were paid affected their spending at the market over the course of the eight paydays during each round. Table 2.4 presents estimates for outcomes from the panel of data collected during paydays.

Columns 6 and 7 of Table 4, Panel B indicates that the day of receipt did not matter for same-day market expenditures. If receiving pay in the environment of Saturday's weekend market was tempting for workers then we should expect to see workers in the Saturday group spending more at the market on the day they were paid. The point estimate is close to zero and relatively tightly bounded: the mean of the dependent variable in the Friday group is MK 1244 for the monthly group, the point estimate for the Saturday effect is MK -53.70 with a standard error of MK 118. The estimated null effect is even tighter for the weekly group; there is no evidence of a differential effect by payment frequency.

Table 2.4 columns 1 through 5 reveal that those workers with payments on Friday spend more money at the market on Fridays – the estimate of the Saturday coefficient is negative for Friday expenditures – and those with payments on Saturday spend more on Saturdays. The negative coefficient on the Saturday dummy is larger in absolute value for Friday outcomes than for Saturday outcomes, suggesting that Friday wage receivers spend some of their money on Saturday, while Saturday wage receivers do not have extra funds to spend on Friday – the day before their pay receipt. There is no meaningful evidence of a differential effect by payment frequency: columns 1 and 2 cover a period when the monthly group was not receiving any pay, so no effects are to be expected.

The natural follow-up question is to ask whether total expenditures over the whole weekend and in the days following the payday weekend were different by Saturday vs. Friday payment. Thus, we turn to Table 2.5 column 1 which presents the effects on spending during and after the fourth payday weekend for each of the two rounds, including also non-market expenditures. In Panel A, the point estimate for the Saturday effect is negative for the weekly group and positive for the monthly group, but far from statistically significant in either case. Taken at face value, the point estimate of MK 256.4 for the monthly group would imply a relative effect of ca. 12.3% of the Saturday assignment on total expenditures compared to the Friday assignment (mean of MK 3091). The effect would be about -4.3% for the monthly group. Compared to the market data of Table 2.4 that was available for all payday weekends, the standard errors are higher, and so moderate Saturday effects cannot be rejected with high confidence for this outcome variable.

Column 2 shows a statistically-insignificant but negative estimated effect of the Saturday condition on the amount of cash respondents had received since the Friday before the interview but had not yet spent. The differences of about 50 for the weekly group and MK 165 for the monthly group are large relative to the respective Friday payday condition means of MK 483 and MK 670, and so we cannot reject moderate-sized effects on this outcome.

We have established that there is no detectable Saturday effect on the level of expenditures on market day and beyond. However, if Saturdays are tempting, being paid on Saturdays could also affect the composition of expenditures. To explore this we look at the two sets of outcome variables: self-reported wasteful expenditures, in Table 2.5 columns 3 through 6, and the composition of spending in broad expenditure categories, in Table 2.6 columns 1 to 4. Again, we find no robustly-significant Saturday effects on average or in interactions with payday frequency. Lastly, column 5 of Table 2.6 shows that over the course of the entire payment period, Saturday payments did not differentially affect asset accumulation compared to Friday payments. While some of the coefficients in columns 3 to 5 of Table 2.5 are statistically-significant, the overall effect on wasteful spending in column 6 is not. Two of the coefficients in Table 2.6 are statistically-significant, but only at the 10% level, and each for only one of the two variations in payment frequency. For all the outcomes in Tables 2.5 and 2.6, however, the standard errors are relatively large and so we cannot reject the hypothesis that effects are in fact economically significant.

Overall, we find no strong evidence that receiving one's pay during the Saturday market affected expenditure or savings behavior. The implied confidence bands around many of our sets of point estimates given the standard errors are, however, not very narrow. We therefore cannot reject relatively large differences – compared to the mean in the control group – with confidence for any of the outcomes except for expenditure levels.

2.5 Discussion and Conclusion

Markets for financial intermediations in developing countries are imperfect. Besides the "external" constraints this creates for households, these market imperfections may exacerbate "internal" constraints such as time-inconsistent preferences and limited attention. In such a setting the exact timing of income streams can matter for spending and savings decisions. Spending may be higher, or skewed towards unplanned or wasteful expenditures in environments that are tempting, and spending may be different depending on the frequency of payments. If the timing of income receipt matters, this may have implications for the payment policies of employers and cash transfer programs, who may be interested in structuring payments to maximize benefits to income recipients.

In the specific context of this study, and in developing countries in general, there are two concerns about how wage payments are structured across time. First, when income is received in tempting environments, recipients may end up spending more, or may spend more on different items than they had planned ex ante, or than they deem prudent ex post. Second, when income is received in small installments, people may find it harder to generate meaningful sums that can be used for large-installment expenditures such as durable goods purchases, buying in bulk to receive quantity discounts, or high-return investments. In order to determine if these concerns are empirically relevant we designed a field experiment that varied the degree of temptation people faced when receiving payments, as well as whether payments were received in small installments or as a lump sum.

Based on ample qualitative evidence suggesting that spending – in particular frivolous spending – might be higher if income is received on market days, our experiment used the day of the week that workers were paid to vary the level of temptation workers faced when receiving income. Half of our sample received their income during the major local market day, which happened on Saturdays; the other half received their income at the same site on Fridays. However, we do not find evidence, for the sample of casual workers in Malawi that were part of our study, that the specific day of the receipt of income is an important driver of expenditures. Observed spending and savings behavior had no statistically-significant differences between those paid on Fridays and those paid on Saturdays, and we can rule out moderate-sized effects. This pattern does not depend on whether people are paid in a single lump sum or in small installments.

These findings do not reject the general idea that the environment in which people are paid matters. We worked in seven villages around one particular trading center in Malawi. In this setting, other trading centers with complementary market days – e.g. ones that take place on Fridays, when the payday trading center's market was not occurring – are within 30

minutes' travel. In other settings in which there are no complementary nearby market days, the day of payment may matter more. However, the setting of our study is fairly typical for many rural areas in Malawi and other countries in the region, where there are very often trading centers with a market day covering most days of the week, located within distances that can be traveled in reasonable times. Thus, the findings of our study should imply that the specific day of income receipt is not a major driver of spending decisions in a broad range of settings in rural Africa.

We also investigate the impact of paying workers in one lump sum compared to weekly payments. Our findings suggest that organizations can help income recipients overcome savings constraints by providing income in larger installments rather than smaller ones. Workers in the lump sum payment group spend relatively less of it immediately on receipt. Since they also receive more money on the last payday weekend – the full amount of wages compared to the weekly group that is receiving only the fourth of four equal installments – lump sum payment group members remain with more cash in the week after the last payday. In general, receiving income in a lump sum does not appear to affect the composition of expenditure, only the level. This mitigates concerns that lump sums "burn a hole in workers' pockets". Moreover, we find evidence that lump-sum income receipt promotes saving: people in the lump sum payment group show a higher propensity to save in a high interest, relatively short-term asset that was offered to all respondents and required a large minimum investment. We argue that the differential investment is largely a function of the weekly payment group workers' inability to have cash available at the time of the investment offer (the timing of which was known to all workers before any payments were made).

The findings suggest that it is preferable for recipients that organizations pay at least part of wages or cash transfers in lump sums as a form of pre-committed savings. There is a trade-off between the desire to smooth consumption and the ability to generate lump sums; and so in an environment with borrowing constraints and generally high costs of risk coping, receiving all household income infrequently is unlikely to be desirable for households. In the context of this study, however, almost all households had some other means – besides the income from this project – of securing basic levels of consumption. Furthermore, a majority of households reports that they prefer to receive this additional income as a lump sum. This supports the idea that projects designed to generate income for people in developing countries, such as GiveDirectly, should provide income in strategically-timed lump sums (or at least offer this option) in order to maximize benefits to recipients.

The investment opportunity was artificially provided to study participants as part of this project in order to improve measurement of investment behavior in a small sample observed over a short time horizon and in a context where absolute income differentials across

treatment groups were small. In addition, overall take-up of the investment opportunity was low. As such, the observed effects mainly support the overall conceptual point. However, the implied magnitudes are also interesting: we provided both the weekly and the lump group households with identical total additional income of MK 4000 (MK 3200 wages + 8 x MK 100 show up fees) over the course of the second round of this project. The point estimates imply that on average each member of the lump sum group was able to increase household income by an additional MK 65[16] – about 1.6% of income from the project's employment – within two weeks of the last payday via the investment opportunity, solely because of the changed timing of payments.

Practically speaking, the effect of changing the payments from small installments to lump sums will depend on the return to the relevant investment. We can get a sense of this by considering an example of an investment that is conceptually similar to the one we offered: secondary school fees, which are approximately MK3000 ($7.50) per year in Malawi, and which generally must be paid in total at the beginning of the school year, rather than in installments. If people do think about education as an investment, we would expect that a project that pays respondents' total wages of MK3000 in a single lump sum timed for the beginning of the school year, rather than in small installments, to increase school fees payments by as much as 9 percentage points. This could have significant social benefits: if school fees are the only barrier to attending secondary school (and they are commonly cited as a reason teenagers do not go to school in Malawi) then that shift would have similar effects on the rate of school attendance. To get a sense of the total social benefit of this change in timing, note that Malawi has a GNI per capita of $320, and that research on the returns to education generally estimates figures of at least 10% per year in developing countries. Thus the additional 9% of children who are able to attend school would earn an additional $32 per year. Over a 40-year working life, starting 4 years after the investment, and at a social discount rate of 10%, this would raise a child's income by $213, for a net benefit of $206 per person. This is a substantial payoff for a relatively minor change.

School fees also highlight the external validity of our results for the investment good: they are time-sensitive, as are many other investment opportunities in the developing world, such as farm input purchases, which must be timed for the planting season.[17] This exacerbates the savings constraints that people face: it is easier to save up for an investment if you can make the purchase whenever you have the money, as opposed to needing to bring the money on a specific day. There are other important investments that do not have this same

[16] 33% of 196.2, from Table 7, column 4.

[17] While some farm inputs can be bought and stored, others cannot for various reasons. For example, Malawi's government subsidizes fertilizer purchases immediately before the planting season, so farmers must have the cash to purchase the subsidized fertilizer within a fairly tight window.

time-sensitive feature: for example, metal roofing has a large minimum installment size, but can be purchased whenever people have the money for it. Due to the design of the investment option used in this study, we cannot be sure that our results hold for alternative, less time-sensitive goods.

These benefits would come at relatively little cost, and organizing payroll just once a month could even be cheaper for the paying organization. We also see no significant downsides to partial lump sum payments, even when they are received during one of the most tempting environments that people typically experience in rural Africa. However, further research is needed in order to better-establish whether lump-sum payments can potentially backfire in developing countries.

Our results provide several lessons for future research on lump sum payments as well as on the role of self-control problems in driving savings constraints. First, people are aware of the self-control problems they face, and thus survey questions that directly ask people about temptation and wasteful spending are a useful way to measure people's self-control issues. Second, offering study participants a meaningful investment opportunity that bears actual interest can be a helpful way to isolate an intervention's effects on savings constraints. Other outcomes have two kinds of limitations: non-financial investments such as health and education may not be perceived as investments by respondents, and heterogeneity in returns may generate misleading inferences about the extent of savings constraints. Third, to the extent that self-control problems are generating internal savings constraints in rural Africa, they may not be particularly amenable to policy interventions. Receiving one's pay during the market – a location commonly listed as being tempting by the respondents in our study – generated only small variations in their level of self-reported wasteful spending, possibly because people continue to select into other tempting situations. This suggests that other causes of savings constraints may merit further research.

Figure 2.1
Timing of work, payments and data collection

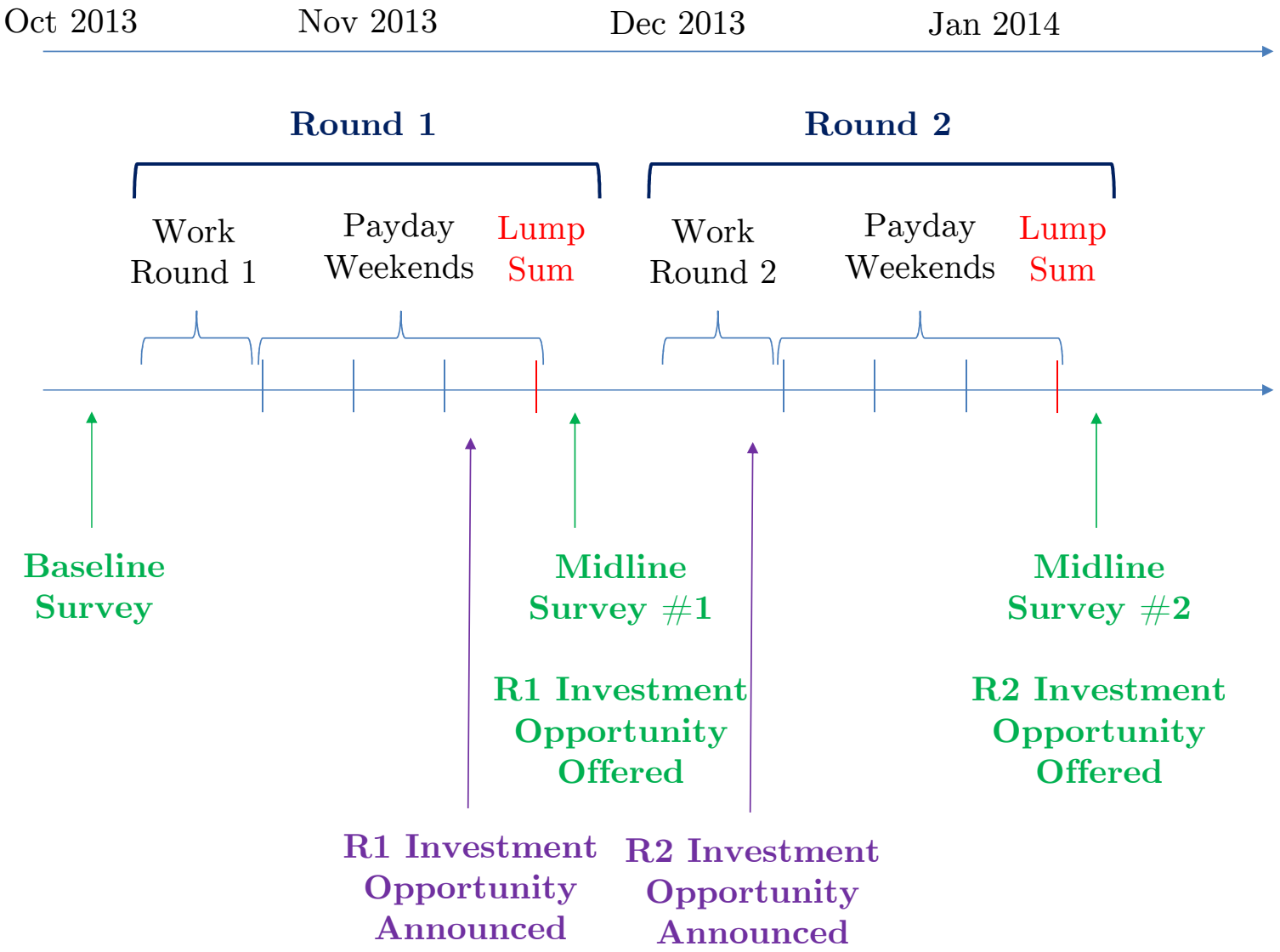


Table 2.1

Distribution of worker-round observations into experimental groups,
(a) pooled across round 1 and 2 and (b) separately for round 1 and round 2

a)

Frequency \ Payday	Friday	Saturday	
Weekly Installment Payments	172	177	349
Single Monthly Payment	178	172	350
	350	349	699

b)

Experimental group	Round 1	Round 2	Total
Weekly Installment Payments, Friday	86	86	172
Weekly Installment Payments, Saturday	89	88	177
Single Monthly Payment, Friday	87	91	178
Single Monthly Payment, Saturday	86	86	172
Total	348	351	699

Table 2.2
Payment schedules by payday group and round (all values in MK)

	Payday weekends							
	#1		#2		#3		#4	
Round 1	Fri	Sat	Fri	Sat	Fri	Sat	Fri	Sat
Payment group								
Weekly Installment Payments, Friday	**800**	100	**800**	100	**800**	100	**800**	100
Weekly Installment Payments, Saturday	100	**800**	100	**800**	100	**800**	100	**800**
Single Monthly Payment, Friday	100	100	100	100	100	100	**2,900**	100
Single Monthly Payment, Saturday	100	100	100	100	100	100	100	**2,900**
Round 2								
Payment group								
Weekly Installment Payments, Friday	**900**	100	**900**	100	**900**	100	**900**	100
Weekly Installment Payments, Saturday	100	**900**	100	**900**	100	**900**	100	**900**
Single Monthly Payment, Friday	100	100	100	100	100	100	**3,300**	100
Single Monthly Payment, Saturday	100	100	100	100	100	100	100	**3,300**

Table 2.3
Summary statistics

	Mean	Std. dev.	10th percentile	Median	90th percentile	Obs.
Baseline variables						
Index of asset ownership	-0.02	2.695	-2.489	-0.713	3.061	342
Total spending since last Friday, inclusive [MK]	2257	3763	200	1000	4600	321
Remaining cash out of received since last Friday, inclusive [MK]	670	2623	0	20	1400	321
Expenditure shares based on itemized elicitation						
Food for consumption at home	0.690	0.214	0.361	0.742	0.937	341
Maize only	0.234	0.260	0.000	0.170	0.605	341
Food for consumption out of home	0.061	0.069	0.000	0.038	0.144	341
Non-Food	0.279	0.235	0.040	0.189	0.655	341
Outcome variables						
Market spending on paydays						
Amount spent on day of wage receipt	1645	1151	200	1500	3200	683
Amount spent at market on Fridays 1, 2, & 3	651	685	200	300	1895	690
Amount spent at market on Saturdays 1, 2, & 3	829	759	200	480	2300	691
Amount spent at market on Friday 4	524	761	50	120	1500	675
Amount spent at market on Saturday 4	823	939	60	500	2300	689
Follow-up survey measures						
Total spending since last Friday, inclusive [MK]	2509	2395	800	2300	4000	689
Remaining cash out of received since last Friday, inclusive [MK]	529	996	0	0	2000	689
Expenditure shares based on itemized elicitation						
Food for consumption at home	0.698	0.212	0.371	0.751	0.930	689
Maize only	0.359	0.266	0.000	0.371	0.709	689
Food for consumption out of home	0.051	0.056	0.000	0.034	0.125	689
Non-Food	0.251	0.206	0.043	0.188	0.572	689
Value of net asset purchases since last interview	2154	7486	0	0	5300	689
Self-reported wasteful spending on weekend 4 of round 2						
Total since last Friday, inclusive [MK]	306	685	0	25	800	346
Friday [MK]	164	462	0	0	400	346
Saturday [MK]	73	256	0	0	150	346
Sunday and after [MK]	66	281	0	0	90	346
Round 2 investment opportunity take-up						
Bought any shares [0/1]	0.108	0.311				351
Total spent on shares [MK]	265	798	0	0	1500	351

Notes: Sample includes 359 respondents who participated in at least one round of the work program and have data from at least one data source for that round (either the payday data, the survey, or both). All money amounts are in Malawian Kwacha (MK); during the study period the market exchange rate was approximately MK400 to the US dollar, and the PPP exchange rate was approximately MK160 to the US dollar. See Appendix B for variable definitions.

Table 2.4
Effects of treatment assignment on market spending

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable:	**Total spent at market on Fridays 1, 2, 3**	**Total spent at market on Saturdays 1, 2, 3**	**Amount spent at market on Friday 4**	**Amount spent at market on Saturday 4**	**Total spent at market on Fri and Sat 1 - 4**	**Amount spent on day of income receipt**	**Ratio amount spent over received on day of income receipt**
Panel A - Lump sum vs. weekly							
Lump sum payment	-604.6***	-697.9***	318.2***	495.0***	-488.5***	-938.8***	-0.242***
	(49.03)	(53.25)	(55.24)	(72.61)	(126.6)	(81.73)	(0.0252)
Mean dep. var., weekly payment group	980.7	1201	365.6	576.3	3129	2142	0.631
Number of observations	696	696	696	696	696	696	696
Panel B - Saturday vs. Friday							
i) Weekly study arm only							
Saturday payday	-1,203***	588.0***	-372.4***	192.6**	-795.6***	-3.884	0.00214
	(64.08)	(88.79)	(35.99)	(75.80)	(151.9)	(113.7)	(0.0336)
Mean dep. var., Fri payment group	1595	899.8	555.9	474.5	3532	2151	0.631
Number of observations	347	347	347	347	347	347	347
ii) Monthly study arm only							
Saturday payday	16.26	28.08	-1,105***	228.0**	-843.0***	-53.70	-0.0177
	(29.10)	(49.93)	(82.62)	(113.8)	(175.3)	(118.0)	(0.0381)
Mean dep. var., Fri payment group	365.9	503.4	1244	959.1	3082	1244	0.401
Number of observations	349	349	349	349	349	349	349

Notes: Stars indicate significance at 10% (*), 5% (**), and 1% (***) levels. Regressions are run on pooled data from round 1 and round 2 (see Empirical Strategy for details). Standard errors are clustered at the individual level in parentheses. USD 1 is ca. MK 400 for study period. All regressions include stratification cell fixed effects and an index of baseline asset ownership based on first principal components, difference in days between date of interview and the preceding weekend, baseline total spending. For complete variable definitions, see Appendix B, and Table 3 for summary statistics.

Table 2.5
Effects of treatment assignment on total spending
and cash saving and wasteful spending

	(1)	(2)	(3)	(4)	(5)	(6)
			Self-reported wasteful spending, ***round 2 only***			
Dependent variable:	Total spending since last Fri, inclusive [MK]	Remaining cash out of received since last Fri, inclusive [MK]	Friday [MK]	Saturday [MK]	Sunday and after [MK]	Total since last Fri (cols 3+4+5) [MK]
Panel A - Lump sum vs. weekly						
Lump sum payment	1,451***	139.2*	66.54	32.28	-7.165	92.77
	(159.1)	(71.40)	(47.43)	(27.63)	(31.06)	(70.53)
Mean dep. var., weekly payment group	1836	468.5	132.3	58.36	67.60	261.8
Number of observations	689	689	346	346	346	346
Panel B - Saturday vs. Friday						
i) Weekly study arm only						
Saturday payday	-80.14	-50.32	-113.2**	14.74	-28.25	-131.0
	(221.3)	(114.8)	(56.42)	(28.55)	(51.81)	(83.46)
Mean dep. var., Fri payment group	1881	483.5	189.1	52.59	73.53	322.3
Number of observations	344	344	171	171	171	171
ii) Monthly study arm only	256.4	-165.2	-126.4	79.94*	14.49	-0.125
Saturday payday	(222.3)	(110.9)	(79.62)	(45.47)	(41.94)	(117.5)
Mean dep. var., Fri payment group	3091	670.6	250.4	34.89	55.22	326.1
Number of observations	345	345	175	175	175	175

Notes: Stars indicate significance at 10% (*), 5% (**), and 1% (***) levels. Regressions of columns 1 and 2 are run on pooled data from round 1 and round 2 for which standard errors are clustered at the individual level; remaining columns use only round 2 data since outcomes are not available in round 1. All regressions include stratification cell fixed effects and an index of baseline asset ownership based on first principal components, difference in days between date of interview and the preceding weekend, baseline total spending and -if available- the baseline value of the outcome variable. For complete variable definitions, see Appendix B, and Table 2 for summary statistics.

Table 2.6

Effects of treatment assignment
on expenditure composition and asset accumulation

	(1)	(2)	(3)	(4)	(5)
	Expenditure shares based on itemized elicitation				
Dependent variable:	Food for consumption at home	Maize only	Food for consumption out of home	Non-Food	Value of net asset purchases since last interview
Panel A - Lump sum vs. weekly					
Lump sum payment	-0.0153	0.0182	-0.00416	0.0190	19.61
	(0.0162)	(0.0192)	(0.00449)	(0.0161)	(525.7)
Mean dep. var., weekly payment group	0.707	0.352	0.0523	0.240	2271
Number of observations	689	689	689	689	689
Panel B - Saturday vs. Friday					
i) Weekly study arm only					
Saturday payday	0.0124	0.0124	0.0100*	-0.0234	-395.5
	(0.0224)	(0.0278)	(0.00577)	(0.0219)	(848.2)
Mean dep. var., Fri payment group	0.702	0.348	0.0473	0.250	2604
Number of observations	344	344	344	344	344
ii) Monthly study arm only					
Saturday payday	-0.0222	-0.00801	-0.000380	0.0224	-1,230*
	(0.0224)	(0.0246)	(0.00614)	(0.0224)	(718.8)
Mean dep. var., Fri payment group	0.698	0.371	0.0508	0.251	2558
Number of observations	345	345	345	345	345

Notes: Stars indicate significance at 10% (*), 5% (**), and 1% (***) levels. Regressions are run on pooled data from round 1 and round 2 (see Empirical Strategy for details). Standard errors clustered at the individual level in parentheses. USD 1 is ca. MK 400 for study period. All regressions include stratification cell fixed effects and an index of baseline asset ownership based on first principal components, difference in days between date of interview and the preceding weekend, baseline total spending and -if available- the baseline value of the outcome variable. For complete variable definitions, see Appendix B, and Table 2 for summary statistics.

Table 2.7

Effects of treatment assignment
on post-interview risk-free, high-return investment offer

	(1)	(2)	(3)	(4)
Dependent variable:	Total spending since last Fri, inclusive [MK]	Remaining cash out of received since last Fri, inclusive [MK]	Bought any shares [0/1]	Total spent on shares [MK]
Round 1 and 2 pooled				
Lump sum payment	1,451***	139.2*	0.0484*	121.7**
	(159.1)	(71.40)	(0.0247)	(58.81)
Mean dep. var., weekly payment group	1836	468.5	0.106	223.5
Number of observations	689	689	699	699
Round 1 only				
Lump sum payment	1,252***	-4.320	0.00396	52.51
	(245.2)	(109.6)	(0.0381)	(79.20)
Mean dep. var., weekly payment group	2036	543.0	0.149	274.3
Number of observations	343	343	348	348
Round 2 only				
Lump sum payment	1,658***	274.0***	0.0949***	196.2**
	(190.6)	(96.82)	(0.0327)	(84.80)
Mean dep. var., weekly payment group	1634	393.1	0.0632	172.4
Number of observations	346	346	351	351

Notes: Stars indicate significance at 10% (*), 5% (**), and 1% (***) levels. Regressions in Panel A are run on pooled data from round 1 and round 2 (standard errors clustered at the individual level in parentheses); Panels B & C are run separately on round 1 and round 2, respectively (robust standard errors in parentheses) . USD 1 is ca. MK 400 for study period. All regressions include stratification cell fixed effects and an index of baseline asset ownership based on first principal components, difference in days between date of interview and the preceding weekend, baseline total spending and -if available- the baseline value of the outcome variable. For complete variable definitions, see Appendix B, and Table 2 for summary statistics.

CHAPTER III

Making the Grade: Understanding What Works for Teaching Literacy in Rural Uganda

From a work with Rebecca Thornton.

3.1 Introduction

One of the major development successes of the past several decades has been the increased access to primary education. Primary school enrollment and completion rates have grown worldwide, and particularly in sub-Saharan Africa, which had the world's highest increase in primary school enrollment – up 42 percent from 1999 to 2006 (UNESCO 2011). However, successes in getting students to school have not been accompanied by improvements in learning or increases in basic metrics such as literacy. Governments and policy organizations have now shifted their focus to raising the quality of education, rather than just its quantity, and translating years of education into improved learning. A large body of research has shed light on the effectiveness of various education interventions on learning. However, the majority have shown relatively small effects. A meta-analysis of 77 randomized trials of primary education programs in developing countries found the average mean effect size was an increase in 0.14 standard deviations (McEwan 2014). This paper evaluates a primary literacy program in rural Uganda for Primary 1 students, using a randomized experiment. The literacy program that we evaluate combines multiple educational components including a mother-tongue-first instructional approach, a revised curriculum, locally-appropriate teaching materials, extensive teacher support and training, and parent engagement. In contrast to previous studies, we find large, precisely measured effects of the program on learning: letter name knowledge, improves by 1.04 SDs of the control-group score distribution. Taking the average across an index of all six components of a standardized reading test, the effect is still 0.80 SDs. The experiment also studies a more-scalable, lower-cost version of the pro-

gram in order to help shed light on issues of scalability and cost-effectiveness. The second variant is implemented at significantly lower cost, by conducting teacher training and monitoring through the existing Coordinating Centre Tutors, government employees charged with training and supporting primary school teachers in Uganda. It also provides fewer teaching materials, in particular omitting the writing slates provided to the full-cost version of the program. This reduced-cost version of the program has smaller effects, improving letter name knowledge scores by 0.42 SDs and the index of all reading test components by just 0.15 SDs, with the latter not reaching conventional levels of statistical significance. We examine other outcomes to shed light on the possibly mechanisms for the large effects. We find through student surveys that students increase their confidence in their ability and there is suggestive evidence that they increase their enthusiasm – although not effort – in school. We also find differences in teachers behavior in the classroom where they shifted to mother-tongue instruction and activities, and spent less time bringing students back on task. A cost-effectiveness comparison of the two programs reveals the low-cost version to be slightly more cost-effective than the full-cost one, at 0.09 SDs of letter name knowledge per dollar as opposed to 0.07 for the full-cost variant. However, focusing on the "headline" measure of letter name knowledge hides significant drawbacks to the low-cost version of the program: the cost-effectiveness result is reversed when considering the overall reading score index, and the low-cost version of the program causes a small (but statistically-insignificant) decline in students' English speaking ability, whereas the full-cost version improves performance on the subtests of the English exam that are free-form and open-ended. Most concerningly, the low-cost program causes large and statistically-significant reductions in several aspects of writing ability – of about 0.3 SDs – relative to the control group. These reductions are despite the fact that on the writing test the "headline" measure (in this case the ability to write one's name) once again improves. In contrast, the full-cost version of the program improves writing scores across the board, with the effects on several exam components being statistically significant. The remainder of this paper proceeds as follows. In Section 3.2, we describe the details of the literacy intervention. Section 3.3 describes the research design and Section 3.4 the sources of data we use. Section 3.5 outlines our empirical strategy. Our results, including the effects of the two program variants on test scores, and their effects on intermediate outcomes that shed light on the mechanisms at work, are presented in Section 3.6. Section 3.8 concludes.

3.2 NULP Primary Literacy Program

3.2.1 Background

We evaluate a primary literacy-promotion program called the Northern Uganda Literacy Project (NULP), developed by Mango Tree Educational Enterprises Uganda.[1] Mango Tree, a private, locally-owned education company, has been operating in northern Uganda in the Lango Sub-region since 2009. Within this area there are over two million people, mostly of the Langi tribe, who speak Leblango. A civil war led by the Lord's Resistance Army from 1987-2007 had a devastating impact on the region, which to date suffers severe infrastructure shortages, extreme poverty and poor access to quality education. In addition to these challenges, the region's schools show extremely poor learning outcomes, especially in terms of literacy. An assessment of early grade reading conducted by RTI in 2009 showed that over 80 percent of students in the Lango Sub-region were nonreaders at the end of P2, meaning that they could not read a single word out of a chosen paragraph. Another assessment from November 2010 found that almost none of students in the study could recognize and read a single letter by the end of P1.

3.2.2 Mango Tree Model of Instruction

To address this challenge, Mango Tree began working with teachers, local language boards, and government officials in 2009, to develop an innovative new educational paradigm, the NULP. The NULP focuses on P1 to P3 students, employing a mother-tongue-first instructional approach and extensive teacher support and training. We outline the main features of the program below.

Mother-Tongue Instruction

The basis of the NULP model is mother tongue instruction, which means that children are taught in the language they grew up speaking, rather than a different language that they first encounter in school. It is common across the world, and especially in Africa, for children to enroll in school and immediately begin learning in a language that they do not understand. This other language is frequently a colonial language; English is used as the de facto language of instruction in primary schools throughout Uganda. Learning may happen through complete immersion, where all subjects are taught in English, or where some subjects are taught in the students' mother tongue while students are also immersed

[1] Uganda's primary school system numbers the levels from P1 up to P7. P1 is the first grade level offered in government schools, and the official minimum age for enrollment is 6.

in English speaking, reading, and writing from the first day of school. Bilingual education has numerous benefits, and parents and teachers often have strong preferences for students to learn English. However, full immersion in reading and writing a language that students do not yet know can also have powerful drawbacks. Children often simply learn to memorize and copy words, letters, and numbers, without gaining any understanding of what they are doing or how it connects to spoken words or meaning. This works against research that finds that students learn best by building on what they already know and working from simple concepts to more complex ones. Previous research suggests that education systems that use a language unfamiliar to children in school, and simply hope that children will pick up that language, are failing (Webley 2006). Despite the common practice of immersing students in a national language for literacy class, several countries including Uganda have explicit policies mandating "mother-tongue instruction" for primary schools, which means that the primary language of instruction should be students' native language. In Uganda, this policy is not entirely enforced by schools, and teachers on not trained in local orthographies. The Mango Tree program teaches literacy in P1 entirely in the students' mother tongue. Oral English is given as a subject, but no English is written on the board or for students to read.

Teacher training and on-going support

The NULP provides extensive training and support for teachers in the program's classrooms. Mango Tree's training approach focuses on the uptake of practical and appropriate classroom skills. The first teacher training module involves a five day residential workshop on the Leblango orthography, including grammatical features and letter names and sounds. Teachers also undergo three additional intensive, residential trainings on literacy methods (both whole language and phonics approaches) during the school holidays. Teachers also participate in six Saturday in-service training workshops throughout the school year.

Teaching Materials

Mango Tree developed NULP materials continuously since 2010 in partnership with teachers and local government education officials. Mango Tree's primers and readers are small and easy to store in the classroom. Classrooms are provided with slates that allow each student to practice writing individually, and to assist the teacher to review their work effectively in classes of over 100 students with limited walking space (children can hold up their slates to show their work).

Pace and Repetition

The NULP model introduces content slowly, providing time for repetition and revision. This slower instructional pace allows for students to develop necessary pre-and early literacy skills and gives more time to prepare teachers for phonics instruction. Every teacher is also provided with teachers' guides that provide a script for each literacy lesson. Four literacy lessons are taught each day in the same order. This provides teachers, who have hugely varying and underdeveloped capacities and experiences creating effective literacy lesson plans, with easy-to-remember steps that become routine over time.

Parent and Community Engagement

Part of the NULP model involves engaging with parents and the local community to communicate the benefits of mother tongue instruction. Three parent meetings are held each year to discuss language of instruction, as well as how to assess and support children's learning and literacy development at home. This involves parent training on how to interpret their child's literacy report card, and how to use a simple reading assessment tool at home. These tools are developed by the program; the assessment allows parents to know their child's performance in key literacy skills.

