Essays on Education and Health in Sub-Saharan Africa

by

James Allen IV

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Public Policy and Economics)
in The University of Michigan
2023

Doctoral Committee:

Professor Dean Yang, Chair
Professor Lauren Falcao Bergquist, Yale University
Professor David Lam
Professor Tanya Rosenblat

James Allen IV
alleniv@umich.edu
ORCID ID 0000-0002-3084-7785

This dissertation is dedicated to the people of Simona and Yorosso, Mali,
especially the incredibly patient Dorro Goita and Achata Samaké,
and to the memory of Rubé Dao, Drissa Goita, Khalifa Goita, and Sonata Goita.

ACKNOWLEDGEMENTS

Thank you to my advisors, professors, benefactors, research assistants, university staff, friends and family for their support during my doctoral studies and dissertation research.

First, I am forever grateful to my committee: Dean Yang (chair), Lauren Falcao Bergquist, David Lam and Tanya Rosenblat. Dean Yang has been an exceptional and dependable advisor and mentor, pushing me to go beyond what I thought possible for myself. By bringing me onto his ongoing projects and treating me as true co-author, I was given the opportunity to learn research by doing and consequently have learned more than I had ever dreamed about how to be a development economist. Thank you Dean for investing your time, effort and talent into me on a daily basis. Lauren Falcao Bergquist mentored me as a graduate student instructor and a research assistant and provided great advice on navigating the job market. David Lam took me on as a research assistant in my second year while he was Director of the Institute for Social Research, and yet somehow always made time to teach me about economic demography, research, and fatherhood. Tanya Rosenblat's passion for behavioral economics and doing good also greatly influenced me and, moreover, working with her on my second and third essays (along with co-authors Arlete Mahumane, James Riddell IV, Dean Yang, and Hang Yu) was a joy and a privilege during the straining COVID-19 pandemic. Thank you all for your life-changing guidance!

Second, I appreciate the insightful feedback on my dissertation research from professors around the world. In Michigan's development field sequence, thank you to Hoyt Bleakley for his eye-opening views on education, Sara Heller for her incredible course on randomized controlled trials (and assistance with randomization inference in my dissertation), and Achyuta Adhvaryu for serving as my third-year paper reader. Thank you also to Yuehao Bai, Natalie Bau, Brian Dillon, Esther Duflo, Isaac Mbiti, Yusuf Neggers, Anant Nyshadham, Emma Riley, Bryce Steinberg, Hang Yu, Basit Zafar and job interview committees for proving their insights as well. A special thank you to participants of Michigan's Economic Development Seminar (EDS), Health, History, Development, and Demography (H2D2) Seminar and the University of Cape Town's Southern Africa Labour and Development Research Unit (SALDRU) Seminar for invaluable feedback on what eventually became the essays in this dissertation. Finally, thank you to participants of NEUDC 2022, PacDev 2023, and MWIEDC 2023 for being the best part of the job market–I look forward to returning to each conference.

Third, I am indebted for the support of institutions and individuals who enabled the work in this dissertation. I enormously benefited from a traineeship implemented by Michigan's Population Studies Center (PSC) that gave me a research community and enabled work with my PSC mentor Dean Yang with fellowship support by the National Institute on Aging of the National Institutes of Health (award number T32AG000221). PSC also administered several small grants generously provided by Marshall Weinberg, whose small grant support elevated each essay in my dissertation. For the second and third essays, I am also grateful for support from supported by the Abdul Latif Jameel Poverty Action Lab (J-PAL) Innovation in Government Initiative through a grant from The Effective Altruism Global Health and Development Fund (grant number IGI-1366), the UK Foreign, Commonwealth & Development Office awarded through Innovations for Poverty Action (IPA) Peace & Recovery Program (grant number MIT0019-X9), the Michigan Institute for Teaching and Research in Economics (MITRE) Ulmer Fund (grant number G024289), the United States Agency for International Development (USAID) awarded through the Feed the Future Innovation Lab for Markets, Risk and Resilience (MRR) Innovation Lab (award no. A20-1825-S007), and the National Institutes of Health Eunice Kennedy Shriver National Institute of Child Health & Human Development (award no. 1-R01-HD102382-01A1).

Fourth, I have been fortunate to work with amazing research managers, assistants and administrators. On my first essay, I thank Danielle-Andree Atangana, Noelle Seward and Laston Manja for their amazing help. On the second and third essays, I thank Faustino Lessitala for providing top-notch leadership and management of fieldwork in Mozambique, particularly during the tumultuous COVID-19 pandemic (and for hosting me just before the pandemic!); Patricia Freitag, Ryan McWay, Rita Neves, and Maggie Barnard for fantastic research assistance; and Julie Esch, Laura Kaminski, and Lauren Tingwall for world-class grant management. Finally, I have significantly benefited from Michigan's amazing administrative staff, including the ever-patient PhD Coordinator for the Ford School Kathryn Cardenas and incredibly helpful staff supporting the Economics department.

Fifth, thank you to life-long friends. I was lucky to be in the 2016 Economics PhD cohort, who made first year fun. I am particularly grateful to my prelim study buddies Jon Denton-Schneider and Jennifer Mayo, and to the other members of our Covid pod Emily and Michael (and then baby!) Murto, Ben Pyle, Gillian Wener, and Zach Yetmar. I greatly appreciate the support and friendship from other Michigan development students Alexander Fertig, Maximilian Huppertz, Russell Morton, and Emir Murathanoglu and the incredible up-and-coming PhDs cohorts who are forming the DevLab community. Thank you also to friends from the University of Kentucky, Peace Corps (including those in the dedication), Michigan State University, Alma College and Harbor Springs Michigan, including life-long friend John O'Neill (i.e., @HotStuffMcTottl) for being my most loyal Twitter follower despite having a PhD in Mechanical Engineering.

Saving the best for last, my family. My wife Joye Allen is my rock; I am eternally indebted and overwhelmingly grateful for your unwavering support during these tumultuous seven years. Thank you for everything. To our children Eleanora and Jameson, who made the PhD harder but oh so much better. To my parents Debra and Jim and brother Ric for their three-decades of support of my academic goals, including periodic babysitting so that I can get work done (or take a break). To my incredible in-laws Scott, Cindy, Scott Peter, (the amazing Auntie) Hope, and David Grace for their support of me and my family. We are so very lucky that I got to do my doctoral studies here in Michigan, close to all of you.

Thank you all. I hope my career will prove worthy of your time, effort and sacrifices.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

My dissertation aims to ameliorate global poverty through the study of development economics with a focus on education and health. I identify novel interventions for improving human capital in sub-Saharan Africa, where poverty is most dire, and offer lessons for both scholars and policymakers.

My first essay observes that, across sub-Saharan Africa, countries with a greater percentage of overlapping days in their school and farming calendars also have lower primary school survival rates, as greater overlap between these calendars presumably reduces the time available for both schooling and farm-based child labor. I causally identify such effects by leveraging a four-month shift to the school calendar in Malawi that differentially affected communities based on their pre-policy crop allocations. I find that a 10-day increase in school calendar overlap during peak farming periods decreases school advancement by 0.34 grades (one lost grade for every three children) and decreases the share of children engaged in peak-period household farming by 11 percentage points after four years. Secondary analyses reveal stronger negative schooling impacts for girls and poorer households driven by school's overlap with the labor-intensive sowing period. Policy simulations illustrate that adapting the school calendar to minimize overlap with peak farming periods should increase school participation by better accommodating farm labor demand.

My second essay implements a randomized experiment to promote learning about COVID-19 among Mozambican adults–implemented over the phone during the pandemic lockdown–via a "supply-side" teaching intervention that provided targeted feedback on knowledge questions, a "demand-side" financial incentives intervention, and a joint treatment. The paper sets up a framework for how to evaluate whether supply- and demand-side educational interventions are substitutes or complements by estimating a "complementarity parameter". Between the treatments, we find significantly more complementary than predicted by experts in a forecasting survey, with the joint treatment performing better than the combined effect of each standalone treatment (increasing COVID-19 knowledge by 0.5 standard deviations), and evidence of the complementarity persisting 9 months after the intervention. The paper's framework can also be used to estimate complementary between supply- and demand-side educational interventions in more complex settings.

My third essay tests randomized treatments aimed at accelerating social norms on an emerging preventive health behavior, motivated by early survey work in Mozambique during the COVID-19 pandemic revealing that respondents underestimated community support for social distancing. In theory, updating social norms upwards on a publicly beneficial health behavior can have ambiguous impacts: encouraging free-riding if "everyone else is doing it", or encouraging good behavior if the norm correction also updates the perceived infectiousness of the disease. Indeed, we find that the effect of correcting individuals' underestimates of community support (or affirming accurate estimates) is heterogeneous: decreasing social distancing where COVID-19 cases were low and free-riding dominated, but increasing it where cases were high and the perceived-infectiousness effect dominated. The findings highlight that correcting misperceptions of health behavior norms may have heterogeneous effects depending on disease prevalence–an important lesson for policymakers on how to schedule public health interventions for maximum efficacy.

CHAPTER I

Double-Booked: Effects of Overlap between School and Farming Calendars on Education and Child Labor

"Different constraints are decisive for different situations, but the most fundamental constraint is limited time." -Gary Becker

1.1 Introduction

Understanding how households in low-income settings allocate their children's time between schooling and child labor has vital implications for economic development. Child labor increases household income and consumption in the short run, but can be physically hazardous (Edmonds, 2007; UIS and UNICEF, 2015). Time in school is an investment toward educational attainment, which is linked to higher earnings in adulthood (Duflo, 2001; Psacharopoulos and Patrinos, 2018) and is widely recognized as a key driver of economic growth (Mankiw et al., 1992; World Bank, 2017). Thus these household allocations can heavily impact a child's and even a country's future economic well-being.

In sub-Saharan Africa (SSA), where over half the workforce is employed in agriculture (World Bank, 2021), child labor often takes the form of household agricultural production. Indeed, SSA has the highest prevalence rate of any region for child labor at 27%, with family work accounting for between 40–80% of all out-of-school child laborers (UIS and UNICEF, 2015).[1] Such household-farm work, especially needed during the labor-intensive sowing and harvest periods, often occurs when school is in session. Across SSA, the average school calendar lasts 192 days (excluding weekends and holidays)—over 50% of total days

[1]Notably, this likely underestimates child labor as the UNICEF definition excludes children ages 7–14 years who performed less than 28 hours of household chores or children 12–14 performing less than 14 hours of economic activity in the week prior to the survey.

in the year.[2] When the school calendar overlaps with the farming calendar, schooling and household-farm work compete for children's time. Moreover, SSA school calendars do not change often and remain highly correlated by region and former colonial power, suggesting that they are not already adapted to seasonal farm labor demand. In this context, this paper asks: how does overlap in the school and farming calendars affect children's schooling and child labor outcomes?

Looking across SSA countries, Figure 1.1 presents some correlational evidence that overlap between the school and farming calendars may impede primary school completion. The figure shows a significant negative correlation between a country's survival rate to the fifth grade (on the vertical axis) and the percent of school and sowing/harvest days that overlap (on the horizontal axis). For every additional percentage point increase in overlap, the survival rate is 2.39 percentage points lower, on average. With overlap's percent of the school and farming calendars ranging from 15% to 32% across SSA countries, overlap may help to explain large differences in advancement within primary school.

In theory, greater overlap between the school and farming calendars should indeed reduce schooling investments, and it should reduce child labor too. This is because, when these calendars overlap at time t, a child wanting to do both must choose one activity—schooling or farm work. Therefore, holding fixed the school calendar and farming calendar, the overlap between calendars defines the child's stock of total time available for schooling and household-farm work. When there is no overlap between the calendars, children are available to attend every day of scheduled school and work every day during the sowing and harvest on the household farm. But when there is overlap, a time "budget constraint" limits the time allocation opportunity set. For households with positive allocations to schooling and household-farm work on that frontier, a further increase in overlap forces households to adjust with reductions to both activities (assuming each have positive but diminishing returns to lifetime utility). This makes overlap theoretically unique from interventions that change either the marginal benefit or marginal cost of schooling, including its opportunity cost. Conceptually, overlap does not change the economic returns to school or outside activities at time t, but rather determines the number of time periods during which these returns are competing in the household problem.

To test the theory, I study a four-month shift to the school calendar in Malawi—one of only six major school calendar changes across SSA from 1997-2019. Implemented to better align with the government's, universities' and neighboring countries' calendars, Malawi's school calendar change provides a plausibly exogenous shock to overlap between school and crop calendars. Malawi is a setting in which both schooling and farm work are common

[2] Author's calculations from daily pre-Covid public school calendars for 78% of SSA's 46 countries.

activities. At the time of the school calendar shift, 81.3% of children aged 6–13 years in my sample were enrolled in school, and 78.1% lived in a farming household. In Figure 1.3, I compare the net change in school days resulting from the school calendar shift to the typical farming calendar in Malawi, which shows how the policy transferred school days into the rainy-season planting (i.e., sowing)—a period of high farm labor demand. Thus, at a high level, it appears that the policy increased overlap between the school and farming calendars, reducing the total time available for schooling and household-farm work. Supported by the theory, I hypothesize that this likely caused decreases in time allocation to both economic activities.