3.2.3 Lower-Cost Model of Instruction

To reach scale, an educational program must be both cost-effective, and sustainable in the rural African setting. In terms of cost, the most expensive inputs of the Mango Tree program are the materials (readers, teacher manuals and slates) and teacher training and support. In addition to measuring the effect of the full Mango Tree program, we also tested the mode of delivery of the program with a scaled down model of the program. The lower-cost model of instruction was explicitly designed to realistically demonstrate how the program might be scaled up for adoption by a larger set of schools. This involved cutting the per-school cost of implementation in two ways. First, the set of materials provided, and the intensity and cost of the trainings and support provided, was reduced relative to the standard Mango Tree Program. Second, the trainings and support for teachers were provided through the employees of the Ministry of Education and Sports (MoES) who are ordinarily tasked with training and supervising teachers in Ugandan primary schools. These employees are known as Coordinating Centre Tutors (CCTs), because each one manages a set of schools near an administrative office known as a Coordinating Centre (CC). We refer to this low-cost version of the program as the CCT Program. In this study we compare the Standard Mango Tree Program and the Government Administered Program to a control group. The details of the

inputs of each program are found in Table 3.1 and Appendix J.

3.3 Research Design

In this section, we describe the research design that underlies this study. Figure 3.1 illustrates the selection and randomization.

3.3.1 Sample

Selection of Schools

The evaluation was conducted among 38 eligible schools located in the five Coordinating Centres with existing Mango Tree-supported schools. Schools were eligible for the study if they they met specific Mango Tree program criteria including: having two P1 classrooms and teachers, having desks and lockable cabinets for each P1 class, a student-to-teacher ratio of no more than 135 during the 2012 school year in grades P1 to P3, being located less than 20 km from the CC headquarters, being accessible by road year round, having a head teacher regarded as "engaged" by the coordinating centre tutor (CCT), and not having previously received Mango Tree-support. These criteria were deemed important by Mango Tree to support the specific aspects of the NULP instructional model. In addition, head teachers agreed to assign the two best early primary teachers in the school to the P1 classrooms. To determine eligibility, school-level data were collected from each school in late 2012. Out of 99 total schools, 38 met these criteria. Each head teacher signed a contract with Mango Tree outlining the guidelines for participation in the evaluation. These contracts had credibility: Mango Tree had used them in previous years in schools where it was piloting the NULP, and schools that did not adhere to the contracts lost Mango Tree support. All schools adhered to the contracts in 2013, so the contracts did not lead any of them to be removed from the study.

Selection of Students

During the first two weeks of the 2013 academic year, enumerators collected enrollment rosters from the P1 classrooms of each school in the study. From these rosters, we generated an ordered list of 70 randomly-selected students, stratified by classroom and gender. Baseline exams were conducted during the third and fourth weeks of school (described below). The first 50 students on the list from each school who were present in the school on the day of

the baseline exams were selected into the sample.[2] These 1900 students from the 38 study schools comprise our baseline sample.

3.3.2 Randomization

The 38 schools in the study were assigned to one of three study arms via public lottery: control schools, Mango Tree-administered program schools, and Government-administered program schools. Prior to the lottery, the schools were grouped into stratification cells by the researchers based on the schools' CC, total P1 enrollment, and distance to the CC. The lottery – held publicly at a stakeholder meeting – proceeded separately for schools in each stratification cell with representatives drawing tokens indicating treatment status from an urn. We discuss tests for balance of baseline sample characteristics across treatment arms below.

3.4 Data

Our primary learning outcomes are measured by a set of examinations conducted at the beginning and end of the school year to assess student performance in reading and writing Leblango, and in speaking English. These data – as well as surveys among students and their parents – were collected among our baseline sample of 1900 students. In addition, we use data from teachers surveys, and classroom visits that collected attendance, enrollment, and conducted classroom observations. The remainder of this section first describes the data sources and then presents summary statistics from the baseline exams.

3.4.1 Student Examinations

Baseline tests were conducted in the third and fourth week of the school year among the baseline sample of 1900 students. Endline tests were conducted during the last two weeks of the school year, in late November 2013. Of the students tested at the baseline, 78 percent were also found for endline exams. This gives us a longitudinal sample of 1481 students, which we use in our main student analysis (attrition across treatment arms is discussed below). Exams were administered by trained examiners hired specifically for the testing process. Examiners were not otherwise affiliated with Mango Tree, and were blinded to the study arm assignments of the schools they visited. Two of the tests, the EGRA and the Oral English Test, were conducted one-on-one by examiners sitting with individual students, making use of visual aids. The examiners marked each question correct or incorrect during

[2] If this process did not yield at least 50 pupils, research assistants proceeded through the list of all remaining pupils and selected every seventh one.

the exam. The third test, the Writing Test, was conducted in a group setting with a single examiner handing out materials and instructing pupils to write a story. We describe each of the tests in detail below.

Early Grade Reading Assessment (EGRA)

Our main outcomes of interest come from the Early Grade Reading Assessment (EGRA). The EGRA is an internationally recognized exam designed to serve as an "assessment of the first steps students take in learning to read: recognizing letters of the alphabet, reading simple words, and understanding sentences and paragraphs" (RTI International 2009). It has been adapted to dozens of languages and implemented in nearly 70 countries around the world (Dubeck and Gove 2015). In 2009, it was adapted to Luganda and Lango and used in Uganda to assess the reading ability of 2000 students in 50 schools across the country. We use this same adaptation of the EGRA to Lango, which covers six components of reading ability: letter name knowledge, initial sound identification, familiar word recognition, invented word recognition, oral reading fluency, and reading comprehension. The first four components involve students attempting to read letters, sounds, and both real and invented words from tables that are shown to them. The last two have students attempt to read a simple passage aloud and then answer comprehension questions about it. Because Mango Tree's main teaching objective in P1 is for students to learn the names of the letters of the alphabet, the letter name knowledge component of the test is of particular interest in evaluating the success of the program.

Oral English

The eventual goal of both the standard government curriculum and the NULP is for students to successfully transition to English by P5. One potential question about local language-first education is the extent to which it increases or inhibits students' progress in learning to speak, and eventually to read and write, in English. We therefore administered a simple oral examination – designed by Mango Tree – that asks students to answer basic English vocabulary questions based on pictures. The oral English examination has three sections. The first focuses on vocabulary and counting skills, asking students to point to a specific object in a picture named in English, and count how many there are. The second section evaluates students on their vocabulary and sentence structure abilities, asking them what a specific person in a picture is doing and what the name of a particular object is. The third section is more open-ended – it presents students with a picture of a scene and asks

them what objects and which people they can see in the picture.[3] In addition to measuring students' ability to speak English, we also wanted to capture the effects of the program on students' ability to read English words. The endline exams therefore added an additional test which asked students to read a list of eighteen words commonly taught in P1 (in the standard government curriculum). Rote memorization of how to read basic words in English aloud is a common technique in P1 classrooms in the Langa sub-Region. The NULP contrasts sharply with that practice, and does not teach any English reading during P1.

Writing

To capture improvements in students' ability to write, we made use of a writing test designed by Mango Tree and previously used to monitor writing skill acquisition in their pilot-testing of the NULP. Students completed the tests at the schools and were scored off-site by an expert in writing acquisition among children in the Langa sub-Region. The test has two broad sections. In the first section, students are asked to write their names.[4] Langi names are divided into an African surname, typically written first, and an English given name, typically written second. Surnames come from a small set of names that are passed down within extended families, with a known spelling in the Leblango orthography. Given names also come from a small list of names with known spellings. Each name was score separately in two categories: spelling and capitalization. Ability to write one's name is a major goal that Mango Tree sets for P1 students in terms of writing acquisition. In the second section of the test, students were asked to write a story about what they like to do with their friends, and to draw a picture to illustrate the story. The picture was unscored, but served to keep children occupied who could not write anything. The story was scored in seven categories: ideas, organization, voice, word choice, sentence fluency, conventions, and presentation.[5]

Combined Exam Score Indices

Our main learning outcomes are measured by the endline exams: reading, using the EGRA, English speaking, using the Oral English Test, and Leblango writing, using the Writing Test. Each of these exams has several modules, designed to test distinct but aspects of a child's ability rather than to produce a single overall score. The modules differ in

[3] The beginning instructions for the test are explained in Langa, and the tests themselves are conducted in English, with the examiner asking, for example, "What can you see?" (for subtest 3). As with the EGRA, the oral English examinations were conducted one-on-one with the students by trained examiners (they immediately followed the EGRA for each student).

[4] This is a purely evaluative exercise; exams were matched to students using pre-printed ID numbers.

[5] Presentation was added as a scoring category for endline and was not included at baseline.

their number of questions and some are scored based on a student's speed while others are untimed. We present the effects on each module separately, but a key question is whether the program has overall effects on each test – and how large those effects are. One challenge is that while there are guidelines for scoring each section of the EGRA, there is no defined system for combining the scores. The same issue holds for the other two tests. To measure the effect of the program on students' overall exam performance, we construct a principal components score index by normalizing each of the test modules against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006). Our results are robust to alternative methods of index construction.[6]

3.4.2 Surveys

Our analysis also makes use of two surveys, one for students and the other for teachers. Both surveys were conducted at the same time as the endline student examinations. The student surveys were a brief set of age-appropriate questions that asked them about their attitudes toward school, their effort, and their perceptions of their own ability and performance. Teacher surveys were designed to capture basic demographic details, as well as attitudes towards school and local language education. The teacher surveys also included details about teaching history, duties at the school, and time use.

3.4.3 Classroom Visits

Attendance and Enrollment

In addition to the baseline and endline examinations at each school, enumerators were also sent to each school three times during the school year to collect additional supporting data on the intervention. These visits took place in July, August, and October, so two visits occurred during the second term of the school year, and one occurred during the third (and last) term of the year. During these visits, enumerators collected data on attendance for all students in P1, as well as data on any new student enrollment. Attendance data was collected using the enrollment rosters. Enumerators noted whether each student on the list was present.

[6] Our estimated effects for the EGRA and the Writing Test are still statistically significant, and slightly larger, for an alternative index that takes the unweighted mean across test modules, following Kling, Liebman and Katz 2007. The estimated effect on the Oral English Test is nearly unchanged. To make these alternative indices, we normalize each module's endline score against the control-group endline score distribution for that module. We then take the simple average of the normalize scores across all the modules.

Classroom Observations

During the same visits at which they collected the attendance and enrollment data, enumerators also conducted classroom observations. These were detailed observations of two lessons in each of the school's two classrooms. These observations captured information about teaching strategies, student behavior and engagement, discipline, language of instruction, and a breakdown of the focus of each lesson on different topics. Enumerators were sent to the schools with paper forms with check boxes to note basic details about the school and classroom, as well as detailed information on each 30-minute lesson. School and classroom details included the teacher's name, number of students in the class, teaching and learning materials that were in the classroom, and which lesson was observed. The details about the lesson were broken up into three 10-minute blocks. For each block, the enumerator captured the start and end time, and ticked boxes to indicate that a teacher had engaged in a range of actions during the block such as referring to the teaching guide and ignoring off-task students. They also noted the share of time the teacher spent speaking English and Leblango. In addition to capturing details about teacher behavior, the enumerators also recorded student actions in three categories: reading, writing, and speaking/listening. Enumerators indicated the number of minutes (out of the 10 in the block) spent on each category and the share of students participating in the activity. They then ticked boxes to note whether they saw students do various actions, such as doing the activity in a group or on their own, using a specific material such as a slate for writing or a reader for reading, and whether English or Leblango was used.

3.4.4 Baseline Characteristics

Tables 3.2 and 3.3 presents baseline summary statistics. We focus on the first column of Table 3.2, which presents the mean of each variable among the control group, and column 2 of Table 3.3, which shows the share of students who got any answers right on each component of the EGRA. The sample is slightly less than half male and the mean age at the beginning of P1 is 7. Very few students got any correct answers on the baseline EGRA – just 40% got a single question right on the entire exam. Looking to the individual components, only 15% could identify a single letter of the alphabet, and even lower proportions scored any points on the more-advanced reading skills.[7] One notable exception to this pattern is the

[7] The maximum raw score on the letter name-knowledge section of the EGRA is 100 letter names correct (some letters are repeated). However, consistent with the EGRA protocol students who did not get any answers right in the first ten letter names were skipped ahead to the next section to minimize embarrassment and discomfort. Thus a zero score on this section of the exam indicates that the student got no answers correct out of the first ten.

Reading Comprehension questions, which have the highest proportion of students getting a question right at 30%.[8] Students were even less successful on the Writing Test: more than three quarters scored zero points on the entire exam. Scores were higher on the Oral English Test, probably because it involved no reading and thus relied on skills that students might have already begun to develop before beginning school.

3.5 Empirical Strategy

3.5.1 Main Econometric Approach

Our main outcomes of interest are student performance on three exams: the EGRA, the Oral English Test, and the Writing Test. For each exam, we examine effects on each component separately, as well as estimating the overall impact of the program on performance using combined outcome index measure. Our empirical strategy relies on the randomized assignment of schools to the three study arms for identification: randomization guarantees that the students in the three study arms will be balanced, in expectation, on observed and unobserved pre-treatment variables, allowing us to attribute any post-treatment differences in outcomes to the effect of the program the school received. While the treatment was assigned at the school level, our main analyses focus on student-level outcomes. We run regressions of the form:

$$y_{is} = \beta_0 + \beta_1 \text{MTSchool}_s + \beta_2 \text{GovtSchool}_s + \mathbf{L}_s'\gamma + \eta y_{is}^{\text{baseline}} + \epsilon_{is} \tag{3.1}$$

$$\tag{3.2}$$

Here i indexes students and s indexes schools. y_{is} is a student's outcome at endline – typically his or her score on a particular exam or exam component. $\mathbf{L}_s$ is a vector of indicator variables for the stratification group that a school was in for the public lottery that assigned schools to study arms; we control for them, following Bruhn and McKenzie (2009), to increase the precision of our estimates. MTSchool_s and GovtSchool_s are indicators for the school being in the Mango Tree- or Government-administered version of the program, with the omitted category being in the control group. ϵ_{is} is a mean-zero error term. To account for the fact that the treatment was randomized at the school level rather than at

[8] This is higher than the share who were able to correctly read any of the words from the passage aloud. This may be because students are better able to make words out on the page than to correctly pronounce them out loud, and also may the result of lenient scoring by the examiners. This pattern is identical across study arms. The same pattern also exists for earlier administrations of the EGRA by Mango Tree in its piloting of the NULP.

the student or teacher level, we uniformly report standard errors that are clustered by school. β_1 and β_2 are our estimates of the effects of the MT and CCT programs, respectively, on exam scores. To restate the identification assumption above in terms of the variables in our estimating equation, consistent estimation of β_1 and β_2 requires that MTSchool_s and GovtSchool_s are independent of the error term ϵ once we condition on the other controls in the regression. This is guaranteed by process that assigned schools to study arms, which was random conditional on stratification cell. We next discuss baseline balance in further detail. Our preferred specifications also control for the baseline value of the outcome variable, y_{is}^{baseline}, whenever possible. We do this for two principal reasons. First, we stated that this would be our preferred specification in our pre-specified analysis plan.[9] Second, it helps address the potential baseline imbalance on some of the test score outcomes described in Section 3.4.1 above. In practice, baseline values for the outcome variables are available only for the student test scores. Therefore, we include this control only in our test score regressions. We also show that our results are not materially affected by the exclusion of this control. In addition to using equation 3.1 to estimate the effects of the two NULP variants on test scores, we also use the same specification to study its effects on student aspirations.

3.5.2 Baseline Balance

Table 3.2 provides evidence of balance across the study arms. The three sets of columns present means by study arm for three different samples of students: the baseline sample, the longitudinal sample, and the set of students who were lost to followup. We formally test for differences between study arms by estimating

$$y_{is} = \beta_0 + \beta_1 \text{MTSchool}_s + \beta_2 \text{GovtSchool}_s + \mathbf{L}_s'\gamma + \tau T_s + \epsilon_{is} \tag{3.3}$$

$$\tag{3.4}$$

Here we control for $\mathbf{L}_s$ for the same reasons noted above. We also control for the date of the baseline exams, τT_s, because it is not balanced across study arms, and because there is evidence of a time trend in scores on Oral English Test and the Writing Test, possibly because the examiners gained experience administering the tests. Statistically significant differences are indicated by stars next to the Mango Tree Program and Government Program means. A comparison of the first three columns shows that the baseline sample is relatively well-balanced across study arms. There are no significant differences in demographics: the sample is slightly less than half male and seven years old on average at the beginning of P1.

[9] See INSERT WEBSITE WHEN PUBLIC for details.

The PCA indices for the exam scores show that overall test performance is roughly the same across study arms. Looking at the detailed list of test components, however, there is evidence of a small degree of imbalance. The Government Program performs slightly worse than the control group on the Reading Comprehension (p¡0.05) of the EGRA, while both versions of the program score somewhat lower than the control group on two of the Oral English Test components. Students in the Mango Tree program score significantly better on the portion of the Writing Test that asks them to write their African names. Columns 4 through 6 replicate columns 1 through 3, but for the longitudinal sample that we actually use to analyze the NULP's effects. Comparing the coefficients and statistically-significant p-values, we see that the same patterns hold for this sample as for the baseline sample: it is balanced on demographics and overall test performance, but with some significant differences in the individual test components. Columns 7 through 9 present variable means by study arm for the set of students who were lost to followup – members of the baseline sample who are not in the longitudinal sample. This sample uniformly performs worse on the baseline tests than the longitudinal sample does. This pattern is balanced across study arms in terms of the overall test score indices, but there is some evidence of differences in performance among attritters on certain test components. However, these differences are not large enough to lead to change the pattern of imbalance for the longitudinal sample relative to the baseline sample. The small degree of imbalance in baseline test scores could have arisen from three sources. First, the random assignment of schools to study arms, which generates balance on all observed and unobserved variables in expectation, could led to an imbalanced sample in realization. Second, the same applies to the random samples of students within schools. Militating against these possibilities somewhat is the fact that the sample looks balanced on demographic factors. A third possible source of imbalance is that the baseline exams took place after the school year had begun, and so they may have picked up some initial, short-run effects of the treatment. The direction of the differences across study arms is consistent with what we would expect from the NULP's emphasis on the use of Leblango instead of English and is focus on teaching students beginning writing skills. The small amount of baseline imbalance in our sample motivates our choice to control for baseline values of the outcome variable in all our test-score regressions.

3.5.3 Additional Specifications

We supplement the student-level analyses in equation 3.1 above with several others. First, we use the set of classroom observations. In these, each school in the study was visited three times; during each visit, both classrooms in the school were observed during two separate lessons. To analyze these data we estimate:

$$y_{lrcs} = \beta_0 + \beta_1 \text{MTSchool}_s + \beta_2 \text{GovtSchool}_s + \mathbf{L}'_s\gamma + \mathbf{R}'_r\delta + \mathbf{E}'_{rcs}\rho + \mathbf{D}'_{rcs}\mu + \epsilon_{lrcs} \tag{3.5}$$

$$\tag{3.6}$$

Here s indexes schools, c indexes classrooms, r indexes the round of the visit and l indexes the lesson being observed. In addition to the variables that appear in equation 3.1 above, equation 3.5 adds as controls vectors of indicator variables for the round of the observation ($\mathbf{R}_r$), the enumerator conducting the observation ($\mathbf{E}_{rcs}$), and the day of week of the observation ($\mathbf{D}_{rcs}$).[10] ϵ_{lrcs} is a mean-zero error term. Enrollment data is collected as total numbers at the school level, so we analyze it at the school level as well:

$$y_s = \beta_0 + \beta_1 \text{MTSchool}_s + \beta_2 \text{GovtSchool}_s + \mathbf{L}'_s\gamma + \epsilon_s \tag{3.7}$$

$$\tag{3.8}$$

Here s indexes schools, ϵ_s is a mean-zero school-level error term, and all other variables are defined in the same way as in equation 3.1. We also examine the sensitivity of our results to using the log of enrollment instead of its level.

We use information from the endline student and teacher surveys to study how the program affected effort (time use, interactions with parents) beliefs and attitudes, and participation in training. To study these, we estimate program effects at the student- and teacher-level by estimating:

$$y_{is} = \beta_0 + \beta_1 \text{MTSchool}_s + \beta_2 \text{GovtSchool}_s + \mathbf{L}'_s\gamma + \epsilon_{is} \tag{3.9}$$

$$\tag{3.10}$$

where i indexes either students or teachers; all other variables are defined as in equation 3.1.

3.6 Results

Our analysis first focuses on the effects of the two program variants on student exam scores. First, as a benchmark, we discuss the performance of P1 students under the status

[10] The classroom observation results are nearly identical in magnitude but less precisely estimated when we omit the enumerator and day-of-week fixed effects.

quo government curriculum – that is, student performance at the endline in control schools. We then turn to impacts on the EGRA, the Oral English Test, and the Writing Test.

subsection*Status Quo* Performance in Literacy at the end of P1

In addition to its use in measuring the impact of the NULP on literacy, the exam data we collected allows us illustrate the gains P1 students in the Lango sub-region make in terms of reading ability in the absence of the program. The blue bars in Panel B of Figures 3.2 and 3.3 show how students in the control schools performed on the EGRA at the end of P1; these changes are also summarized numerically in columns 5 and 6 of Table 3.3. At the end of one year of school, roughly 50% of students could not recognize a single letter of the alphabet (Figure 3.2 Panel B). Just over 20% could recognize between one and five letter names, and a similar fraction could recognize between six and twenty. Fewer than 10% of pupils could correctly identify more than twenty letters out of a total of 100 chances. The NULP sets learning the names of letters as a key goal for P1 students, arguing that it is a critical building block for more-advanced reading skills. Consistent with this claim, overall reading performance mirrors the performance on letter-name recognition. The blue bars in Panel B of Figure 3.3 show that 40% of all students could not answer a single question correctly on the entire EGRA. The remainder of Figure 3.3 Panel B confirms that overall EGRA performance is largely driven by letter name recognition in P1. A comparison between the first and second panels of Figures 3.2 and 3.3, focusing on the blue bars, reveals a staggering lack of improvement in reading over the course of P1. Over 80% of students enter P1 unable to recognize a single letter of the alphabet, and the majority of those students leave P1 having made no progress whatsoever. Overall EGRA scores do not look much better: 40% of students get at least one correct answer across the six components of the exam at the beginning of the school year, but that number rises to just 60% by the end of the year. A small number of highly-performing readers do much better than the typical student: the fraction of students, answering more than twenty questions right rises from negligible at the beginning of the year to 10% by the end of the year. But these top students leave the preponderance of their classmates far behind. The measured increases in exam scores in the control group form a natural basis for comparison for the effects of the two variants of the NULP on exam scores: we can compare the gains from the program to the typical gains experienced by a child during P1. We now turn to the impacts of the program on the EGRA, performance on which is our main outcome of interest.

3.6.1 Program Effects on EGRA Scores

The impacts of the two versions of the NULP on EGRA scores are shown in Table 3.4, which estimates equation 3.3. Column 2 presents the impact on students' knowledge of

letter names, the principal learning goal that Mango Tree sets for P1 students. The Mango Tree-administered version of the program has a very large impact on letter name knowledge: scores increase by 1.01 standard deviations. The government-administered program improves performances in recognizing the names of letters by 0.41 SDs, which is still a significant gain but less than half as much as the full-cost version of the program. Examining the effects of the two versions of the program on the other EGRA components reveals a more nuanced picture. The Mango Tree-administered program has strong effects on all six components that are uniformly significant at the 0.05 level. The government-administered program, however, has no statistically-significant effect on any EGRA component other than letter name knowledge. The low-cost version of the program, then, improved only the headline measure of literacy emphasized by Mango Tree, with no benefits to other, more advanced aspects of literacy. This finding is verified by Column 1 of Table 3.4, which presents estimates for the combined score index described in Section 4.1 above. The Mango Tree-administered program raises this index by 0.63 SDs, confirming that the large effect of the program on exam scores is not merely an artifact of focusing on knowledge of letter names. Even taking 0.63 SDs as our best estimate of the program's impact on reading ability, the effect of this program would be among the largest ever measured in a randomized trial of an education program (McEwan 2014). Moreover, we can reject gains smaller than 0.37 SDs at the 0.05 level; in the few cases where large effect sizes have been found in primary education programs, those effects have had wide confidence intervals that do not exclude much smaller impacts. The government-administered program's effect on the EGRA index is just 0.13 SDs and is statistically indistinguishable from zero. The estimated effects on EGRA performance are virtually unchanged when we omit the baseline exam score controls; see Appendix B.1 for a detailed discussion and tables. The huge magnitude of the benefits of the program for reading is evident from Panel B of Figure 3.2. It shows the distribution of endline letter name knowledge scores by study arm. The full-cost version of the NULP cuts the share of students that cannot recognize a single letter in nearly half, and nearly quadruples the share that can recognize 21 or more letters. The effects are similarly clear-cut in Panel B of Figure 3.3, which shows the distributions of the total number of points scored on the EGRA. The low-cost variant of the NULP achieves smaller improvements in both letter name recognition and overall EGRA performance. It shifts the score distribution to the right, but does so by a smaller degree than the full-cost variant.

3.6.2 Program Effects on English Speaking and Word-Recognition Ability

Since the NULP focuses on promoting the use of the local language, Lango, in classrooms, one area where the program could potentially have effects is on students' English

speaking skills. One concern parents and other stakeholders in the Lango sub-Region have expressed with mother-tongue curriculum is that it would crowd out English skills. Table 3.5 presents the effects of the two program variants on students' scores on the oral English examination, estimated using equation 3.1. Neither the Mango Tree-administered nor the government-administered version of the program had a robustly statistically-significant effect across the different examination components. Column 1 shows that the overall effect of the NULP on the combined score index is statistically insignificant for both program variants. The Mango Tree-administered version raises this index by 0.14 SDs, and the Government-administered version lowers it by 0.09 SDs. Although the overall effect of the program on English speaking ability is not statistically significant, the point estimates in the table still represent our best estimate of the effect of the program; these are uniformly negative for the government-administered program but mostly positive for the Mango Tree-administered version. Moreover, Columns 8 and 9 show that the Mango Tree-administered program had statistically-significant benefits for the third subtest, expressive vocabulary, which uses relatively open-ended questions about a scene ("What do you see?" and "Who do you see?") as opposed to the naming of specific objects and actions ("What is this?" "What is she doing?"). This is noteworthy because the status quo in P1 classrooms in the Lango sub-Region is to focus on the rote memorization of English words, as opposed to actual usage; while control-school students might have an automatic advantage on the closed-ended questions, NULP students are more likely to have gained on open-ended questions. The estimated effect of the Mango Tree-administered version of the program on students' expressive vocabulary is roughly 0.3 SDs for each of the two subtests, which provides suggestive evidence that, in addition to reading Lango, the program also improved students' actual English speaking ability. This argument is also buttressed by Column 10, in which the outcome is a separate test in which students were asked to read a set of 18 printed English words aloud. This is a task that the NULP does not have teachers spend any time on in P1, because English reading does not commence until P2. However, it is common in status quo classrooms in the Lango sub-Region. The test was designed to use words that are commonly used in English curricula in P1 classes; it thus captures the extent to which students have either actually learned to read these words in English or have memorized by rote what to say when they are pointed to. NULP students perform substantially worse on this task, by 0.21 SDs under the Government-administered version and by 0.29 SDs under the Mango Tree-administered version. The latter estimate is significant at the 0.05 level. This result, along with the results from the Oral English Test, suggest that there is no evidence that the NULP harms students' progress in learning English. While they do worse on a simple rote memorization task, they actually improve substantially in their ability to use English in an expressive and

open-ended manner.

3.6.3 Program Effects on Writing

We examine the effect of the two versions of the program on writing ability in Table 3.6, which shows impacts on Mango Tree's writing test, estimated using equation 3.1. Columns 2 and 3 show that both versions of the program have large effects on the first section of the exam, which asks students two write their first and last names. Learning to write one's name is the main goal of the NULP for P1 students. The Mango Tree-administered program also has positive effects on the second section, in which students are asked to write a short story (Columns 4 to 10). The combined writing test index rises by 0.42 SDs (Column 1), which is statistically significant at the 0.05 level. The government-administered program, however, has uniformly negative effects on the story-writing component of the exam, with the negative effects on Voice, Word Choice, and Presentation reaching significance at the $p = 0.05$ level.[11] The combined Writing Test score index falls by 0.17 SDs, although this drop is not statistically significant. This suggests that the government-administered version of the program significantly boosted the headline measure of writing ability – name writing – at the cost of progress in overall writing skills, and in particular the ability to actually write a passage.

3.7 Mechanisms of the NULP's Effects

Tables 3.4 through 3.6 illustrate that the full-cost version of the Mango Tree program has significant benefits for pupil literacy, with some evidence of ancillary benefits for English-speaking ability, while the reduced-cost version seems to achieve gains on only the most basic outcomes that are targeted as goals for P1 students – letter recognition and name writing, with no gains in other areas and statistically-significant losses on more advanced aspects of writing ability. The two variants of the program were randomly allocated as complete packages, so we cannot causally separate which parts of the program had the most benefits or where the downsides of the low-cost version are coming from. However, we can approach the question of why the program worked, and why the lower-cost version backfired in some areas, by looking at evidence on intermediate outcomes that may shed light on the program's mechanisms. In this section we discuss each set of intermediate outcomes in turn: the student surveys, the classroom observations, attendance and enrollment, and teacher

[11] One of the 12 control schools was mistakenly instructed to complete the Writing Test in English instead of Leblango. Our results include this school, with the test marked in English. Our findings are robust to dropping the stratification cell for this school from our sample – see Appendix B.2 for a detailed discussion.

surveys. We then draw general conclusions about what all these data sources tell us about the mechanisms behind the NULP's impacts on learning.

3.7.1 Changes in Student Effort, Beliefs, and Attitudes

To do this we begin by looking at students' responses on the age-appropriate surveys that we conducted during the endline exams. The effects of the two program variants are shown in Table 3.7. The effects are estimated using equation 3.1, but without controlling for baseline values of the outcome because no data was collected on these outcomes at baseline. Students in both versions of the program show evidence of increases in perceived ability. They are more likely to report that they think they will pass the PLE (primary leaving examination), a high-stakes test that determines secondary school admissions, at the end of primary school. The estimated increase is 2.2 percentage points for the Mango Tree-administered program and 1.5 percentage points for the government-administered program (column 1), over a very high base rate of 95%.[12] Likewise, students' perceived class rank improves by 0.15 SDs in the Mango Tree-administered program (no effect is seen for the government-administered version). We find mixed results on enthusiasm for school and future aspirations. No effects are evident on students preferring school to other activities or preferring literacy class to math (columns 2 and 3); the estimated effects are not just statistically insignificant but nearly zero in magnitude. However, we do see evidence of admiration for teachers and an appreciation for education: students in the Mango Tree-administered program are seven percentage points more likely to want to go into a career in education (column 4). This is offset by an eight percentage-point drop in desire to become a doctor or nurse (column 5). Since students could list only one career, and the NULP does not affect how amibitious of a career students want (column 9), this suggests that the most ambitious students in class now want to go into education instead of healthcare. Finally, a roughly zero effect is also seen for our measure of effort, practicing writing at home (column 6) This suggests that changes in student effort in literacy are not important drivers of the observed effects. Overall, the results from the survey suggest that there was some increase in student confidence and enthusiasm for school, and these effects are larger for the Mango Tree-administered program than for the government-administered version. This gap may help explain part of the gap between the impacts of the full-cost and reduced-cost versions of the program on student test scores.

[12] The actual pass rate is much lower: in most Ugandan schools, fewer than half of students who begin P1 even complete P7 and take the PLE, and a small fraction of those pass it.

3.7.2 Changes in Teacher and Student Behavior in the Classroom

The most likely mechanism for the program's effects is that it changes how teaching actually takes place in the classroom. To explore this, we examine data from a set of classroom observations that measured teacher (Table 3.8) and pupil behaviors (Tables 3.9, 3.10, and 3.11) during class. These four tables use regressions of the form specified in equation 3.5. Table 3.8 reveals that both variants of the program induced teachers to spend more of their time speaking in Lango, by twelve percentage points for the full-cost NULP and nine percentage points for the reduced-cost variant (Column 2). Teachers in the full version of the program were also more likely to move around the classroom – they were twelve percentage points less likely to simply remain at the front of the class (significant at the $p = 0.05$ level) and nine percentage points more likely to move freely throughout the classroom (not statistically significant). Teachers in both NULP variants were 6 percentage points more likely to be observed ignoring off-task students (Column 7), with no statistically-significant changes in the other outcomes. This is somewhat surprising, but it may reflect the establishment of a better overall classroom environment: in an ideal classroom full of readily-participating, on-task students, teachers will never have to bring students back onto task. Also, teacher training courses often encourage teachers to ignore off-task students rather than call attention to them. Table 3.9 shows differences across study arms in student behavior while working on reading tasks. Students in both versions of the NULP are more likely to be observed reading sounds, and students in the full-cost version are more likely to be seen reading full sentences. Both variants of the program are more likely to be reading out of readers or primers. The proportion of reading done in Leblango rises by 22%. Classes also spend a higher proportion of time on reading: an additional 0.7 minutes per ten-minute observation window for the full-cost version of the program, and 0.5 minutes for the reduced-cost version. This represents an increase of roughly 15% over the control-group mean of 3.7 minutes. In Table 3.10, we examine the changes in student behavior while writing. Students in the full-cost version of the NULP are 8 percentage points more likely to be observed drawing pictures, and 6% more likely to spend time writing their names. The Government-administered program shows a 6 percentage-point increase in the chance students will be seen air-writing, but there is no comparable effect for the Mango Tree-administered program. This may reflect the fact that the Government-administered version of the program did not include the writing slates. If students lacked their own exercise books to write in, this would force teachers to improvise if they want to their students to be able to practice writing. Another difference that is asymmetric across program variants is a 9% rise in the chance that students in the Mango Tree-administered version will be seen writing their own text. This is a gain of more than 100% over the control group mean, and helps explain

the large improvements in passage writing in that version of the program. The change in the amount of time spent on writing is not statistically significant, but is comparable in magnitude to the increase in time spent on reading: about 16% of the control-group mean of 1.2 minutes for the Mango Tree-administered version of the program, and 29% for the Government-administered version. Finally, Table 3.11 turns to changes in student behavior while speaking and listening. Students more than double the chance that they speak or listen in small groups, and the chance that students will be observed speaking and listening to the teacher falls by a comparable magnitude. This is consistent with a drop in the amount of rote memorization "call and response"-style learning that is typical in status quo schools in the Lango sub-Region. The share of speaking and listening done in Lango rises, which would fit into a story where students especially spend less time doing rote call-and-response in English, to memorize English words. Finally, the amount of time spent on speaking and listening falls by 16% of the control-group mean in the full-cost version of the program (significant at $p = 0.05$) and by 7% in the reduced-cost version (not statistically significant). This also matches a story in which the teacher engages in less call-and-response repetition of words and phrases as a way to memorize them.

3.7.3 Changes in Attendance and Enrollment

Teacher and student behavior during class can be thought of as variation at the intensive margin of effort. Another important factor is changes at the extensive margin: whether students and teachers show up for class at all. Table 3.12 shows estimated differences in student attendance and enrollment and teacher attendance across study arms. Columns 1 to 4 are estimated using equation 3.1 on the full sample of students enrolled in the schools at baseline. Column 5 is estimated at the school level using equation 3.7. Column 6 is estimated at the teacher level, using equation 3.9. There is no evidence of any differential changes in enrollment across study arms, nor of differences in teacher attendance. There is some evidence of a limited increase in attendance for students in the Mango Tree-administered version of the program (it rises by 5 percentage points, with p¡0.1), concentrated in the first visit to schools which happened early in the second term of the school year. Students in the Government-administered version of the program are 4 percentage points less likely to attend than control-group students. Though the p-value on this difference exceeds 0.1, the difference from the full cost version of the program is statistically significant, and is 9 percentage points over a base of 42% attendance. The lower attendance is concentrated toward the end of the school year. Part of the improvement in performance in the Mango Tree-administered version the NULP may be due to the simple fact that students are exposed to more teaching because they were in class for longer. The smaller gains in the low-cost

variant of the program can be ascribed in part to students spending less time in class than in the full-cost variant.

3.7.4 Changes in Teacher Effort, Beliefs, Attitudes, and Training

Our final ancillary data source for examining the mechanisms of the NULP's benefits is the endline teacher surveys, which were done at the same time as the endline exams. We estimate effects on the survey outcomes using equation 3.9, and the results are shown in table 3.13. The outcomes are grouped into three categories: columns 1 to 5 measure teacher effort; columns 6 to 10 measure teacher beliefs and attitudes, and columns 11 to 14 measure the main human capital input the NULP provides for teachers – training. Changes on teacher effort as a result of the program are fairly muted. The full-cost version of the program shows a marginally-significant increase in the amount of time spent on helping students outside of the classroom, but it is large in magnitude – 2 hours more per week, nearly as much as the control-group mean. There are no appreciable changes in interactions with parents: the number of parents the teacher met with during the school year is essentially unchanged, and this result is consistent with other outcome measures that we omit for space reasons. The one margin of effort where we detect effects is a significant increase in the chance a teacher has taught literacy classes (reading and writing), which rises from 61% in the control group to 80% in the Government-administered version of the program 92% in the Mango Tree-administered version. The NULP appears to reduce the division of labor across the two P1 teachers, which in the control group are more likely to split the literacy and non-literacy parts of class. While effort changes very little, we observe large shifts in beliefs and attitudes. Both variants of the NULP cause teachers to be 20 percentage points more likely to say they would still want to teach if they could go back and re-pick their career. Though this is significant only for the Government-administered NULP, and only at $p = 0.10$, pooling the two study arms for this outcome generates the same coefficient and significance at the $p = 0.05$ level. Teachers in the Mango Tree-administered program are less likely to blame teachers for students' failure to learn, which could mean they feel less frustrated when their students struggle. Teachers in the Government-administered program rate themselves 0.3 points lower than control teachers do on a 1-3 scale of relative performance. Both versions of the program sharply reduce teachers' satisfaction with the reading performance of higher-year students in their schools, suggesting an elevation of standards. Consistent with this, and in contrast with the students' self-perceptions, there is no change in teachers' beliefs about their students' ability to eventually pass the PLE. The overall pattern is one of higher standards for students, and some increase in satisfaction with teaching as a career. The effects of the program on training are interesting primarily because there is evidence of substitution of

the NULP's training opportunities for other ones that teachers might do instead. There are increases in the rate of attending any training and the total days of training attended, which is sensible because the NULP invests heavily in training teachers. This is reflected by an approximately 50 percentage-point increase in having attended a training provided by an NGO (Mango Tree is perceived locally as an NGO despite its status as a private-sector business). But there is a compensating decline of roughly half that magnitude in attending other training. This may mean that some of the training Mango Tree provides for the NULP simply substitutes for other valuable human capital investments that teachers would be making anyway. However, trainings for public sector workers are often seen not as ways to invest in skills but as opportunities to earn extra income through the per diem payments that are provided. Thus declines in attending other training may actually reflect increased effort put toward the broader job of teaching students.

3.7.5 Overview of Potential mechanisms

In this section we summarize the findings from our four ancillary datasets to address two key questions about the mechanisms of the NULP's effects on student performance. First, how exactly does the program achieve such enormous gains in student performance in reading? Second, why did the low-cost version of the program backfire in terms of writing, leading to decreases. Our ancillary data sources allow us to identify two broad mechanisms help us answer the first question: changes in beliefs and attitudes and changes in how class time is spent. The NULP causes marked changes in beliefs and attitudes: students become significantly more positively-inclined toward school, and teachers become marginally more positively-inclined toward teaching. Students believe more in their own ability, and teachers have higher standards for student performance. These attitudinal factors could improve learning in two ways. The first way is that they could reduce the cost of effort, leading to higher effort and better performance. This could operate in our setting through changes in effort that we do not observe or do not measure well – how closely students pay attention in class, for example, or how much of official class time teachers actually spend teaching (since teachers are likely to teach for the whole period while actually being observed). The second way is by making learning easier for psychological reasons that do not involve any changes in effort. We see no evidence of effects on student or teacher effort, which we mostly measure at the extensive margin – time spent on educational activities. Likewise, attendance is affected only marginally by the program. However, at the intensive margin of student and teacher effort – choices about how time is allocated within the fixed class periods – we observe large changes in behavior. More time is spent on reading and writing, and less on speaking/listening activities that probably reflect rote memorization through call-and

response. Students spend more reading time on making out sounds, which helps develop a key basic skill on which literacy is built. Much more time across all lessons is spent speaking Leblango instead of English. Broadly, teachers spend more of their class working on actual reading skills and focusing on Leblango, and less time having their students repeat English words they can see on the board but have trouble attaching meanings to. In addition to contributing to the large gains in literacy the program causes, the effects of this channel are also evident in performance in English. Students in the full-cost version of the program do much worse at reading common English words aloud but much better at actually speaking English. Our analysis of the four ancillary data sources also helps us address the second question. The larger gains in the full-cost version of the program can be ascribed partially to attendance. While the full-cost NULP did not change attendance significantly relative to the control schools, it did have significantly higher attendance than the reduced-cost version. This difference was particularly sharp toward the end of the year, which helps explain why more advanced reading did not improve in the reduced-cost version of the program, and also why writing might have actually gotten worse. The potential role of attendance raises the question of why attendance suffered in the schools that received the Government-administered version of the program. We cannot answer this question definitively, but we can raise a couple of possibilities. One is that students have gotten lost and stopped bothering to come to school. A second is that teachers have engaged in the practice, common in the Lango sub-Region, of chasing away the worse-performing children so they can focus on the better-performing students. A second potential contributor to lower performance in the Government-administered version of the program is reduced inputs. In particular, students in the Mango Tree-administered NULP were given slates and the ones in the Government-administered version were not. In simple terms, this could be thought of as an input x into an education production function $L = L(x, y)$ that takes x (and also other factors, y) as inputs, and has positive and diminishing marginal returns to x. Schools that do not get slates ($x = 0$) should then have lower levels of L, than schools with positive values of x, but there is no reason that removing the slates from the Government-administered NULP should lead to worse performance than in the control schools. What could explain the worsening in performance in writing is that the NULP actually alters the production functions for various writing outcomes. The NULP provides tightly-organized lesson plans, with specific ways of teaching different skills. In the absence of the slates, Mango Tree assumed that schools would simply substitute the students' own exercise books for writing practice. What happens when those are also not available? The evidence from the classroom observations suggests that teachers substitute classroom time toward the parts of the curriculum that are more manageable: students in the Government-administered program are more likely to practice

"air writing", where they practice tracing out words and letters with their fingers. They do not see the increases in practicing writing their own text experienced by the students in the Mango Tree-administered program. The conclusion we draw is that resource-strapped teachers may have focused the time they spent on writing on the more-manageable parts of the NULP curriculum, and ended up spending less time in the aggregate on actual useful writing skills. We conclude the results section with a discussion the cost-effectiveness of each variant of the program and the implications of our findings for the use of cost-effectiveness comparisons.

3.7.6 Cost-effectiveness

The large effects of the program naturally raise the question of its cost-effectiveness. While few other programs have shown such large gains, can the NULP compete on a value-per-dollar-spent basis? We examine this question in Table 3.14, which presents the cost per 0.2-SD gain and the SD gain per dollar spent for three different measures of the program's effectiveness. We begin with letter name knowledge, the most important outcome emphasized by Mango Tree for P1 students. The full-cost version of the program shows a gain of 0.7 SDs in this measure for each dollar spent, which trails the 0.9 SDs per dollar figure for the reduced-cost version of the program. Based on this outcome, it would cost an extra 56 cents per student to raise scores by 0.2 SDs. A more detailed analysis tells a different story. The second and third panels of the table present the same analysis for the overall indices of reading and writing ability. Relying on overall reading ability instead of just letter-name knowledge reverse the conclusions in terms of cost-effectiveness: the Mango Tree-administered version of the program yielded over twice the gains in performance per dollar compared to the government-administered version. The writing ability index shows an even starker pattern: because the government-administered version of the program actually reduced writing performance, the cost per 0.2-SD gain from that version of the program is undefined. Instead, each dollar spent on the government-administered version of the program will decrease writing performance by 0.04 SDs. This finding raises general questions about the use of cost-effectiveness measures in comparing the effects of education programs: they may mask considerable heterogeneity in program impacts across educational domains, leading to relatively cheap gains that come at potentially large hidden costs.

3.8 Conclusion

The educational challenges facing the Lango sub-Region of Northern Uganda typify those present across rural Africa. Literacy rates are low, little learning is achieved in schools

(despite recent successes in increasing enrollment), few students finish primary school, and the broader context is characterized by limited resources and a wide range of constraints on policymakers, educators, and parents. These challenges have helped lead to an increased call for cost-effective ways to promote learning in Africa. We evaluate one approach, developed by a Uganda-based company called Mango Tree Educational Enterprises, that focuses on promoting literacy through native language-first instruction in first-grade classrooms in the Lango sub-Region. We measure the impact of two variants of the program: a full-cost version, implemented by Mango Tree, and a reduced-cost version, which was implemented by government officials from Uganda's Ministry of Education and Sports. The full-cost version of the program causes large improvements in students' reading and writing ability across all measures of each, and we find suggestive evidence of gains in English speaking ability as well. The reduced-cost version is less effective: it shows improvements in the headline measures of student reading and writing that are the basic benchmarks for first-grade students in Uganda. Our analysis suggests that the gains in both versions of the program may be partly attributable to increased student confidence and enthusiasm, and to increased use of the students' native language in class. The larger improvements in the full-cost version of the program may arise in part from teachers having better control of their classroom and encouraging more interactive and participatory lessons. While the government-administered version of the program is less effective at improving literacy, it is much lower-cost and hence cheaper in terms of value-per-dollar for the headline measure of reading. However, this result hides significant variation in the impact of the low-cost version of the program on different measures of student performance. Students show no gains in more advanced aspects of reading and actually do worse than control schools on the advanced aspects of writing. The cost-effectiveness result is completely reversed when a more comprehensive measure of performance is used: it is the full-cost, Mango Tree-administered version of the program that provides more value per dollar in improving student performance. The cost-effectiveness of the Mango Tree-administered program is very high: at $2.76 per 0.2 SD gain in the benchmark component of the literacy exam for first-graders (and $4.41 per 0.2 SD gain for a comprehensive reading ability index) it is among the most cost-effective educational interventions to be measured in a randomized experiment (JPAL 2014). However, our findings indicate that these comparisons are highly sensitive to the outcome measure used, leading to not just small shifts in the exact figures but also total reversals in the sign of the measured gain per dollar (a switch from gains into losses). Our results also suggest that attempting to reach more students with an intervention by reducing monetary and physical inputs can backfire in specific ways. The low-cost version of the program substantially increases scores on the headline measures of reading and writing

ability for first-graders – the exact outcomes emphasized by Mango Tree in their internal assessments of how well the program is going. These gains come at a cost to other, less-prioritized measures: no gains in more-advanced reading skills were seen, and more-advanced aspects of writing actually got worse. One potential reason for this is that due to constrained resources, teachers in the reduced-cost version of the program may reduced the effort and inputs that would have gone toward the lower-priority aspects of reading and writing, in order to make sure they achieve the basic benchmarks. To the extent that this happened, it was without any high-stakes test to speak of: the results of the EGRA exams were not used in evaluating any of the teachers and were not even communicated back to them. Teachers' own intrinsic motivations, perhaps spurred by the program, were enough to cause unintended drawbacks from the program. Future research should explore the role of teacher effort and motivation to further document and understand this pattern; in addition, more research is needed to understand which components are critical to achieving the large across-the-board gains of the NULP, and which can be reduced or cut in order to deliver results in a truly cost-effective fashion.

Figure 3.1
Randomization of Schools to Study Arms

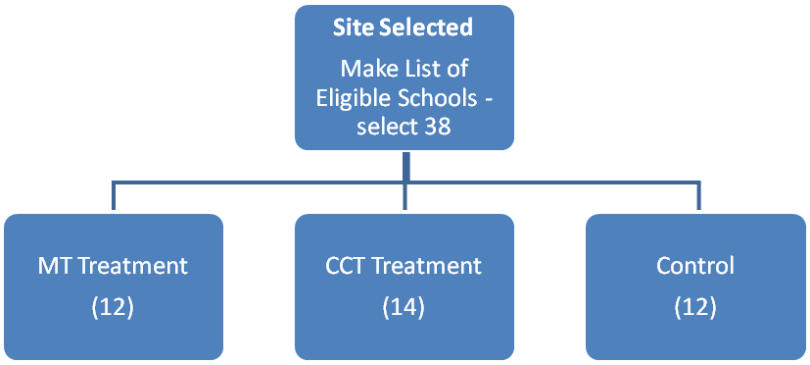
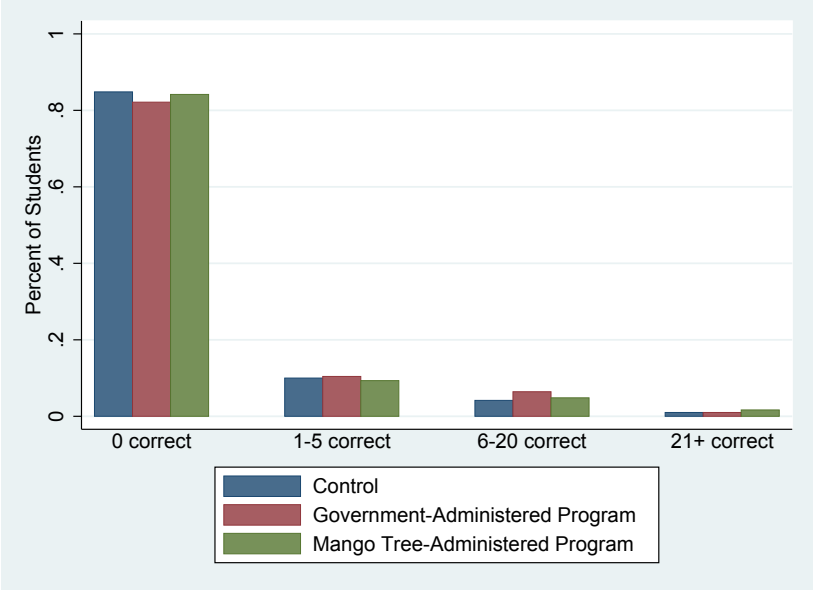
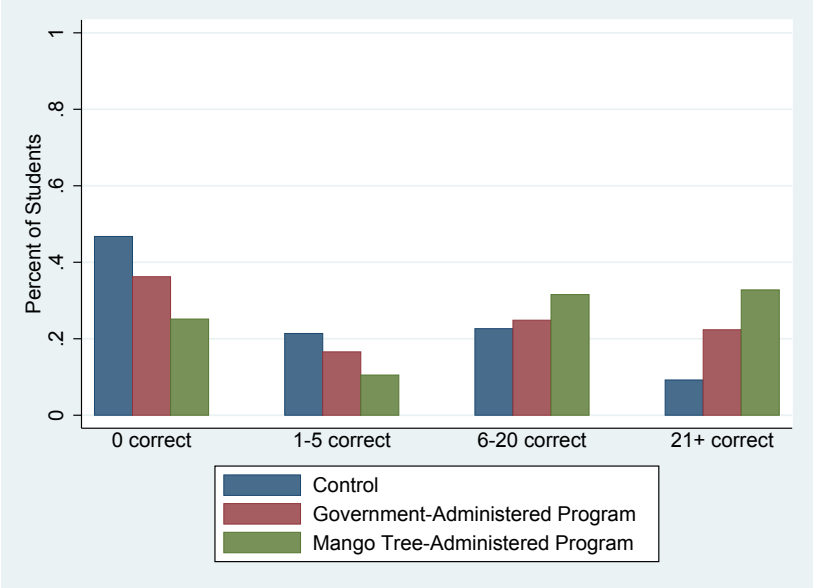


Figure 3.2
Performance on Letter Name Recognition by Study Arm
(Number of Letters Correctly Recognized)

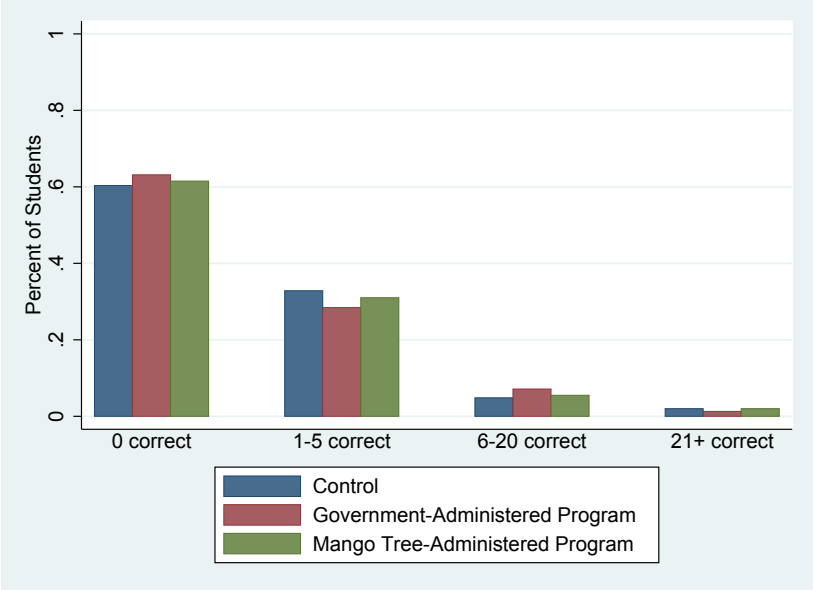


Panel A: Baseline

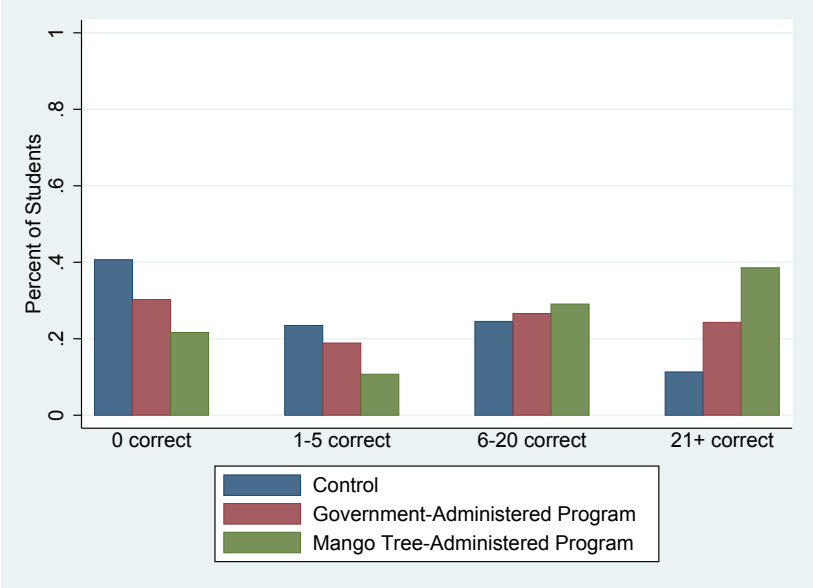


Panel B: Endline

Figure 3.3
Performance on Overall EGRA by Study Arm
(Total Questions Answered Correctly)



Panel A: Baseline



Panel B: Endline

Table 3.1
NULP Components by Study Arm

	NULP Components Received						
Study Arm	**Slates and Wall Clocks**	**Textbooks and Primers**	**Teachers Guides**	**Training and Support**	**Parent Meetings**	**Take a Book Home Activity**	**Monthly Radio Program**
MT Program (12)	X	X	X	X (MT)	X (MT-Run)	X	X
CCT Program (14)		X	X	X (CCT)	X (CCT-Run)		X
Control (12)							X

Table 3.2
Baseline Covariate Balance, Longitudinal Sample

	Baseline Sample			Longitudinal Sample			Lost to Followup		
	Control Mean (1)	MT Program Mean (2)	Govt. Program Mean (3)	Control Mean (4)	MT Program Mean (5)	Govt. Program Mean (6)	Control Mean (7)	MT Program Mean (8)	Govt. Program Mean (9)
Panel A: Students									
Present at Endline	0.795	0.808	0.741	1.000	1.000	1.000	0.000	0.000	0.000
Male	0.486	0.509	0.474	0.488	0.524	0.479	0.475	0.447	0.460
Age	7.018	7.078	7.017	7.013	7.052	7.000	7.041	7.191	7.066
EGRA									
PCA EGRA Score Index	0.000	0.006	-0.084	0.001	0.046	-0.100	-0.003	-0.160	-0.038
Letter Name Knowledge (Letters per Minute)	1.150	1.190	1.274	1.180	1.377	1.206	1.033	0.400*	1.469
Initial Sound Identification (Sounds Identified)	0.153	0.123	0.070	0.161	0.148	0.046	0.122	0.017	0.138
Familiar Word Reading (Words per Minute)	0.169	0.182	0.044	0.168	0.225	0.025	0.171	0.000	0.099
Invented Word Reading (Words per Minute)	0.094	0.132	0.029	0.084	0.163	0.008	0.130	0.000	0.088
Oral Reading Fluency (Words per Minute)	0.503	0.552	0.126	0.508	0.684	0.037	0.480	0.000	0.382
Reading Comprehension (Questions Correct)	0.327	0.318	0.266**	0.327	0.342	0.272*	0.325	0.217	0.249
Oral English Test									
PCA Oral English Score Index	-0.000	-0.326	-0.265	0.084	-0.284	-0.244	-0.327	-0.501	-0.325
Test 1 (Vocabulary)	1.645	1.122	1.254	1.774	1.212	1.274	1.146	0.739	1.199
Test 1 (Count)	0.452	0.177**	0.276*	0.501	0.181**	0.279**	0.260	0.157	0.265
Test 2a (Vocabulary)	0.637	0.240**	0.360**	0.669	0.245**	0.391*	0.512	0.217*	0.271***
Test 2a (Phrase Structure)	0.723	0.460	0.496	0.801	0.487	0.538	0.423	0.348	0.376
Test 2b (Vocabulary)	1.328	0.797*	1.091	1.400	0.866	1.106	1.049	0.504***	1.050
Test 2b (Phrase Structure)	1.378	1.197	0.941	1.520	1.285	0.992	0.829	0.826	0.796
Test 3 (Vocabulary, Expressive - Objects)	2.188	1.657	1.763	2.365	1.724	1.802	1.504	1.374	1.652
Test 3 (Vocabulary, Expressive - People)	1.392	1.347	1.223	1.505	1.414	1.206	0.951	1.061	1.271
Writing Test									
PCA Writing Score Index	0.000	-0.024	-0.165	0.067	0.001	-0.144	-0.259	-0.130*	-0.226
African Name (Surname) Spelling & Capitalization	0.180	0.323***	0.181	0.201	0.348***	0.193	0.098	0.217**	0.149
English Name (Given name) Spelling & Capitaliza	0.127	0.043	0.054*	0.145	0.043*	0.058*	0.057	0.043	0.044
Ideas	0.005	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.000
Organization	0.002	0.002	0.000	0.002	0.002	0.000	0.000	0.000	0.000
Voice	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Word Choice	0.057	0.023	0.016	0.069	0.023	0.019*	0.008	0.026	0.006
Sentence Fluency	0.005	0.000*	0.001	0.006	0.000*	0.002	0.000	0.000	0.000
Conventions	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Panel B: Teachers									
Present at Endline	0.870	0.917	0.929	1.000	1.000	1.000	0.000	0.000	0.000
Male	0.435	0.333	0.357	0.400	0.364	0.346	0.667	0.000	0.500
Age	39.826	39.125	44.857***	39.750	38.227	44.692***	40.333	49.000	47.000
Married	0.783	0.833	0.679	0.750	0.864	0.654	1.000	0.500	1.000
Monthly Household Income (PPP USD)	51.729	48.246	50.590	52.068	47.485	50.794	49.470	56.625	47.948
# Other People in Household	5.217	6.167	6.607*	5.200	6.227	6.846	5.333	5.500	3.500
Lives in Same Village as School	0.435	0.667	0.571	0.500	0.682	0.538	0.000	0.500	1.000
Lives in Government-Provided Teacher Housing	0.304	0.458	0.429	0.350	0.500	0.423	0.000	0.000	0.500
Taught P1 Last Year	0.783	0.583	0.821	0.800	0.545	0.846	0.667	1.000	0.500
Years of Teaching Experience	13.478	16.167	20.643***	13.000	15.636	21.000***	16.667	22.000	16.000
# Previous Schools Taught At	2.391	2.708	2.714	2.400	2.636	2.692	2.333	3.500	3.000
Highest Education is Diploma or Higher	0.478	0.333	0.500	0.500	0.273	0.500	0.333	1.000	0.500
Total Score on Ravens Progressive Matrices (0-3)	2.304	2.167	2.214	2.250	2.182	2.269	2.667	2.000	1.500
Panel C: Schools									
# of students who passed PLE in 2012	58.083	48.333	51.500						
Overall enrollment in 2012	1133.750	1219.417	1223.000*						
P1 enrollment in 2012	167.500	186.250	191.500**						
P1-P3 student-to-teacher ratio in 2012	78.333	68.750	79.786						
Overall student-to-teacher ratio in 2012	58.748	61.517	65.010*						
Distance to Coordinating Centre (CC) in km	9.083	7.417	8.857						
P1 classrooms can be locked	0.917	0.833	0.929						
Head Teacher is "highly" engaged	0.250	0.167	0.071						

Notes: Baseline Sample includes 1,900 students who were tested at baseline. Longitudinal Sample includes 1,481 students who were tested at baseline as well as endline. Lost to Followup includes 419 students who were tested at baseline but not at endline. Stars indicate cluster-adjusted p-values for a test of the null hypothesis of no difference between each NULP variant and the control group, conditioning on stratification cell indicators and the date of the baseline exam: * p<0.05, ** p<0.01, *** p<0.001.

Table 3.3
Control Group Baseline Attributes
and Improvements in Test Performance Over the School Year

		Baseline			Change from Baseline to Endline	
Variable	N (1)	% Any Correct (2)	Mean (3)	SD (4)	Mean (5)	SD (6)
Male						
Age						
EGRA						
Letter Name Knowledge (Letters per Minute)	476	15.3%	1.180	4.424	4.857	9.349
Initial Sound Identification (Sounds Identified)	477	2.9%	0.161	1.028	0.455	2.011
Familiar Word Reading (Words per Minute)	476	1.3%	0.168	1.617	0.165	2.588
Invented Word Reading (Words per Minute)	474	0.6%	0.084	1.191	0.275	2.309
Oral Reading Fluency (Words per Minute)	474	1.9%	0.508	4.537	0.102	5.012
Reading Comprehension (Questions Correct)	477	30.0%	0.327	0.559	-0.111	0.703
English Oral Assessment						
Test 1 (Vocabulary)	477	58.5%	1.774	1.993	0.275	2.089
Test 1 (Count)	477	32.9%	0.501	0.771	-0.208	0.813
Test 2a (Vocabulary)	477	36.9%	0.669	1.008	-0.168	1.068
Test 2a (Phrase Structure)	477	36.3%	0.801	1.169	0.006	1.343
Test 2b (Vocabulary)	477	54.9%	1.400	1.655	0.426	2.079
Test 2b (Phrase Structure)	477	48.4%	1.520	1.892	0.572	2.512
Test 3 (Vocabulary, Expressive - Objects)	477	67.1%	2.365	2.436	-0.038	2.490
Test 3 (Vocabulary, Expressive - People)	477	52.2%	1.505	1.789	0.080	2.177
Writing Test						
African Name (Surname) Spelling & Capitalization	477	20.1%	0.201	0.401	0.392	0.654
English Name (Given name) Spelling & Capitalization	477	14.5%	0.145	0.352	0.193	0.499
Ideas	477	0.6%	0.006	0.079	0.135	0.360
Organization	477	0.2%	0.002	0.046	0.284	0.589
Voice	477	0.0%	0.000	0.000	0.164	0.393
Word Choice	477	6.9%	0.069	0.254	0.099	0.374
Sentence Fluency	477	0.6%	0.006	0.079	0.261	0.584
Conventions	477	0.0%	0.000	0.000	0.116	0.339

Notes: Statistics are for the 477 control-group members of the Longitudinal Sample, which includes students who were tested at baseline as well as endline. Change from Baseline to Endline is the student's endline score on the component minus his or her baseline score.

Table 3.4
Program Impacts on Early Grade Reading Assessment Scores
(in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	PCA EGRA Score Index†	Letter Name Knowledge	Initial Sound Recogniton	Familiar Word Recognition	Invented Word Recognition	Oral Reading Fluency	Reading Comprehension
Mango Tree-Administered Program	0.634***	1.014***	0.647***	0.374***	0.215**	0.476***	0.445***
	(0.136)	(0.168)	(0.131)	(0.094)	(0.100)	(0.129)	(0.113)
Government-Administered Program	0.133	0.407**	0.076	-0.002	0.031	0.071	0.045
	(0.103)	(0.179)	(0.094)	(0.075)	(0.067)	(0.082)	(0.085)
Number of Students	1438	1475	1481	1471	1467	1450	1481
Number of Schools	38	38	38	38	38	38	38
Adjusted R-Squared	0.153	0.219	0.103	0.067	0.076	0.075	0.058
Control Group Mean§	0.002	5.977	0.616	0.335	0.360	0.615	0.216
Control Group SD§	1.005	9.374	1.920	2.209	2.770	4.176	0.437

Notes: Longitudinal sample includes 1,478 students who were tested at baseline as well as endline. All regressions control for stratification cell indicators and baseline values of the outcome variable. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

† PCA EGRA Score Index is constructed by normalizing each of the 6 test modules (columns 2 through 7) against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); estimated effects are comparable but slightly larger for an alternative index that uses the unweighted mean across test modules instead.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations except for the index (column 1), where they are the normalized values.

Table 3.5
Program Impacts on Oral English Test Scores & English Word Recognition
(in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCA Oral English Score Index†	Test 1 (Vocab.)	Test 1 (Count)	Test 2a (Vocab.)	Test 2a (Phrase Structure)	Test 2b (Vocab.)	Test 2b (Phrase Structure)	Test 3 (Vocab., Expressive - Objects)	Test 3 (Vocab., Expressive - People)	Recognition of Printed English Words‡
Mango Tree-Administered Program	0.141	0.157	-0.118	-0.034	0.045	0.025	-0.114	0.306***	0.295**	-0.290**
	(0.100)	(0.099)	(0.097)	(0.095)	(0.114)	(0.100)	(0.113)	(0.105)	(0.117)	(0.135)
Government-Administered Program	-0.089	0.001	-0.115	-0.020	-0.113	-0.154	-0.213*	-0.023	-0.099	-0.209
	(0.091)	(0.082)	(0.091)	(0.103)	(0.092)	(0.095)	(0.119)	(0.095)	(0.086)	(0.140)
Number of Students	1481	1481	1481	1481	1481	1481	1481	1481	1481	1481
Number of Schools	38	38	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.346	0.164	0.163	0.205	0.186	0.279	0.0920	0.238	0.188	0.274
Control Group Mean§	0	2.048	0.294	0.501	0.807	1.826	2.092	2.327	1.585	1.792
Control Group SD§	1	1.888	0.620	0.911	1.209	1.928	2.217	2.133	1.839	4.184

Notes: Longitudinal sample includes 1,478 students who were tested at baseline as well as endline. All regressions control for stratification cell indicators and baseline values of the outcome variable except for Recognition of Printed English Words (column 10), which was not administered at baseline. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA EGRA Score Index is constructed by normalizing each of the 8 test modules (columns 2 through 9) against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); estimates are comparable but slightly larger in magnitude for an alternative index that uses the unweighted mean across test modules instead. ‡ Recognition of Printed English Words is not part of the Oral English examination (and is not included in the computation of the overall PCA index), but it is a skill that is commonly practiced in *status quo* schools in the Lango sub-Region. § Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They are raw means and SDs except for the indices, where they are the normalized values.

Table 3.6
Program Impacts on Writing Test Scores
(in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCA Writing Score Index†	African Name (Surname) Writing	English Name (Given Name) Writing	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions	Presentation
Mango Tree-Administered Program	0.422***	0.922***	1.312***	0.163	0.441**	0.152	0.175	0.383*	0.221	0.139
	(0.146)	(0.107)	(0.143)	(0.171)	(0.207)	(0.156)	(0.153)	(0.207)	(0.173)	(0.150)
Government-Administered Program	-0.172	0.435***	0.450***	-0.274*	-0.316*	-0.313**	-0.262**	-0.330*	-0.253	-0.330**
	(0.125)	(0.119)	(0.147)	(0.144)	(0.177)	(0.134)	(0.124)	(0.177)	(0.156)	(0.129)
Number of Students	1373	1447	1374	1475	1475	1474	1474	1475	1475	1475
Number of Schools	38	38	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.356	0.240	0.236	0.174	0.304	0.177	0.200	0.302	0.164	0.171
Control Group Mean§	0	0.593	0.350	0.141	0.286	0.164	0.166	0.267	0.116	0.175
Control Group SD§	1	0.685	0.533	0.372	0.594	0.393	0.416	0.590	0.339	0.396

Notes: Longitudinal sample includes 1,478 students who were tested at baseline as well as endline. All regressions control for stratification cell indicators and baseline values of the outcome variable except for Presentation (column 10), which was not one of the marked categories at baseline. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

† PCA EGRA Score Index is constructed by normalizing each of the 9 test modules (columns 2 through 10) against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); with an alternative index that uses the unweighted mean across test modules instead, estimated effects are larger in magnitude and more statistically significant for the Mango Tree-Administered Program and closer to zero for the Government-Administered Program.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations except for the indices, where they are the normalized values.

Table 3.7

Program Impacts on Student Aspirations, Preferences, and Effort from Endline Survey

Variable	(1) Pupil Thinks He/She will Pass PLE at End of P7	(2) Preference for School over Other Activities†	(3) Prefers Literacy to Math Class	(4) Wants a Career as a Doctor/Nurse	(5) Wants a Career as a Headmaster/ Teacher	(6) Practices Writing at Home	(7) Thinks He/She is a Good Student	(8) Perceived Rank in Class‡	(9) Career Ambition Rating††
Units	Percentage Points	Control Group SD	Percentage Points	Percentage Points	Percentage Points	Percentage Points	Percentage Points	Control Group SD	Control Group SD
Mango Tree-Administered Program	0.022**	-0.114	-0.000	-0.078**	0.071***	0.006	0.002	0.148**	-0.059
	(0.009)	(0.112)	(0.023)	(0.033)	(0.023)	(0.025)	(0.013)	(0.063)	(0.068)
Government-Administered Program	0.015*	-0.097	-0.021	-0.030	0.035	-0.002	0.006	0.018	-0.085
	(0.009)	(0.087)	(0.021)	(0.027)	(0.024)	(0.020)	(0.015)	(0.076)	(0.056)
Number of Students	1330	1470	1457	1427	1427	1420	1371	1333	1417
Number of Schools	38	38	38	38	38	38	38	38	38
Adjusted R-Squared	-0.002	0.003	0.005	0.024	0.008	0.009	0.004	0.027	0.026
Control Group Mean§	0.947	4.614	0.544	0.396	0.154	0.900	0.971	2.245	2.837
Control Group SD§	0.225	0.657	0.499	0.490	0.361	0.300	0.169	0.666	0.886

Notes: Longitudinal sample includes 1,478 students who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† Preference for School over Other Activities is a 5-point scale based on a list of questions that compared school activities to other activities, capturing the number for which the student expressed a preference for school (and omitting those where she provided no response or could not answer).