To obtain causal estimates, I use Malawi's Integrated Household Panel Survey (IHPS) to compare pre- and post-policy schooling and time use outcomes for children of primary-school age (6-13) in the pre-policy year, who are fully exposed to the shock over the four-year study period. My primary outcomes are completed grade level—a key educational outcome—and an indicator for working on the household farm during sowing and harvest periods. To measure the policy-induced change in overlap between school and farming calendars, I construct a novel shift-share variable. The "shift" is a crop-level shock: the policy-induced change in the number of days a crop's calendar overlaps with the school calendar. The identification assumption is that these crop-level overlap shocks are as-good-as-randomly assigned. This assumption is supported by the non-agricultural rationale for implementing the policy, balance tests of pre-policy characteristics, and a "pretrend test" of a primary outcome. Crop-level shocks are then weighted by a community's exposure to the policy as determined by its pre-policy "share" of labor devoted to producing each crop. I then regress outcomes on shift-share overlap controlling for the sum of community crop shares, among other controls. To assuage recent concerns about over-rejection in shift-share estimation, I perform a randomization inference procedure following Borusyak and Hull (2021) that perturbs the plausibly exogenous shock to overlap between school and crop calendars, while holding fixed the endogenous community crop shares.

Consistent with my theoretical predictions, I find that increases in overlap between the school and farming calendars reduces both schooling and household-farm labor. An increase of 10 days of overlap during the rainy-season sowing and harvest in the average sample community—estimated at a 1.21 standard deviation increase in shift-share overlap—leads to a significant reduction in school advancement by 0.34 grades (one lost grade for every 3 children) and an 11 percent decline in children doing household-farm work during peak periods (a 44% decline from baseline levels) after four years. Secondary analyses reveal stronger negative schooling impacts for girls and poorer households. Channels vary between the two outcomes: school participation is most impacted by sowing-period overlap along its intensive margin (i.e., not via enrollment), while household-farm child labor reacts more to harvest-period overlap along its extensive margin and is then substituted with increased household expenditure on hired labor without affecting farm profits. Further analysis of a sub-sample surveyed 10 years later suggests persistent negative effects on school advancement and incidence of household-farm work.

This paper contributes to the literature on trade-offs between schooling investments and child labor by obtaining causal estimates from a unique natural experiment that negatively shocks both activities equally without changing their nominal returns. Previous studies of positive shocks to expected returns of schooling find they increase schooling—e.g., via correcting mis-perceived returns (Nguyen, 2008; Jensen, 2010), merit-based scholarships (Kremer et al., 2009a), or access to skilled labor opportunities (Jensen and Miller, 2017). Other studies have looked at positive shocks to returns to wage labor and find they reduce schooling—e.g., via new factory openings (Atkin, 2016), booms in gold mining (Santos, 2014), or positive rainfall shocks (Shah and Steinberg, 2017)—or at how improved healthcare access can differentially change returns to both (Adhvaryu and Nyshadham, 2012). Rather, I study a plausibly exogenous decrease to children's stock of total time available for both school and farming that forces households to reduce both activities significantly. First, the magnitude of the effects signify the important role that farm labor demand plays in competing for schooling investments (and vice versa) in low-income agrarian settings. Second, heterogeneity analyses of this trade-off reveal for whom the perceived value of schooling is lower relative to household-farm labor: younger girls compared to younger boys, and children of low-income households compared to middle-income households. I also compare overlap impacts between sowing and harvest periods and find greater inelasticity in labor demand during the sowing period, which is when farm labor demand is most concentrated.

Second, my paper is the first, to my knowledge, to analyze the effects of a plausibly exogenous shock to a household time constraint and conceptualize it as such, providing yet another application of Becker (1962)'s seminal work on incorporating time allocation in household models.[3] Since then, household models typically constrain time allocation at some fixed amount of total time. However, in this paper, time allocation is further restricted by some tighter constraint: the school calendar and the farming calendar each constrain one time allocation—time in school and farm work, respectively—and then the overlap between

[3]Studies that perhaps come close include Gibson and Shrader (2018)'s analysis of sunset time on time allocated for sleep, Montero and Yang (2022)'s look at religious festivals limiting time allocated for agricultural production, and Baker et al. (2008)'s study of universal child care on maternal labor supply via increasing time available for work (if some time in child care is assumed to be exogenous and fixed). However, none of these studies quite frames their mechanisms in these terms. See Heckman (2015) for a brief review of Becker (1962) and its influence.

these calendars jointly constrains these two time allocations. With regards to improving household welfare in the developing world, this is important as previous works have analyzed a multitude of inputs to the household's income budget constraint: household income, costs of schooling, savings, credit, insurance, seasonal liquidity constraints, and the like. My paper suggests that activity-specific time constraints may too limit household decision-making and warrant future research with regards to their impact on education and other topics.

Third, in addition to identifying time constraints as a problem, my paper also suggests a solution: the school calendar itself is a feasible policy lever to increase time available for schooling, which in turn can increase educational attainment. The results validate at least 40 years of descriptive evidence (e.g., Schiefelbein (1987); Admassie (2003)) and recommendations from international educational development practitioners (e.g., Bustillo (1989); Kadzamira and Rose (2003)) about the benefits of adapting the school calendar around agricultural labor demand. They also expand evidence that overlapping end-of-year examinations with the harvest has detrimental effects (Ito and Shonchoy, 2020) by showing negative schooling effects can persist across sowing and harvest periods regardless of timing within the school year. Additionally, school calendars can affect school participation by changing the length of the school year (Watkins, 2000) and making it easier for households to finance school fees (Dillon, 2021). To identify the ideal school calendar, I run a policy simulation that approximates counterfactual effects of other potential school calendars. I find that the pre-policy school calendar was actually ideally situated in that it minimized overlap with the labor-intensive sowing period, and also that overlap falls further when communities adopt their own overlap-minimizing school calendar rather than the one calendar that minimizes overlap across all communities on average.

This paper proceeds as follows. Section 1.2 presents a theoretical framework that conceptualizes overlap as an added time constraint in a household model. Section 1.3 describes the setting, Malawi's school calendar change, and data on individuals and crop calendars. Section 1.4 details the estimation strategy, including construction of the shift-share measure, the randomization inference technique, and falsification tests. Section 1.5 presents the primary and secondary analyses. Section 1.6 describes policy impacts and insights from simulating alternative school calendars. Section 1.7 concludes.

1.2 Theoretical Model

I model the school and farming calendars and the overlap between them as additional constraints on time that limit the amount of available time that a child can spend on either school or farm work. Using a simple household model, I describe under what conditions an increase in overlap (holding fixed the length of each calendar) could be expected to decrease allocations to both time in school and farming.

1.2.1 Overlap in a Household Model

Consider a household consisting of *adults* and a *child*.

Child's Time Endowments. A child has a total time endowment T which is spent on schooling s, farm work h, and leisure ℓ such that:

$$s + h + \ell = T \tag{1.2.1}$$

The school year has $S \in (0, T)$ days and the farming season has $H \in (0, T)$ days. There is also an overlap between school and farming days captured by Θ: the number of days in which school is scheduled during the farming season. On a day when the school year and the farming season overlap, a child can either attend school or farm, but not both. Therefore, an "overlap constraint" limits the amount of available time that a child can spend on either school or farm work, as follows:

$$s + h \leq S + H - \Theta \tag{1.2.2}$$

The following also holds:

$$\begin{aligned} s &\leq S \\ h &\leq H \\ \Theta &\leq min(S, H) \\ S + H - \Theta &< T \end{aligned} \tag{1.2.3}$$

The first two conditions specify that the child will attend at most S days to school and spend at most H days on the farm (though not both given Equation 1.2.2 unless $\Theta = 0$). Further, overlap Θ in schooling and farming days can be no more than the minimum value between S and H. Finally, the total number of school and farming days accounting for overlap is less than total time T, meaning that there days in which leisure ℓ is the only available activity.

Household Utility. Household income consists of adult income n and child income $w \cdot h$ where w is the child's wage. The household will spend income on consumption c.

$$c \leq w \cdot h + n \tag{1.2.4}$$

The household maximizes a utility function which is additive in the utility from consumption, schooling and child leisure, respectively:

$$U(c,s,l) = \underbrace{u_C(c)}_{\substack{\text{consumption}\\ \text{utility}}} + \underbrace{u_S(s)}_{\substack{\text{schooling}\\ \text{utility}}} + \underbrace{u_L(\ell)}_{\substack{\text{leisure}\\ \text{utility}}} \tag{1.2.5}$$

Each of the three sub-utilities is monotonic, concave and has infinite derivative at 0, which ensures an interior solution for the household's maximization problem.

I make the following two assumptions to simplify the analysis, and then discuss the implications of relaxing each assumption following the main theoretical result:

Assumption 1. *The household always prefers schooling and farming to leisure at the minimum allocation to leisure:*

$$\begin{aligned} u_S'(S) &> u_L'(T-S-H+\Theta) \\ wu_C'(w\cdot H+n) &> u_L'(T-S-H+\Theta) \end{aligned} \tag{1.2.6}$$

The maximum amount of time the child can spend in school and on the farm is $S+H-\Theta$. Therefore, $T-S-H+\Theta$ is time that a child must spend on leisure ℓ. Assumption 1 says that at this minimum level of leisure, the household prefers the child to spend additional time either in school or on the farm (earning wage income) rather than on even more leisure, and thus $s+h=S+H-\Theta$. This is a reasonable assumption because, in many contexts, the minimum level of leisure is still a significant amount of time. Given Assumption 1, the only conflict in time allocation arises between schooling and farm labor: the household will allocate $S+-\Theta$ days to farming and schooling and the remaining $T-S-H+\Theta$ to leisure (which cannot be further reduced).

Assumption 2. *The household always prefers some positive level of both schooling and farming:*

$$\begin{aligned} u_S'(S) &< w\cdot u_C'(w(H-\Theta)+n) \\ w\cdot u_C'(w+n) &< u_S'(S-\Theta) \end{aligned} \tag{1.2.7}$$

Given that households allocate $S+H-\Theta$ days to farming and schooling (Assumption 1), a household can either choose to maximize school by devoting S to schooling and $H-\Theta$ to farm labor; maximize farm work by devoting H to farm work and $S-\Theta$ to schooling; or somewhere in-between by allocating $s<S$ to schooling and $h<H$ to farm work such that $s+h=S+H-\Theta$. Assumption 2 ensures that the latter scenario prevails by stating that

the household prefer farm work at the maximum level of schooling S, and prefer schooling at the maximum level of farm work H (via earning wages for consumption).

Given Assumption 2, the household has an *interior* maximizing solution s and h (i.e., S and H are non-binding). This can be solved by taking the first-order condition of the household problem with respect to schooling: $u'_S(s) = w \cdot u'_C(w \cdot h + n)$.[4] This says that households optimize schooling and farm work where the marginal gains from schooling s equal the marginal gains of new consumption purchased with wage w from household-farm work h. The household will send the child to school for $S - \Theta < s < S$ days and will let them farm on $H - \Theta < h < H$ days such that the first-order condition is satisfied.

1.2.2 Analysis of an Increase in Overlap

I am mainly interested in comparative statics on the parameter Θ which is under the control of the policymaker who determines the placement of the school calendar relative to the known farming season.

Proposition 1. *Suppose the policy-maker increases overlap Θ in the school and farming calendar holding fixed the number of days in the farming calendar H and school calendar S. Then households respond by decreasing their child's time allocation to both schooling s and farming h.*

Proof. Recall that household preferences for a child's time spent schooling s and farming h are monotonic, concave and have infinite derivatives at 0. By Assumption 1, $s + h = S + H - \Theta$. Given that S and H are fixed, an increase in Θ must decrease $s + h$ by the same magnitude, which requires that at least s or h to decrease. Suppose a household decreases schooling from s to $\hat{s}$ to satisfy the new constraint. Concave preferences increase the marginal utility of schooling so that now it is greater than the marginal utility of farm work: $u'_S(\hat{s}) > w \cdot u'_C(w \cdot h + n)$. By Assumption 2, households optimize allocations to schooling and farming where $u'_S(s) = w \cdot u'_C(w \cdot h + n)$. So, if $u'_S(\hat{s}) > w \cdot u'_C(w \cdot h + n)$, then households would allocate time toward schooling s reducing its marginal utility, and time away from h raising its marginal utility, until both Assumption 1 and Assumption 2 hold. Therefore, an increase in Θ will reduce allocations to both s and h. □

Relaxing Assumption 2 allows households to either allocate time in school at its maximum S or time in farming at its maximum H. In this case, an increase to overlap from Θ_0 to Θ_1 does one of two things: 1) if a household still prefers the maximum allocation of either

[4] To calculate, I substitute the overlap constraint into the budget constraint via h, then substitute the combined constraint into the total utility function, and take its partial derivative with respect to s.

schooling or farming at Θ_1, then it will only reduce its allocation to the other activity; or 2) if a household now prefers an interior solution at Θ_1, then it will no longer maximize either time in school or farming and will instead reduce its allocation to both activities.

Relaxing Assumptions 1 and 2 allows households to allocate additional time to leisure ℓ greater than $T - S - H + \Theta$. This means that households can choose time allocations to schooling, farm work and leisure that fall below the schooling-farming time allocation frontier such that $s + h < S + H - \Theta$. Again, an increase to overlap from Θ_0 to Θ_1 does one of two things: 1) if optimal leisure remains greater than minimum leisure at Θ_1 (i.e., $\ell^* > T - S - H + \Theta_1$), then the "overlap constraint" remains non-binding and the optimal allocation to schooling, farm work and leisure will remain the same; or 2) if optimal leisure is greater than minimum leisure at Θ_0 but is less than or equal to minimum leisure at Θ_1 (i.e., $T - S - H + \Theta_0 < \ell* \leq T - S - H + \Theta_1$), then the new "overlap constraint" is binding (i.e., $s + h = S + H - \Theta_1$) and households will have to reduce allocations to schooling and farm work in the same manner shown above but with smaller expected reductions since the initial allocation fell below the schooling-farming time allocation frontier.