‡ Perceived Rank in Class is a 1-3 scale, with 1 being the bottom of the class, 2 being the middle of the class, and 3 being the top of the class.

†† Career Amibition Rating is a subjective 1-5 scale where 1 is the least ambitious and 5 is the most ambitious; the ratings for each career were done by an evaluator who was blinded to the treatment status of the pupils.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations.

Table 3.8

Classroom Observations – Teacher Behavior

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Refers to Teacher's Guide	% time Speaking Leblango	Observe/ Record Performance	Moves Freely	Remains at Front of Class	Encourages Participation	Brings Pupils back on Task	Ignores Off-Task Students
Mango Tree-Administered Program	0.035	11.513***	0.047	0.087	-0.121**	-0.004	0.007	0.056**
	(0.041)	(3.524)	(0.052)	(0.067)	(0.053)	(0.018)	(0.038)	(0.027)
Government-Administered Program	0.041	8.907**	-0.025	-0.007	-0.048	-0.001	-0.070	0.062**
	(0.036)	(3.592)	(0.048)	(0.045)	(0.044)	(0.018)	(0.043)	(0.025)
Number of Observations	441	438	441	441	441	441	441	441
Number of Schools	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.166	0.121	0.256	0.032	0.061	-0.004	0.061	0.006
Control Group Mean§	0.802	67.210	0.237	0.733	0.237	0.962	0.870	0.031

Notes: All regressions control for stratification cell indicators, the round of the observations, and enumerator and day-of-week fixed effects. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

§ Control Group Mean is computed using the pooled data for control-group across all three rounds of classroom observations.

Table 3.9
Classroom Observations – Student Behavior While Reading

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Students are Reading:								
	Sounds	Letters	Words	Sentences	On Board	Reading From Primer	Reading From Reader	% of Reading Done in	Minutes Spent on Reading
Mango Tree-Administered Program	0.113***	-0.004	0.050	0.124***	-0.053	0.121***	0.064***	0.219***	0.669***
	(0.034)	(0.043)	(0.043)	(0.042)	(0.043)	(0.025)	(0.023)	(0.050)	(0.242)
Government-Administered Program	0.067**	0.054	-0.025	0.019	0.021	0.069***	0.031	0.165***	0.523**
	(0.028)	(0.044)	(0.045)	(0.051)	(0.039)	(0.022)	(0.020)	(0.051)	(0.212)
Number of Observations	441	441	441	441	441	441	441	441	441
Number of Schools	38	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.015	0.007	0.047	0.018	0.039	0.041	0.165	0.039	0.083
Control Group Mean§	0.061	0.206	0.649	0.282	0.672	0.023	0.038	0.466	3.687

Notes: All regressions control for stratification cell indicators, the round of the observations, and enumerator and day-of-week fixed effects. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.
§ Control Group Mean is computed using the pooled data for control-group across all three rounds of classroom observations.

Table 3.10
Classroom Observations – Student Behavior While Writing

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Students are Writing:						Copying		% of Writing	Minutes
	Pictures	Letters	Words	Sentences	Their Names	Air Writing	Text from Board	Writing Own Text	Done in Leblango	Spent on Writing
Mango Tree-Administered Program	0.076**	-0.024	0.044	0.023	0.059**	0.019	-0.024	0.094***	0.108**	0.199
	(0.033)	(0.035)	(0.028)	(0.023)	(0.028)	(0.030)	(0.036)	(0.028)	(0.047)	(0.253)
Government-Administered Program	0.034	0.042	0.059*	-0.028	0.006	0.063**	-0.017	0.019	0.115***	0.294
	(0.031)	(0.034)	(0.029)	(0.017)	(0.023)	(0.029)	(0.038)	(0.022)	(0.041)	(0.227)
Number of Observations	441	441	441	441	441	441	441	441	441	441
Number of Schools	38	38	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.012	-0.012	0.007	0.004	0.055	0.007	0.024	0.034	-0.002	0.001
Control Group Mean§	0.069	0.115	0.084	0.038	0.046	0.076	0.130	0.061	0.168	1.237

Notes: All regressions control for stratification cell indicators, the round of the observations, and enumerator and day-of-week fixed effects. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.
§ Control Group Mean is computed using the pooled data for control-group across all three rounds of classroom observations.

Table 3.11
Classroom Observations – Student Behavior While Speaking and Listening

	(1)	(2)	(3)	(4)	(5)	(6)
	Students are Speaking and Listening:				% of Speaking and Listening Done in Leblango	Minutes Spent on Speaking and Listening
	To Partner	To Small Group	To Whole Class	To Teacher		
Mango Tree-Administered Program	-0.028	0.050*	-0.041	-0.064*	0.080**	-0.786**
	(0.045)	(0.029)	(0.043)	(0.036)	(0.035)	(0.325)
Government-Administered Program	-0.014	0.066**	0.006	-0.094**	0.067*	-0.330
	(0.036)	(0.031)	(0.037)	(0.036)	(0.033)	(0.540)
Number of Observations	441	441	441	441	441	441
Number of Schools	38	38	38	38	38	38
Adjusted R-Squared	0.276	0.025	0.140	0.103	0.068	0.062
Control Group Mean§	0.221	0.038	0.748	0.947	0.802	4.916

Notes: All regressions control for stratification cell indicators, the round of the observations, and enumerator and day-of-week fixed effects. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.
§ Control Group Mean is computed using the pooled data for control-group across all three rounds of classroom observations.

Table 3.12
Attendance and Enrollment

	(1)	(2)	(3)	(4)	(5)	(6)
	Pupil Attendance:				Pupil Enrollment	Teacher Surveys
	Present for Visit 1	Present for Visit 2	Present for Visit 3	Average across all 3 visits	Total Enrollment at Endline	Reports Having Missed School in Past Month
Mango Tree-Administered Program	0.105**	0.020	0.026	0.050*	2.364	-0.067
	(0.044)	(0.031)	(0.035)	(0.029)	(25.008)	(0.167)
Government-Administered Program	0.019	-0.062**	-0.080**	-0.041	2.797	0.109
	(0.046)	(0.026)	(0.034)	(0.031)	(27.545)	(0.169)
Number of Observations	5334	5334	5334	5334	38	71
Number of Schools	38	38	38	38	38	37
Adjusted R-Squared	0.026	0.023	0.032	0.038	0.017	-0.025
Control Group Mean§	0.459	0.406	0.405	0.423	233.3	0.348

Notes: All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

§ Control Group Mean is computed using the endline data for control-group observations in the estimation sample.

Table 3.13
Responses to Teacher Survey by Study Arm

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Weekly Hours Spent on:													
	Teaching	Prep. for Class	Helping Students Outside Class	Taught Literacy this Year	# Parents Met with This Year	Would Choose to Teach if Could Restart Career	Teacher's Fault if Students Don't Learn	Satisfied with P2/P3 Reading at This School	Rating of Own Teaching Compared to Rest of School (1-3)	% of Pupils Teacher thinks will Pass PLE	Attended Any Training this Year	Days of Training Attended This Year	Went to NGO-Provided Training	Went to Other Training
Mango Tree-Administered Program	1.904	-0.623	2.042*	0.176*	61.288	0.199	-0.342**	-0.539***	0.015	0.003	0.319***	3.147*	0.567***	-0.255**
	(2.206)	(2.643)	(1.126)	(0.097)	(45.124)	(0.124)	(0.160)	(0.098)	(0.150)	(0.108)	(0.105)	(1.558)	(0.115)	(0.099)
Government-Administered Program	1.808	1.902	0.547	0.313***	35.918	0.200*	-0.034	-0.434***	-0.324**	-0.123	0.295***	2.348	0.470***	-0.169
	(2.494)	(2.851)	(0.970)	(0.095)	(34.203)	(0.105)	(0.159)	(0.094)	(0.150)	(0.094)	(0.105)	(2.043)	(0.121)	(0.110)
Number of Observations	73	72	69	73	67	70	72	71	71	73	73	70	73	73
Number of Schools	38	38	38	38	36	37	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.101	0.219	-0.0480	0.203	0.0810	0.0370	0.150	0.245	0.146	0.197	0.131	0.0940	0.326	0.0940
Control Group Mean§	14.55	9.601	1.765	0.609	37.86	0.565	0.739	0.727	2.545	0.498	0.652	4.957	0.435	0.348
Control Group SD§	8.780	10.67	2.221	0.499	46.94	0.507	0.449	0.456	0.510	0.291	0.487	6.852	0.507	0.487

Notes: All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.
§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample.

Table 3.14
Cost-Effectiveness Calculations

	Program Administered By:	
	Mango Tree	**Government**
Cost per student	$13.98	$4.47
Letter Name Knowledge		
Effect Size (SDs)	1.01	0.41
Cost per student/0.2 SDs	$2.76	$2.20
SDs per dollar	0.07	0.09
PCA EGRA Index		
Effect Size (SDs)	0.63	0.13
Cost per student/0.2 SDs	$4.41	$6.72
SDs per dollar	0.05	0.03
PCA Writing Test Index		
Effect Size (SDs)	0.42	-0.17
Cost per student/0.2 SDs	$6.63	N/A
SDs per dollar	0.03	-0.04

APPENDIX A

Technical Details of Theoretical Framework

A.1 Proof of Existence of Interior Solutions

Since the object of interest in this analysis is the response of y^* to changes in x, one concern is that all solutions to the problem are trivial, with fatalism representing jumps to some maximal level of risk taking. In this section I show that as long as each risky act has some cost, interior solutions are guaranteed to exist. The only exception is if risk-taking is not beneficial at all, in which case the agent chooses to take zero risky acts. The other analyses of optimal risk-taking that admit fatalistic responses (O'Donoghue and Rabin 2001; Sterck 2014) have shown fatalism only as a corner case, in which the individual pursues the maximum feasible level of risk-taking. While corner solutions are a fairly intuitive response – they align with the reasoning that once one is doomed, one might as well indulge as much as possible – they are not empirically relevant: there is little evidence that individuals ever truly seek out the *maximal* level of available risk-taking. Moreover, the reason for this is exactly that given above – taking additional risky acts, whether that means smoking more or seeking out sex partners, carries pecuniary costs so that there are tradeoffs with other goods an individual might desire.

The optimization problem in Section 1.2.1 admits many conceivable forms for the benefit function $B(y)$, including some that make little intuitive sense. To restrict the discussion to reasonable benefit functions, I assume that at some point taking additional risks yields no utility gains.

Assumption A.1.1.

$$\lim_{n \to +\infty} B'(y^*) = 0$$

As the number of risky acts chosen approaches infinity, the marginal benefit from an additional risky act approaches zero.[1]

Under Assumption A.1.1 (and the assumptions in Section 1.2 the problem still admits trivial corner solutions where $y^* = 0$. In order to discuss interior solutions, I impose one additional assumption.

Assumption A.1.2.

$$B'(0) > q + P_2(x, 0 + m)c$$

Risk-taking is desirable: given the stochastic and non-stochastic costs of risky acts, agents will choose a non-zero level of risk-taking.

Assumption A.1.2 seems reasonable in many applications: for example, a large proportion of people have had unprotected sex at some point in their lives. It is also empirically appropriate for my sample, as nearly 9 out of 10 sex acts are unprotected and this is essentially unchanged by the randomized information treatment. If the converse of Assumption A.1.2 holds, agents will (weakly) prefer to set $y = 0$, and the problem becomes trivial. Given Assumption A.1.2, however, the model allows a fairly powerful statement to be made:

Proposition A.1.1.

$$\exists y^* \epsilon (0, \infty) : y^* = \arg\max_{y \geq 0} \{U(y; x, m, q, c)\} \text{ if } q > 0$$

An interior solution to the optimization problem described in Equation 1.1 is guaranteed whenever the non-stochastic cost (e.g. the price) of a risky act is not zero.

Proposition A.1.1 follows because $\lim_{y \to +\infty}[B'(y^*) - q - P_2(x, y^* + m)c] = -q < 0$ by Assumption A.1.1 and Assumption A.1.2, and because $B'(0) - q - P_2(x, 0 + m)c > 0$. This, along with the continuity of U, allows me to use the extreme value theorem to state that U has at least one optimum where $y^* \epsilon (0, \infty)$, as long as $q > 0$. This eliminates the possibility of trivial corner solutions, in which the optimal response to an increase in risk is always to either choose $y^* = 0$ or $y^* = y_{max}$ (where y_{max} is some upper bound on y that prevents it from reaching infinity). Conversely, if $q = 0$, then given the other conditions the optimal y^* can be arbitrarily large: U is initially upward-sloping and its slope never becomes negative,

[1] This assumption is substantively identical to the sixth Inada condition used to guarantee the stability of neoclassical growth models.

so additional risk-taking is always weakly beneficial. Proposition A.1.1 guarantees that the optimum will be non-trivial if the price of risk-taking is positive. It does not rule out interior optima in other cases; O'Donoghue and Rabin do have an interior optimum in their model's non-fatalistic case, for example. However, it is a fairly intuitive economic result: people are constrained by resources from pursuing the high extreme in risk-taking. The results in Section 1.2 hold for the commonly-seen case in which people pursue some intermediate level of risk-taking irrespective of their perception of the per-act risk x. In the following section I will show that fatalism can occur even for these interior solutions.

A.2 Proof of Tipping Point in Cross-Partial Derivative of $P(x, y + m_0 + m_1)$

In this section, I show that any well-behaved function $P(x, y+m)$ – that is, any function that satisfies the conditions laid out in Section 1.2.1 – will have the property that its cross-partial derivative changes from positive to negative when x crosses a threshold value defined by $y + m_0 + m_1$. First, I prove some intermediate results, Lemma A.2.1 and Lemma A.2.2. I then use these to prove the proposition in question.

Lemma A.2.1. *Partial derivatives of P asymptote to zero*

1. $\lim_{x\to 1} P_1 = 0$
2. $\lim_{y+m_0+m_1\to+\infty} P_2 = 0$

The effect of increasing the per-act risk on the total probability of the bad outcome is zero if the per-act risk is one. The effect of the total number of risky acts approaches zero as their sum approaches infinity.

Part 1 of Lemma A.2.1 holds trivially if $y+m = 0$, and likewise for part 2 if $x = 0$. To see why they must hold in the non-trivial case, assume they do not hold. Then P is unbounded. But by assumption P is bounded above at 1, so we have a contradiction. Therefore Lemma A.2.1 must hold in general. Note that because P is continuously differentiable, Lemma A.2.1part 1 also implies that $P_1(1, y + m) = 0$. Conceptually, Lemma A.2.1 says that increasing the riskiness of each act high enough, or taking a sufficiently high number of risks, pushes the likelihood of the bad outcome to 100%. Once it has reached that point, additional risk-taking does not increase the probability any further. To prove this I first show that the cross-partial is initially positive:

Lemma A.2.2. *Cross-partial derivative of P is initially positive*

$$P_{21}(0, y+m) > 0 \text{ and } P_{21}(y, 0) > 0$$

The cross-partial derivative of the total probability of a failure with respect to riskiness and number of risky acts chosen is positive when the number of risky acts or the per-act riskiness is zero (or both)

This follows straightforwardly from Assumption A.1.2 P_1 is zero if y, m_0, and m_1 are all zero and positive if at least one of them is positive, so the initial cross-partial is positive; a symmetric analysis holds for P_2.

Given Lemma A.2.2, we can therefore prove that this cross-partial changes sign in general, for all functions P that meet the conditions laid out above.

Proposition A.2.1.

$$\exists \tilde{x} = x(y+m) \text{ with } y+m < +\infty \text{ s.t. } \begin{cases} P_{21}(x, y+m) > 0 & \text{if } x < \tilde{x} \\ P_{21}(x, y+m) < 0 & \text{if } x > \tilde{x} \end{cases}$$

For sufficiently high values of the per-act risk, increasing the per-act risk actually diminishes the marginal impact of additional risk-taking

To prove this, I consider two functions $P_{x_L}(y+m) = P(x_L, y+m)$ and $P_{x_L}(y+m) = P(x_H, y+m)$ with $x_L < x_H$. By Lemma 2,

$$P'_{x_L}(0) < P'_{x_H}(0)$$

Assumption 1 also gives us

$$P_{x_L}(0) = P_{x_H}(0) = 0$$

and

$$\lim_{y+m \to +\infty} P_{x_L}(y+m) = \lim_{y+m \to +\infty} P_{x_H}(n+m) = 1$$

Then these two continuous functions begin at the same value and converge to the same value, but the slope of P_{x_H} is initially higher than that of P_{x_L}. This implies that there must be some point at which the slope of P_{x_L} exceeds that of P_{x_H}. If not then the value of P_{x_L} can never catch up with that of P_{x_H}.

Formally, consider a point y_1 sufficiently close to zero that $P_{x_L}(y_1) < P_{x_H}(y_1)$, which must be possible because the second function's slope is initially higher. Then the average slopes of the two functions between y_1 and some higher point y_2 are $\frac{P_{x_L}(y_2)-P_{x_L}(y_1)}{y_2-y_1}$ and $\frac{P_{x_H}(y_2)-P_{x_H}(y_1)}{y_2-y_1}$, so the ratio of the two slopes is $\frac{P_{x_L}(y_2)-P_{x_L}(y_1)}{P_{x_H}(y_2)-P_{x_H}(y_1)}$. Taking the limit as y_2 approaches infinity, this ratio approaches $\frac{1-P_{x_L}(y_1)}{1-P_{x_H}(y_1)}$, which is greater than one. This implies that there is a point above which the average slope of P_{x_L} exceeds that of P_{x_H}. Figure 1.2 illustrates why this must be the case. The solid blue line gives the known initial shape of P_{x_L} and likewise the dashed red line for P_{x_H}. Above the breakpoint at infinity, the two-colored line shows their common value of 1. The middle range shows the implied average slopes in the intermediate region; because P_{x_L} is initially shallower, it must be steeper on average over this range.

The higher average slope of P_{x_L} over this later range implies, by the mean value theorem, that there must be at least one point where the instantaneous slope is also higher, that is ${P_{x_L}}' > {P_{x_H}}'$. Specifically, I can pick a point y_3 sufficiently close to infinity that the average slope of P_{x_L} between y_1 and y_3 is greater than the average slope of P_{x_H}. Then we have that $\frac{P_{x_L}(y_3)-P_{x_L}(y_1)}{y_3-y_1} - \frac{P_{x_H}(y_3)-P_{x_H}(y_1)}{y_3-y_1} > 0$. Define a new function $H(y) = P_{x_L} - P_{x_H}$. Then $\frac{H(y_3)-H(y_1)}{y_3-y_1} > 0$, and the mean value theorem requires that there is at least one point $\tilde{y}$ between y_1 and y_3 where $H'(\tilde{y}) > 0$ and hence $P'_{x_L}(\tilde{y}) - P'_{x_H}(\tilde{y}) > 0$, or equivalently $P'_{x_L}(\tilde{y}) > P'_{x_H}(\tilde{y})$.

This ensures that a tipping point must exist in any valid risk-aggregation function $P(x, y+m_0+m_1)$. It does not rule out multiple tipping points, which could conceivably arise from sophisticated curvature of the risk-aggregation function, but the number of such tipping points must be odd. I ignore the possibility of multiple tipping points, motivated by the fact that for the true risk-aggregation function Φ the cross-partial derivative changes sign only once.

A.3 Other Risk-Aggregation Functions

The results in Section 2 hold for a broad range of possible risk-aggregation functions that satisfy a minimal set of conditions, including the true function $\Phi(x, y+m_0+m_1)$. However, the central point – that behavior will swing from self-protection to fatalism for sufficiently high values of x – is driven by a tipping point in impact of riskiness on the marginal cost of riskiness. This kind of tipping point may exist even for far simpler heuristic risk aggregation functions that agents might employ, in particular ones that are not differentiable and therefore not amenable to the calculus techniques employed in that section. I therefore cannot prove that an interior optimum exists for such functions, or that optimal risk-taking will switch from self-protective to fatalistic. Instead, I demonstrate that two very simple

heuristic risk aggregation functions exhibit this tipping point phenomenon.

It might seem that this sort of tipping point is an esoteric mathematical feature of how probabilities add up that people cannot be expected to understand, but in fact such tipping points arise naturally and in a comprehensible way from some fairly basic heuristic risk aggregation functions. Consider the simple linear function used in much of the literature, where the assumption is made that levels of risk-taking and per-act risks are sufficiently low that the probability never approaches 1. Agents might use a similar rule, but also assume that if the probability does reach 1 then it stays there forever:

$$P(x, y + m_0 + m_1) = \begin{cases} \gamma x(y + m_0 + m_1) & : \gamma x(y + m_0 + m_1) < 1 \\ 1 & : \gamma x(y + m_0 + m_1) \geq 1 \end{cases} \tag{A.1}$$

This function might appear to lack a tipping point as defined in Proposition A.2.1, but the same basic behavior actually obtains. Consider two agents, one who believes $x = 0$ and one who believes $x = 1/\gamma(y + m) - \varepsilon$. If both agents increase their risk belief by 2ε, the marginal cost of increasing x rises for the first agent and falls for the second. Any shift in x that increases its value to at least $1/\gamma(y + m)$ will induce fatalism, with further increases having no additional effect on behavior.

An even simpler alternative is the "exposed enough" heuristic discussed in MacGregor, Slovic and Malmfors (1999), wherein people think they are totally safe as long as they stay below some level of activity, and then doomed with certainty if they take too many risks:

$$P(x, y + m_0 + m_1) = \begin{cases} 0 & : \gamma x(y + m_0 + m_1) < 1 \\ 1 & : \gamma x(y + m_0 + m_1) \geq 1 \end{cases}$$

In this case only the act that shifts an agent over the threshold, $y = 1/(\gamma x) - m$, has a direct marginal cost – all other acts carry no cost at all. Increasing x will in general push agents closer to the margin of being "sufficiently exposed" to suffer harm, thus carrying an indirect marginal cost. But if x reaches or crosses $1/\gamma(y + m_0 + m_1)$, the agent believes he or she is already sure to suffer the bad outcome and hence this decreases the marginal cost of an additional act to zero.

Despite not being amenable to analysis through standard optimization techniques, these functions both exhibit the crucial tipping-point phenomenon, implying that the results of Section 1.2 could hold even if agents handle the addition of risks in a very simple and heuristic way.

A.4 Extension of the Model to the Dynamic Case

The theoretical results that I derive in Section 1.2 for the static case can also be extended to a discrete-time dynamic setting. This can be done in two ways. First, it is possible to solve the model numerically if one imposes a number of functional-form and parameter-value assumptions. Sterck (2014) does this under the assumption that agents impose the true risk aggregation function, $P(x, y + m_0 + m_1) = 1 - (1 - x)^{y+m_0+m_1}$. Second, without imposing functional-form assumptions, one can derive the same tipping point for the comparative static $\partial y^*/\partial x$ in both a two- and three-period model. The push toward fatalism is even stronger in the three-period case than in the single-period model, because in addition to the tipping point in the sign of P_{21}, rises in the per-act risk x also increase the chance that the agent will die in the future no matter what. The fact that this result holds for three periods, along with the estimates of Sterck (2014), suggests that the intuitive result that sufficiently-high risks drive the marginal cost of risk-taking to zero also holds for rational choices made in an infinite-time dynamic framework. In this section I show that the one-period result extends to two and three periods.

A.4.1 Preliminaries

Define the total stock of sex acts in period t to be $M_t = \sum_{i=0}^{t}(y_i^* + m_i)$ where $y_0^* = 0$. m_0 is the number of past risky acts that agent has engaged in and does not yet know the outcome of. m_1 is the number of unavoidable future risky acts in this period. These can be thought of as safe acts that spontaneously become risky without benefiting the agent (e.g. condom breakage, or a perceived-safe sex partner turning out to have HIV). For $t > 1$, m_t is the number of unavoidable future risky acts that enter in each future period. From this definition, we have that $\partial M_t/\partial y_i^* = 1$ if $i <= t$ and 0 otherwise.

Agents have a discount factor β. Having HIV increases mortality, so that the probability of survival from one period to the next falls to γ. So the total discount factor on future utility is the probability of not having HIV times the discount factor, plus the probability of having HIV, times the chance that HIV kills you, times the discount factor:

$$\begin{aligned}
&\beta(1 - P(x, M_t)) + \beta\gamma P(x, M_t) \\
&= \beta - (\beta - \beta\gamma)P(x, M_t) \\
&= \beta - \delta P(x, M_t)
\end{aligned}$$

where $\delta = \beta - \beta\gamma > 0$ is the total discount factor conditional on having HIV.

A.4.2 Two-period Model

The solution to the model with two periods is fairly trivial. The two-period version of the utility function is just

$$U(y_1, y_2, x, m_0, m_1, m_2, \beta, \delta) = B(y_1) - qy_1 + [\beta - \delta P(x, y_1 + m_0 + m_1)] \{B(y_2) - qy_2\}$$

The maximization with respect to y_2 is unaffected by the choice of y_1, so I define $B(y_2^*) - qy_2^* = u_2 > 0$. u_2 is just a positive constant, so the problem can be re-written as

$$U(y_1, y_2, 8x, m_0, m_1, m_2, \beta, \delta) = B(y_1) - qy_1 + \beta u_2 - \delta u_2 P(x, y_1 + m_0 + m_1)$$

The agent then maximizes utility ignoring the βu_2 term, so if we define the cost of contracting HIV $c = \delta u_2$ then the model reduces to the one-period version studied in Section 1.2.

A.4.3 Three-period Model

Extending the model to three periods gives the following optimand:

$$\begin{aligned} U =& B(y_1) - qy_1 + [\beta - \delta P(x, y_1 + m_0 + m_1)]\big\{B(y_2) - qy_2 \\ &+ [\beta - \delta P(x, y_1 + y_2 + m_0 + m_1 + m_2)](B(y_3) - qy_3)\big\} \end{aligned}$$

We immediately see that the third-period choice y_3 is independent of the previous-period choices, so the problem reduces to

$$\begin{aligned} U =& B(y_1) - qy_1 + [\beta - \delta P(x, y_1 + m_0 + m_1)]\big\{B(y_2) - qy_2 \\ &+ [\beta - \delta P(x, y_1 + y_2 + m_0 + m_1 + m_2)]u_3\big\{ \end{aligned}$$

Assuming that an internal solution exists, it has the following first-order conditions:

$$
\begin{aligned}
G_1(y_1^*, y_2^*, m_0, m_1, m_2) =& B'(y_1^*) - q - \delta P_2(x, M_1) \left\{ B(y_2^*) - q y_2^* + [\beta - \delta P(x, M_2)] u_3 \right\} \\
& - \delta [\beta - \delta P(x, M_1)] P_2(x, M_2) u_3 \\
=& 0 \\
G_2(y_1^*, y_2^*, m_0, m_1, m_2) =& [\beta - \delta P(x, M_1)] \left\{ B'(y_2^*) - q - \delta P_2(x, M_2) u_3 \right\} \\
=& 0
\end{aligned}
$$

The implicit function theorem for two choice variables gives us the comparative static:

$$
\begin{aligned}
\frac{\partial y_1^*}{\partial x} &= - \frac{\det \begin{bmatrix} \frac{\partial G_1}{\partial x} & \frac{\partial G_1}{\partial y_2^*} \\ \frac{\partial G_2}{\partial x} & \frac{\partial G_2}{\partial y_2^*} \end{bmatrix}}{\det \begin{bmatrix} \frac{\partial G_1}{\partial y_1^*} & \frac{\partial G_1}{\partial y_2^*} \\ \frac{\partial G_2}{\partial y_1^*} & \frac{\partial G_2}{\partial y_2^*} \end{bmatrix}} \\
&= - \frac{\frac{\partial G_1}{\partial x} \frac{\partial G_2}{\partial y_2^*} - \frac{\partial G_1}{\partial y_2^*} \frac{\partial G_2}{\partial x}}{\frac{\partial G_1}{\partial y_1^*} \frac{\partial G_2}{\partial y_2^*} - \frac{\partial G_1}{\partial y_2^*} \frac{\partial G_2}{\partial y_1^*}} \\
&= - \frac{A}{B}
\end{aligned}
$$

For the denominator, I assume that we are at an interior solution, so the second-order condition holds. The denominator is just the determinant of the Hessian of the utility function, so it is negative, canceling out the leading negative sign in the expression. Thus the sign of the comparative static is just the sign of the numerator A.

$$
\begin{aligned}
\frac{\partial G_1}{\partial x} =& -\delta P_{21}(x,M_1)\left\{B(y_2^*)-qy_2^*+[\beta-\delta P(x,M_2)]u_3\right\}\\
&+\delta^2 P_2(x,M_1)P_1(x,M_2)u_3+\delta^2 P_1(x,M_1)P_2(x,M_2)u_3\\
&-\delta[\beta-\delta P(x,M_1)]P_{21}(x,M_2)u_3\\
=& -\delta\langle P_{21}(x,M_1)\left\{B(y_2^*)-qy_2^*+[\beta-\delta P(x,M_2)]u_3\right\}\\
&-\delta u_3\left\{P_2(x,M_1)P_1(x,M_2)+P_1(x,M_1)P_2(x,M_2)\right\}\\
&+[\beta-\delta P(x,M_1)]P_{21}(x,M_2)u_3\rangle\\
\frac{\partial G_2}{\partial y_2^*} =& [\beta-\delta P(x,M_1)]\left\{B''(y_2^*)-\delta P_{22}(x,M_2)u_3\right\}\\
\frac{\partial G_2}{\partial x} =& -\delta P_1(x,M_1)\left\{B'(y_2^*)-q-\delta P_2(x,M_2)u_3\right\}\\
&+[\beta-\delta P(x,M_1)]\left\{-\delta P_{21}(x,M_2)u_3\right\}\\
=& [\beta-\delta P(x,M_1)]\left\{-\delta P_{21}(x,M_2)u_3\right\}\\
\frac{\partial G_1}{\partial y_2^*} =& -\delta P_2(x,M_1)\left\{B'(y_2^*)-q-\delta[\beta-\delta P(x,M_2)]u_3\right\}\\
&-\delta[\beta-\delta P(x,M_1)]P_{22}(x,M_2)u_3\\
=& -\delta[\beta-\delta P(x,M_1)]P_{22}(x,M_2)u_3
\end{aligned}
$$

For $\partial G_1/\partial y_2^*$, the first term on the first line is zero because the portion in the curled brackets must be set to zero by the choice of y_2^* according to the F.O.C.)

$$
\begin{aligned}
A =& -\big(\delta\langle P_{21}(x,M_1)\left\{B(y_2^*)-qy_2^*+[\beta-\delta P(x,M_2)]u_3\right\}\\
&-\delta u_3\left\{P_2(x,M_1)P_1(x,M_2)+P_1(x,M_1)P_2(x,M_2)\right\}\\
&+[\beta-\delta P(x,M_1)]P_{21}(x,M_2)u_3\rangle\frac{\partial G_2}{\partial y_2^*}\big)\\
&-\big(-\delta[\beta-\delta P(x,M_1)]P_{22}(x,M_2)u_3[\beta-\delta P(x,M_1)]\left\{-\delta P_{21}(x,M_2)u_3\right\}\big)\\
=& -\big(\delta\langle P_{21}(x,M_1)u_2-\delta u_3\left\{P_2(x,M_1)P_1(x,M_2)+P_1(x,M_1)P_2(x,M_2)\right\}+\\
&+[\beta-\delta P(x,M_1)]P_{21}(x,M_2)u_3\rangle\times\frac{\partial G_2}{\partial y_2^*}\big)\\
&-\big(\delta^2[\beta-\delta P(x,M_1)]^2u_3^2P_{22}(x,M_2)P_{21}(x,M_2)\big)\\
=& C+D
\end{aligned}
$$

where C is defined as the value of the first line and D as the value of the second line.

$u_2 = B(y_2) - qy_2 + [\beta - \delta P(x, M_2)]u_3 > 0$ is the maximized value of the utility function in period 2 and hence negative. $\frac{\partial G_2}{\partial y_2} < 0$ and $P_{22} < 0$, $\delta > 0$, and all squared terms are positive. So

$$\begin{aligned}
\text{sign}(C) =& \text{sign}\big(P_{21}(x, M_1)u_2 - \delta u_3 \{P_2(x, M_1)P_1(x, M_2) + P_1(x, M_1)P_2(x, M_2)\} \\
&+ [\beta - \delta P(x, M_1)]P_{21}(x, M_2)u_3\big) \\
\text{sign}(D) =& \text{sign}\big(P_{21}(x, M_2)\big)
\end{aligned}$$

Suppose that $P_{21}(x, M_1)$ is beyond the tipping point and hence negative. Then D is negative. To determine the sign of C, first note that $P_{21}(x, M_2)$ will be negative as well (since the stock of risky acts is weakly higher), so the first and third terms in C are both negative. The second term in C is always negative, irrespective of the sign of $P_{21}(x, M_1)$. So $C < 0$ and $D < 0$, and $A < 0$. Therefore if x and $y_1^* + m_0 + m_1$ are sufficiently high, then the sign of the comparative static $\partial y_1^*/\partial x$ will be positive.

The conceptual reason for this result is similar to that for the one-period model. The $P_{21}(x, M_1)u_2$ term in C captures the effect of changes in x on the marginal cost of risk-taking, in terms of a decreased probability of being alive in the next period. This change, while intuitively positive, is negative if the risk x and number of unavoidable acts m_0+m_1 are sufficiently high, since the probability of having HIV is pushed so close to 1 that additional risk-taking is very low-cost. The $P_{21}(x, M_2)$ terms capture the fact that the same reasoning applies to acts chosen in future periods; this applies to both my current choice, and the choice of my *future* self, so that term appears twice.

The $-\delta u_3 \{P_2(x, M_1)P_1(x, M_2) + P_1(x, M_1)P_2(x, M_2)\}$ term captures the fact that the risks are linked across periods. Looking at the first part of this term, the marginal cost of risk-taking in the first period depends on the chance that you will survive the future period. When the per-act risk x rises, the increases the chance that you will die in the second period no matter what, so in addition to the direct effect of x on the marginal cost of risk-taking, $P_{21}(x, M_1)$, there is an offsetting effect $-\delta P_1(x, M_2)$ which is multiplied by the slope of the risk-aggregation function $P_2(x, M_1)$. The second part, $-\delta P_1(x, M_1)P_2(x, M_2)$, captures the same effect but in reverse: the marginal cost of second-period risk-taking is diminished by the fact that a rise in x increases the chance of contracting HIV in the first-period irrespective of the first-period choice. This term indicates that the push toward fatalism in the dynamic setting will be even stronger than in a one-period model, because of the linkage of risks across periods.

APPENDIX B

Data Details

This Appendix includes four tables that present details on my sample and data that were omitted from the main text for brevity. Appendix Table B.1 presents the composition of the sample by study arm and sampling stratum. Appendix Tables B.2 and B.3 present attrition regressions.[1] Appendix Table B.4 conducts balance tests for baseline values of risk beliefs, both with and without correcting for enumerator-knowledge contamination of measured beliefs.