1.2.3 Graphical Description of Overlap

Figure 1.2 summarizes the main takeaways of the model. The figure depicts the time allocation trade-off as a budget constraint diagram, where the area from the origin to the frontier represents a child's time allocation opportunity set in household-farm work h and schooling s given the overlap constriant $s + h \leq S + H - \Theta$, while leisure $\ell = T - (s + h)$ represents an unseen third dimension. On the horizontal axis, the school-farming time frontier begins at H, the total number of days in the farming calendar, and continues linearly upwards. On the vertical axis, the school-farming time frontier begins at S, the total number of days in the school calendar, and continues extends horizontally rightwards. If there is no overlap in the school and farming calendars (i.e., $\Theta = 0$), then the time opportunity set extends to the point (H,S). However, when overlap in the calendars exists—e.g., at Θ_0—an "overlap line" with a slope of -1 cuts into the opportunity set, limiting possible allocations of schooling and farm work (and leisure). Further, when overlap increases from Θ_0 to Θ_1, the "overlap line" draws closer to the origin.

Household preferences for s and h (and ℓ) are represented by the convex indifference curves U_0 and U_1. Since both s and h provide positive diminishing returns to marginal utility, households have strictly convex preferences for s and h characterized by a diminishing marginal rate of substitution between the two inputs. Assumption 1 ensures that households will choose allocations on the school-farming time frontier where the overlap constraint is binding—i.e., $s+h = S+H-\Theta$. Assumption 2 ensures that households will choose allocations on the "overlap line" where the marginal utility of time in school is equal to the marginal utility of time spent farming, and not "at the kinks" where either s is bound by S or h is bound by H.

When overlap increases from Θ_0 to Θ_1, households must reallocate h and s to the newly constrained time frontier. As an example, a household may have to reallocate from point A to point B, decreasing utility from U_0 to U_1. The preferences are also such that households will choose to reduce allocations to both h and s when overlap increases, though the magnitudes of the reductions will depend on the shape of the marginal utility curves.

1.3 Data

To test the model's predictions, I require data to represent children's time in school s and on the household farm h that are plausibly affected by an exogenous shock to overlap between the school and farming calendars. I look in Malawi, where a four-month shift to the school calendar between 2009 and 2011 coincided with the first wave of the World Bank's Integrated Household Panel Survey (IHPS). In addition to outcome data, I also require measures of community exposure to the school calendar policy change, which I construct using data on community crop shares and the overlap between the school calendar and crop-specific calendars on sowing and harvest periods.

In this section, I first describe this study's setting and then the school calendar change. Next, I describe the individual- and household-level data taken from the IHPS. Finally, I describe the crop calendars that collectively comprise the farming calendar in the overlap measure.

1.3.1 Setting

Malawi is a landlocked country in southeast Africa with an estimated population of 13 million people as of 2008 (World Bank (2010) for all of Section 1.3.1). At the time of the first IHPS in 2010, 63% lived on less than US$2 a day, and 82% of the population lived in rural areas where most engaged in subsistence, smallholder, rain-fed agriculture. Youth malnutrition was estimated at 49%, and adult literacy was 69%. In 2008, 37% of the population were children ages 5-16—the highest in Southern Africa. Relative to other countries in sub-Saharan Africa (SSA), Malawi had the fifth lowest GDP per capita.

Malawi's formal education system consists of 8 grades (or "Standards") of primary school, 4 grades (or "Forms") of secondary school, and 4 years at the university level. While private schools exist, 99% of students in primary and 77% in secondary attend public institutions. Between 2000 and 2010, school enrollment increased but fell slightly relative to population growth. School enrollment is heavily concentrated in primary school, particularly grades 1-4, due to high rates of initial enrollment into primary school and grade repetition.

High initial enrollment and late entry in primary school are often attributed to Malawi's Free Primary Education (FPE) program in 1994 that abolished tuition fees for primary school. As a result, households in 2007 paid less than US$2 on average annually per student in primary school, and primary school expenditures only represented 17% of households' total education budgets.[5] Thus, primary school fees do not majorly prohibit enrollment.

Additionally, grade repetition increased between 1999 and 2006, reaching 20% in primary education. A 2004 survey found that 47% of grade 1 students were repeating it, with repeaters of other primary school grades ranging from 13% to 30%. High repetition rates worsen student-teacher ratios, schooling costs and dropout rates. As such, dropout is common, and only 35% of primary school students complete grade 8. School principals reported family responsibilities as the main reason for dropout (44% for boys and 41% for girls) in a 2007 survey, with marriage, pregnancy, and employment stated as other reasons.

1.3.2 School Calendar Shift

School calendars in sub-Saharan Africa (SSA) do not change often and are highly correlated with former colonial power. According to UNESCO, Institute for Statistics (2022) (UIS) data, in 45 SSA countries from 1997-2019 (over 900 country-years), there are only 6 instances of countries permanently shifting the start or end months of their school calendar by 2 months or more: Angola and Ghana both lengthened their calendar, and Malawi, Rwanda, South Sudan and Tanzania have shifted their calendar once. Of the latter group, I examine Malawi's school calendar change due to the availability of a detailed record of the school calendar change and contemporaneous household panel data.

In Malawi, the school calendar shifted from an early January start date to an early September start date over the course of two school years. The 2009 school calendar started in early January and ended in mid-to-late November. Then, the 2010 school calendar served as a transition year, starting school one month prior in early December 2009 and ending in early August 2010. Finally, the 2010/11 school calendar began the new schedule, starting school an extra three months prior in early September and ending in early-to-mid July.[6] In this

[5]While the FPE program is often credited with a 51% increase in enrollment (e.g., Kattan and Burnett (2004)), the 1994 educational reform package also included Malawi's first school calendar change that effectively *reduced* overlap between school and periods of peak farm labor demand. Thus, my results suggest that some of the FPE enrollment boost may have been due to a reduction in overlap with the farming calendar as well as the reduction in its direct costs.

[6]Interestingly, this change actually reversed Malawi's first school calendar change in 1994 (prior to the earliest year in the UIS dataset), moving the start date from early September to early January, as part of a large education reform package that also abolished fees for primary school. By contrast, the 2009-11 school

study I am primarily interested in comparing the change in overlap between the pre-policy 2009 school calendar and the post-policy 2011 school calendar.

As shown in Figure 1.3, this school calendar change reduced school days in July and August and increased school days in November and December, largely due to the shift in the end-of-year school break. The exact dates of these school calendars, including breaks, are well documented by Frye (2011). Using this record, I construct $SchoolCalendar_{d,t}$, an indicator for if school was scheduled for day $d \in [1, 365]$, where $d = 1$ is January 1st and $d = 365$ is December 31st, in time period t, either 2009 (pre-policy) or 2011 (post-policy).

Reasons for the school calendar change as reported by local and district officials (Frye, 2011) include: 1) alignment with the university calendar in Malawi; 2) alignment with the university and international school calendar of neighboring countries; 3) closer alignment with the government fiscal calendar, which begins in July; 4) reversing the policy of the previous administration for politically symbolic reasons; and 5) alignment with the initiation schedule for the Yao people, a Bantu ethnic and linguistic group. Reasons (1)-(4) build confidence in the argument that the school calendar change was made without consideration of the farming calendar. Reason (5) may be correlated with the Yao farming calendar; however, in Table 1.1, I do not find any correlation between shift-share overlap and the pre-policy share of Yao households in the community—only 11% of households, on average, across communities–though I include it as a control in my main specification nonetheless to err on the side of caution.

1.3.3 Panel Survey Data

Outcome data and household agricultural data come from Malawi's Integrated Household Panel Survey (IHPS) 2010-2013. This section summarizes key variable attributes while additional details are provided in Appendix A.2.

Outcome data include highest grade level completed and variables describing household-farm work during peak periods. First, $Grade_i$ is the highest grade level completed for individual i in the referenced academic year. Additionally, I have measures of enrollment and extended absences that I use to examine channels affecting overlap's effect on grade level. Second, $Farmed_i$ is an indicator equal to one if individual i worked on household plots during peak periods in the past year, and zero otherwise, and $Farm\ Hours_i$ is the corresponding number of hours worked, both of which are reported in the household agricultural survey. Peak periods are defined as rainy-season sowing and harvest periods, when 93.6% of the average household's cultivated acres are under production. Table 1.1 presents summary statistics of baseline values of these outcomes for the sample of interest.

calendar change was not accompanied with other well-documented educational reforms.

Household agricultural data from the IHPS include plot-level production and member-specific time use data, both of which serve as inputs to calculating community crop shares. For agricultural production, I use pre-policy levels of cultivated acres of each crop type recalled for the rainy season 2008/09, dry season 2009, and permanent crops. With member-specific time use data, I estimate off-farm labor as the sum of annual hours of formal work and ganyu (day) labor and on-farm labor via plot-specific reports on household labor, both of which are aggregated to the community level.

As controls in the regression, I use individual sex and age, household size, and a household asset index as the first principal component of a vector of indicator variables for ownership of 12 assets.[7] Additionally, from a community-level survey, I use estimates of a community's share of Yao households (given possible heterogeneity discussed in the previous section) and an indicator for if the community experienced a drought in the prior five years.

1.3.4 Crop Calendars

Data on community crop production are matched to crop calendars from the Food and Agriculture Organization (FAO) Crop Calendar Tool, which provides start and end months for the sowing and harvest periods for 45 major crops in Malawi, which match to 83% of pre-policy cultivated acres in the IHPS data. For remaining 17% of cultivated acres, I use the modal sowing month and harvest month reported by households in the first IHPS. Additional details are provided in Appendix A.2. Using these data, I generate $CropCalendar_{d,c}$, an indicator for if crop c is either being sown or harvested on day $d \in [1, 365]$, where $d = 1$ is January 1st and $d = 365$ is December 31st.

1.4 Empirical Approach

1.4.1 Constructing Overlap

I construct overlap as a shift-share variable that captures a community's exposure to crop-level shocks of changes to overlap between the school and crop calendars. First, I estimate the crop-level shock as the "shift" (or change) in the number of days during which both school is scheduled and crop c is being sown or harvested (times of peak farm labor demand). Then, I estimate the "share" of annual labor devoted to producing crop c in the community, which measures the community's exposure to crop c's overlap shock. Crop-level

[7] Following Yang et al. (2021), the 12 assets are car, motorcycle, bicycle, radio, television, sewing machine, refrigerator, iron, bed, table, clock, and solar panel. Missing are two assets not reported in the data: freezer and mobile phone.

shocks weighted by the community crop shares are then summed across all crops to collectively capture the community's exposure to the set of crop-level overlap shocks.

I estimate the crop-level shock using the indicator variables characterizing school and crop calendars defined in the previous section. I define $overlap_{c,t}$ as the number of days in year t in which both school is scheduled and crop c is being sown or harvested. Formally,

$$overlap_{c,year} = \sum_{d=1}^{365} (CropCalendar_{d,c} * SchoolCalendar_{d,t}) \tag{1.4.1}$$

where time $year$ takes on the values of the pre-policy year 2009 or the post-policy year 2011. I then estimate the shock change in overlap $\Delta overlap_{c,2011-2009}$ as the difference in $overlap_{c,year}$ for crop c in 2011 relative to 2009. Across the 135 unique crops in the data, $\Delta overlap_{c,2011-2009}$ has a mean of 1.7 days, standard deviation of 10.6, minimum of -27, and maximum of 21.

I estimate the community's exposure to crop c's overlap shock via the community crop share $share_{c,\ell}$, which captures the relative importance of crop c in community ℓ based on the labor and land resources devoted to it. Specifically, I estimate $share_{c,\ell}$ using IHPS data on on-farm vs off-farm labor and pre-policy levels of cultivated acres by crop by season, as shown in Equation 1.4.2. Defining a community by each 16-household enumeration area in the IHPS, I start by calculating the on-farm share of community ℓ's total annual hours worked as $farmshare_\ell$ and the share of cultivated acres devoted to crop c in community ℓ as the variable $acreshare_{c,\ell}$. Next, I estimate the community ℓ's share of crop c as the product of these two shares for each crop:

$$\begin{aligned} farmshare_\ell &= hours\_onfarm_{c,\ell}/(hours\_onfarm_{c,\ell} + hours\_offfarm_{c,\ell}) \\ acreshare_{c,\ell} &= acres_{c,\ell}/\sum_C acres_{c,\ell} \\ share_{c,\ell} &= farmshare_\ell * acreshare_{c,\ell} \end{aligned} \tag{1.4.2}$$

All shares are $\in [0,1]$. Note that the sum of $share_{c,\ell}$ across all crops is equal to $farmshare_\ell$ since the sum of $acreshare_{c,\ell}$ across all crops is equal to 1.

Putting these together, I estimate shift-share overlap for each community by weighting the crop-level shock by community crop share $share_{c,\ell} \in [0,1]$ and summing across crops:

$$ssoverlap_\ell = \sum_c (share_{c,\ell} * \Delta overlap_{c,2011-2009}) \tag{1.4.3}$$

Finally, I normalize $ssoverlap_\ell$ to unit variance across all IHPS communities surveyed in 2010. Across the 135 communities in the sample, this variable *before normalizing* has a mean of 8.0, standard deviation of 4.5, minimum of -3.9, and maximum of 17.8, and *after normalizing* has a mean of 1.8, standard deviation of 1.0, minimum of -0.9, and maximum of 4.0. To put $ssoverlap_\ell$ into perspective, I use a simulation to estimate that a 1.21 standard deviation increase in normalized $ssoverlap_\ell$ is roughly equivalent to adding 10 days of overlap during sowing and harvest of rainy-season maize in the average sample community.[8]

Moreover, for secondary analyses, I construct versions of shift-share overlap that refer to specific periods of the farming and school calendar. First, I distinguish between sowing and harvest periods in $CropCalendar_{d,c}$ and use the process described above to construct measures of shift-share overlap during sowing $ssoverlap\_sow_\ell$ and harvest $ssoverlap\_harv_\ell$, which together sum to $ssoverlap_\ell$. Second, I identify the first and last four weeks of school in $SchoolCalendar_{d,t}$ to construct measures of shift-share overlap between peak farming periods and the first month of school $ssoverlap\_admis_\ell$ and last month of school $ssoverlap\_exams_\ell$, which represent the important admissions and exam periods. All new shift-share overlap regressors are normalized to unit variance of $ssoverlap_\ell$ to have comparable coefficients.

1.4.2 Sample

I define my sample as individuals surveyed about their schooling outcomes for the pre-policy school year 2009 when they were 6-13 years old. In the IHPS 2010, this includes households that were interviewed prior to the last scheduled day of the 2010 school year (i.e., August 7, 2010), so that schooling responses about the previous academic year refer to the pre-policy 2009 school year. The age range of 6-13 years old ensures that the sample is fully exposed to the shock over the four-year study period. In Malawi, children typically start school at age six and, if they complete one grade each school year, can complete primary school by age 13 and secondary school by age 17. Thus, from 2009 to when they are ages 10-17 in 2013, the sample are eligible school-aged children—the population of interest for this study.[9]

This yields a sample of 2,287 individuals at baseline, 2,142 of which are still present in the IHPS 2013, producing a sample retention rate of 94% that Table 1.1 shows is not correlated with shift-share overlap.

[8]See Appendix A.3.1 for details on the simulation used to make this comparison.

[9]This age range is also supported by pre-policy data of enrollment by age showing over half of children being enrolled in school from ages 6 to 17 but not otherwise.

1.4.3 Regression

1.4.3.1 Primary Analysis

I seek to estimate the causal effect β of $ssoverlap_\ell$ in a linear model of:

$$Y_{i,\ell} = \alpha + \beta ssoverlap_\ell + \delta Y_{base,i,\ell} + \rho farmshare_\ell + w_{i,\ell}{}'\gamma + \epsilon_{i,\ell} \tag{1.4.4}$$

where $Y_{i,\ell}$ is the outcome for individual i in community ℓ; $ssoverlap_\ell$ is the shift-share regressor calculated in Equation 1.4.3; $Y_{base,i,\ell}$ is the baseline value of the outcome; $farmshare_\ell$ is the on-farm share of total annual hours worked (i.e., the sum of crop shares); $w_{i,\ell}$ is a vector of controls defined below; and $\epsilon_{i,\ell}$ is an error term.

The control $farmshare_\ell$ represents the sum of community crop shares across crops, as recommended by Borusyak et al. (2022) in the case of incomplete shares. It ensures that the regression compares individuals from communities with comparable shares of on-farm labor while estimating the effect of $ssoverlap_\ell$.

The vector of pre-policy controls $w_{i,\ell}$ includes the location-level controls and crop-level controls. Location-level controls include individual sex and age; household size and asset index; an indicator for three sample communities containing no farming households, the community's share of Yao households, and an indicator for if the community experienced a drought in the prior five years. Crop-level controls include: seasonal dummies (rainy, dry, permanent) and a dummy for grain crops as share-weighted location-level variables (as in Borusyak et al. (2022)) and altitude zone. Crop season and altitude are included as controls as both are correlated with crop calendar length and thus crop-level overlap as well, while the grain dummy captures the community share of maize—Malawi's dominant cash and food crop. Finally, out of caution and to control for baseline imbalance in the $Farm\ Hours_i$ (described in Section 1.4.5), I include the baseline values of all three main outcomes $Grade_i$, $Farmed_i$ and $Farm\ Hours_i$ in each regression of post-policy outcomes.

Given the theoretical model's predictions that additional overlap in the school and farming calendars reduces both time in school and hours worked on the household farm, I hypothesize that $\beta < 0$ for the main outcomes $Grade_i$, $Farmed_i$ and $Farm\ Hours_i$.

1.4.3.2 Period-Specific Analysis

Most secondary analyses will use the primary specification in Equation 1.4.4 but with alternative subsamples or outcomes. However, one secondary analysis will regress different specifications to estimate the causal effect $ssoverlap_\ell$ at specific periods in the farming and school calendar. To compare overlap between the sowing and harvest periods, I regress:

$$Y_{i,\ell} = \alpha + \beta_1 ssoverlap\_sow_\ell + \beta_2 ssoverlap\_harv_\ell + X'_{i,\ell}\phi + \epsilon_{i,\ell} \tag{1.4.5}$$

where $ssoverlap\_sow_\ell$ and $ssoverlap\_harv_\ell$ represent shift-share overlap during the sowing and harvest periods, respectively, and controls $X_{i,\ell} = \delta Y_{base,i,\ell} + \rho farmshare_\ell + w_{i,\ell}{}'\gamma$ are the same as those described in Equation 1.4.4. Here, I expect the impact of overlap in both the sowing period β_1 and harvest period β_2 to be negative, but do not make predictions about their relative effect size. Note that because $ssoverlap\_sow_\ell$ and $ssoverlap\_harv_\ell$ sum to $ssoverlap_\ell$, the latter does not serve a purpose in the regression.

Moreover, to measure overlap's effects during critical periods in the school year, I regress:

$$Y_{i,\ell} = \alpha + \beta_1 ssoverlap\_\ell + \beta_2 ssoverlap\_admis_\ell + \beta_3 ssoverlap\_exams_\ell + X'_{i,\ell}\phi + \epsilon_{i,\ell} \tag{1.4.6}$$

where $ssoverlap\_admis_\ell$ and $ssoverlap\_exams_\ell$ represent shift-share overlap during the admissions and exam periods, respectively, and other variables are as described in Equations 1.4.4. Here, coefficients on the new regressors β_2 and β_3 are interpreted as interaction terms (i.e., overlap's effect during the first month of school is $\beta_1 + \beta_2$) and test how overlap during the first and last month of school differ from the rest of the school year. I do not make a formal prediction on β_2 and β_3. If β_2 and β_3 estimates have different signs across the two primary outcomes, then this reveals that overlap during this period affects the trade-off between schooling and household-farm work. Meanwhile, if β_2 and β_3 estimates are both positive (negative) across the primary outcomes, then an interpretation is that overlap during this period is a less (more) binding constraint on time available for both activities.

Finally, to check the robustness of the secondary results to the full specification, I regress:

$$\begin{aligned} Y_{i,\ell} = \alpha + &\beta_1 ssoverlap\_sow_\ell + \beta_2 ssoverlap\_harv_\ell + \\ &\beta_3 ssoverlap\_admis_\ell + \beta_4 ssoverlap\_exams_\ell + X'_{i,\ell}\phi + \epsilon_{i,\ell} \end{aligned} \tag{1.4.7}$$

where variables are defined as described above.

1.4.4 Inference

Conventional standard errors in shift-share regressions do not account for unobserved correlation between observations with similar exposure shares, and tend to over-reject when this correlation is positive (Adão et al., 2019; Borusyak et al., 2022). Therefore, I test the hypotheses by employing a randomization inference (RI) procedure that compares my actual effects to those estimated from counterfactuals of shift-share overlap. Following insights from Borusyak and Hull (2021), I specify a shock assignment process aligning with my assumption that the crop-level change-to-overlap shocks are plausibly exogenous to my outcomes, while holding fixed the community crop shares that are likely endogenous. By using shock counterfactuals to simulate an empirical distribution of test statistics, I can test the sharp null hypothesis that shift-share overlap has no effect for any observation by checking if actual test statistics are in the tails of the null distribution (Fisher, 1935). This RI test remains valid despite the high concentration of rainy-season maize in the data that could hinder asymptotic approximation.

The RI procedure is as follows. First, I randomly draw (with replacement) crop-level shocks $\Delta overlap_{c,2011-2009}$ from its actual distribution. Second, I weight the redrawn shocks by actual crop shares $share_{c,\ell}$, and sum across crops to generate a counterfactual shift-share overlap measure for each location (as in Equation 1.4.3). Third, I run the regression specified in Equation 1.4.4, replacing only the actual shift-share overlap measure with the counterfactual one, and collect the counterfactual β, called $\tilde{\beta}$. Then, I repeat these three steps 1000 times. Finally, I calculate Fisher exact p-values as the fraction of $\tilde{\beta}$ for which $|\tilde{\beta}| \geq |\hat{\beta}|$.

In Appendix A.4, I describe alternative, and in some cases more conservative, inference procedures that I use to test the robustness of my results. Alternative RI procedures include drawing shocks from a normal distribution defined by the first and second moment of the actual overlap distribution, redrawing shocks with replacement within season, and estimating counterfactual shocks based on simulated school calendar changes. I also run a share-weighted shock-level regression that produces exposure-robust standard errors (Borusyak et al., 2022). Further, I detail how the high concentration of rainy-season maize in my data possibly violates a key assumption for their procedure, which explains why I chose the RI procedure for inference ex ante. As it turns out, the exposure-robust p-values are comparable to those estimated from RI.

1.4.5 Falsification Tests

To evaluate the key identifying assumption that shocks are quasi-randomly assigned, I implement falsification tests outlined in Borusyak et al. (2022). First, I conduct traditional "balance tests" on baseline values of key outcomes and location-level characteristics that proxy for the unobserved residual evaluated via the RI procedure. Second, I conduct a district-level "pre-trend" analyses on the schooling outcome: highest completed grade level. Collectively, results support the assumption that shift-share overlap is pseudo-randomly assigned to observations in Equation 1.4.4.

First, I check for balance in the retention between the eligible 2010 IHPS sample and the 2013 IHPS sample to rule out sample selection issues confounding the analysis. Then, I check for balance on the baseline values of $Grade_i$, $Farmed_i$ and $Farm\ Hours_i$ and location-level controls. Balance regressions are estimated on the following:

$$Y_{base,i,\ell} = \alpha + \beta ssoverlap_\ell + \rho farmshare_\ell + w_{i,\ell}{}'\gamma + \epsilon_{i,\ell} \tag{1.4.8}$$

where $Y_{base,i,\ell}$ is the outcome and the control vector does not include baseline values of $Grade_i$, $Farmed_i$ and $Farm\ Hours_i$ (as this was motivated by the balance test), but all other regressors from Equation 1.4.4 are included except when serving themselves as the dependent variable.

Table 1.1 presents summary statistics and balance test results. Retention between the 2010 and 2013 IHPS sample is 94% and is balanced with respect to shift-share overlap, as are most other variables. However, there are two incidences of imbalance: First, a one standard deviation increase in shift-share overlap is correlated with a 3.17 hours *increase* in the baseline measure of $Farm\ Hours_i$. This imbalance must lie along the intensive margin of $Farm\ Hours_i$, given that its extensive margin $Farmed_i$ appears balanced. While the imbalance in the opposite direction of the hypothesized effect, this motivates prioritizing $Farmed_i$ as a measure of household-farm work and also controlling for baseline values of $Farm\ Hours_i$ as well as $Grade_i$ and $Farmed_i$ in the main specification. Second, communities with greater exposure to the school calendar change were *less likely* to experience a drought in the prior five years, possibly making these communities relatively better off in the pre-policy period. However, when controlled for, the table shows that the baseline values of the $Grade_i$ and $Farmed_i$ are balanced, alleviating concern of these imbalances driving the results for these outcomes. Also, in regards to the Yao ethnic group's initiation schedule as a possible reason for the school calendar change, note that community share of Yao households is balanced across the sample.

Additionally, Table 1.2 presents a district-level "pre-trend" analysis of the change in highest completed grade level between the pre-policy data from 2009 and those collected in Malawi's 2004 Integrated Household Survey (IHS2). Without individual panel data before 2009, I examine pre-trends by averaging the individual cross-sectional data from 2004 and 2009 to the district level, and then matching observations across 26 (of 28) formal districts plus 4 distinct urban areas. Outcomes include $\Delta Grade$ 2004-2009: the change in the district's average grade completion between 2004 and 2009. For comparison, I also examine $\Delta Grade$ 2009-2013: the change in the district's average grade completion between 2009 and 2013. Unfortunately the IHS2 agricultural modules do not collect time use data for household individuals, so I cannot examine household-farm work. Adapting main specification Equation 1.4.4 into a district-level regression, I regress outcomes on district-averaged $ssoverlap_\ell$ (called $ssoverlap_d$)

and district-averaged controls and then infer using a RI procedure with district-averaged counterfactual shift-share overlap measures. For robustness, I study two district-averaged samples: the 6–13 age group comparing the study sample's age range across surveys; and a cohort of school-aged individuals who are 6–9 years old in 2004, 11–14 in 2009 and 15–18 in 2013.

Regressions of $\Delta Grade$ 2004-2009 in Table 1.2 do not detect a significant correlation with $ssoverlap_d$, suggesting "parallel pre-trends" between different levels of $ssoverlap_d$ in the main specification. Moreover, regressions of $\Delta Grade$ 2009-2013 estimate significant negative effects in the "Age 6–13" regressions (RI p-value = 0.026) and marginally significant negative effects in the "Cohort" comparison (RI p-value = .108), suggesting that some of the hypothesized negative effects of overlap on highest completed grade level are even detectable at the district level.

1.5 Results

1.5.1 Primary Results

Table 1.3 presents results testing the paper's primary hypotheses with conventional standard errors in parentheses and randomization inference (RI) p-values in square brackets. I estimate a statistically significant negative effect of shift-share overlap on both $Grade_i$ and $Farmed_i$. These results support the model's predictions that the added time constraint imposed by additional overlap between the school and farming calendars decreases both time in school and work on the household farm. In Table 1.4, I show that these results are also robust to other and often more conservative inference procedures.[10]

These negative effects are substantial across the four year period. Recall from Section 1.4.1 that a 10-day increase in overlap during the rainy-season sowing and harvest in the sample-average community increases shift-share overlap by 1.21 standard deviations. Then, multiplying 1.21 by the coefficients in Table 1.3, I estimate that a 10-day increase in overlap during peak production decreases $Grade_i$ by 0.34—equivalent to one grade lost for every three children. Additionally, a 10-day increase in overlap during peak production decreases $Farmed_i$ by 11 percentage points. With the share of the sample engaged in household-farm work increasing by 22 percentage points over the four-year panel (as the children age and physically mature), this represents a sizeable 50% reduction in children working on their household farm.