[1] The results in Appendix Table B.2 are substantively identical when the regressions are instead run as logits or probits.

Table B.**1**
Sample Selection and Randomization

	Overall	Control	Treatment
Villages	70	35	35
Sampling Stratum†			
0-2 km from a trading center	24	12	12
2-5 km from a trading center	24	12	12
5+ km from a trading center	22	11	11
Respondents			
With Complete Baseline Survey	1503	759	744
With Complete Endline Survey	1292	645	647
Successful Followup Rate	0.86	0.85	0.87

Notes: † Sampling strata are defined by the combination of gender and three categories of distance to the nearest trading center. Villages were selected randomly from three sampling strata defined by the distance to the nearest trading center, and assigned randomly (within strata) to treatment or control status. This stratification was based on qualitative evidence suggesting sexual activity and fatalistic behavior were concentrated around trading centers. 1/3 of the sample was drawn from each stratum, oversampling the villages closest to trading centers; the population distribution of villages was 10% from the closest stratum, 40% from the middle stratum, and 50% from the farthest one. Respondent selection within villages was further stratified by gender; no more than one respondent per household was selected.
Baseline survey was the initial contact with each respondent, prior to the information treatment. The information about HIV transmission risks was provided to treatment-group respondents at the end of the baseline survey. Endline survey was the second contact with a respondent, and took place 6-12 weeks after the baseline.

Table B.**2**
Treatment-Control Differences in Attrition Rates

	Present in Final Sample†	
	(1)	(2)
Treatment	0.02	0.02
	(0.02)	(0.02)
Constant	0.85***	0.81***
	(0.02)	(0.13)
Other controls‡		X
Observations	1503	1484

Notes: † Present in Final Sample denotes the set of respondents who were contacted at baseline, had a complete baseline survey, and were subsequently found for the endline survey.
‡ Other controls include (baseline-observed) marital status, age, age squared, whether respondent grew up in current village, education, total household size, number of living and desired future children, total number of media sources respondent uses at least once per month, enumerator-rated respondent attractiveness, total common assets owned by household, logged spending in the past month, logged income in the past month, and categorical indicators for sampling stratum, ethnic group, religion, and the week of the followup survey.
Sample is 1503 sexually-active adults who were successfully interviewed at baseline; 19 of these have missing data for one of the controls. Heteroskedasticity-robust standard errors, clustered by village, in parentheses. * p< 0.1; ** p< 0.05; *** p<0.01.

Table B.**3**
Treatment-Control Differences in Attrition Rates by Baseline Covariates

	Present in Final Sample						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment Group [T]	0.020	0.014	0.037	-0.009	-0.066	0.067	0.009
	(0.020)	(0.029)	(0.025)	(0.022)	(0.072)	(0.053)	(0.022)
T*(Overall Sexual Activity Index)	-0.007						
	(0.020)						
T*(HIV Risk Belief [0-1])		0.011					
		(0.056)					
T*(Sex Acts in Past Week)			-0.012				
			(0.008)				
T*(Gender==Male)				0.054			
				(0.036)			
T*(Age)					0.003		
					(0.002)		
T*(Married==Yes)						-0.063	
						(0.062)	
T*(Ethnicity==Lomwe)							0.024
							(0.034)
T*(Ethnicity==Yao)							-0.068
							(0.077)
T*(Ethnicity==Chewa)							0.167*
							(0.092)
T*(Ethnicity==Other)							-0.147
							(0.141)
Observations	1,483	1,480	1,543	1,543	1,543	1,541	1,543
Adjusted R-squared	0.029	0.028	0.037	0.035	0.037	0.053	0.038

Table B.4

Baseline Balance in Subjectively-Assessed Risk of HIV Infection
from Different Sex Acts, with and without Correcting for Contamination

	N	Control	Treatment	C-T
	(1)	(2)	(3)	(4)
Panel A – Only Treatment Group Measured after Enumerators Learned Risk Information† (C and T Both Measured at Baseline)				
HIV Transmission				
One Act				
Unprotected	1289	0.83	0.74	0.09***
If using a condom	1291	0.12	0.10	0.03**
One Year				
Unprotected	1284	0.93	0.88	0.05***
If using a condom	1284	0.24	0.16	0.08***
HIV Prevalence				
All local people of opposite sex	1279	0.53	0.51	0.01
People respondent finds attractive	1276	0.55	0.52	0.03*
Panel B – Both Treatment and Control Group Measured after Enumerators Learned Risk Information (C Beliefs at Endline, T Beliefs at Baseline)				
HIV Transmission				
One Act				
Unprotected	1284	0.74	0.74	0.01
If using a condom	1286	0.08	0.10	-0.01
One Year				
Unprotected	1281	0.91	0.88	0.03**
If using a condom	1282	0.18	0.16	0.02
HIV Prevalence				
All local people of opposite sex	1270	0.48	0.51	-0.03
People respondent finds attractive	1269	0.46	0.52	-0.06***
Panel C – Baseline-Measured Beliefs, Using Regression Adjustment to Correct for Enumerator-Knowledge Contamination of Belie				
HIV Transmission				
One Act				
Unprotected	1289	0.79	0.79	-0.00
If using a condom	1291	0.14	0.15	-0.00
One Year				
Unprotected	1284	0.93	0.93	-0.00
If using a condom	1284	0.29	0.28	0.00
HIV Prevalence				
All local people of opposite sex	1279	0.54	0.53	0.01
People respondent finds attractive	1276	0.52	0.53	-0.00

Notes: Panel A presents unadjusted baseline belief variables for both groups, without correcting for contamination of treatment-group respondents' beliefs due to enumerator knowledge. Panel B compares the control group's beliefs at endline to the treatment group's beliefs at baseline; because the survey enumerators were taught the HIV risk information after the control-group baseline surveys and before the treatment-group baseline surveys, this panel thus compares the groups when the information sets of both respondents and enumerators are equivalent. Panel C adjusts the baseline belief variables for time trends with a structural break on the date that the enumerators were taught the HIV risk information (allowing for a jump at that point and also for different slopes on either side); the regression adjustment is described below Figure 6. Sample is 1292 people from 70 villages for whom both baseline and endline surveys were successfully completed. Cluster-adjusted significance tests: * p< 0.1; ** p< 0.05; *** p<0.01.

APPENDIX C

Ethical Dimensions

In this appendix I address the ethical dimensions of the study related to providing the respondents in the treatment group with information about the true risk of HIV infection. In Appendix C.1 I describe the ethical considerations that went into the design of the information treatment. Appendix C.2 then examines the data to estimate the effect the study may have had on the number of HIV infections in the treatment group, as well as other consequences of having more unprotected sex. I also discuss the ethical implications of people choosing to be exposed to more risks as a result of more accurate information.

C.1 Ethical Considerations in Designing the Information Intervention

The key potential ethical concern about the design of this study was that people may respond self-protectively to HIV infection risks on average. In this case the information treatment would increase the average amount of risky sex people have, leaving people in the treatment group worse off. This concern is mitigated by four factors. First, to the extent that we believe responsible adults can be trusted to make their own choices with the information they have, it is appropriate to provide people with better information rather than worse. The de facto policy in Malawi is to overstate HIV transmission risks. This strategy is potentially at odds with the first ethical principle emphasized in the Belmont Report, which is that individuals should be respected as autonomous persons:

> To respect autonomy is to give weight to autonomous persons' considered opinions and choices while refraining from obstructing their actions unless they are clearly detrimental to others. To show lack of respect for an autonomous agent is

> to repudiate that person's considered judgments, to deny an individual the freedom to act on those considered judgments, or to withhold information necessary to make a considered judgment, when there are no compelling reasons to do so. (Office of the Secretary 1979)

Hence the policy of denying people information about the true risks they face is potentially unethical, given that there is very little empirical evidence that would provide compelling reasons to withhold that information. Second, the information provided to the treatment group is medically-accurate, publicly available information. It is also the same information provided by the Malawi National AIDS Commission (NAC) in their policy documents, which state that the annual risk of HIV transmission is 10% (Malawi National AIDS Commission 2003, p. 11). NAC's official policy is also that HIV information and education programs should provide accurate information about safer sex:

> Government, through the NAC, undertakes to do the following:
> • Ensure that all people have equal access to culturally-sound and age-appropriate formal and nonformal HIV/AIDS information and education programmes, which shall include free and accurate information regarding mother-to-child transmission, breastfeeding, treatment, nutrition, change of lifestyle, safer sex and the importance of respect for and nondiscrimination against PLWAs [people living with AIDS].
> (Malawi National AIDS Commission 2003, p. 6)

Hence the additional information provided to the treatment group is completely consistent with Malawi government policy, and can be seen as a test of what would happen if HIV information and education campaigns actually provided HIV transmission risk information that is consistent with what NAC provides on its website.

A third mitigating factor is that previous estimates of responses to HIV risks in Africa are very small in magnitude (e.g. Oster 2012), and the *ex ante* expected impact of the information treatment was small, limiting any potential harm. The reason that the experiment was still interesting was that the responses were not expected to be uniform. There is reason to believe that many people in Malawi may react fatalistically to HIV risks. As mentioned above, cross-sectional data from elsewhere in Zomba District shows suggestive evidence that the response of sexual behavior to HIV infection is positive for people with high risk beliefs (Kerwin 2012). Kaler (2003) documents that men from rural Southern Malawi employ fatalistic reasoning - saying that it is sometimes not worthwhile to use condoms, because the risk of contracting the virus is so high:

> And then I asked my in-law, "What do people do after noticing that his/her partner seems to have AIDS?" He said, "Some couples come to an end and for

> others the marriage continues." And I asked, "Do they use condoms then?" He said "I don't think they use [them] because it will just be a waste of time since both of them have contracted the disease." (Simon, journal May 3 2002)

For people who respond fatalistically, learning that their assessment of the risk is an overestimate will actually reduce sexual risk-taking, rather than increasing it. This experiment was designed to capture heterogeneity in responses around a mean response that is small in magnitude.

Finally, this concern is mitigated because excessively high risk beliefs may have negative long-term effects independent of any direct effects on sexual behavior. As people realize that it is possible for sexually active married couples to remain serodiscordant for a long time, they may lose trust in the medical and science community or the education system, and may also promulgate false rumors about HIV transmission and immunity. Since most people believe that the transmission rate of HIV is 100%, they may instead falsely assume that continued serodiscordance means that a specific person or group is immune to the virus. There is already evidence that the latter is going on: 42% of my respondents said that they believed people with type-O blood were immune to HIV, an idea which has no basis in scientific fact.

A separate potential concern is that the information presented is about the approximate overall average risk, but transmission risks actually vary by demographic groups. For example, the transmission rate is 3 to 5 times higher for women than for men, and about 60% lower for circumcised men than for uncircumcised men. However, this concern is mitigated by the fact that baseline beliefs are very high (93% per year on average for the control group). Hence virtually all respondents in the treatment group have more-accurate beliefs after the information treatment than they did beforehand.

To ensure that respondents' well-being was protected, ethics oversight for this study was provided by both an in-country IRB (The University of Malawi's College of Medicine Research and Ethics Committee, or COMREC) and one at my home institution (The University of Michigan's IRB-Health Sciences and Behavioral Sciences, or IRB-HSBS). The final study protocol, including the information treatment, was reviewed and approved by both IRBs. The approved protocol also included a management plan under which preliminary results were provided to the two IRBs in order to manage any possible rise in HIV transmissions as a result of the information treatment.

C.2 Direct Effects on HIV infections

One question that can be addressed directly is the impact of this specific information intervention on additional HIV infections. To compute an effect on HIV infections I would ideally rely on HIV test results, but these were not collected as part of the study. Instead, I use two proxies. First, respondents' self-reported beliefs about their own, and their primary sex partners', serostatus, which are measured on the survey. Second, data from the 2010 DHS, which measured the prevalence of HIV in Zomba District, but cannot be directly tied to my respondents.

For the first proxy, I consider as serodiscordant couples in which the respondent reports no likelihood of being HIV infected him- or herself, but some likelihood for his or her partner, or vice-versa. I compute the total change in weekly sex acts for all 77 respondents in such couples by computing the treatment effects as in Figure 8 and summing over the changes for each individual respondent. The total is is 4.39, and respondents in this group used condoms just 3.7% of the time, so this would mean 4.23 more unprotected acts per week. This would correspond to an additional 0.004 HIV infections per week total, or 0.2 per year. This would mean the effects of the information treatment would need to persist for five years before we would expect an additional HIV infection to be induced.

To use the second proxy, I first note that the 2010 DHS measured an HIV prevalence of 18% for Zomba District. The DHS data has too few instances of multiple partners within a household to estimate the degree of matching on HIV status among couples in Zomba District, so instead I assume that people pick their sex partners randomly with respect to HIV status. The heterogeneity in people's responses to the treatment will tend to work against an increase in HIV transmissions, because people with higher risk beliefs are more important for the spread of the virus. I can therefore find an upper bound by ignoring any heterogeneity and assuming the reduced-form effect of the information treatment (10.1% more sex acts per week, from Column 2 of Table 6) as constant, so the treatment group has an additional 0.17 sex acts per week total. Under these assumptions, the 18% of the treatment group that are HIV-positive have an additional (0.17)X(0.82) = 0.14 sex acts each week with HIV-negative partners, while the 82% of my sample that is HIV-negative has an additional (0.20)X(0.18) = 0.03 sex acts each week with HIV-positive sex partners. Taking the weighted sum, and multiplying by the fraction of sex acts that are unprotected (88%), the average person in the treatment group would have an additional 0.05 sex acts per week where an HIV transmission was possible. That would mean a total of 34 such sex acts per week across the 647 people in the treatment group. This corresponds to 0.03 additional HIV transmissions per week in the absence of any other changes. This upper bound indicates

that the treatment effect would need to persist for at least eight months before we would expect it to cause any additional HIV infections.

Any potential rise in the number of HIV infections due to the information treatment was offset by the fact that respondents were sold heavily-discounted condoms as part of the study. We would expect these to decrease the number of HIV infections observed in this population: although condoms are available for free at local clinics, the Chishango-brand condoms sold as part of the study are higher in quality. The subsidized sales also removed the travel and time cost of getting free condoms. On average people in the treatment group bought 5.22 condoms as part of the study. At baseline, my respondents had used about 20% of the condoms they had gotten in the past 30 days. Making the assumption that the same will hold for these condoms, that would mean an additional 675 condom-protected sex acts among treatment group respondents. If the population mixes randomly, 15% of all sex acts involve one HIV-positive and one HIV-negative sex partner, so 100 sex acts would be switched from unprotected to protected. This would almost totally offset the increase in unprotected sex among the treatment group. Moreover, a similar calculation holds for the control group, averting more HIV transmissions among that group and their sex partners. Thus over the short term, HIV infections were unlikely to be increased by the intervention – even if we assume the upper bound number of new HIV infections was generated.

Whether the treatment group experienced an increase in HIV infections over the longer term depends on how long the effects of the information intervention persist, and how that varies across fatalistic and self-protective individuals. It also depends on how the rest of the stock of over 3000 condoms that were distributed to the treatment group get used over time. Due to the short-run nature of the followup survey, which was conducted just 6 to 12 weeks after the baseline, little can be said about the longer-term dynamics of either the effect of the information treatment or the distribution of condoms.

A similar line of reasoning can be extended to other consequences of unprotected sex. Additional unprotected sex among the treatment group could lead to more pregnancies, for example, or additional infections with HSV-2. Concerns about the spread of HSV-2 are mitigated by its extremely high prevalence in Malawi: Kenyon, Colebunders and Hens (2013) estimate that 78% of women in Malawi have HSV-2 by age 44. All the consequences of unprotected sex should be considered in light of the Belmont Report's admonishment that we should respect the considered judgments of autonomous individuals: to the extent that people are exposed to risks by their choice to engage in unprotected sex, and those risks lead to negative outcomes, those were the results of choices they decided to make given the information they had at hand. The potential for an increase in pregnancies is an important case in point. People in the treatment group altered their choices about how much

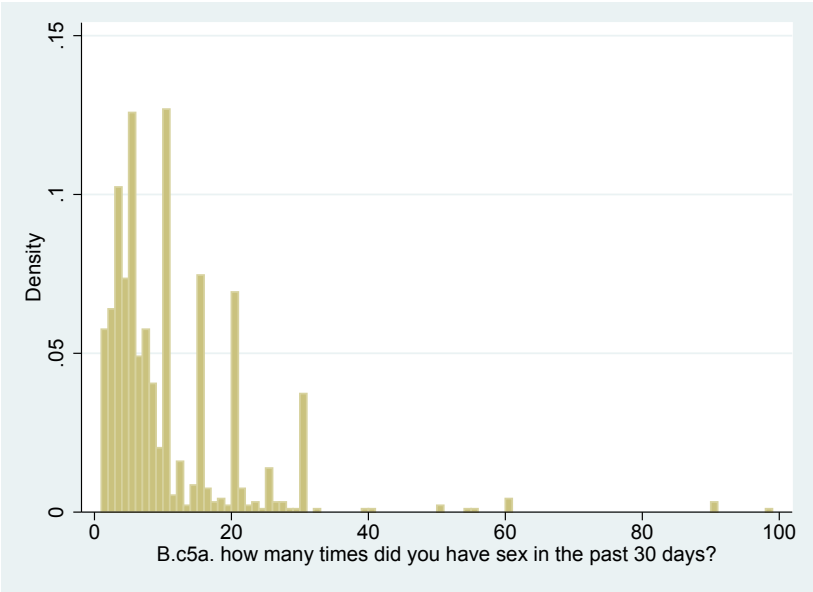
unprotected sex to have based on better information about HIV; they realized it was less of a risk than they had thought. While some pregnancies are surely unwanted, children are often a desired outcome of intercourse. By choosing to have more unprotected sex, people in the treatment group were also choosing to potentially have more children – children they might have been afraid to have ex ante, due to fears about contracting HIV.

APPENDIX D

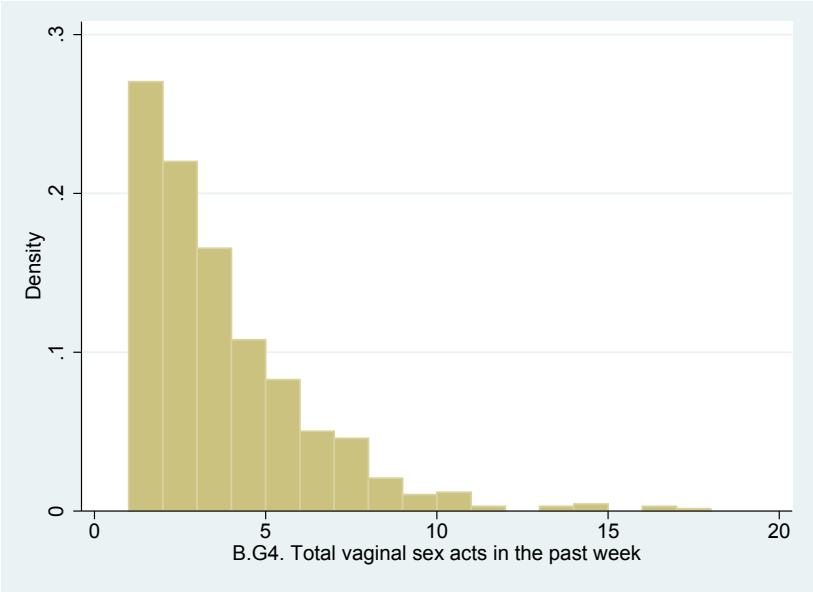
Comparison of Outcome Measures for Sexual Activity

Panel A of Appendix Figure D.1 shows the distribution of sex acts in the past 30 days from the single-question recall variable. It demonstrates substantial "heaping" at multiples of 5, with large spikes at 5, 10, 15, etc. In contrast, the distribution of sex acts in the past 7 days for the sex diary question (Panel B) has no appreciable heaping whatsoever. Unlike classical measurement error, this kind of heaping in the dependent variable may bias the point estimates from a linear regression. In a set of simple simulations that took smoothly measured data and progressively added more heaping, I find that sufficiently high levels of heaping bias the estimated coefficients toward zero (results available upon request). The reason for this is clear from considering the extreme case, where the heaping is so extreme that all values are collapsed to a single point. While one could not run an OLS regression in this case, it is clear that the effect of any variable on this (mismeasured) outcome is zero. This lends further support to my decision to focus on the sex diary outcomes in the majority of my analysis.

Figure D.1
Histogram of sex acts reported conditional on any sex



Panel A: Single-question recall, past 30 days



Panel B: Retrospective sex diary, past 7 days

Notes: Data for Panel A comes from a single question that asked respondents how many times they had had sex in the previous 30 days. Data for Panel B is the total number of sex acts reported on a 7-day retrospective "diary" that walked respondents through the previous 7 days, asking about a range of activities including sex and recording details about each sex act. Both histograms omit a large mass point at zero for readability. Panel A exhibits a far greater degree of heaping, suggesting that it is a lower-quality measure of sexual behavior than Panel B. Sample is 1292 people from 70 villages for whom both baseline and followup endline surveys were successfully completed.

APPENDIX E

Proof that Controlling for Baseline Values of the Outcome Variable Minimizes the Bias in Estimated Treatment Effects

Consider estimating the effect of a randomly-assigned treatment T on outcome y. The typical econometric strategy for analyzing experiments is to estimate

$$y_i^e = \alpha + \beta_{POST} T_i + e_i \tag{E.1}$$

That is, regress endline values of the outcome on an indicator for treatment status plus a constant. $\hat{\beta}_{POST}$ will consistently estimate the causal effect of T on y due to the random assignment of the treatment. When baseline data is available, it is also common to use difference-in-difference or "value-added" specifications which utilize first differences of the outcome and treatment status as the dependent and independent variable respectively (e.g. Card and Giuliano 2013):

$$Dy_i = \alpha + \beta_{DIFF} DT_i + e_i \tag{E.2}$$

Here $Dy_i \equiv y_i^e - y_i^b$ and $DT_i \equiv T_i^e - T_i^b = T_i$, and β_{DIFF} also consistently estimates the parameter of interest. McKenzie (2012) and Frison and Pocock (1992) show that both β_{POST} and β_{DIFF} have higher variance than a third alternative, which includes baseline values of the outcome of interest as a control in a regression of endline outcomes on treatment status:[1]

[1] This is also referred to as the "ANCOVA" (analysis of covariance) estimator in the medical literature, where the relevant alternatives were variants of analysis of variance ("ANOVA") methods.

$$y_i^e = \alpha + \beta T_i + \gamma y_i^b + e_i \tag{E.3}$$

$\hat{\beta}$ is also consistent for the effect of T on y; as it is more efficient, it is preferable on those grounds alone. However, $\hat{\beta}$ has a further advantage in the case of (even slight) baseline imbalance in an outcome variable: it is also less biased than either other option.

Let $d^b = \bar{y}_T^b - \bar{y}_C^b$ be the baseline difference in the outcome of interest, and σ^2 be the variance of the error term. The variance of the error can be decomposed into a component due to measurement error (σ_e^2), and a remaining component $\sigma^2 - \sigma_e^2$. Frison and Pocock (1992) show that for a single baseline and followup the bias due to baseline imbalance is given by:

1. $Bias_{POST} = \frac{\sigma^2 \rho}{\sigma^2 - \sigma_e^2} d^b$ for the POST estimator,
2. $Bias_{DIFF} = \frac{\sigma^2(\rho-1)+\sigma_e^2}{\sigma^2 - \sigma_e^2} d^b$ for the DIFF estimator, or
3. $Bias_{OPTIMAL} = \frac{\sigma_e^2 \rho}{\sigma^2 - \sigma_e^2} d^b$ for the optimal estimator.

It is important to note that although the size of the bias term will diminish as d^b falls, it will be nonzero unless d^b is identically zero. Thus these finite-sample bias terms are potentially relevant even if the outcome is balanced in the sense of not having statistically-significant differences at baseline. Frison and Pocock show that the relative size of $Bias_{POST}$ and $Bias_{DIFF}$ depends on whether ρ is greater or less than 0.5, and note that in most cases σ_e^2 will be very small relative to $\sigma^2 - \sigma_e^2$ so that $Bias_{OPTIMAL}$ is nearly zero. However, it is also possible to show the intuitive result that, in addition to having lower variance than the alternatives, $\hat{\beta}$ is also uniformly less biased in the presence of baseline imbalance in a finite sample. Consider the relative size of the bias terms,

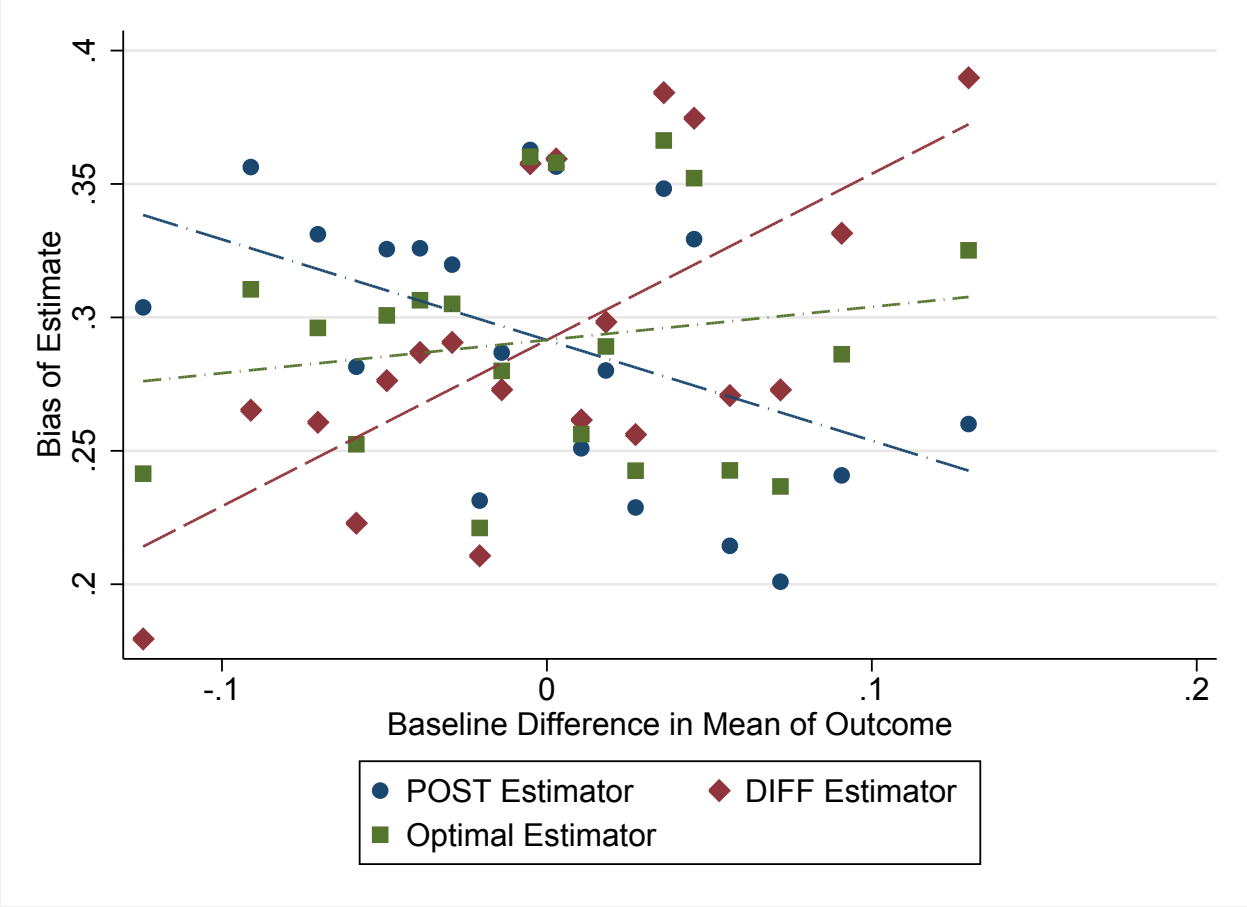
$$\frac{Bias_{DIFF}}{Bias_{OPTIMAL}} = \frac{\sigma^2(\rho-1)+\sigma_e^2}{\sigma^2 - \sigma_e^2} \frac{\sigma^2 - \sigma_e^2}{\sigma_e^2 \rho} = \frac{\sigma^2(\rho-1)+\sigma_e^2}{\sigma_e^2 \rho} \tag{E.4}$$

And

$$\frac{Bias_{POST}}{Bias_{OPTIMAL}} = \frac{\sigma^2 \rho}{\sigma^2 - \sigma_e^2} \frac{\sigma^2 - \sigma_e^2}{\sigma_e^2 \rho} = \frac{\sigma^2}{\sigma_e^2} \tag{E.5}$$

Each of these ratios approaches infinity as the portion of variance due to measurement error approaches zero, and reaches a minimum value of 1 if $\sigma_e^2 = \sigma^2$. This is equivalent

Figure E.1
Bias of Different Estimators
as a Function of the Baseline Treatment-Control Difference in Outcomes



to saying that 100% of the residual variance of y is due to measurement error; we can rule that out in the case of sexual activity since our regression model will logically predict only a small portion of the true variation in patterns of sex. Thus, when the baseline mean of the outcome of interest is not identical across the treatment and control groups, $\hat{\beta}$ will be less biased than $\hat{\beta}_{POST}$ or $\hat{\beta}_{DIFF}$.

This derivation is confirmed by a simple simulation of the DGP described above. Appendix Figure E.1 shows the results of simulating the DGP 1000 times and computing the bias of each estimator. The green squares show the binned average of estimates from the optimal estimator, while the red diamonds show the binned average bias for the DIFF estimator and the blue circles show the binned average bias for the POST estimator. The optimal estimator's bias always lies between that of the DIFF and POST estimators, and in expectation it is less than that of the other two estimators when the treatment-control difference is not zero.

APPENDIX F

Sensitivity Analysis

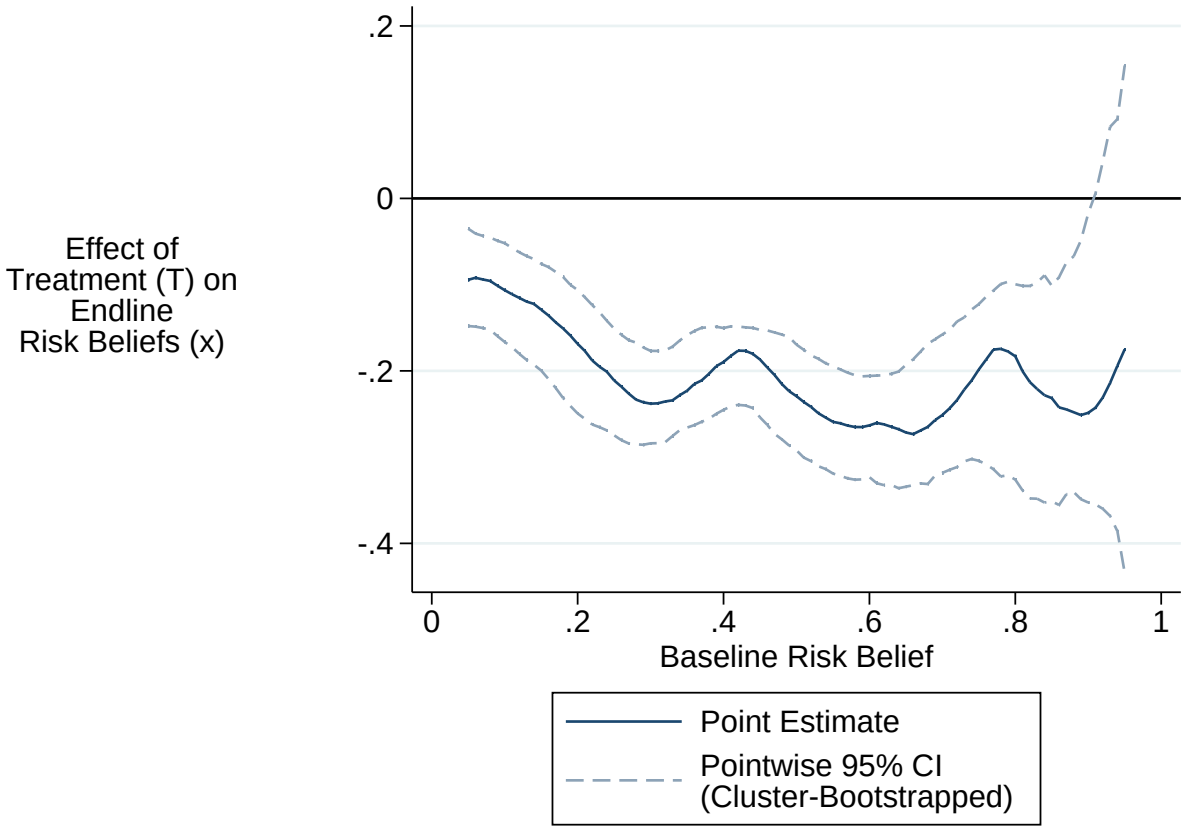
The results presented in the main body of the paper focus on my preferred specifications, which I argue give the best estimates of the causal effect of the information treatment and how it varies across the distribution of initial risk beliefs. In this section, I explore the sensitivity of those specifications to a number of alternative approaches. I begin by showing that the results presented in Figures 1.7, 1.8 and 1.9 are robust to halving all the bandwidths used to estimate the regressions. I then show that they are robust to running the semi-parametric regressions using the bracketed method described in Section 1.4.5 (creating indicators for 1/8 ranges of the baseline beliefs and interacting them with the treatment indicator) reproduces the same qualitative results.

I then use bracketed parametric regressions to conduct a range of other sensitivity analyses. Because the focus of my analysis is on the heterogeneity in responses by baseline risk beliefs, I focus my analysis here on alternative methods of constructing Figure 1.8, which shows the reduced-form effect of the information treatment on sexual activity by people's baseline risk beliefs. I focus on Figure 1.8, rather than the elasticity estimates in Figure 1.9, because constructing the confidence intervals for Figure 1.9 is computationally intensive, and because Figure 1.9 can be constructed by simply dividing Figure 1.8 by the first-stage graph.

F.1 Robustness to Smaller Bandwidth Choices

The Loader (2004) GCV-minimizing bandwidths can sometimes have issues due to over-smoothing. In order to rule out that my results arise from an excessively-large bandwidth choice, I re-run all my Robinson-based estimators dividing the bandwidths by 2 for all variables (the final estimates as well as all the underlying regressions involved in residualizing out the controls. Appendix Figures F.1 to F.3 confirm that all my qualitative results are

Figure F.1
First-Stage Effect of Treatment (T) on Endline Risk Beliefs (x),
by Baseline Risk Belief



robust to dividing the bandwidth by 2: both the reduced-form estimates and the elasticity estimates exhibit statistically-significant fatalism for people with the highest risk beliefs, although the confidence intervals widen at the very end. The jagged shapes of the curves estimated with the smaller bandwidths also suggest that this smaller bandwidth is smoothing the data too little, and supports the original bandwidth choice as my preferred specification.