In Table 1.3 column (3), I also show that shift-share overlap has a significant negative

[10]Alternative inference procedures detailed in Appendix A.4.

effect on overall household-farm hours worked during peak periods.[11] While this effect is driven by changes along the extensive margin in column (2), it is useful to estimate the overall time lost to household-farm work *on average*. The results imply that a 10-day increase in overlap during peak production decreases $Farm\ Hours_i$ by 6.26 hours on average, which, to put into perspective, is almost one school day for the median grade level in primary school.[12]

1.5.1.1 Comparison of Primary Effects

Taken together, the results on $Grade_i$ and $Farm\ Hours_i$ suggest that households appear to make larger reductions to their children's time in school relative to time on the household farm, despite both allocations facing the same constraint. To explore this comparison further, this section uses the theoretical model to discuss possible effects of overlap on time in school and then performs some "back-of-the-envelope" calculations to obtain first-order approximations of households' perceived value of time in school.

First, while data limitations prevent the testing of overlap's direct effect on time in school (e.g., via daily attendance), I can use my theoretical model to infer some likely possibilities. Recall Assumption 1 of the theoretical model: households always prefer schooling and farming to leisure given some number of calendar days in which only leisure (and neither school nor farming) is available. If Assumption 1 holds, then a child's time allocation to schooling and farming is constrained by overlap both before and after the policy change; hence a 10-day increase to overlap must decrease the total allocation to school and farming by exactly 10 days. Then, given the empirical finding that a 10-day increase in overlap during peak production decreases household-farm work by about 1 day annually, the same 10-day increase in overlap must decrease time in school by 9 days annually (almost two weeks of school). However, if Assumption 1 does not hold, then overlap's effect on time in school may be smaller, though not so much smaller that it cannot feasibly explain the observed reductions in highest completed grade level. Given this, it seems likely that overlap has a larger negative effect on time in school than on time in household-farm work, since a 1-day decrease in schooling is likely too small to explain the 0.34 reduction in grade advancement. If so, this suggests that household demand for child labor during peak farming periods is relatively more inelastic than demand for time in school, on average, though pinning down precise estimates for overlap's effect on time in school must be left to future research.

Second, I approximate the perceived value of time in school for the average sample household, following calculations detailed in Appendix A.5. As a starting point, I assume that

[11] Alternative inference procedures in Table 1.4 reveal p-values ranging from 0.054 to 0.163, with most following below 0.10.

[12] Generally, Malawian public primary school starts at 7:30 for all standards and ends after 4.5 hours for Standards 1-2, 6.5 hours for Standards 3-4, and 7 hours for Standards 5-8.

the average overlap-constrained household will reallocate time to each activity until the point where the expected utility lost from decreases in schooling is equal to the expected utility lost from decreases in farm work (given monotonic and concave preferences for schooling and farm work, as assumed in the model). If so, then the results suggest that, on the margin, households equate 0.34 grades to 3 days of household-farm work (i.e., one day annually for three post-policy years) in terms of their expected impact on household utility. In my reference estimate, I multiply the 3 lost days of household-farm work by the average under-15 daily wage for hired farm labor, and divide by 0.34 to estimate the perceived value per completed grade. I also obtain an upper-bound estimate intended to approximate a more generous perceived valuation of time in school. Compared to the reference estimate, the upper bound estimate assumes 1) that the positive baseline imbalance is muting the same magnitude of negative effect in the main analysis, and 2) the average *adult* daily wage for hired farm labor. The reference and upper-bound estimates of the perceived value of a completed grade of school for the average sample household are \$21 and \$33 in 2009 USD, respectively.

How do these estimates of the perceived value of one completed grade level compare to potential returns to education? Consider Montenegro and Patrinos (2014)'s "Mincerian" estimates for the average rate of return for another year of schooling in Malawi of 5.2% in 2004 and 9.8% in 2010. In this range, the annual return for another year of schooling for an non-educated worker is between \$3.75–\$7.07 USD (would be higher for educated workers), which—if added to annual income for all working-age years—has an approximate present discounted value to an individual of between \$94–\$179 USD.[13] Even comparing the upper-bound perceived value of one completed grade level of \$33 to the more conservative estimated lifetime return of \$94, the household's revealed valuation of an additional year of school is only about one-third of its potential contribution to the child's lifetime income. This finding is surprising but also consistent with the literature on households underestimating the returns to education (Jensen, 2010; Nguyen, 2008) and specifically Dizon-Ross (2019)'s 2012 estimates of Malawian households' perceived returns to secondary school relative to primary-school earnings of 3.2% (SD of 3.8%). Other possible explanations include households strategically under-investing in schooling to benefit the household (Jensen and Miller, 2017), or simply underestimating the negative effect of lost time in school on grade advancement. Regardless of the reason, this "back-of-the-envelope" comparison highlights how households underestimate the value of time in school when making marginal decisions about how to best allocate their child's time under constraints.

[13] See Appendix A.5 for calculation details.

1.5.2 Secondary Results

Secondary analyses further characterize the significant negative effects of overlap between the school and farming calendars. In Section 1.5.2.1, I start by comparing testing for heterogeneity by a child's age, sex and household wealth, focusing on the primary outcomes with the clearest $Grade_i$ and $Farmed_i$. In Section 1.5.2.2, I examine the impacts of overlap during different farming and schooling periods and schooling impacts by grade level. In Section 1.5.2.3, I analyze possible channels for the school effect. In Section 1.5.2.4, I assess other adjustments to child's time allocation and household farm decisions. Finally, in Section 1.5.2.5, I present a useful but somewhat limited long-run analysis.

1.5.2.1 Heterogeneity

In this section, I test for heterogeneity in the primary analysis by individual sex and age and by household asset wealth. Doing so reveals who is most affected by overlap between the school and farming calendars: girls, younger boys, and households in the lower two-thirds of asset wealth.

Table 1.5 examines heterogeneity in the main results by individual's sex and age. Results of the main specification are presented for subsamples determined by an individual's reported sex and age in 2009. In columns (1)-(4), results for individuals age 6-9 in 2009 (pre-policy) who become age 10-13 in 2013 (when outcome data were collected post-policy), with distinct results for boys and girls. In columns (5)-(8), results for individuals age 10-13 in 2009 who become age 14-17 in 2013, with results for boys and girls.

First, looking at results for the younger age 6-9 group, Table 1.5 columns (1)-(4) reveal that shift-share overlap has a significant negative effect on both younger boys' and girls' highest grade level completed, but only affects boys' likelihood of working on the household farm despite similar levels at baseline. Further, if we accept the point estimates and assume that boys and girls at this age are equally constrained by overlap, then the results suggest that the school-farming trade-off is different by sex: in response to an overlap shock, girls see a larger reduction in grade level advancement presumably because they make larger reductions to time in school,[14] while boys see a larger reduction in household-farm work. However, I make this interpretation cautiously given that point estimates under each outcome are not significantly different between boys and girls.

Second, looking at results for the older age 10-13 group, Table 1.5 columns (5)-(8) reveal that shift-share overlap maintains a significant negative effect on girls but not boys. For

[14] This interpretation does require the additional assumption that the function of time in school to highest completed grade level is the same for boys and girls.

girls, point estimates on grade level are very near to those estimated in the younger age group while estimates on the likelihood of farming are much larger. For boys, I estimate insignificant estimates in both measures. A possible theory-driven explanation may be that a large proportion of boys in this age group do not allocate time near the school-farming time frontier, either because they only do one of these two activities (i.e., schooling or farming but both) or allocate a relatively large amount of time to other activities including leisure or other forms of work.

Next, Table 1.6 examines heterogeneity in the main results by household asset wealth. Results of the main specification are presented for subsamples of individuals belonging to the bottom, middle or top tercile of their households' pre-policy asset index.[15]

First, overlap's effects appear weaker for wealthier households, as coefficients corresponding to the top tercile of household asset wealth are much smaller in magnitude and generate weaker RI p-values than in other sub-samples. This would be the case if wealthier households are less vulnerable to time constraints on the joint allocation to school and household-farm work—for example, if wealthier households are less likely to have their children work on a household farm. Indeed, 27% of top-terciles households do not even have a farm, compared to 9% of households in the lower two terciles.

Second, I compare the relative trade-off between overlap-induced declines in school and farming between the lower two terciles. While the poorest households in the bottom tercile make larger reductions in schooling, households in the middle tercile make larger reductions to household farming, consistent with the idea that poorest households are most reliant on child labor. However, I make this interpretation cautiously given these point estimates are not significantly different for each other.

1.5.2.2 Period-Specific Effects

To test which calendar periods are most sensitive to overlap, Table 1.7 presents variations of the main specification with regressors representing overlap during different farming and school periods. The results suggest that labor demand is relatively inelastic during the labor-intensive sowing period, leading to greater reductions in schooling and smaller reductions in farming.

First, columns (1) and (2) estimate how the impact of overlap varies between sowing and harvest periods. Column (1) reveals negative effects on grade completion for both sowing and harvest overlap (though only the coefficient on sowing overlap is significant with an RI p-value of 0.056), with sowing overlap having a larger coefficient but not significantly so. Column (2) reveals negative point estimates on the indicator for working on the household-farm, where

[15] Recall that the asset index is estimated as the first principal component of a vector of indicator variables for ownership of 12 assets in 2010.

the coefficient on harvest overlap is a significantly larger (p-value = 0.014) and the only one that is significantly different from zero. Taken together, the results suggest that labor demand is relatively inelastic during the sowing period, with households less willing to reduce their children's household-farm work as compared to the harvest period at the expense of time in school.

One possible explanation is that the rainy-season sowing period is generally more concentrated in the calendar year, meaning that sowing is a period of peak labor demand even relative to the harvest. Indeed, in Appendix A.3, I find that a 10-day increase during rainy-season sowing increases shift-share overlap by 1.35 standard deviations, while a 10-day increase during rainy-season harvest increases it by only 1.07 standard deviations—the difference caused by the fact that sowing periods across all crops are more concentrated from mid-November through December, whereas harvest periods across all crops vary from February through July depending on the length of the crop's growth cycle. Comparing the results through this framing only exacerbates sowing overlap's negative effect on schooling: a 10-day increase in sowing overlap decreases $Grade_i$ by 0.49 (i.e., almost one lost grade for every two children).

Second, columns (3) and (4) estimate how the impact of overlap during school's admissions and exam periods (i.e., the first and last four weeks of school, respectively) varies from the remainder of the school year. Results of admissions-period overlap have near-zero point estimates and are not statistically significant, suggesting overlap's effect does not vary during this time. Results of exam-period overlap on $Grade_i$ are also insignificant, perhaps because most grade levels (particularly in primary school) do not have qualifying examinations to advance, and thus households may not perceive the value of time in school any differently during the exam period. However, exam-period overlap has a significant positive effect on the household-farm work indicator, countering the negative effect of overall overlap but not enough to make the coefficient statistically different from zero.

Finally, columns (5) and (6) present results for the full specification. Only minor changes to the point estimates and statistical significance builds confidence in the robustness of the prior results. The most notable change is column (5)'s significant difference (at the 90% confidence level) between sowing overlap's larger negative effect and harvest overlap's diminished negative effect on $Grade_i$, further emphasizing the conclusion that sowing overlap is more detrimental to grade advancement while harvest overlap has larger impacts on household-farm labor.

1.5.2.3 Channels of Schooling Effect

In this section, I test for possible channels for overlap's negative effect on schooling. To start, I analyze enrollment outcomes and find that overlap primarily affects grade advancement among those who remain enrolled in school. Then, to understand which grades in the school cycle are most affected by overlap, I estimate the effect of $ssoverlap_{\ell}$ on grade-specific outcomes. The results show that grade completion is affected through grades 1–6 while grade survival (i.e., the likelihood of reaching a grade conditional on having started school) is largely impacted in grades 2–7.

In Table 1.8, I test different ways in which overlap may reduce schooling. First, I test for overlap's effect on a cascading set of outcomes that represent three possible extensive margins for schooling. Column (1) is whether an individual started school by 2013 (i.e., initial enrollment). For those who had started school by 2013, column (2) is whether an individual was enrolled in school in 2013. And for those who were enrolled in 2013, column (3) is whether an individual is reported by the household respondent to have missed more than two consecutive weeks of school in the previous school year (i.e., extended absence). Effects of overlap on these outcomes are statistically insignificant, suggesting that these extensive margins are not driving overlap's effect on grade advancement.

Second, I test for overlap's effect on the intensive margin of schooling by testing on a sub-sample that was least likely to be affected by these extensive margin channels. Table 1.8 column (4) regresses $Grade_i$ on shift-share overlap, as in Table 1.3, but only for the sub-sample of individuals enrolled in school in both 2009 and 2013 and not reported to miss more than two consecutive weeks of school in 2013. Even after ruling out these possible extensive margins, the negative effect on $Grade_i$ is estimated with an RI p-value of 0.052, suggesting that overlap's effect predominately occurs at the intensive margin. One may speculate about whether this intensive margin affect is due to reduced school attendance for shorter durations, less time spent after school on studying or homework, or a reduced ability to learn caused by physical labor (e.g., exhaustion or injury); however, data limitations push testing of these intensive margin mechanisms beyond the scope of the study.