Figure F.2
Reduced-Form Effect of Treatment (T) on Log Sex Acts in Past Week ($\ln(y)$), by Baseline Risk Belief

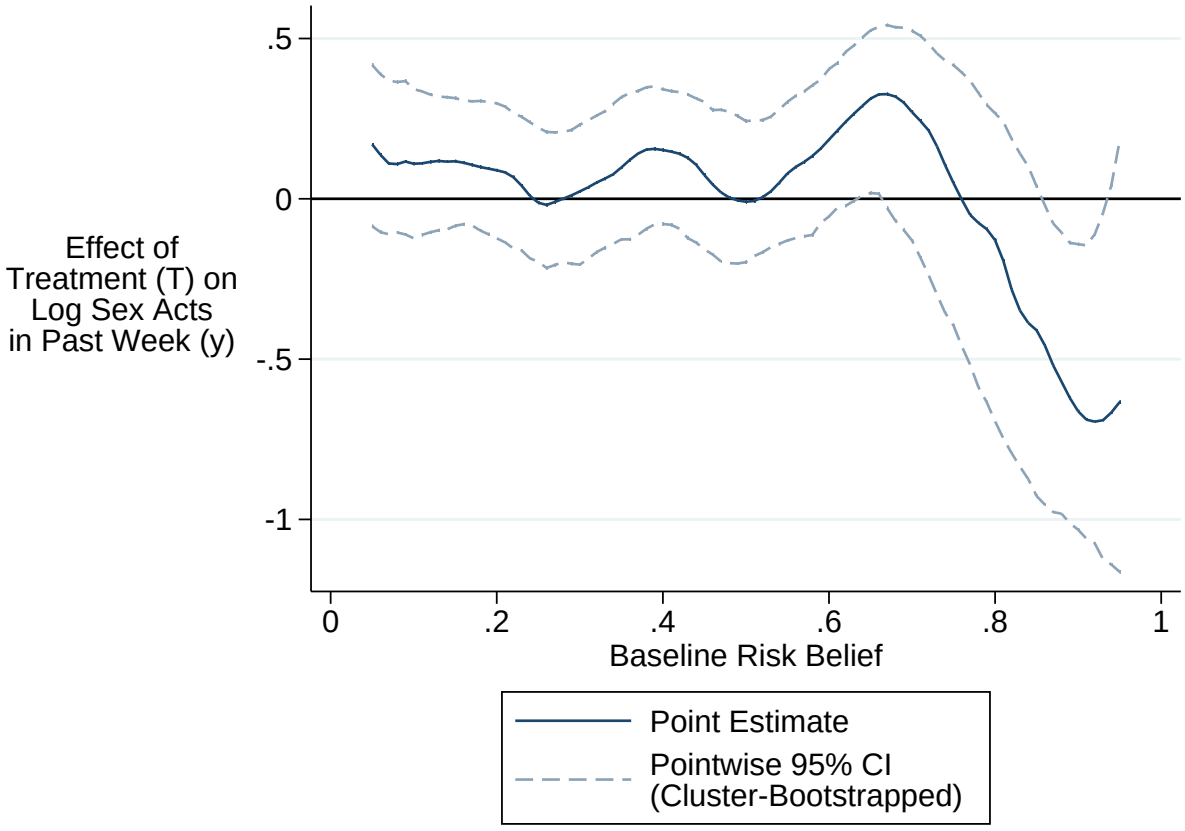
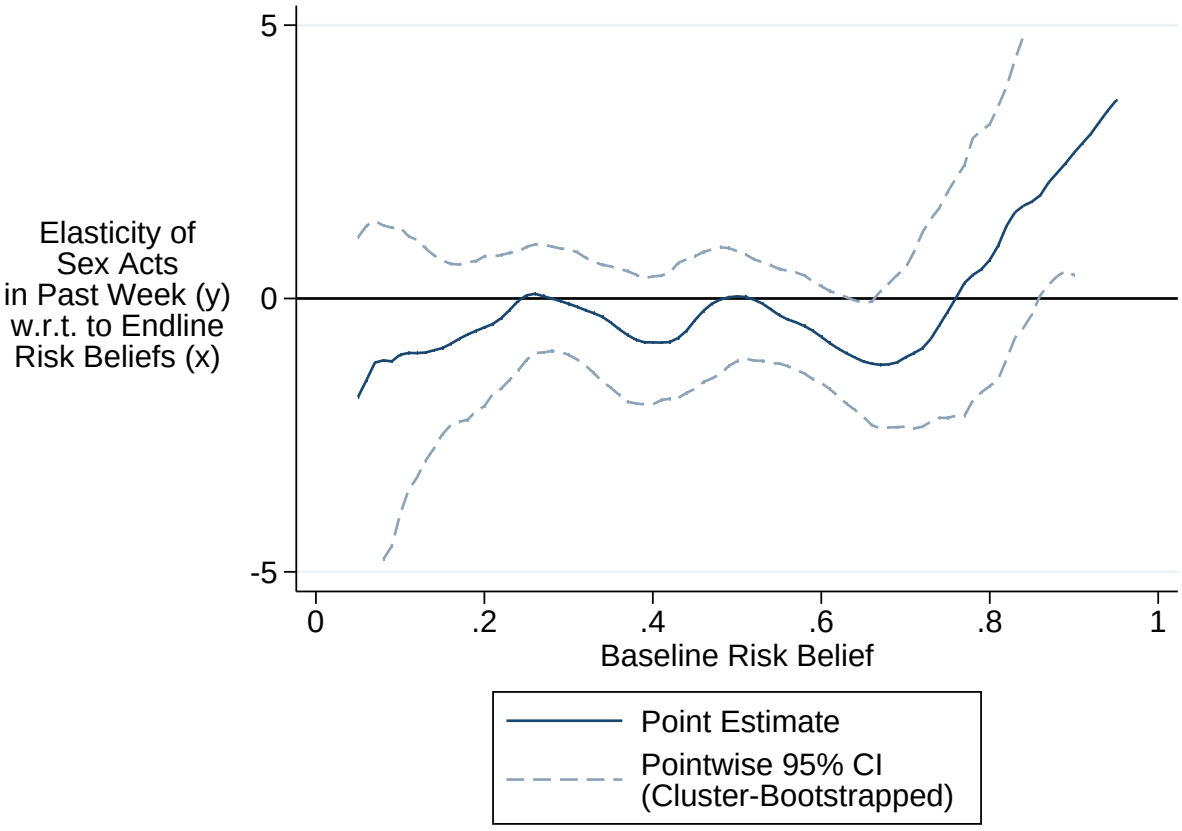


Figure F.3
IV Estimates of the Elasticity of Sex Acts in Past Week (y) w.r.t Endline Risk Beliefs (x), by Baseline Risk Belief



F.2 Bracketed Semi-Parametric Regression Estimates

As described in Section 1.4.5, I can also estimate my regressions by constructing indicator variables for ranges of the baseline risk belief distribution, interacting those with the treatment indicator, and running linear regressions of the outcome on the indicators, the interactions, and the controls. This alternative estimator has several attractive features that make it a useful supplement to the Robinson double-residualized local linear regressions. First, it does not suffer from boundary bias issues. Second, there is no bandwidth choice to make. Third, the nature of my estimation strategy makes constructing simultaneous (as opposed to pointwise) confidence intervals difficult, but the bracketed approach is amenable to standard techniques such as the Bonferroni correction. Fourth, it is much less computationally intensive, so I use it to conduct the sensitivity analyses later in this section.

Appendix Figures F.4 through F.6 reconstruct Figures 1.7 through 1.9 using the bracketed approach. The results are qualitatively identical to the Robinson method. The p-values for the highest category of baseline risk beliefs are well below 0.01; running the conservative Bonferroni correction on them yields a p-value below 0.02, so I can rule out the possibility that my results are arising from multiple-comparisons issues.

The confidence intervals in Appendix Figures F.4 through F.6 use cluster-bootstrapped standard errors with the belief adjustment process repeated within each bootstrap sample. The resulting confidence intervals are virtually identical to those that result from using analytic standard errors and ignoring the generated regressor problem from the adjustment procedure (not shown). Therefore, to reduce the computational complexity of the remainder of the sensitivity analyses, I will henceforth use simple analytic CIs with no adjustment for generated regressors.

Figure F.4
First-Stage Effect of Treatment (T) on Endline Risk Beliefs (x),
by Baseline Risk Belief

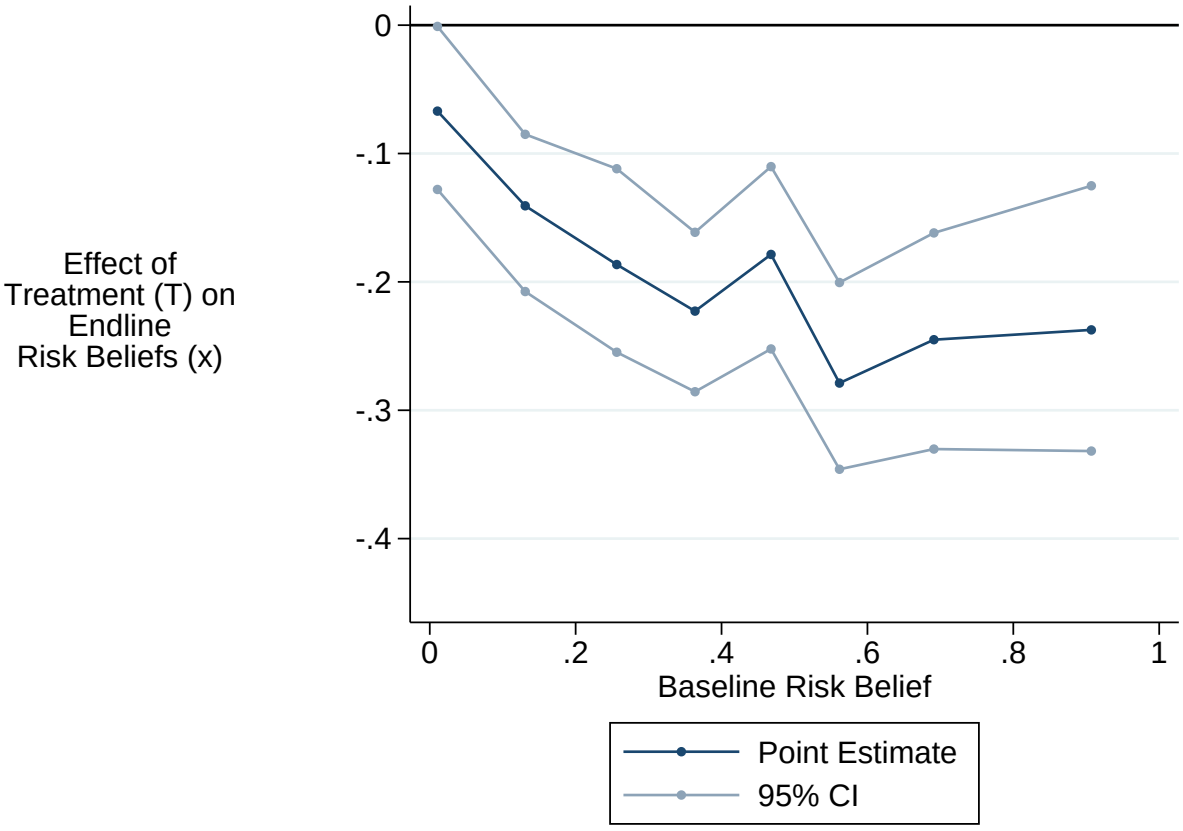


Figure F.5
Reduced-Form Effect of Treatment (T) on Log Sex Acts in Past Week ($\ln(y)$), by Baseline Risk Belief

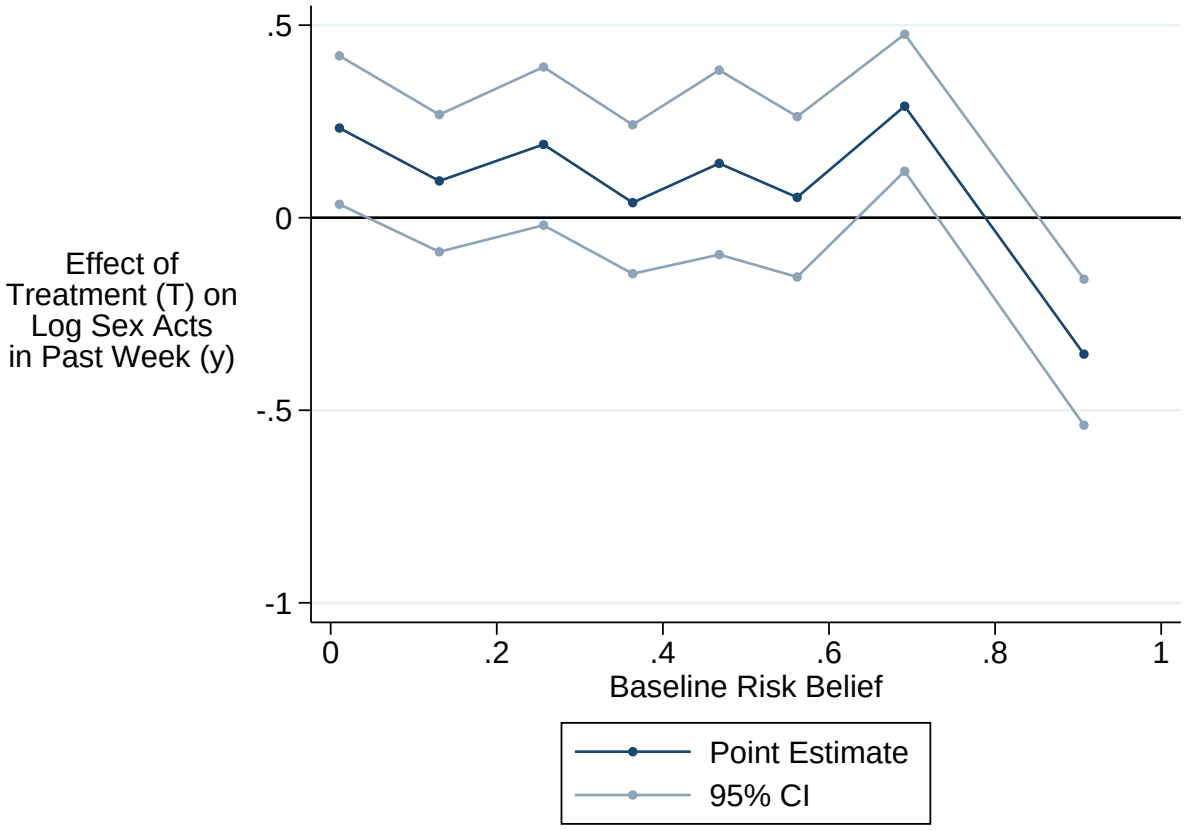


Figure F.6
IV Estimates of the Elasticity of Sex Acts in Past Week (y) w.r.t Endline Risk Beliefs (x), by Baseline Risk Belief

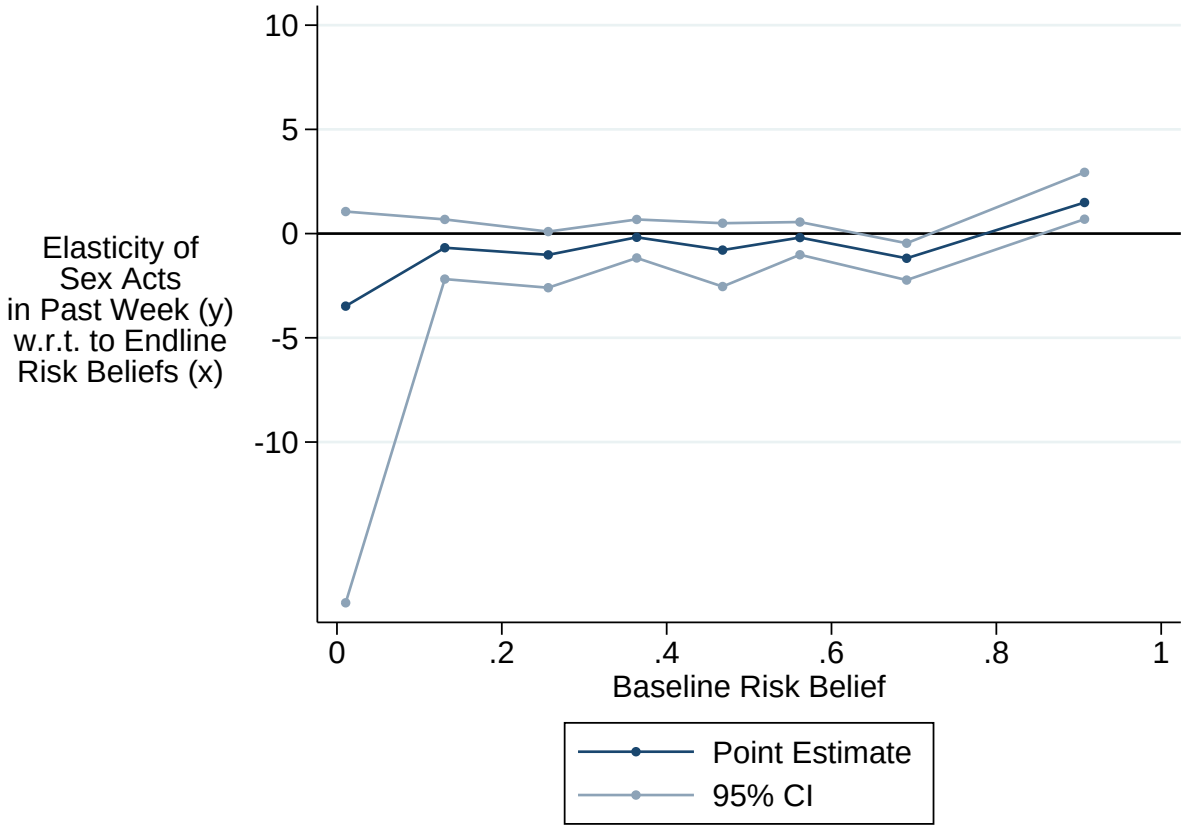
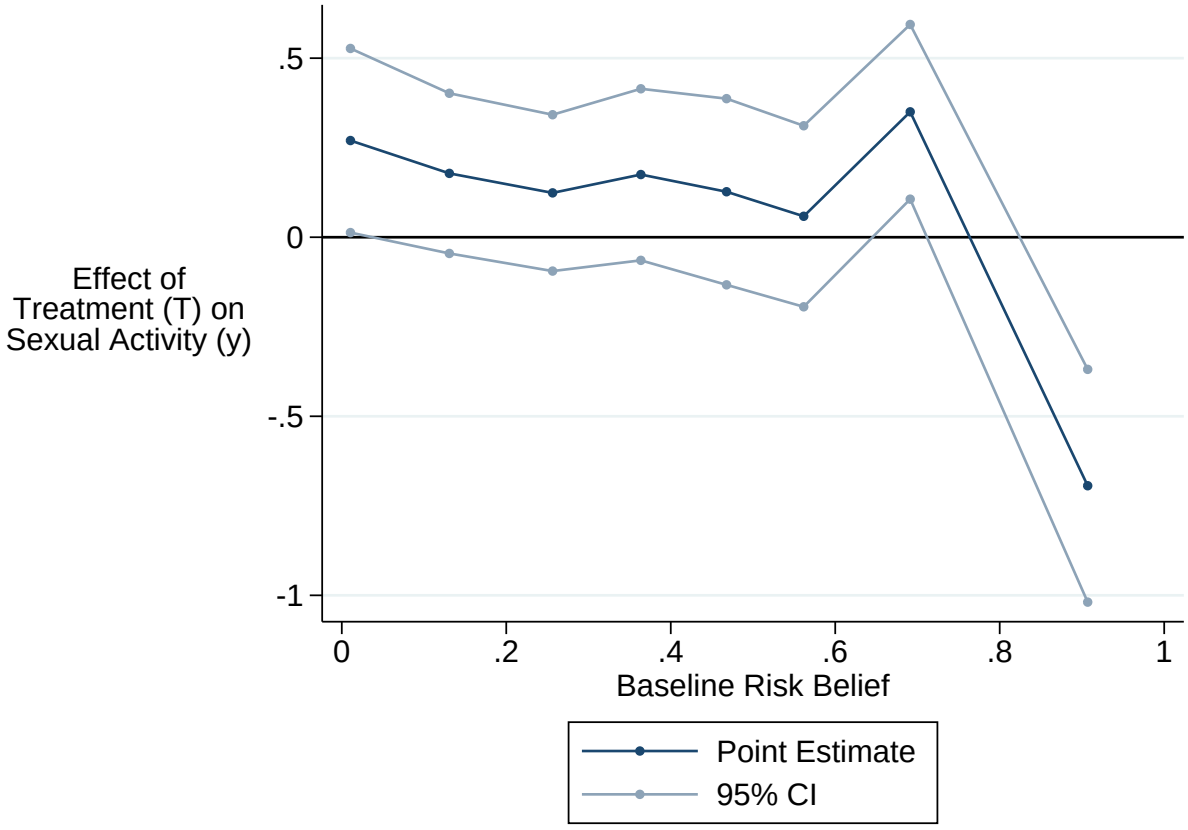


Figure F.7
Reduced-Form Effect of Treatment (T) on Log Sex Acts in Past Week ($\ln(y)$),
by Baseline Risk Belief
Without Adjusting Beliefs



F.3 Variations in Handling Baseline Risk Beliefs

My preferred specification adjusts baseline risk beliefs for contamination due to enumerator knowledge, as described in Section 1.3.5. Here I present alternative ways of handling enumerator-knowledge contamination: using the raw (unadjusted) beliefs; using the endline values of the belief variable for respondents whose baseline data was collected prior to the enumerators learning the HIV risk information; and using the within-group rank of beliefs (with ties broken at random) for respondents surveyed before the enumerator training, and after the enumerator training, with ranks normalized to lie within 0-1.

Figure F.8
Reduced-Form Effect of Treatment (T) on Log Sex Acts in Past Week ($\ln(y)$),
by Baseline Risk Belief
Using Endline Beliefs for Respondents with Baseline Survey Before Training

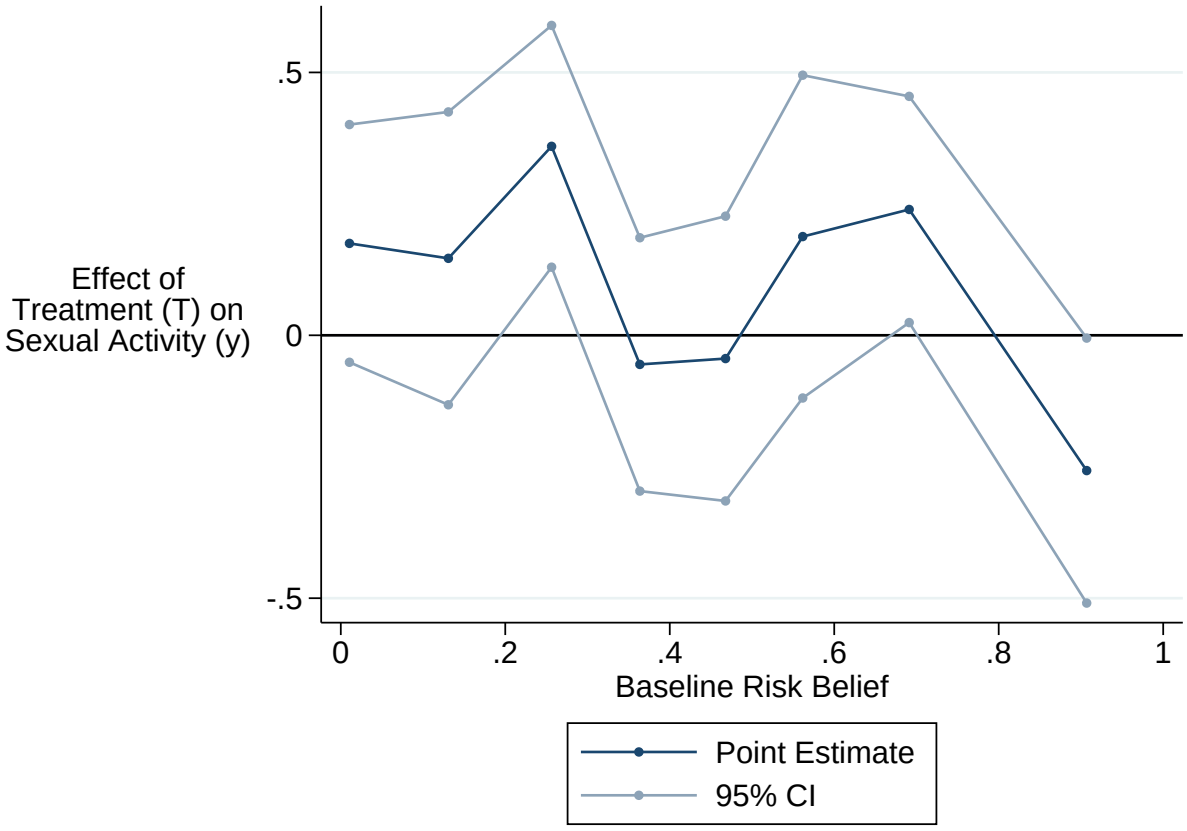
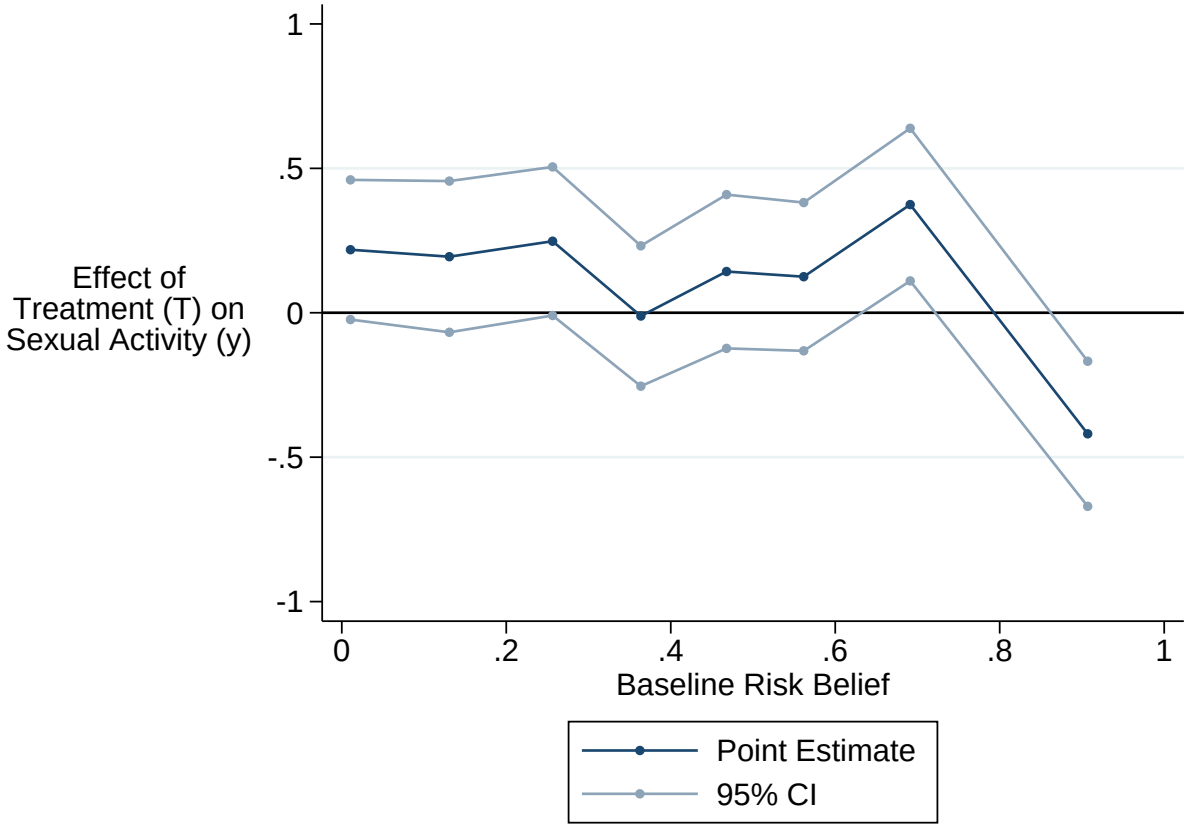


Figure F.9
Reduced-Form Effect of Treatment (T) on Log Sex Acts in Past Week ($\ln(y)$),
by Baseline Risk Belief
Using Normalized Within-Group Rank of Beliefs for Respondents Surveyed Before & After Training Session
as Belief Measure



One specific concern is whether the particular HIV risk variable I am using matters for the pattern of effects I find in my results. My choice of risk variable (the per-act risk of contracting HIV from a single unprotected sex act with a randomly-selected attractive person from the local area) is motivated by the literature and is the same one I used in a working paper I wrote prior to running the field experiment. However, there are several underlying risk variables that I could have used, and many conceivable ways of combining them. Rather than explore all potential options for constructing HIV risk variables, I simply use principal components analysis as an automated way to take a weighted average of the four underlying variables. Appendix Figure F.10 shows the results for this weighted average of the four variables.[1] As with all my other specifications, I estimate a negative treatment effect for the highest category of beliefs, and positive or zero effects for all lower categories. The estimates are noisier, but I continue to reject a zero effect for the highest category at the p=0.1 significance level.

[1] The four underlying variables collected on the survey are the subject's perceived (1) per-act risk of contracting HIV from unprotected sex with an infected partner, (2) annual risk of contracting HIV from unprotected sex with an infected partner, (3) prevalence of HIV in the local area among all people of the opposite sex, and (4) prevalence of HIV in the local area among attractive people of the opposite sex.

Figure F.10
Reduced-Form Effect of Treatment (T) on Log Sex Acts in Past Week ($\ln(y)$),
by Baseline Risk Belief
Using First Principal Component of all HIV Risk Beliefs as Belief Measure

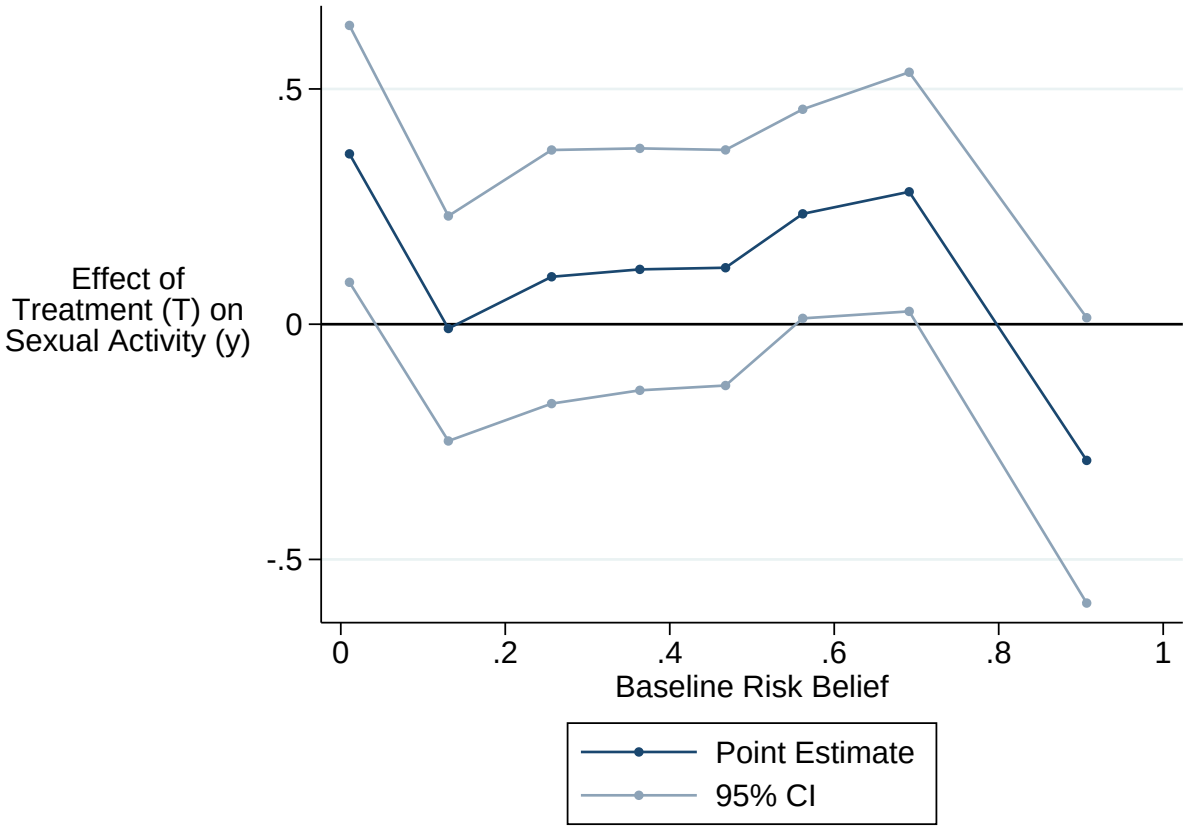
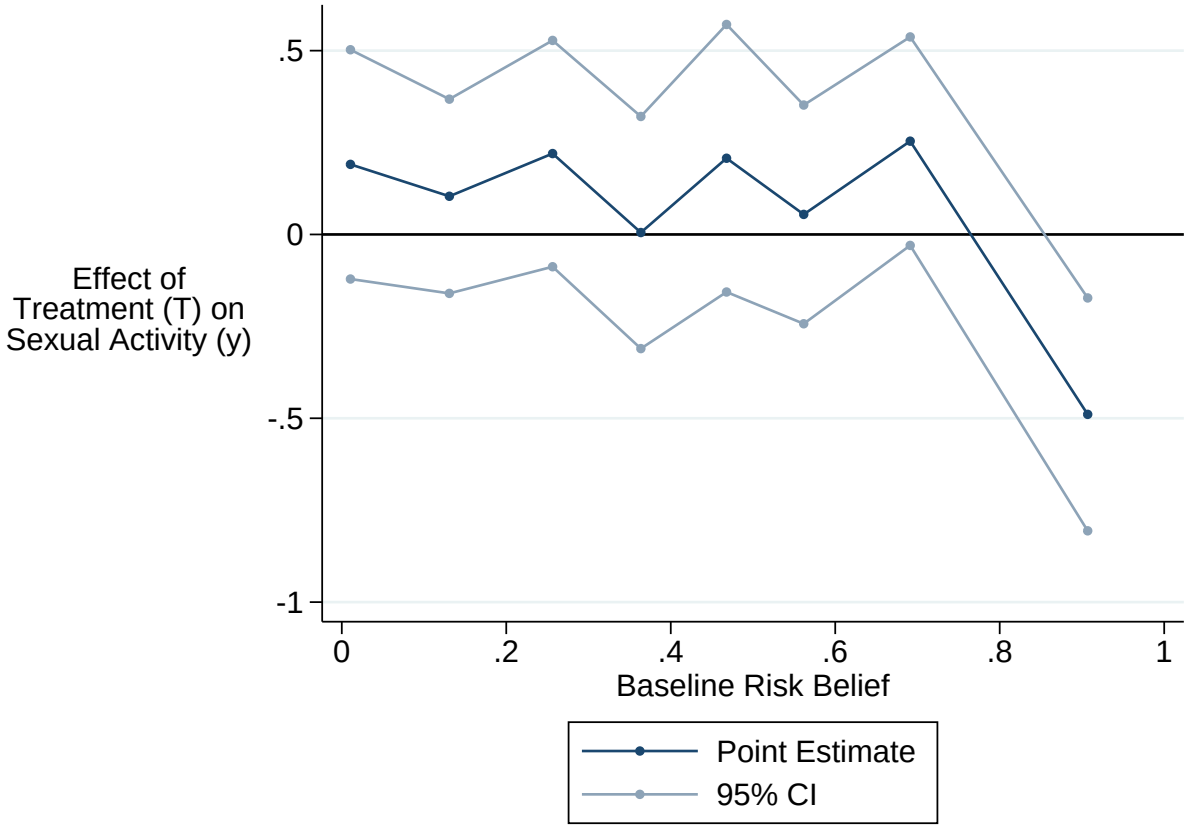


Figure F.11
Reduced-Form Effect of Treatment (T) on Log Sex Acts in Past Week ($\ln(y)$),
by Baseline Risk Belief
No Controls



F.4 Variations in Regression Specification

In Appendix Figures F.11 through F.14 I explore various alternative regression specifications for my main outcome.

Figure F.12
Reduced-Form Effect of Treatment (T) on Log Sex Acts in Past Week ($\ln(y)$),
by Baseline Risk Belief
Controlling for Sampling Strata Only

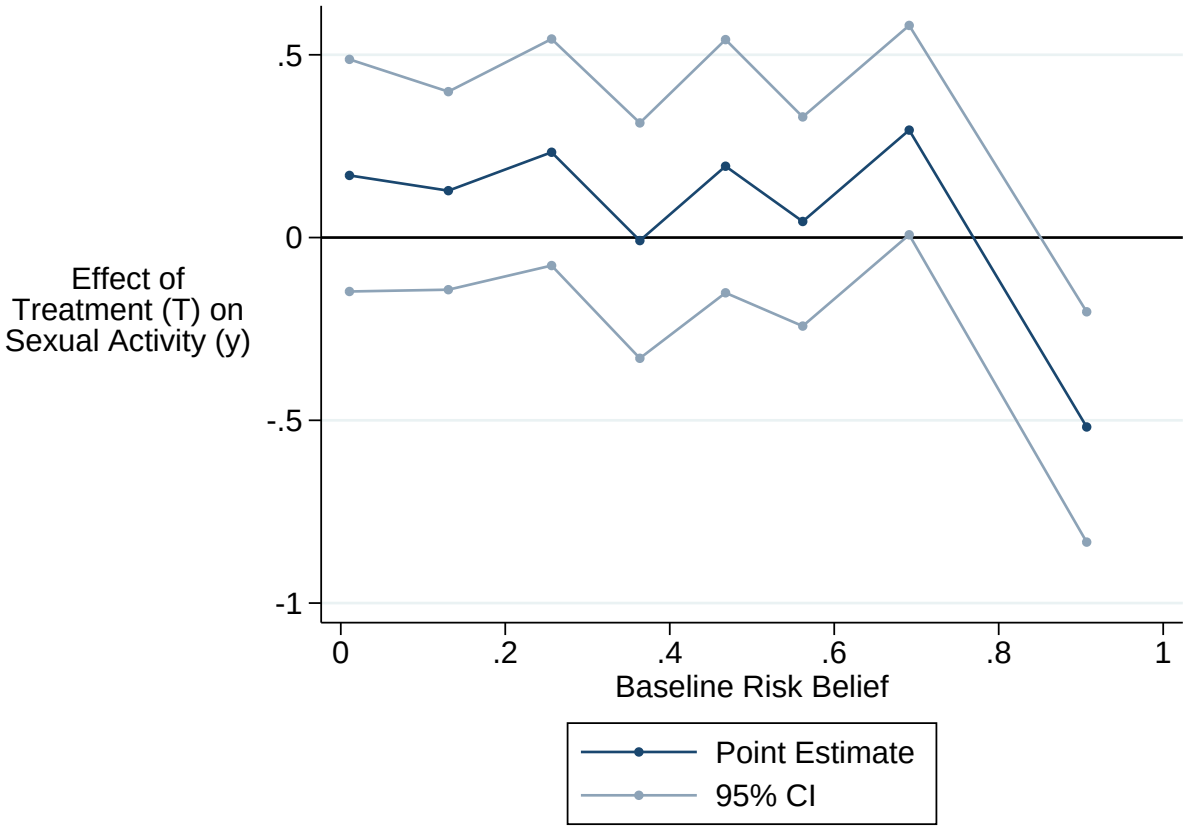


Figure F.13
Reduced-Form Effect of Treatment (T) on Sex Acts in Past Week (y),
by Baseline Risk Belief
Without Logging y

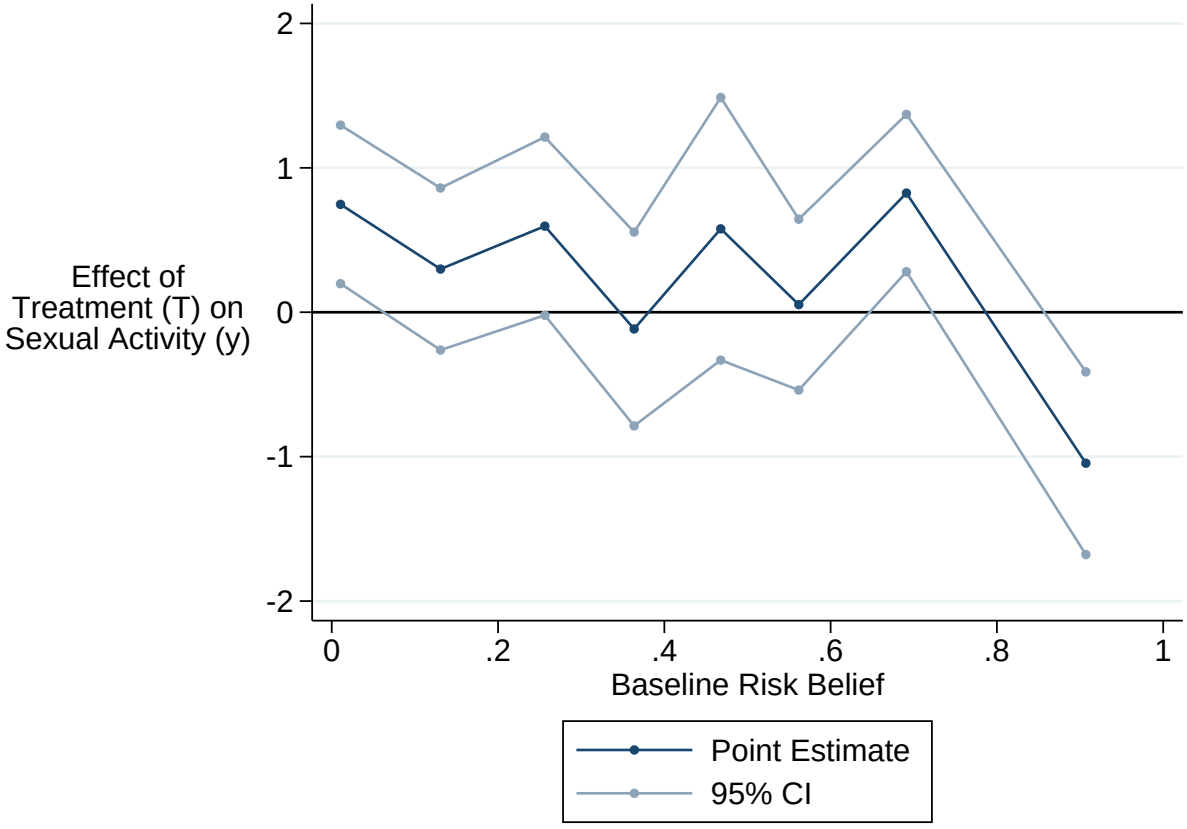


Figure F.14
Reduced-Form Effect of Treatment (T) on Sex Acts in Past Week (y),
by Baseline Risk Belief
Without Logging y, Zero-Inflated Negative Binomial Regression (Marginal Effects)

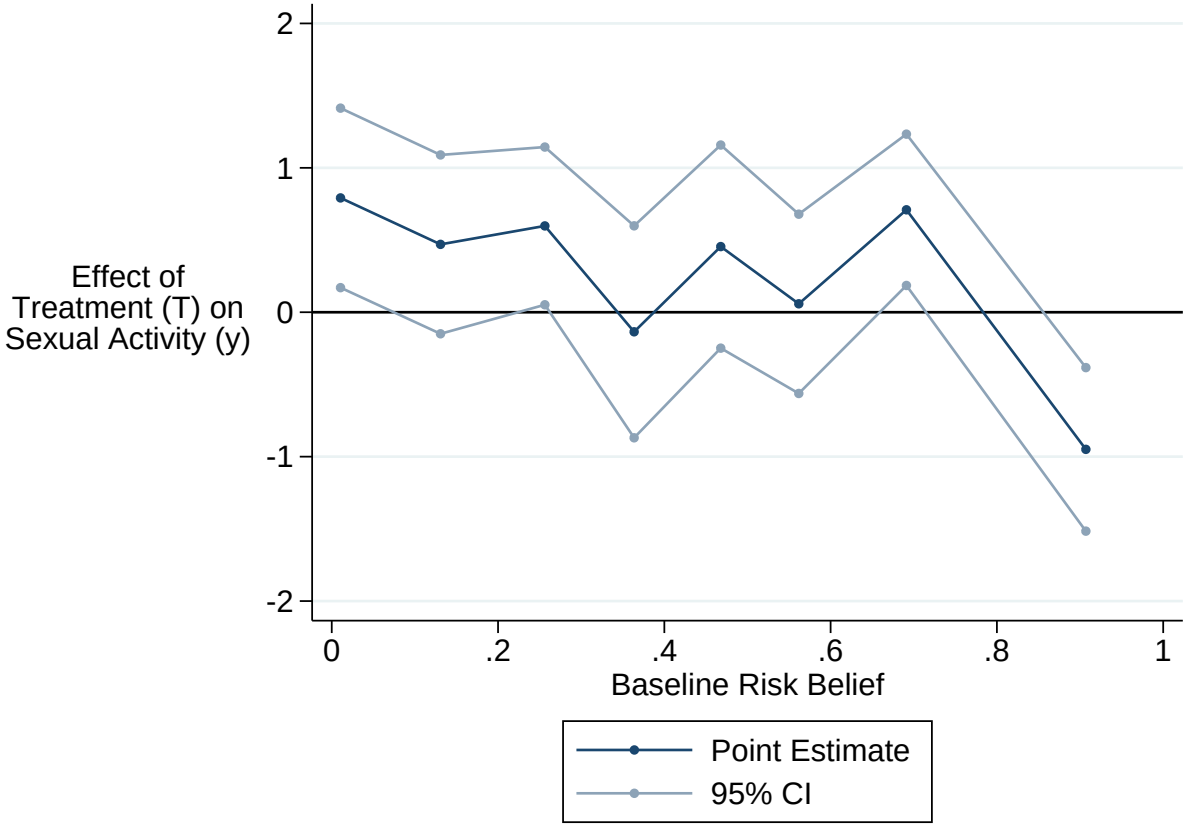
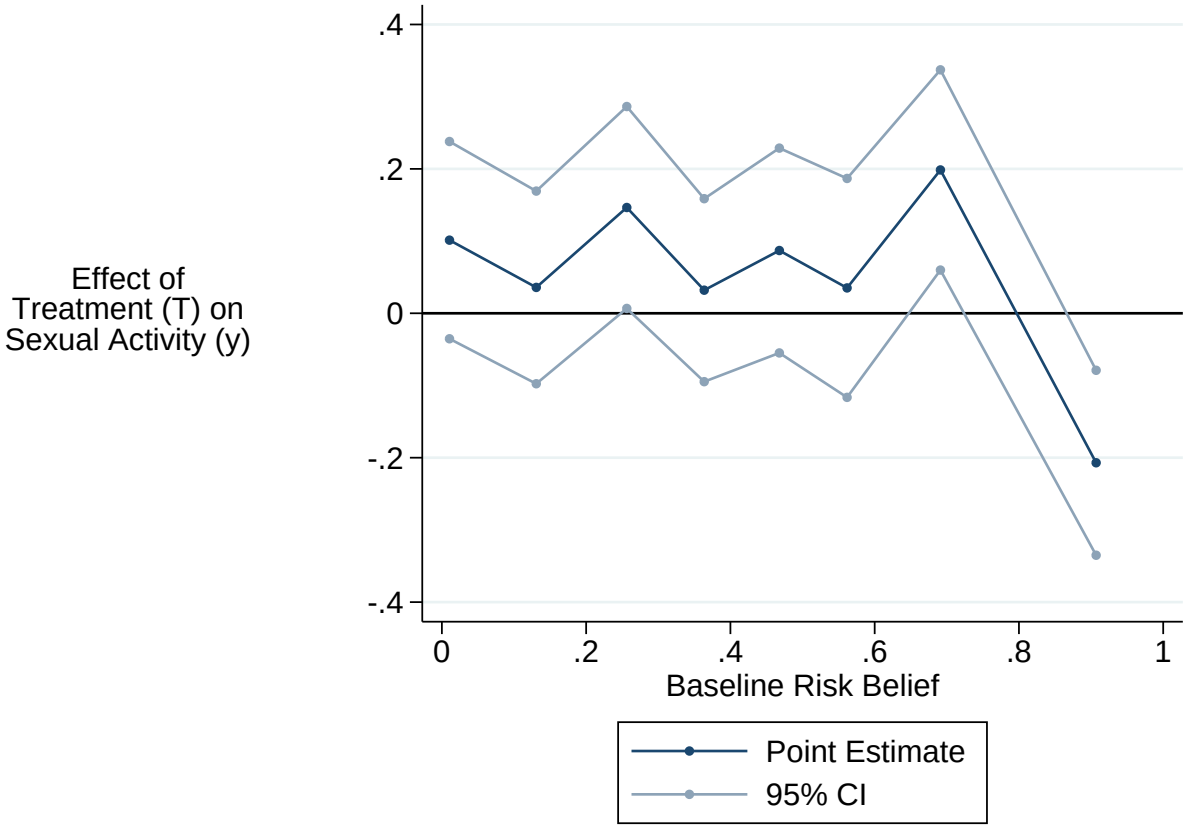


Figure F.15
Reduced-Form Effect of Treatment (T) on Any Sex Acts in Past Week (y),
by Baseline Risk Belief
LPM Results



F.5 Alternative Outcome – Any Sex in Past Week

The following regressions, presented in Appendix Figures F.15 through F.17 look at any sex in the past week as the outcome instead of total sex. They report marginal effect estimates from LPM, Logit, and Probit regressions.

Figure F.16
Reduced-Form Effect of Treatment (T) on Any Sex Acts in Past Week (y),
by Baseline Risk Belief
Logit Marginal Effects

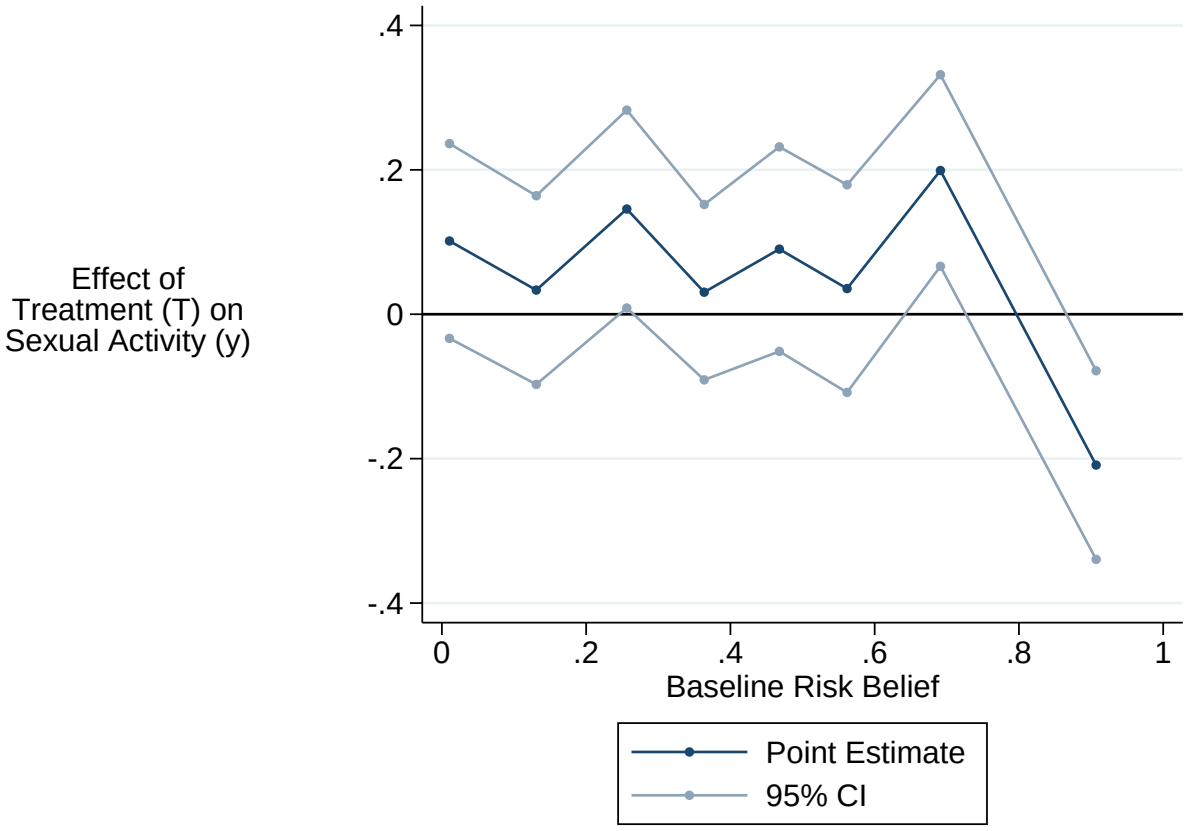


Figure F.17
Reduced-Form Effect of Treatment (T) on Any Sex Acts in Past Week (y),
by Baseline Risk Belief
Probit Marginal Effects

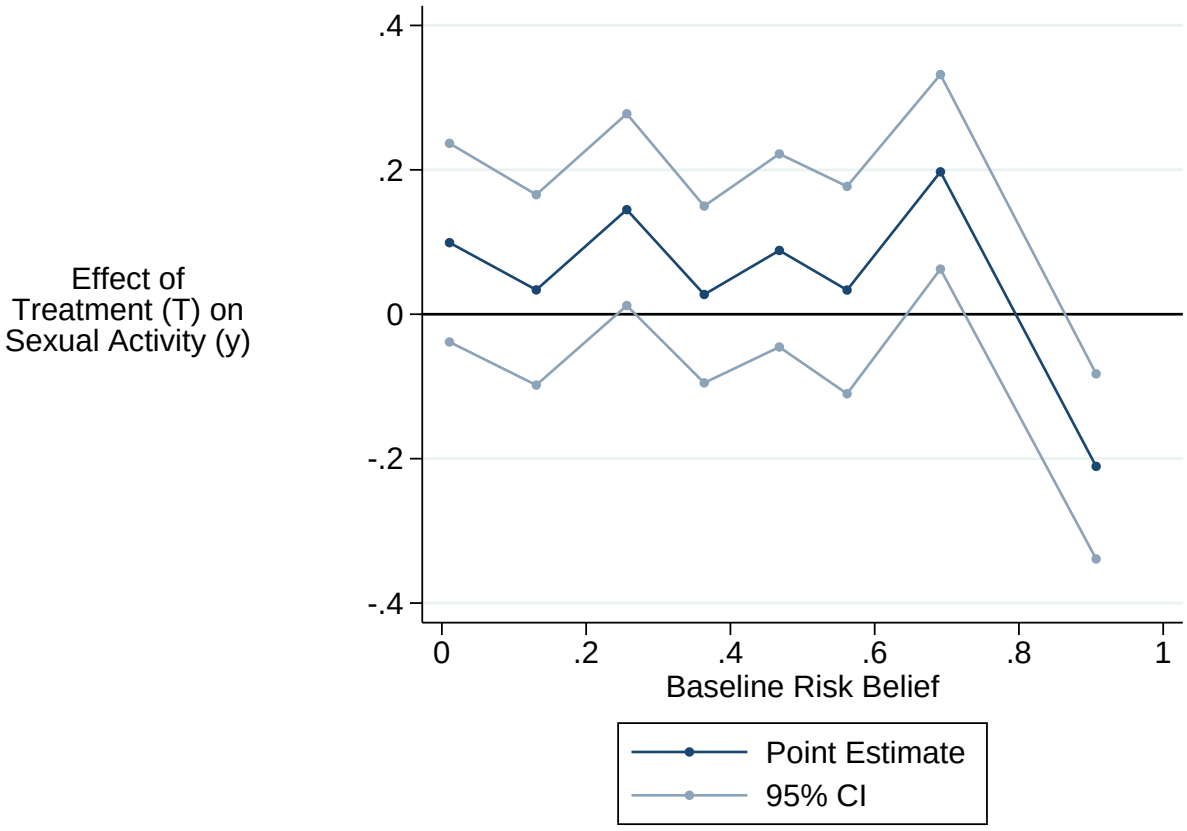
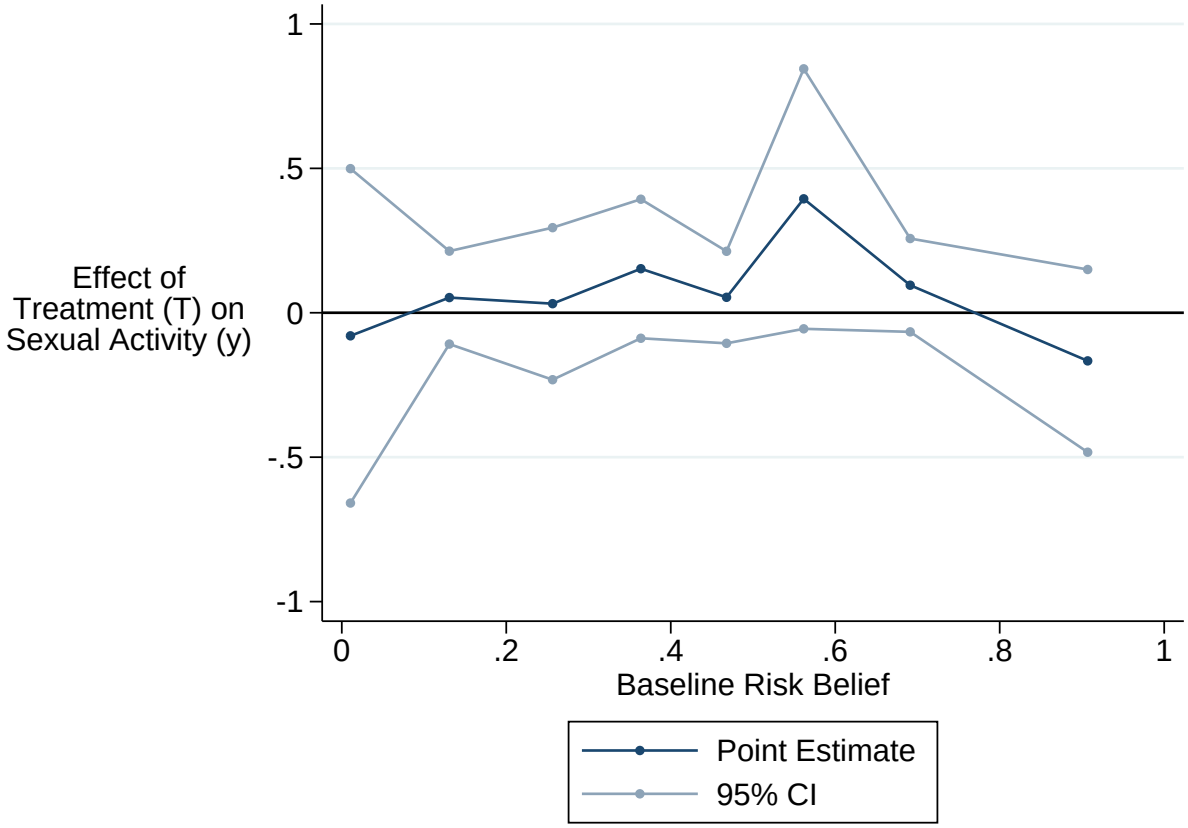


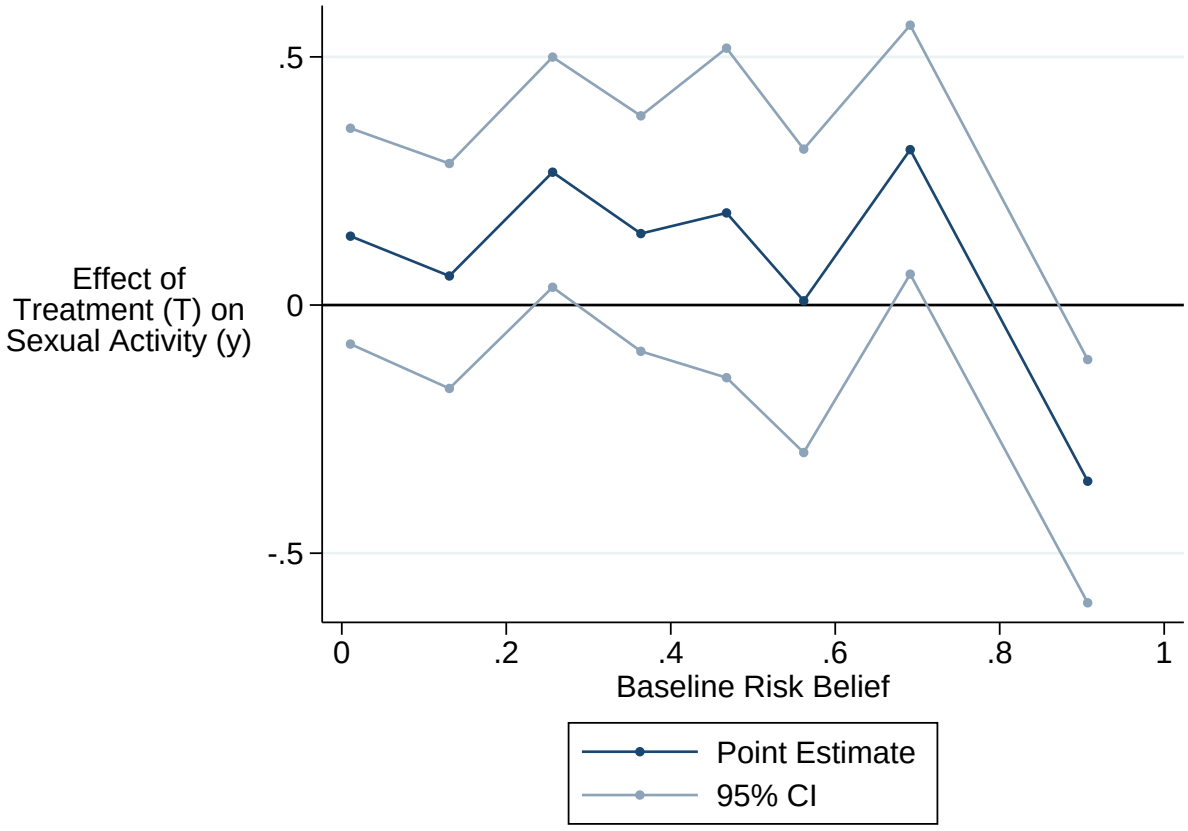
Figure F.18
Reduced-Form Effect of Treatment (T) on Log Diary Sexual Activity Index ($\ln(y)$), by Baseline Risk Belief



F.6 Alternative Outcomes – Sexual Activity Indices

The following graphs (Appendix Figures F.18 and F.19) present estimates using the (logged) sexual activity indices (both overall and sex diary-only) as outcome variables.

Figure F.19
Reduced-Form Effect of Treatment (T) on Log Overall Sexual Activity Index ($\ln(y)$), by Baseline Risk Belief



APPENDIX G

Relationship between overall and covariate-specific LATEs

The w^k-specific LATEs estimated by the procedure in Section 1.4.6 form the components of the overall LATE for the entire sample, but the overall LATE is an unequally-weighted average of these components, not a simple mean. In this subsection I show that their weights in forming the overall LATE are given by the share of the data with each w^k times the degree of compliance with the instrument (the extent to which the instrument shifts x) for each w^k.

For the sake of exposition, begin with the Wald IV estimator of the local average treatment effect for the whole population, given by

$$\hat{\delta}_{Wald} = \frac{\mathbb{E}[y_i|T_i = 1] - \mathbb{E}[y_i|T_i = 0]}{\mathbb{E}[x_i|T_i = 1] - \mathbb{E}[x_i|T_i = 0]} \tag{G.1}$$

This is asymptotically equal to the ILS and 2SLS estimators because all three are consistent. Assume the baseline covariate w_{i0} is discrete (or measured discretely due to the data collection process) with values $w_1, ..., w_K$. Define $y_i^0 = \mathbb{E}[y_i|T = 0]$, where the expectation is taken over the support of i for the given value of the treatment indicator, and likewise for y_i^1, x_i^0, and x_i^1. Using the Law of Total Expectation, one can rewrite $[\mathbb{E}[y_i|T_i = 1]$ as $\sum_{k=1}^{K} \mathbb{E}[y_i|T_i = 1, w_{i0} = w^k]\mathbb{P}(w^k)$. Then

$$\hat{\delta}_{Wald} = \frac{\sum_{j=1}^{m} \mathbb{E}[y_i|T_i = 1, w_{i0} = w^k]\mathbb{P}(w^k) - \sum_{k=1}^{K} \mathbb{E}[y_i|T_i = 0, w_{i0} = w^k]\mathbb{P}(w^k)}{\sum_{k=1}^{K} \mathbb{E}[x_i|T_i = 1, w_{i0} = w^k]\mathbb{P}(w^k) - \sum_{k=1}^{K} \mathbb{E}[x_i|T_i = 0, w_{i0} = w^k]\mathbb{P}(w^k)} \tag{G.2}$$

$$= \frac{\sum_{k=1}^{K} \mathbb{E}[y_i^1 - y_i^0]|w_{i0} = w^k]\mathbb{P}(w^k)}{\sum_{k=1}^{K} \mathbb{E}[x_i^1 - x_i^0|w_{i0} = w^k]\mathbb{P}(w^k)} \tag{G.3}$$

Let $\hat{\beta}^y(w^k)$ be a consistent estimate of the effect of the treatment on y given $w_{i0} = w^k$, $\mathbb{E}[y_i^1 - y_i^0 | w_{i0} = w^k]$. Then the Wald estimator can be rewritten as $\frac{\sum_{j=1}^m \hat{\beta}^y(w^k)\mathbb{P}(w^k)}{\sum_{j=1}^m \hat{\beta}^x(w^k)\mathbb{P}(w^k)}$. The ILS estimator for a w^k-specific slope is $\hat{\alpha}_{ILS,j}(w^k) = \frac{\hat{\beta}^y(w^k)}{\hat{\beta}^x(w^k)}$, so we can rewrite the Wald estimator as:

$$\frac{\sum_{j=1}^m \hat{\alpha}_{ILS}(w^k)\hat{\beta}^x(w^k)\mathbb{P}(w^k)}{\sum_{j=1}^m \hat{\beta}^x(w^k)\mathbb{P}(w^k)} = \sum_{j=1}^m \hat{\alpha}_{ILS}(w^k)\frac{\hat{\beta}^x(w^k)}{\hat{\beta}^x}\frac{n_j}{N} \tag{G.4}$$

where n_j is the number of observations with $w_{i0} = w^k$ and N is the total number of observations in the dataset. Let $\theta_j = \frac{\hat{\beta}^x(w^k)}{\hat{\beta}^x} n_j$. Then we have $\hat{\delta}_{Wald} = \frac{\sum_{j=1}^m \hat{\delta}_{Wald}(w^k)\theta_j}{N}$. The overall estimate of the slope of y with respect to x is the weighted average of the w^k-specific slope estimates.

The weights θ_j have two components. The first part is the number of observations with $w_{i0} = w^k$. This is multiplied by the second part: the ratio of the impact of the treatment on x at $w_{i0} = w^k$ to the overall treatment effect, which is a continuous measure of compliance with the categorical instrument T. Observations where the treatment shifts x more have greater weight in determining the overall mean marginal effect estimate – in other words, the overall LATE is the compliance-weighted average of the baseline covariate-specific LATEs.

APPENDIX H

Balance and comparison demographic characteristics of sample to census data

H.1 Balance and comparison demographic characteristics of sample to census data

Appendix Table H.1 shows summary statistics for important baseline characteristics: all available baseline measures corresponding to outcome variables used in the results tables as well as an index of asset holdings (used as a control in the main results tables). Columns 4 and 5 present formal statistical tests of the null hypothesis that pre-program characteristics have equal means across all four study arms. For each covariate, the test is conducted by running two linear regressions as seemingly-unrelated regressions (SURs) of the variable on a saturated set of categorical indicator variables for study arm, one regression for each round. We then run a joint test of the null hypothesis that all the coefficients are zero. Column 5 shows the p-values, which are uniformly above 0.3. The last row shows the test statistic and p-value for a joint test of the hypothesis that all the coefficients equal zero across all 26 regressions. We fail to reject the null of no differences (p-value of 0.81). The sample is similarly balanced on demographic covariates; see the analogous test statistics in Appendix Table H.2.

Table H.1
Balance of baseline variables

Variable	Worker Sample Summary Statistics (1) Mean	(2) SD	(3) N	Test for Difference Across Study Arms (4) Chi^2	(5) p-value
Income and Spending					
Total spending since last Friday, inclusive [MK]	2271.04	3728.39	329	3.84	0.70
Cash remaining out of total received since last Friday, inclusive [MK]	683.81	2618.13	329	7.00	0.32
Expenditure Composition					
Food for consumption at home	0.66	0.23	349	3.05	0.80
Maize only	0.23	0.26	349	2.24	0.90
Food for consumption out of home	0.06	0.07	349	1.77	0.94
Non-Food	0.28	0.23	349	3.83	0.70
Assets					
Baseline Asset Ownership Index (First Principal Component)	0.00	2.68	350	5.53	0.48
Combined Test Across All Variables				28.50	0.81

Notes: Sample includes 359 respondents who participated in at least one round of the work program and have data from at least one data source for that round (either the payday data, the survey, or both). All money amounts are in Malawian Kwacha (MK); during the study period the market exchange rate was approximately MK400 to the US dollar, and the PPP exchange rate was approximately MK160 to the US dollar. Tests for any difference in means across study arms use seemingly-unrelated regressions of a variable on a full set of categorical indicator variables for study arm, clustered by respondent, to do pooled tests of the null hypothesis that all study arms have equal means in both rounds; the test statistics are chi-square distributed with 6 degrees of freedom. The Combined Test Across All Variables is a combined SUR of all 8 covariates in both rounds; its chi-square test statistic has 36 degrees of freedom due to collinearity between some of the equations estimated.

Table H.2, columns 1 to 3, presents summary statistics of demographic characteristics for the 350 workers from our sample for whom baseline data is available. As a basis for comparison, we also present statistics for Mulanje District as a whole, taken from the IPUMS-International 10% sample of the 2008 Malawi Population and Housing Census. A comparison of our sample with the rest of the district suggests that it is generally representative of the local area, with differences that are likely due to the criteria used by the Village Development Committee (VDCs) to select workers for the program. Our sample is 69% female, which is substantially higher than the district average of 55%. It also has a larger share of people from the Lomwe ethnic group, at 90% compared with 75%. It is otherwise quite similar to the district as a whole, with similar rates of marriage (70%) and Christian religion (90%). The differences in the other variables are fairly small, and consistent with the VDCs selecting people of lower socioeconomic status for the program. For example, our sample averages 3.5 years of completed schooling, compared with 4.4 years for the district as a whole, and has a mean age of 40 compared with 37 for Mulanje District. Our workers are also more likely to be divorced and less likely to be single.