To present results on grade-specific outcomes, Figure 1.4 plots the $\hat{\beta}$ coefficient on $ssoverlap_{\ell}$ (on the vertical axis) from regressing the main specification on a vector of indicators corresponding to grades 1 through 8 (on the horizontal axis). Panel (a) plots the effect of completing grades 1 through 8 by 2013. Panel (b) plots the effect of "surviving" to grades 1 through 8 by 2013, which refers to reaching the grade conditional on having started school. Each point estimate is accompanied by error bars of the 95% confidence interval from conventional standard errors and RI p-values in adjacent boxes. I focus on the RI p-values when discussing significance below.

Panel (a) reveals that the negative effect of overlap hinders grade completion throughout most of primary school, with negative point estimates for all grades, statistically significant at the 90% confidence level for grades 1–2 and 4–6. The largest negative effect is estimated for grade 2: a one standard deviation increase in shift-share overlap decreases the probability of an individual completing grade 2 by 5.0 percentage points. Grade 4 completion is also notably affected with a one standard deviation increase in shift-share overlap decreasing the outcome by 4.4 percentage points. As grade 4 is the final grade that a new student could potentially pass in the four-year period between surveys, the estimate suggests that overlap's negative effect on time in school is most concentrated in first half of primary school.

Panel (b) reveals that overlap also negatively affects primary school survival rates, with the negative point estimates for all grades (except grade 1, which is automatically reached conditional on having started school). Moreover, estimates are statistically significant at the 90% confidence level for grades 2–7. The largest negative effect is estimated for grade 3: a one standard deviation increase in shift-share overlap decreases the probability of an individual who started school reaching grade 3 by 4.6 percentage points. Survival to grade 5 is also substantially impacted with a one standard deviation increase in shift-share overlap decreasing the outcome by 4.4 percentage points, representing a substantial decline in the survival rate to mid-level grades in primary school as well. The fact that panel (b)'s findings on reaching a grade roughly correspond to results for completing the previous grade in panel (a) is additional evidence that overlap does not seem to affect enrollment decisions.

1.5.2.4 Other Adjustments

Next, I examine if households made other adjustments in response to the school calendar change that may attenuate the primary results. First, I look at whether overlap affects children's time allocation to work activities that may complement or substitute rainy-season farm labor, for which there is no evidence. Then I check if households made adjustments to the household farm in order to reduce demand for or substitute away from child labor, which finds increased expenditures to hired labor and seeds but not enough to affect total farm costs or profits.

Table 1.9 tests for overlap's effect on whether or not a child was engaged in other forms of child labor in the prior 12 months: dry-season household-farm work, tending to household livestock, working as a day laborer (i.e., ganyu), and working unpaid for another household.[16] However, note that incidence of all activities falls below 5% at baseline and at or below 10% post-policy, so the lack of 12-month recall data on common alternative activities may be a

[16] Formal work is excluded as it employs less than 0.15% of the sample post-policy.

limitation of the analysis. No significant effects are identified, suggesting that school-farming calendar overlap does not have any observable effects on other types of child labor.

Table 1.10 regresses household-level farm outcomes on a household-averaged version of Equation 1.4.4 for all 1,174 households containing sample individuals. Dependent variables include the number of acres cultivated on the household farm throughout the year; major cost categories: labor hired from outside the household, expenses on seeds, and expenses on fertilizer, pesticide and herbicide; total farm costs including smaller expenses like land rent, coupon purchases, and transportation; total farm revenue including crop sales, land rent or coupon sales; and total farm profits. All outcomes are winsorized at their 95th percentile due to positive outliers, and farm profits is also winsorized at its 5th percentile due to negative outliers.

The results in Table 1.10 provide evidence suggesting that households respond to the additional constraint on their children's time by *substituting away* from child labor. A one standard deviation increase in shift-share overlap leads households to purchase significantly more hired labor and seeds—both increases worth over 70% of average household's expenses on these categories at baseline. By comparison, there is no evidence that households reduce their demand for child labor by reducing their cultivated acres. Moreover, these adjustments appear to result in positive but statistically insignificant increases to total farm costs, farm revenues and consequently farm profits. That overlap's reductions in child labor do not negatively affect farm profits is an important finding, suggesting that the average household either places so little value of time in school that they sometimes give up schooling for child labor despite its zero returns to farm profits, or that households overestimate their child's productivity prior to tighter time constraint enacted by the policy.

1.5.2.5 Long-Run Analyses

To analyze the potential long-run effects of overlap on schooling outcomes, I use data from the 2016 and 2019 wave of Malawi's Integrated Household Panel Survey (IHPS). Unfortunately, due to a reduced target sample size and additional participant attrition, 2016 and 2019 data are only available for 44% and 39% of my original sample, respectively, and retention from the 2010 IHPS is weakly positively correlated with shift-share overlap suggesting slight over-sampling from locations that experienced greater overlap due to the school calendar change.

To address both issues, I broaden my sample to include those ages 0-5 pre-policy who become school-aged in later years (ages 7-12 in 2016 and 10-15 in 2019). This sample includes 1,918 and 1,714 individuals in 2016 and 2016, respectively, and retention of the sample that includes this group is slightly less imbalanced at the 90% confidence level in 2016 and producing an RI p-value of 0.141 in 2019.[17] Still, I interpret the results with caution.

I look at the paper's two main outcomes in the long-run analysis, $Grade_i$ and $Farmed_i$, regressing the 2016 and 2019 values of these outcomes using Equation 1.4.4. Table 1.11 presents results of this analysis. The results suggest persistent negative effects on grade completion through 2016 and on household-farm labor. In columns (1) and (2), a one standard deviation increase in shift-share overlap is shown to decrease $Grade_i$ by 0.38 and 0.13 in 2016 and 2019, respectively, though only the 2016 coefficient is statistically significant with an RI p-value of 0.014. The pattern of the results suggest that overlap's negative effect persists for at least three additional years but then may diminish as individuals eventually "catch up" to their desired level of schooling. In columns (3) and (4), overlap's significant negative effects on $Farmed_i$ in 2016 and 2019 are statistically similar in magnitude than in the primary analysis suggesting a persistent effect of overlap on reducing incidence of working the household farm.

In sub-Appendix A.6.2, I show these results broken down by age groups: ages 0-5 in 2009 (pre-policy) who become ages 7-12 in 2016 and 10-15 in 2019; ages 6-9 in 2009 who become ages 13-16 in 2016 and 16-19 in 2019; and 10-13 in 2009 who become ages 17-20 in 2016 and 20-23 in 2019. For $Grade_i$, the age breakdown suggests persistent negative effects on grade level across age groups through 2016, but then effects diminish and become statistically significant in 2019 with a near-zero coefficient in the middle age group (aged 13-16 in 2019) and even a positive point estimate in the upper age group (aged 20-23 in 2019). For $Farmed_i$, the age breakdown suggests larger declines in household farming for younger age groups, with a one standard deviation increase in shift-share overlap inducing significant decreases in 2019 household farming by 21.8 percentage points for the youngest age group (aged 10-15 in 2019) and a 12.0 percentage points for the middle age group (aged 16-19 in 2019).

While they should be interpreted with care due to the limitations outlined above, the results suggest that changes to overlap between the school and farming calendars are potentially persistent in the medium-run and may have long-run repercussions as well that warrant future research.

1.6 Policy Implications

Given the primary results, how potentially harmful to educational attainment was Malawi's policy to shift the school calendar vis-à-vis increasing overlap?[18] I estimate that the policy

[17] Appendix section A.6.1 provides more details on the 2016 and 2019 IHPS dataset.

[18] To be clear, I do not assign blame on the Government of Malawi (GoM) for two reasons: 1) these results were obviously unknown; and 2) there are other potentially beneficial reasons for realignment of a school calendar, which are not captured in this analysis. Rather, I commend the GoM for their ongoing support of household data collection that make this research possible.

increased shift-share overlap by 1.8 standard deviations for the average community in the sample. Hence, I estimate that the policy reduced schooling after four years by an average of 0.50 grades per youth in my sample (roughly one grade lost for every two children). So while there are potentially beneficial reasons for changing the school calendar, this paper's results suggest that the policy potentially had detrimental effects as well.

So now what? In this section, I aim to address the next question a Malawian policymaker might ask: "What school calendar then is best for minimizing overlap with the farming calendar?". Then I discuss more general implications for sub-Saharan African (and other developing) settings similar to Malawi.

1.6.1 Simulating Alternative School Calendars

To identify Malawi's overlap-minimizing school calendar, I simulate 52 other potential school calendars that could have been used for the 2011 school year and estimate their effects on shift-share overlap and grade advancement. The simulation's main findings are summarized in Figure 1.5, which depicts the simulated policy impacts of alternative 2011 school calendars relative to the actual 2009 school calendar for the average sample community. The 52 simulated calendars are denoted by the week in which they start, where week 1 begins the first week of January and week 52 begins in the last week of December. Each calendar's counterfactual change in shift-share overlap (measured in standard deviations of $ssoverlap_\ell$) is measured using drop lines via the left vertical scale, while the projected effects on $Grade_i$ (approximated as a linear extrapolation of the primary results) are measured using bars via the right vertical scale.[19]

Starting with Panel (a), Figure 1.5 shows the counterfactual measure of $ssoverlap_\ell$ and the corresponding projected effects on $Grade_i$ given the causal results in Table 1.3 column (1). The simulation reveals that a Malawian school calendar starting in January is typically best to minimize overlap with the farming calendar. Assuming the structure of the 2011 calendar, a start date of January 10, 2011 (Week 2) would have minimized overlap. Overall, this suggests that Malawi's pre-policy school calendar was ideally situated.[20] Indeed, assuming the structure of the 2009 calendar, a start date of Monday, December 29, 2008 would have minimized overlap, though the actual start date of January 3, 2009 (one week later) was second best. Moreover, the post-policy 2011 school calendar that started in September was far from the worst calendar that could have been chosen. In terms of maximizing overlap

[19] Additional details on the simulation and projected effects on $Farmed_i$ are depicted in Appendix A.3.2.

[20] Given this, one might have expected the simulated change in shift-share overlap to be near zero in early January instead of negative, as seen here. This is primarily because the 2011 school year had 5 fewer days of school (one less week) than the 2009 school year, which is just over a one standard deviation change in shift-share overlap if occurring during a peak farming period.

with the farming calendar, March and October would be the worst times to start the school calendar, increasing shift-share overlap by an additional 78% and 63% (Week 13 and 40, respectively) more than the actual 2011 school calendar did.

In Panel (b), Figure 1.5 shows counterfactual measures of $ssoverlap\_sow_\ell$ and $ssoverlap\_harv_\ell$ and the aggregated projected effects on $Grade_i$, which was calculated by first estimating separate sowing and harvest effects given results from Table 1.7 column (1), and then summing them together. The simulation reveals that, when it comes to minimizing overlap's harmful effect on grade level, there are more gains by minimizing school calendar overlap with the sowing period relative to the harvest period. Note that in early January—the start date for the "ideal school calendar"—sowing overlap is at its minimum and harvest overlap is near its maximum, and the projected effect size is positive. On the contrary, in early July when sowing overlap is at its maximum and harvest overlap is near its minimum, the projected effect size is still solidly negative. Then, to minimize sowing overlap, policymakers should attempt to align school's end-of-year break (typically 1–2 months long) with the sowing season. For school calendars starting in January, the end-of-year break falls during November and December—the designated sowing season for most rainy season crops. Thus, school calendars starting in January free up students during their end-of-year break for the most concentrated period of peak labor demand. Overlap with the harvest period is also important, but harvest dates for rainy season crops are spread out from March through June, so labor demand is less concentrated.

Furthermore, deeper analysis of the simulation shows that the ideal school calendar varies by community because crop bundles (and hence the "farming calendar") vary by community. For the 2011 school calendar, while a January 10th start date would have minimized overlap across all communities on average, it actually turns out that the optimal start date is January 17th for 44% of communities, January 10th for 29% of communities, and other start dates for the remaining 27%. This suggests that while there may be an optimal school calendar for Malawi *on average*, there are potential gains to granting communities some flexibility in setting the school calendar. The simulation reveals that shift-share overlap improves (falls) by an additional 12% when communities adopt their own overlap-minimizing school calendar rather than the one calendar that minimizes overlap across all communities *on average*.

1.6.2 Implications for Sub-Saharan Africa

For educational policymakers in sub-Saharan Africa (SSA) and similar settings, the primary results provide evidence of a clear causal link between policies that constrain the time available for schooling and household-farm work and reductions in time spent on these activities. To reinforce this point, I identify a Malawi-based causal estimate to compare with the SSA cross-sectional analysis. Recall that Figure 1.1 shows across SSA countries that a one percentage point increase in the overlapping percent of school and sowing/harvest days is significantly correlated with a 2.39 percentage point decrease in a country's survival rate to grade 5 (with a standard error of 0.53). By comparison, in Figure 1.4, Malawi-based causal estimates on survival rate to grade 5 suggest that a one percentage point increase in the percent of overlapping school and sowing/harvest farm days should decrease the survival rate to grade 5 by approximately 1.72 percentage points.[21] The similarity in the effect's magnitude and significance between the cross-sectional and causal estimation strategies suggest that Figure 1.1's correlation likely captures some causality. With overlap's percent of the school and farming calendars ranging from 15% to 32% across SSA countries, Figure 1.1 suggests that overlap may indeed help explain large differences in survival rates within primary school. For countries with similarly large shares of children participating in school and household agriculture, such time constraints likely apply as well.

Further, for policymakers outside of Malawi, the policy simulation's findings suggest scheduling school outside the labor-intensive sowing and harvest periods if they wish to foster more school participation (while also factoring in the net effects of additional household-farm work on child welfare). One approach is to shift the start of the existing school calendar to best align school breaks, especially between end-of-year exams and the start of the next school year, with periods of peak farm labor demand like the sowing period.[22] Another approach is to allow for more flexibility and adaptation in the school calendar at the local level—for example, by allowing local school boards to declare up to two weeks of school holidays that can be made up at the end of the year. Analysis of such decentralized school scheduling policies is left as an avenue for future research.