Table H.2

Demographic characteristics of sample - balance and comparison to census

	Worker Sample Summary Statistics			Test for Difference Across Study Arms		Mulanje District 2008 Census Summary Statistics	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Variable	Mean	SD	N	Chi-square	p-value	Mean	SD
Male	0.31	0.46	344	3.79	0.70	0.45	0.50
Religion							
Christian	0.90	0.30	341	5.93	0.43	0.91	0.28
Muslim	0.10	0.30	341	5.93	0.43	0.05	0.22
Marital Status							
Married	0.69	0.46	338	5.46	0.49	0.71	0.45
Divorced/Widowed	0.25	0.44	338	5.44	0.49	0.17	0.37
Single	0.05	0.21	338	8.74	0.19	0.12	0.33
Ethnic Group							
Lomwe	0.89	0.31	344	1.66	0.95	0.75	0.43
Yao	0.07	0.26	344	2.29	0.89	0.05	0.22
Mang'anja	0.02	0.13	344	6.15	0.41	†	
Other	0.02	0.15	344	8.27	0.22	0.20	0.40
Years of Education Completed	3.54	3.15	341	3.47	0.75	4.45	3.91
Age (Years)	40.03	15.40	344	4.24	0.64	37.35	17.27
Combined Test Across All Variables				61.42	0.73		

Notes: Pooled Sample includes 359 respondents who participated in at least one round of the work program. Tests for any difference in means across study arms use seemingly-unrelated regressions of a variable on a full set of categorical indicator variables for study arm, clustered by respondent, to do pooled tests of the null hypothesis that all study arms have equal means in both rounds; the test statistics are chi-square distributed with 6 degrees of freedom. The Combined Test Across All Variables is a combined SUR of all 13 covariates in both rounds; its chi-square test statistic has 69 degrees of freedom due to collinearity between some of the equations estimated. † The 2008 Malawi Census does not report Mang'anja ethnicity as a separate category, so it is included in "other".

APPENDIX I

Variable definitions

I.1 Variable definitions

Data used in this paper come from three rounds of "full length" surveys (a baseline and two follow-up interviews), from two- to four-question surveys during paydays as well as from administrative records of the project. We conducted a baseline survey from 4 Oct 2013 to 19 Oct 2013 and two follow-up surveys after the last payday weekend of each round, once from 2 Dec 2013 to 7 Dec 2013 and once from 27 Jan 2014 to 31 Jan 2014. All variables that are created from survey data are Winsorized at the 1st and 99th percentile. All figures in money terms are in local currency units, Malawi Kwacha (MK).

I.1.1 Variables from payday surveys

Amount spent on same day as income receipt is total market spending on all days that workers received their wages (sum of all four payday Fridays or Saturdays for the weekly payment group; the fourth payday Friday or Saturday for the lump sum payment group).

Money spent at market on Fridays 1, 2, 3 is the sum of total market spending on the first three payday Fridays.

Money spent at market on Saturdays 1, 2, 3 is the sum of total market spending on the first three payday Saturdays.

Money spent at market on Friday 4 is the total market spending on the fourth payday Friday.

Money spent at market on Saturday 4 is the total market spending on the fourth payday Saturday.

I.1.2 Variables from follow-up surveys

Total spending since last Friday, inclusive [MK] is the total household spending starting from the fourth payday Friday until the day of the survey interview in the week after the fourth payday. The variable is derived from the difference of the answers to the questions "Since last Friday, how much cash have you received?" and "How much of that cash do you have left?", respectively.

Remaining cash out of received since last Friday, inclusive [MK] is the household's remaining cash holdings out of money received starting from the fourth payday Friday until the day of the survey interview.

Self-reported wasteful spending on weekend 4 of round 2 variables ask for money that respondents report as "wasted" or spending which the respondent was tempted into spending that he/she should not have spent:

- *Total since last Friday, inclusive [MK]* is the sum of total wasteful spending starting from the fourth payday Friday until the day of the survey interview in the week after the fourth payday.
- *Friday [MK]* is total wasteful spending on the fourth payday Friday.
- *Saturday [MK]* is total wasteful spending on the fourth payday Saturday.
- *Sunday and after [MK]* is the sum of total wasteful spending starting from the fourth payday Sunday until the day of the survey interview in the week after the forth payday.

Expenditure shares based on itemized elicitation is the sum of itemized expenditures, grouped into different categories as a share of total expenditures across all items based on an large listing of possible items (with items derived from Malawi's Integrated Household Survey; a select number of items was consolidated or omitted but each category had an "other" option to capture items that were left out; total number of 105 items in 12 categories).

- *Food for consumption at home* includes eight categories of food items typically used for home consumption.
- *Maize only* includes only maize flour and maize grain.
- *Food for consumption out of home* includes all items from the categories "cooked foods from vendor" and "Beverages" which are typically consumed away from home.
- *Non-Food* includes all non-food items.

Value of net asset purchases since last interview is the sum of the difference between the value of assets bought and assets sold from an itemized list of common assets (as well as an "other" category) considering purchases and sales since the last interview, i.e. since baseline interview for follow-up 1 and since follow-up 1 for follow-up 2.

I.1.3 Variables from baseline surveys

Assets index is an index based on the first principal component of the number of items owned out of 64 common non-financial, non-livestock assets and the number of animals owned out of 9 common types of livestock.

Total spending is defined similarly to "Total spending since last Friday, inclusive" described under follow-up variables above, covering the last Friday prior to the interview until the day of the survey interview.

I.1.4 Variables from project records

Bought any shares is an indicator for whether the respondent bought at least one "share" of the investment opportunity offered after the follow-up interviews (see details in main text in Data Collection section).

Total spent on shares is the total amount spent on the investment opportunity offered and equals the number shares bought times the price of one share (MK 1,500).

APPENDIX J

Intervention Inputs

The Mango Tree and Government Administered Programs differ in terms of the materials, training, and other support provided to schools; we specify the differences for each below, and also show them in Table 3.1.

J.1 Materials

The NULP provides the following materials to each MT and CCT school:

- One Leblango Teacher's Guide for each teacher
- Three term-specific Leblango primers for each student (up to 200 students per class)
- Three term-specific Leblango readers for each student (up to 200 students per class)
- One English Teacher's Guide for each P1-P3 teacher
- Three term-specific English primers for each student (up to 200 students per class)

In addition, the MT Program provides additional materials to each school:

- One slate for each student (up to 200 students per class)
- Two wall clocks per school

J.2 Teacher Training

The NULP's teacher training comprises the following:

- One residential five-day training in the Leblango orthography for P1-P3 teachers in December the year before they enter the program (MT Program only)
- Three trainings in literacy methods for P1-P3 teachers during the school holidays each year
  - MT Program: residential trainings held in the district capital, conducted by experienced MT staff
  - CCT Program: non-residential trainings held at the CCs, conducted by CCTs. To facilitate these trainings, Mango Tree CCTs with instructional videos to learn which they play on solar-powered, portable DVD players. The videos also provide examples of instructional practice in real-life classrooms, as well as provide a possible inexpensive alternative to residential training models.
- Special field monitoring and support supervision visits to schools
  - MT Program: 3 times per term by project staff, 2 times per term for CCTs
  - CCT Program: 2 times per term for CCTs

J.3 Other Support

- Parent Interaction. Schools in both the MT Program and CCT Program hold a parent meeting each term. Each meeting has specific content designed by Mango Tree as well as time for other school-related issues to be addressed. These meetings are conducted by the field officers for the MT Program schools and the CCTs for the CCT Program schools. The term 1 meeting focuses on answering parents' questions about literacy and the NULP. It also introduces a specialized report card, which differs from the ones ordinarily used by school, that the NULP uses to provide parents with feedback on their children's performance. The term 2 meeting allows parents to observe classes in session and trains parents in the Parent Assessment Tool. Modeled after one developed in India by Pratham and also used by UWEZO in East Africa, the tool a simple way for parents to assess their children in basic reading skills.[1] At the term 3 meetings,

[1] The tool has 4 parts: 1) letter name knowledge; 2) familiar word reading; 3) reading fluency test; and 4) reading comprehension test.

students demonstrate what they've learned during the school year for their parents and are awarded prizes for a variety of literacy and other academic achievements.

- Monthly Radio Program. Mango Tree sponsors a one-hour monthly radio program (supported by SMS messages and surveys to engage listeners in feedback) that broadcasts literacy and local language education topics to parents, teachers and communities in the Lango Sub-region. This program is available to students, teachers, and parents in all three study arms, and thus we cannot analyze its effects in this study.

- Take a Book Home Activity (MT Program only). Beginning near the end of the first term, children take home books each week that they are expected to read with their parents and other family members. Teachers are given a simple recording sheet to track the movement of books.

APPENDIX K

Robustness Checks

K.1 Effect of NULP on Exam Scores without Controlling for Baseline Scores

Our preferred specification for analyzing the effect of the NULP on exam scores controls for the pupil's baseline score on the test component in question, or when analyzing the effect on the combined exam score indices, controls for the pupil's baseline score on the index. In this section, we show that our results are qualitatively and numerically robust to the exclusion of those controls from our regressions. In this section we replicate Tables 3.4-3.6, but instead of estimating equation 3.1 we estimate:

$$y_{is} = \beta_0 + \beta_1 \text{MTSchool}_s + \beta_2 \text{GovtSchool}_s + \mathbf{L}_s'\gamma + \epsilon_{is} \tag{K.1}$$

$$\tag{K.2}$$

Here i indexes students and s indexes schools. y_{is} is a student's endline score on a particular exam or exam component. L_s is a vector of indicator variables for the stratification group that a school was in for the public lottery that assigned schools to study arms. This specification differs from 3.1 solely in that it omits y_{is}^{baseline}, the student's baseline score on the test component, from the right-hand side.

The results are presented in Appendix Tables K.1 to K.3, which mirror tables 3.4 to 3.6 in the main text. The point estimates and standard errors are nearly unaffected by the exclusion of the controls. For the EGRA (Appendix Table K.1), including the regression

without baseline test score results yields to slightly larger effect sizes for the Mango Tree-Administered Program and slightly smaller effect sizes for the Government-Administered Program.

For the Oral English Test (Appendix Table K.2)[1] and the Writing Test (Appendix Table K.3)[2], omitting the baseline test score controls leads to marginally smaller estimates of the gains for students in the Mango Tree-Administered variant of the program, and marginally larger estimated losses for students in the Government-Administered version. The exception is the two name-writing components of the Writing Test, for which the students receiving the Government-Administered version of the program showed gains rather than losses. For African Name (Surname) Writing, the estimated effect of the Government-Administered program differs only in the third decimal place. For English Name (Given Name) Writing, the estimated effect is somewhat smaller without controlling for baseline performance.

None of the differences affect the statistical significance of any of the point estimates, nor do they alter any of the conclusions we draw in the main text.

[1] Note that Column 10 is identical between Table 3.5 and Appendix Table K.2; no controls were included for this column in Table 3.5 because this test, which is not a component of the Oral English Examination, was not conducted at baseline.

[2] Column 10 is identical between Table 3.6 and Appendix Table K.3 because Presentation was not one of the scored categories at baseline. Columns 6 (Voice) and 9 (Conventions) are also identical because no pupils received any points for those categories at baseline, so the controls were dropped due to collinearity with the constant term.

Table K.1

Program Impacts on Early Grade Reading Assessment Scores, without Controlling for Baseline Scores
(in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	PCA EGRA Score Index†	Letter Name Knowledge	Initial Sound Recogniton	Familiar Word Recognition	Invented Word Recognition	Oral Reading Fluency	Reading Comprehension
Mango Tree-Administered Program	0.654***	1.043***	0.649***	0.382***	0.233**	0.484***	0.449***
	(0.127)	(0.163)	(0.129)	(0.0909)	(0.0967)	(0.121)	(0.110)
Government-Administered Program	0.110	0.418**	0.0639	-0.0116	0.0206	0.0581	0.0337
	(0.102)	(0.181)	(0.0956)	(0.0742)	(0.0692)	(0.0807)	(0.0837)
Number of Students	1460	1476	1481	1474	1471	1467	1481
Number of Schools	38	38	38	38	38	38	38
Adjusted R-Squared	0.118	0.175	0.0965	0.0559	0.0367	0.0629	0.0509
Control Group Mean§	0.000	5.973	0.616	0.334	0.358	0.611	0.216
Control Group SD§	1.000	9.364	1.920	2.207	2.762	4.163	0.437

Notes: Longitudinal sample includes 1,478 students who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA EGRA Score Index is constructed by normalizing each of the 6 test modules against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); estimated effects are comparable but slightly smaller for an alternative index that uses the unweighted mean across test modules instead.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations except for the index, where they are the normalized values.

Table K.2

Program Impacts on Oral English Test Scores & English Word Recognition, without Controlling for Baseline Scores
(in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCA Oral English Score Index†	Test 1 (Vocab.)	Test 1 (Count)	Test 2a (Vocab.)	Test 2a (Phrase Structure)	Test 2b (Vocab.)	Test 2b (Phrase Structure)	Test 3 (Vocab., Expressive - Objects)	Test 3 (Vocab., Expressive - People)	Recognition of Printed English Words‡
Mango Tree-Administered	0.0677	0.122	-0.133	-0.072	0.016	-0.014	-0.120	0.275**	0.291**	-0.290**
	(0.123)	(0.108)	(0.094)	(0.106)	(0.131)	(0.112)	(0.117)	(0.117)	(0.119)	(0.135)
Government-Administered	-0.133	-0.019	-0.124	-0.040	-0.130	-0.165	-0.223*	-0.040	-0.106	-0.209
	(0.102)	(0.086)	(0.088)	(0.108)	(0.099)	(0.102)	(0.120)	(0.099)	(0.088)	(0.140)
Number of Students	1481	1481	1481	1481	1481	1481	1481	1481	1481	1481
Number of Schools	38	38	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.319	0.155	0.162	0.199	0.178	0.268	0.090	0.230	0.183	0.274
Control Group Mean§	0.000	2.048	0.294	0.501	0.807	1.826	2.092	2.327	1.585	1.792
Control Group SD§	1.000	1.888	0.620	0.911	1.209	1.928	2.217	2.133	1.839	4.184

Notes: Longitudinal sample includes 1,478 students who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA EGRA Score Index is constructed by normalizing each of the 8 test modules against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); estimated are comparable but slightly larger in magnitude for an alternative index that uses the unweighted mean across test modules instead.

‡ Recognition of Printed English Words is not part of the Oral English examination, but it is a skill that is commonly practiced in status quo (i.e. control) schools in the Lango sub-Region. This involves reading a set of 18 printed words from a piece of paper. It is not included in the computation of the overall PCA index in column 1.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations except for the indices, where they are the normalized values.

Table K.3

Program Impacts on Writing Test Scores, without Controlling for Baseline Scores
(in SDs of the Control Group Endline Score Distribution)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCA Writing Score Index†	African Name (Surname) Writing	English Name (Given Name) Writing	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions	Presentation
Mango Tree-Administered Program	0.399**	1.015***	1.230***	0.147	0.442**	0.152	0.128	0.377*	0.221	0.139
	(0.186)	(0.116)	(0.148)	(0.178)	(0.207)	(0.156)	(0.178)	(0.210)	(0.173)	(0.150)
Government-Administered Program	-0.232	0.437***	0.393**	-0.288*	-0.317*	-0.313**	-0.308**	-0.334*	-0.253	-0.330**
	(0.163)	(0.127)	(0.152)	(0.150)	(0.178)	(0.134)	(0.151)	(0.179)	(0.156)	(0.129)
Number of Students	1373	1447	1374	1475	1475	1474	1474	1475	1475	1475
Number of Schools	38	38	38	38	38	38	38	38	38	38
Adjusted R-Squared	0.265	0.193	0.217	0.161	0.304	0.177	0.165	0.300	0.164	0.171
Control Group Mean§	0	0.593	0.350	1.141	1.286	1.164	1.166	1.267	1.116	1.175
Control Group SD§	1	0.685	0.533	0.372	0.594	0.393	0.416	0.590	0.339	0.396

Notes: Longitudinal sample includes 1,478 students who were tested at baseline as well as endline. All regressions control for stratification cell indicators. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA EGRA Score Index is constructed by normalizing each of the 11 test modules against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); with an alternative index that uses the unweighted mean across test modules instead, estimated effects are larger in magnitude and more statistically significant for the Mango Tree-Administered Program and closer to zero for the Government-Administered Program.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations except for the indices, where they are the normalized values.

K.2 Effect of NULP on Writing Scores, Excluding Stratification Cell of School that Completed Writing Test in English

Students from one of the 12 control schools were mistakenly asked to complete their writing tests in English. The name-writing components of the test were unchanged, and the tests were scored using the exact same rubric as the Leblango writing test. However, there is still the potential concern that the tests from this school may not be comparable to those from the other 37 schools. To address this possibility we re-estimate equation 3.1 for the writing test, excluding the stratification cell for the school that completed the test in English. This stratification cell includes one school from each of the other two study arms as well, so dropping the cell yields a reduced sample of 35 schools. Since the random assignment of schools to study arms was conducted within stratification cells, the exogeneity assumption that $MTSchool$ and $GovtSchool$ are independent of ϵ_{is} will also hold for this reduced sample. In the presence of treatment effect heterogeneity, however, we would not expect this sample to produce identical treatment effect estimates even if there were no issues with the control school's tests.

Appendix Table K.4 shows the estimated effects of the two program variants on test scores using the reduced sample described above. Excluding this cell changes the magnitude of the estimated effects, but does not change their sign or affect our interpretation of them. The estimated gains from the Mango Tree-administered version of the program are similar but somewhat larger; the combined PCA index shows a 50% larger increase using the reduced sample. For the Government-administered program, the combined index shows a fairly precise zero change. The improvements in name-writing are similar to the full sample, while the declines in the other exam components are smaller. Nevertheless, two of the seven writing components show statistically-significant decreases in performance, as compared with three for the full sample. Overall, the results are not particularly sensitive to the inclusion of this stratification cell.

Table K.4

Program Impacts on Writing Test Scores, Excluding Stratification Cell for School that Completed Exam in English
(in SDs of the Control Group Endline Score Distribution)

	(1) PCA Writing Score Index†	(2) African Name (Surname) Writing	(3) Name (Given Name) Writing	(4) Ideas	(5) Organization	(6) Voice	(7) Word Choice	(8) Sentence Fluency	(9) Conventions	(10) Presentation
Mango Tree-Administered Program	0.594***	0.933***	1.364***	0.372***	0.701***	0.350***	0.351***	0.638***	0.435***	0.328***
	(0.107)	(0.117)	(0.150)	(0.109)	(0.129)	(0.091)	(0.114)	(0.130)	(0.110)	(0.088)
Government-Administered Program	-0.010	0.473***	0.527***	-0.093	-0.079	-0.130**	-0.107	-0.093	-0.050	-0.155**
	(0.075)	(0.125)	(0.149)	(0.078)	(0.088)	(0.060)	(0.078)	(0.085)	(0.082)	(0.060)
Number of Students	1262	1336	1263	1361	1361	1360	1360	1361	1361	1361
Number of Schools	35	35	35	35	35	35	35	35	35	35
Adjusted R-Squared	0.323	0.234	0.241	0.153	0.319	0.165	0.151	0.302	0.146	0.158
Control Group Mean§	-0.261	0.527	0.274	0.061	0.131	0.084	0.075	0.108	0.037	0.098
Control Group SD§	0.585	0.671	0.486	0.239	0.338	0.278	0.264	0.310	0.190	0.298

Notes: Sample includes 1,478 students who were tested at baseline as well as endline. All regressions control for stratification cell indicators as well as baseline values of the outcome variable, except for "Presentation" (column 10) which was not included in the baseline scores. Heteroskedasticity-robust standard errors, clustered by school, in parentheses; * p<0.05, ** p<0.01, *** p<0.001.

† PCA EGRA Score Index is constructed by normalizing each of the 11 test modules against the control group, then taking the (control-group normalized) first principal component as in Black and Smith (2006); with an alternative index that uses the unweighted mean across test modules instead, estimated effects are larger in magnitude and more statistically significant for the Mango Tree-Administered Program and closer to zero for the Government-Administered Program.

§ Control Group Mean and SD are computed using the endline data for control-group observations in the estimation sample. They represent the raw means and standard deviations except for the indices, where they are the normalized values.

BIBLIOGRAPHY

Ahituv, Avner, V. Joseph Hotz, and Tomas Philipson. 1996. "The responsiveness of the demand for condoms to the local prevalence of AIDS." *The Journal of Human Resources*, 31(4): 869–897.

Angelucci, Manuela, Dean S. Karlan, and Jonathan Zinman. 2013. "Win Some Lose Some? Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco." Center for Global Development Working Papers 330.

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

Attanasio, Orazio. 2009. "Expectations and perceptions in developing countries: Their measurement and their use." *American Economic Review*, 99(2): 87–92.

Auld, M. Christopher. 2006. "Estimating behavioral response to the AIDS epidemic." *Contributions to Economic Analysis & Policy*, 5(1): 12.

Baird, Sarah J., Richard S. Garfein, Craig T. McIntosh, and Berk Özler. 2012. "Effect of a cash transfer programme for schooling on prevalence of HIV and herpes simplex type 2 in Malawi: a cluster randomized trial." *The Lancet*, 1–10.

Banerjee, A. V, and S. Mullainathan. 2013. "The shape of temptation: Implications for the economic lives of the poor." *NBER Working Paper*.

Barnett, Tony, and Piers Blaikie. 1992. *AIDS in Africa: Its present and future impact.* London:The Guilford Press.

Becker, Gary S., and Kevin M. Murphy. 1988. "A theory of rational addiction." *Journal of Political Economy*, 96(4): 675–700.

Beegle, Kathleen, Emanuela Galasso, and Jessica Goldberg. 2014. "The design of public works and the competing goals of investment and food security." Extended Abstract prepared for Population Association of America Annual Meetings.

Belli, Robert F., William L. Shay, and Frank P. Stafford. 2001. "Event history calendars and question list surveys: A direct comparison of interviewing methods." *Public Opinion Quarterly*, 65(1): 45–74.

Bennear, Lori, Alessandro Tarozzi, Alexander Pfaff, Soumya Balasubramanya, Kazi Matin Ahmed, and Alexander van Geen. 2013. "Impact of a randomized controlled trial in arsenic risk communication on household water-source choices in Bangladesh." *Journal of Environmental Economics and Management*, 65(2): 225–240.

Black, Dan A., and Jeffrey A. Smith. 2006. "Estimating the returns to college quality with multiple proxies for quality." *Journal of Labor Economics*, 24(3): 701–728.

Bruhn, Miriam, and David McKenzie. 2009. "In pursuit of balance: Randomization in practice in development field experiments." *American Economic Journal: Applied Economics*, 1(4): 200–232.

Brune, Lasse, Xavier Giné, Jessica Goldberg, and Dean Yang. 2015. "Facilitating Savings for Agriculture: Field Experimental Evidence from Malawi." University of Michigan 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.

Caplin, Andrew. 2003. "Fear as a policy instrument." *Time and decision: Economic and psychological perspectives on intertemporal choice*, 441–458.

Card, David, and Laura Giuliano. 2013. "Does gifted education work? For whom?" Working Paper, University of California, Berkeley.

Chinkhumba, Jobiba, Susan Godlonton, and Rebecca Thornton. 2014. "Demand for medical male circumcision." *American Economic Journal: Applied Economics*, 6(2): 152–177.

Cichocki, Mark. 2014. "HIV Reinfection - Positive Prevention - Reinfection."

Collins, D., J. Morduch, S. Rutherford, and O. Ruthven. 2009. *Portfolios of the Poor, How the World's Poor Live on $2 a Day.* New Jersey: Princeton University Press.

Conroy, Amy A. 2014. "Marital Infidelity and Intimate Partner Violence in Rural Malawi: A Dyadic Investigation." *Archives of Sexual Behavior*, 43(7): 1303–1314.

Deaton, Angus S. 2009. "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development."

Delavande, Adeline. 2014. "Probabilistic Expectations in Developing Countries." *Annual Review of Economics*, 6: 1–20.

Delavande, Adeline, and Hans-Peter Kohler. 2009. "Subjective expectations in the context of HIV/AIDS in Malawi." *Demographic Research*, 20(31): 817–875.

Delavande, Adeline, Xavier Giné, and David McKenzie. 2011. "Measuring subjective expectations in developing countries: A critical review and new evidence." *Journal of Development Economics*, 94(2): 151–163.

de Mel, Suresh, David McKenzie, and Christopher M. Woodruff. 2012. "Business Training and Female Enterprise Start-Up, Growth, and Dynamics: Experimental Evidence from Sri Lanka." Social Science Research Network SSRN Scholarly Paper ID 2161233, Rochester, NY.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2008. "Returns to Capital in Microenterprises: Evidence from a Field Experiment." *The Quarterly Journal of Economics*, 123(4): 1329–1372.

de Walque, Damien, William H. Dow, and Erick Gong. 2014. "Coping with risk : the effects of shocks on reproductive health and transactional sex in rural Tanzania." The World Bank Policy Research Working Paper Series 6751.

Donner, Allan, and Neil Klar. 2000. *Design and analysis of cluster randomization trials in health research.* London Arnold Publishers.

Dubeck, Margaret M., and Amber Gove. 2015. "The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations." *International Journal of Educational Development*.

Dupas, Pascaline. 2011. "Do teenagers respond to HIV risk information? Evidence from a field experiment in Kenya." *American Economic Journal: Applied Economics*, 3(1): 1–34.

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.

Frison, Lars, and Stuart J. Pocock. 1992. "Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design." *Statistics in medicine*, 11(13): 1685–1704.

Fudenberg, Drew, and David K. Levine. 2006. "A dual-self model of impulse control." *American Economic Review*, 96(5): 1449–1476.

Giné, Xavier, Jessica Goldberg, Dan Silverman, and Dean Yang. 2012. "Revising commitments: Field evidence on the adjustment of prior choices." National Bureau of Economic Research Working Paper.

Godlonton, Susan, Alister Munthali, and Rebecca Thornton. 2015. "Responding to Risk: Circumcision, Information, and HIV Prevention." *Review of Economics and Statistics*, in press.

Goldberg, J. 2011. "The lesser of two evils: The roles of social pressure and impatience in consumption decisions." *Department of Economics, University of Michigan. December (mimeograph).*

Gong, Erick. 2015. "HIV Testing & Risky Sexual Behaviour." *The Economic Journal*, 125(582): 32–60.

Grant, Monica J. 2012. "Girls' schooling and the perceived threat of adolescent sexual activity in rural Malawi." *Culture, health & sexuality*, 14(1): 73–86.

Gray, Ron, Victor Ssempiija, James Shelton, David Serwadda, Fred Nalugoda, Joseph Kagaayi, Godfrey Kigozi, and Maria J Wawer. 2011. "The contribution of HIV-discordant relationships to new HIV infections in Rakai, Uganda:." *AIDS*, 25(6): 863–865.

Haushofer, Johannes, and Jeremy Shapiro. 2013. "Household Response to Income Changes: Evidence from an Unconditional Cash Transfer Program in Kenya." Massachussetts Institute of Technology Working Paper, Cambridge, MA.

Heckman, James J. 1996. "Randomization as an Instrumental Variable." *The Review of Economics and Statistics*, 78(2): 336–341.

Hollingsworth, T. Déirdre, Roy M. Anderson, and Christophe Fraser. 2008. "HIV-1 transmission, by stage of infection." *Journal of Infectious Diseases*, 198(5): 687–693.

Hudomiet, Peter, Gábor Kézdi, and Robert J. Willis. 2011. "Stock market crash and expectations of American households." *Journal of Applied Econometrics*, 26(3): 393–415.

Jakiela, Pamela, and Owen Ozier. 2012. "Does Africa Need a Rotten Kin Theorem? Experimental Evidence from Village Economies."

Jones, Gareth, Richard W. Steketee, Robert E. Black, Zulfiqar A. Bhutta, and Saul S. Morris. 2003. "How many child deaths can we prevent this year?" *The Lancet*, 362(9377): 65–71.

JPAL. 2014. "Student Learning | The Abdul Latif Jameel Poverty Action Lab."

Juhn, Chinhui, Sebnem Kalemli-Ozcan, and Belgi Turan. 2009. "HIV and fertility in Africa: First evidence from population based surveys." Institute for the Study of Labor Working Paper Discussion Paper No. 4473.

Kaler, Amy. 2003. ""My girlfriends could fill a yanu-yanu bus": Rural Malawian men's claims about their own serostatus." *Demographic Research*, Special Collection(1).

Kaler, Amy, and Susan Watkins. 2010. "Asking God about the date you will die: HIV testing as a zone of uncertainty in rural Malawi." *Demographic Research*, 23.

Karlan, Dean, Aishwarya Ratan, and Jonathan Zinman. 2014. "Savings by and for the Poor: A Research Review and Agenda." *Review of Income and Wealth*, 60(1): 36–78.

Kenyon, C., R. Colebunders, and N. Hens. 2013. "Determinants of generalized herpes simplex virus-2 epidemics: the role of sexual partner concurrency." *International journal of STD & AIDS*, 24(5): 375–382.

Kerwin, Jason T. 2012. ""Rational fatalism": Non-monotonic choices in response to risks." University of Michigan Working Paper, Ann Arbor.

Kerwin, Jason T., Rebecca L. Thornton, and Sallie M. Foley. 2014. "Prevalence of and Factors Associated with Oral Sex among Rural and Urban Malawian Men." *International Journal of Sexual Health*, 26(1): 66–77.

Kerwin, Jason T., Rebecca L. Thornton, Sallie M. Foley, Jobiba Chinkhumba, and Alinafe Chibwana. 2011. *Situational analysis of sexual behaviors and alternative safer sex strategies in-depth interview dataset.* University of Michigan and University of Malawi College of Medicine.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental analysis of neighborhood effects." *Econometrica*, 75(1): 83–119.

Koopman, James S., Carl P. Simon, and Chris P. Riolo. 2005. "When to Control Endemic Infections by Focusing on High-Risk Groups." *Epidemiology*, 16(5): 621–627.

Kremer, Michael. 1996. "Integrating behavioral choice into epidemiological models of AIDS." *The Quarterly Journal of Economics*, 111(2): 549–573.

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

Lillard, Lee, and Robert Willis. 2001. "Cognition and wealth: The importance of probabilistic thinking." University of Michigan Retirement Research Center Working Paper WP 2001-007.

Loader, Catherine. 2004. "Smoothing: local regression techniques." In *Handbook of Computational Statistics*. 571–596. Springer.

Luke, Nancy, Shelley Clark, and Eliya M. Zulu. 2011. "The Relationship History Calendar: Improving the scope and quality of data on youth sexual behavior." *Demography*, 48(3): 1151–1176.

MacGregor, D.G., P. Slovic, and T. Malmfors. 1999. ""How exposed is exposed enough?" Lay inferences about chemical exposure." *Risk Analysis*, 19(4): 649–659.

Malawi National AIDS Commission. 2003. "National HIV/AIDS policy: A call for renewed action." Malawi National AIDS Commission (NAC), Lilongwe.

Malawi National Statistical Office, and ORC-MACRO. 2010. "Malawi demographic and health Survey 2010."

Manski, Charles F. 2004. "Measuring Expectations." *Econometrica*, 72(5): 1329–1376.

McEwan, Patrick J. 2014. "Improving Learning in Primary Schools of Developing Countries A Meta-Analysis of Randomized Experiments." *Review of Educational Research*, (in press).

McKenzie, David. 2012. "Beyond baseline and follow-up: The case for more T in experiments." *Journal of Development Economics*, 99(2): 210–221.

National Health Service. 2013. "Quitting timeline | Why Quit | NHS SmokeFree."

O'Donoghue, Ted, and Matthew Rabin. 2001. "Risky Behavior among Youths: Some Issues from Behavioral Economics." *NBER Chapters*, 29–68.

Office of the Secretary. 1979. "The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research." National Commission for the Protection of Human Subjects of Biomedical Behavioral Research, Bethesda, MD.

Oster, Emily. 2012. "HIV and sexual behavior change: Why not Africa?" *Journal of Health Economics*, 31(1): 35–49.

Ozdenoren, Emre, Stephen W. Salant, and Dan Silverman. 2012. "Willpower and the Optimal Control of Visceral Urges." *Journal of the European Economic Association*, 10(2): 342–368.

Peltzman, Sam. 1975. "The Effects of Automobile Safety Regulation." *Journal of Political Economy*, 83(4): 677–725.

Philipson, Tomas J., and Richard A. Posner. 1993. *Private choices and public health: The AIDS epidemic in an economic perspective.* Cambridge, MA:Harvard University Press.

Robinson, Peter M. 1988. "Root-N-consistent semiparametric regression." *Econometrica*, 56(4): 931–954.

RTI International. 2009. "Early Grade Reading Assessment Toolkit." World Bank Office of Human Development.

Schatz, Enid. 2005. "'Take your mat and go!': Rural Malawian women's strategies in the HIV/AIDS era." *Culture, health & sexuality*, 7(5): 479–492.

Shapiro, Jesse M. 2005. "Is there a daily discount rate? Evidence from the food stamp nutrition cycle." *Journal of Public Economics*, 89(2–3): 303–325.

Smith, Davey M., Douglas D. Richman, and Susan J. Little. 2005. "HIV Superinfection." *Journal of Infectious Diseases*, 192(3): 438–444.

Stephens Jr., M. 2003. "'3rd of tha Month': Do Social Security Recipients Smooth Consumption Between Checks?" *American Economic Review*, 93(1): 406–422.

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

Stock, James H., and Motohiro Yogo. 2005. "Testing for weak instruments in linear IV regression." In *Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas Rothenberg*. 80–108. Cambridge University Press.

Tavory, Iddo, and Ann Swidler. 2009. "Condom semiotics: Meaning and condom use in rural Malawi." *American Sociological Review*, 74(2): 171.

Thaler, Richard H., and H. M. Shefrin. 1981. "An Economic Theory of Self-Control." *Journal of Political Economy*, 89(2): 392–406.

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

UNESCO. 2011. *World Data on Education.*

Viscusi, W. Kip. 1990. "Do smokers underestimate risks?" *Journal of Political Economy*, 98(6): 1253–1269.

Wawer, Maria J., Ronald H. Gray, Nelson K. Sewankambo, David Serwadda, Xianbin Li, Oliver Laeyendecker, Noah Kiwanuka, Godfrey Kigozi, Mohammed Kiddugavu, Thomas Lutalo, Fred Nalugoda, Fred Wabwire-Mangen, Mary P. Meehan, and Thomas C. Quinn. 2005. "Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda." *The Journal of Infectious Diseases*, 191: 1403–1409.

Webley, K. 2006. "Mother tongue first: Children's right to learn in their own languages." Development Research Reporting Service, UK id21.

Weinstein, Neil D., and William M. Klein. 1996. "Unrealistic optimism: Present and future." *Journal of Social and Clinical Psychology*, 15(1): 1–8.

White, Darcy, and Robert Stephenson. 2014. "Correlates of Perceived HIV Prevalence and Associations with HIV Testing Behavior among MSM in the United States." Population Association of America Annual Meeting 2014 Working Paper, Boston, MA.

Wilson, Nicholas L., Wentao Xiong, and Christine L. Mattson. 2014. "Is sex like driving? HIV prevention and risk compensation." *Journal of Development Economics*, 106: 78–91.

Wiswall, Matthew, and Basit Zafar. 2014. "Determinants of college major choice: Identification using an information experiment." *Review of Economic Studies*, in press.