1.7 Conclusion

This paper analyzes a plausibly exogenous change to overlap between the school and farming calendars in Malawi that constrained total time available for schooling and household-farm work. Starting in 2009, Malawi shifted its school calendar by four months, effectively moving school days to a time of higher farm labor demand. Using household panel data from 2009/10 and 2013, I estimate the policy's impact by comparing outcomes between

[21] Figure 1.4 shows that the effect of a one standard deviation increase in shift-share overlap on survival to grade 5 is -0.044. To translate into comparable terms, I first multiply the coefficient by 1.21 to obtain the effect of a 10-day increase to overlap (see Section 1.4.1). Then, given that a 10-day increase in overlap between the school and farming calendars (holding the length of these calendars fixed) corresponds to a 3.1 percentage point increase in the percent of overlapping days, I divide the coefficient by 3.1.

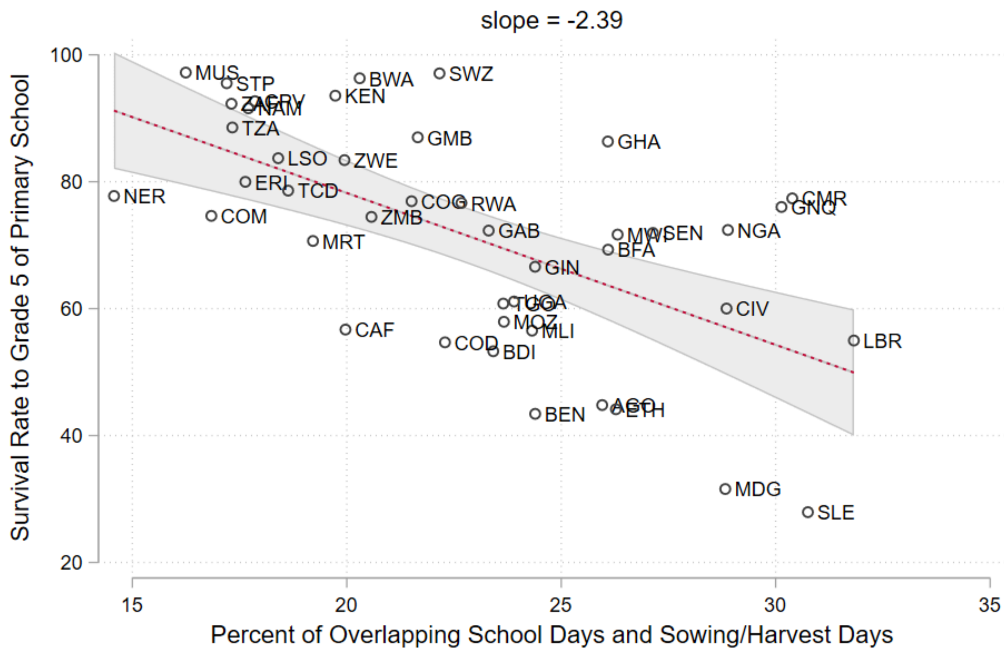
[22] Take caution not to accidentally align end-of-year exams with peak farming periods, as studied in Ito and Shonchoy (2020).

school-aged youth differentially exposed to the shock based on their community's pre-policy crop allocation via a shift-share estimation strategy and frontier randomization inference procedure. I find that, as predicted by theory, overlap between the school and farming calendars reduces both schooling and incidence of child labor on the household farm during peak production periods. Further, I find that schooling effects are most negative during the labor-intensive sowing period. Furthermore, negative education effects are concentrated on girls, poorer households and those in primary school; are likely driven by reductions in school participation along the intensive margin; and occur in new students who enter school after the policy change.

This paper makes several contributions to the study of household decision-making in low-income settings. First, I estimate households' trade-off between schooling investments and child labor by analyzing a natural experiment that shocked a child's time allocation to both activities while keeping the returns to each activity fixed. Second, this paper uniquely identifies and analyzes the impact of a time constraint shock, which may be yet another binding constraint that poor households face in some developing settings. Third, the findings suggest that the school calendar itself may be an effective policy tool for increasing time in school in sub-Saharan Africa by adapting the school calendar to minimize overlap with peak farming periods, as I illustrate with a policy simulation. Rather than conceptualizing the time trade-off between schooling and child labor as a zero-sum game, policymakers have the ability via the school calendar to alleviate the constraint on total time available to both productive activities. Overall, policymakers aiming to increase rural school participation should do more to accommodate farm labor demand.

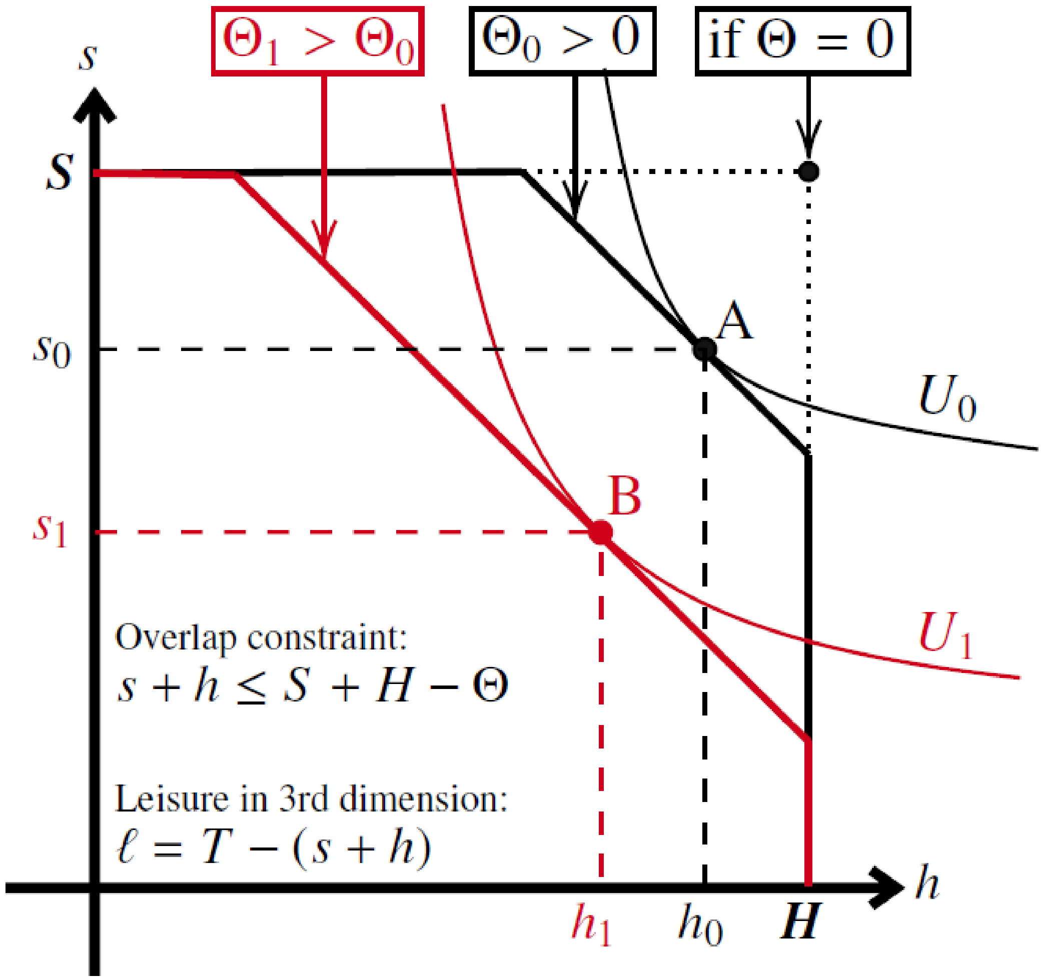
1.8 Tables and Figures

Figure 1.1: Primary School Survival and Overlap between the School and Farming Calendars in Sub-Saharan Africa



Notes: Figure depicts the correlation with the linear fit (and 95% confidence interval) between primary school survival and overlap between the school and farming calendars in 43 sub-Saharan African countries, labeled by their ISO alpha 3 code. The y-axis shows the survival rate to grade 5 of primary school for both sexes, as reported in the most recent year available by UNESCO's Institute for Statistics (UIS). The x-axis shows the overlapping percent of total days in the country's school calendar and total days in the sowing and harvesting periods, as estimated from countries' official primary school calendars (in most cases) and country-level crop calendars from the Food and Agriculture Organization (FAO). Further details are provided in Appendix A.1.

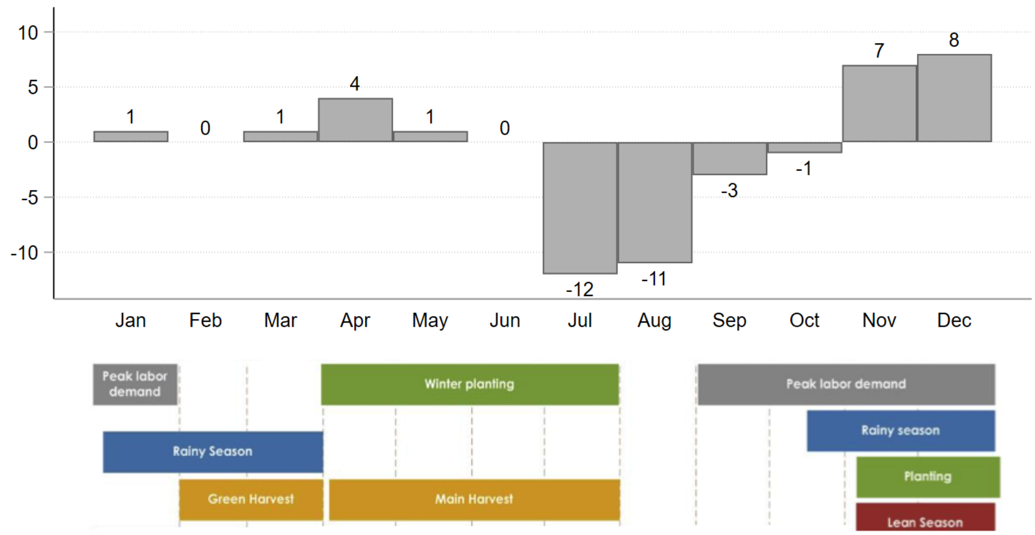
Figure 1.2: A Child's Time Constraint in the Household Problem



Notes: Figure depicts how the time constraint for schooling and farm work $s + h \leq S + H - \Theta$ enters into the household problem, where a household's allocation of a child's time in schooling s and household-farm work h is constrained by the maximum number of school days S and suitable farming days H minus how the number of days that the farm and school calendars overlap Θ. At $\Theta = 0$, the time allocation opportunity set extends to the point (H,S). At Θ_0 and Θ_1, an "overlap line" with a slope of -1 cuts into the frontier. Household preferences define indifference curves U_0 and U_1. As overlap increases from Θ_0 to Θ_1, the "overlap line" draws closer to the origin, potentially forcing reductions of both s and h that reduce household utility.

Figure 1.3: Malawi's School Calendar Change Increased Overlap with Peak Farm Labor Demand

Policy-Induced Change in School Days by Month



Typical Farming Calendar

Notes: The top panel depicts the net change in the number of school days from the 2009 (pre-policy) to the 2011 (post-policy) school calendar, showing that the school calendar change "shifted" school days out of July and August and to November and December. The bottom panel depicts Malawi's typical farming calendar according to the Famine Early Warning System Network (FEWS NET, 2013), and shows that the "shift" increased the number of school days during the rainy-season planting (sowing) period—what FEWS NET designates as "peak labor demand".

Table 1.1: Summary Statistics and Balance Tests

	Summary Statistics			Balance Test	
VARIABLES	N	Mean	SD	Coef.	RI pval
Retention:					
Included in 2013 Sample	2,287	0.94	0.24	-0.008	0.391
Baseline Controls:					
Baseline Values of Primary Outcomes:					
$Grade_i$: Highest Grade Completed	2,142	1.43	1.70	-0.098	0.266
$Farmed_i$: Indicator if Farmed in Peak Rainy Season	2,142	0.25	0.43	0.014	0.492
$Farm\ Hours_i$: Hours Farmed in Peak Rainy Season	2,142	9.82	21.94	3.165	0.031
Other Location-Level Controls:					
Individual Sex: Female Indicator	2,142	0.50	0.50	0.010	0.747
Individual Age	2,142	9.29	2.29	-0.018	0.903
Household Size	2,142	6.64	2.32	0.003	0.999
Household Asset Index	2,142	0.92	1.08	-0.076	0.235
Community had no on-farm labor	2,142	0.02	0.13	-0.009	0.568
Community had drought in prior 5 years	2,142	0.28	0.45	-0.333	0.016
Community share of Yao Hhs	2,142	0.11	0.24	0.021	0.815

Notes: Columns (1)-(3) report sample summary statistics for retention and baseline controls. Columns (4)-(5) report the coefficient and randomization inference p-values (RI pval) from "balance test" regressions of retention and baseline controls on $ssoverlap_\ell$ (defined in Table 1.3). Regressions also include specified controls from Equation 1.4.8: on-farm share of total annual hours worked; individual sex and age; household size and asset index; community-level indicator for containing no farming households, the community's share of Yao households, and an indicator for if the community experienced a drought in the prior five years; crop-level controls seasonal dummies (rainy, dry, permanent) and a dummy for grain crops as share-weighted location-level variables and altitude zone. Excluded from the regression is the baseline value of the control when serving themselves as the dependent variable.

Table 1.2: Pre-trend Analysis on Change in Grade Completion

	Age 6-13 from Each Year		Cohort: Age 6-9 in 2004	
	ΔGrade	ΔGrade	ΔGrade	ΔGrade
VARIABLES	2004-2009	2009-2013	2004-2009	2009-2013
	(1)	(2)	(3)	(4)
ssoverlap$_d$	0.192	-0.348	0.167	-0.863
	(0.268)	(0.218)	(0.173)	(0.659)
	[0.247]	[0.026]	[0.121]	[0.108]
Observations	30	30	30	30
R-squared	0.879	0.834	0.991	0.679
DV Mean	0.04	0.35	2.21	3.19

Notes: Without individual panel data before 2009, I examine pre-trends by averaging individual cross-sectional data from 2004, 2009, and 2013 to the district level, and then matching observations across 30 districts. I regress outcomes on district-averaged shift-share overlap $ssoverlap_d$ and controls, defined at the individual level in Table 1.3. I study two district-averaged samples: "Age 6-13 from Each Year" compares the age group corresponding with the study sample across surveys; and "Cohort: Age 6-9 in 2004" compares a cohort that ages to 11-14 in 2009 and 15-18 in 2013 (ages of intended school participation). Regressions of $\Delta Grade$ 2004-2009 (the change in the district's average grade completion between 2004 and 2009) test $ssoverlap_d$'s correlation with pre-policy movement in the primary outcome $Grade$, and regressions of $\Delta Grade$ 2009-2013 (the change in the district's average grade completion between 2009 and 2013) provide a comparison to post-policy effects. Conventional robust standard errors in parentheses. Randomization inference p-values in square brackets.

Table 1.3: Children's Grade Level and Household-Farm Labor in 2013

VARIABLES	Grade (1)	Farmed (2)	Farm Hours (3)
ssoverlap_ℓ	-0.277	-0.093	-5.170
	(0.111)	(0.035)	(3.704)
	[0.014]	[0.042]	[0.087]
Observations	2,142	2,142	2,142
R-squared	0.637	0.206	0.203
Base DV Mean	1.43	0.25	9.82
Δ DV Mean	2.30	0.22	21.49

Notes: Table presents results of the main specification: $ssoverlap_\ell$ is a shift-share measure of change in overlap between the school and farming calendars for community ℓ, constructed as the policy-induced "shift" in the number of days during which school and a crop's production both occurred, weighted by the pre-policy "share" of community labor devoted to producing each crop, summed across all crops, and normalized to unit variance. Outcomes are from 2013 from school-aged individuals ages 6-13 pre-policy: column (1) is an individual's last completed grade level, column (2) is an indicator equal to one if an individual worked on the household farm during the rainy-season sowing and harvest periods, and zero otherwise, and column (3) is the corresponding number of hours worked. Regressions include specified pre-policy controls, which include the baseline value of the outcome; the on-farm share of total annual hours worked; individual sex and age; household size and asset index; community-level indicator for containing no farming households, the community's share of Yao households, and an indicator for if the community experienced a drought in the prior five years; crop-level controls seasonal dummies (rainy, dry, permanent) and a dummy for grain crops as share-weighted location-level variables and altitude zone; and baseline values of $Grade_i$, $Farmed_i$, and $Farm\ Hours_i$ (when not already included). Conventional robust standard errors in parentheses. Randomization inference p-values in square brackets.

Table 1.4: Main Results: P-values from Alternative Inference Procedures

	VARIABLES	Grade	Farmed	Farm Hours
		(1)	(2)	(3)
(a)	Conventional OLS	0.013	0.008	0.163
(b)	Random draw (w/ replacement)	0.014	0.042	0.087
(c)	Random draw (w/ replacement): correlated by season	0.030	0.084	0.133
(d)	Random Normal shock	0.005	0.029	0.074
(e)	Random Normal shock: correlated by season	0.026	0.062	0.125
(f)	Simulated calendar changes: correlated by season	0.018	0.042	0.084
(g)	Simulated calendar changes: correlated across all crops	0.029	0.061	0.076
(h)	Share-weighted shock-level regression (BHJ)	0.012	0.024	0.054

Notes: Alternative inference procedures test the effect of $ssoverlap_\ell$ in same regressions as in Table 1.3 columns (1)-(3). I report p-values from (a) conventional ordinary least-squares (OLS); randomization inference procedures that perturb the crop-level shock by randomly re-drawing it from (b) the actual distribution of shocks with replacement (as in Table 1.3), (c) the actual distribution of shocks for same-season crops with replacement, (d) a normal distribution defined by the actual distribution's first and second moments, (e) a normal distribution defined by same-season crops' actual distribution's first and second moments, (f) a simulated distribution of shocks corresponding with the universe of possible school calendar changes between 2009 and 2011, assuming school starts on a Monday and a fixed length and structure of the school calendar, drawn for all crops with the same season, and (g) a simulated distribution of shocks as described above but for all crops together; and (h) the share-weighted shock-level regression as described in Borusyak et al. (2022). Further details provided on all alternative inference procedures in Appendix A.4.

Table 1.5: Overlap Effects by Sex and Age

	Pre-Policy Age 6-9				Pre-Policy Age 10-13			
	Girls		Boys		Girls		Boys	
VARIABLES	Grade	Farmed	Grade	Farmed	Grade	Farmed	Grade	Farmed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ssoverlap_ℓ	-0.435	-0.021	-0.179	-0.160	-0.457	-0.151	-0.070	-0.052
	(0.181)	(0.074)	(0.164)	(0.065)	(0.216)	(0.075)	(0.275)	(0.070)
	[0.006]	[0.713]	[0.088]	[0.009]	[0.018]	[0.019]	[0.643]	[0.420]
Observations	580	580	577	577	494	494	491	491
R-squared	0.515	0.136	0.497	0.195	0.605	0.206	0.528	0.174
Base DV Mean	0.56	0.19	0.51	0.17	2.61	0.34	2.35	0.34
Δ DV Mean	2.22	0.18	1.98	0.18	2.48	0.27	2.59	0.28

Notes: Table presents results of the main specification on sub-samples determined by individual sex and age. In columns (1)-(4), results for individuals age 6-9 in 2009 (pre-policy) who become age 10-13 in 2013 (outcome data collection post-policy), with results for boys in columns (1) and (2) and for girls in columns (3) and (4). In columns (5)-(8), results for individuals age 10-13 in 2009 who become age 14-17 in 2013, with results for boys in columns (5) and (6) and for girls in columns (7) and (8). $ssoverlap_\ell$ and included controls defined in Table 1.3. Conventional robust standard errors in parentheses. Randomization inference p-values in square brackets.

Table 1.6: Overlap Effects by Household Assets

	Bottom Tercile		Middle Tercile		Top Tercile	
VARIABLES	Grade	Farmed	Grade	Farmed	Grade	Farmed
	(1)	(2)	(3)	(4)	(5)	(6)
ssoverlap$_\ell$	-0.333	-0.087	-0.185	-0.142	-0.058	-0.041
	(0.221)	(0.055)	(0.157)	(0.063)	(0.195)	(0.078)
	[0.046]	[0.119]	[0.095]	[0.023]	[0.674]	[0.414]
Observations	768	768	708	708	666	666
R-squared	0.528	0.156	0.614	0.237	0.652	0.192
Base DV Mean	0.96	0.33	1.24	0.24	2.17	0.18
Δ DV Mean	2.00	0.25	2.13	0.25	2.82	0.17

Notes: Table presents results of the main specification on sub-samples determined by an individual's household asset index (estimated as the first principal component of a vector of indicator variables for ownership of 12 assets in 2010). Results for the bottom tercile of the asset index are in columns (1)-(2), middle tercile in columns (3)-(4), and top tercile in columns (5)-(6). $ssoverlap_\ell$ and included controls defined in Table 1.3. Conventional robust standard errors in parentheses. Randomization inference p-values in square brackets.

Table 1.7: Heterogeneity by Timing

VARIABLES	Grade (1)	Farmed (2)	Grade (3)	Farmed (4)	Grade (5)	Farmed (6)
ssoverlap_ℓ			-0.257	-0.041		
			(0.126)	(0.040)		
			[0.040]	[0.400]		
$\text{ssoverlap\_sow}_\ell$	-0.360	-0.022			-0.431	0.014
	(0.160)	(0.046)			(0.174)	(0.053)
	[0.056]	[0.763]			[0.044]	[0.867]
$\text{ssoverlap\_harv}_\ell$	-0.211	-0.149			-0.049	-0.108
	(0.119)	(0.042)			(0.162)	(0.056)
	[0.130]	[0.018]			[0.754]	[0.108]
$\text{ssoverlap\_admis}_\ell$			0.018	0.002	0.180	-0.050
			(0.122)	(0.040)	(0.156)	(0.052)
			[0.851]	[0.965]	[0.116]	[0.286]
$\text{ssoverlap\_exams}_\ell$			0.035	0.088	0.123	0.060
			(0.098)	(0.030)	(0.111)	(0.035)
			[0.728]	[0.036]	[0.243]	[0.186]
Observations	2,142	2,142	2,142	2,142	2,142	2,142
R-squared	0.638	0.208	0.637	0.209	0.638	0.210
Base DV Mean	1.430	0.252	1.430	0.252	1.430	0.252
Δ DV Mean	2.300	0.222	2.300	0.222	2.300	0.222
Test pval:Sow=Harv	0.369	0.014			0.088	0.101
Test pval:All+Admis=0			0.179	0.461		
Test pval:All+Exams=0			0.249	0.434		
Test pval:Admis=Exmas			0.912	0.096	0.725	0.043

Notes: Table presents variations of the main specification. Columns (1) and (2) regress Equation 1.4.5: $ssoverlap\_sow_\ell$ and $ssoverlap\_harv_\ell$ are measures of shift-share overlap for the sowing and harvest periods, respectively, which add together to equal $ssoverlap_\ell$ (as defined in Table 1.3). Columns (3) and (4) regress Equation 1.4.6: $ssoverlap\_admis_\ell$ and $ssoverlap\_exams_\ell$ are added to $ssoverlap_\ell$ as measures of shift-share overlap between peak farming periods and the first and last month of school, respectively, representing admissions and exam periods, and can be interpreted as interaction terms (i.e., the additional effect of overlap during admissions and exam periods). Columns (5) and (6) regress Equation 1.4.7 in which $ssoverlap\_admis_\ell$ and $ssoverlap\_exams_\ell$ can also be interpreted as interaction terms since $ssoverlap\_sow_\ell$ and $ssoverlap\_harv_\ell$ sum to $ssoverlap_\ell$. All new regressors are normalized to unit variance of $ssoverlap_\ell$ to have comparable coefficients. Outcomes and controls included in the regression defined in Table 1.3. Conventional robust standard errors in parentheses. Randomization inference p-values in square brackets.

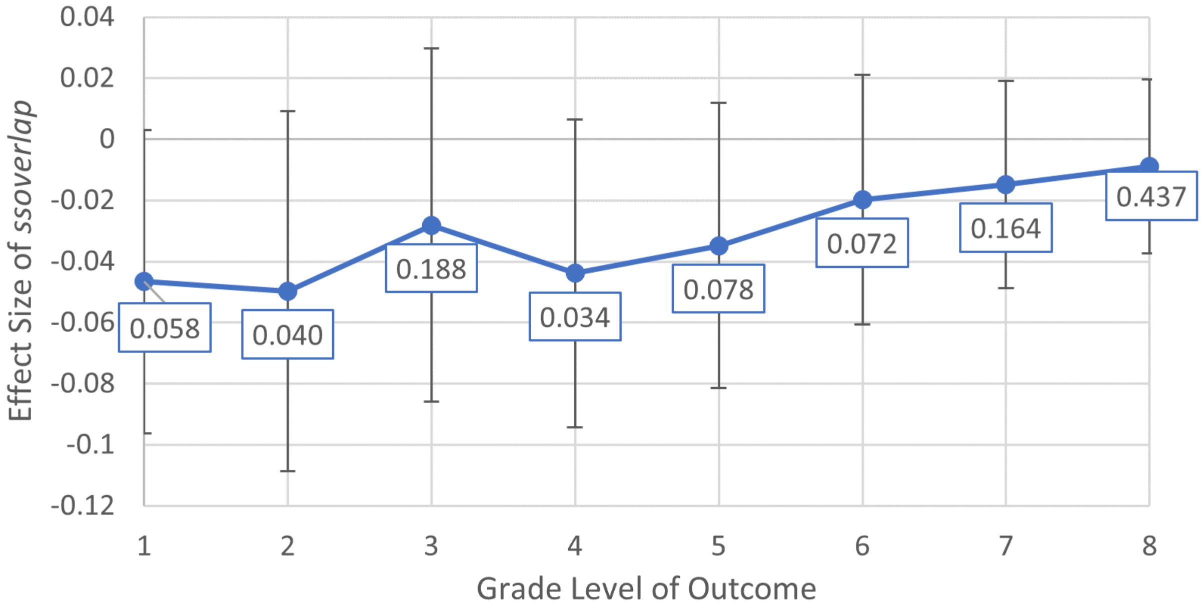
Table 1.8: Possible Channels for the Negative Schooling Effect

VARIABLES	Started School	Enrolled if Started School	Absent 2 Wks if Enrolled	Grade Intensive
	(1)	(2)	(3)	(4)
ssoverlap$_\ell$	-0.008	0.003	-0.019	-0.183
	(0.017)	(0.022)	(0.022)	(0.106)
	[0.557]	[0.790]	[0.338]	[0.071]
Observations	2,142	2,043	1,874	1,515
R-squared	0.299	0.136	0.022	0.671
Base DV Mean	0.83	0.85	0.04	1.61
Δ DV Mean	0.12	0.06	0.02	2.63

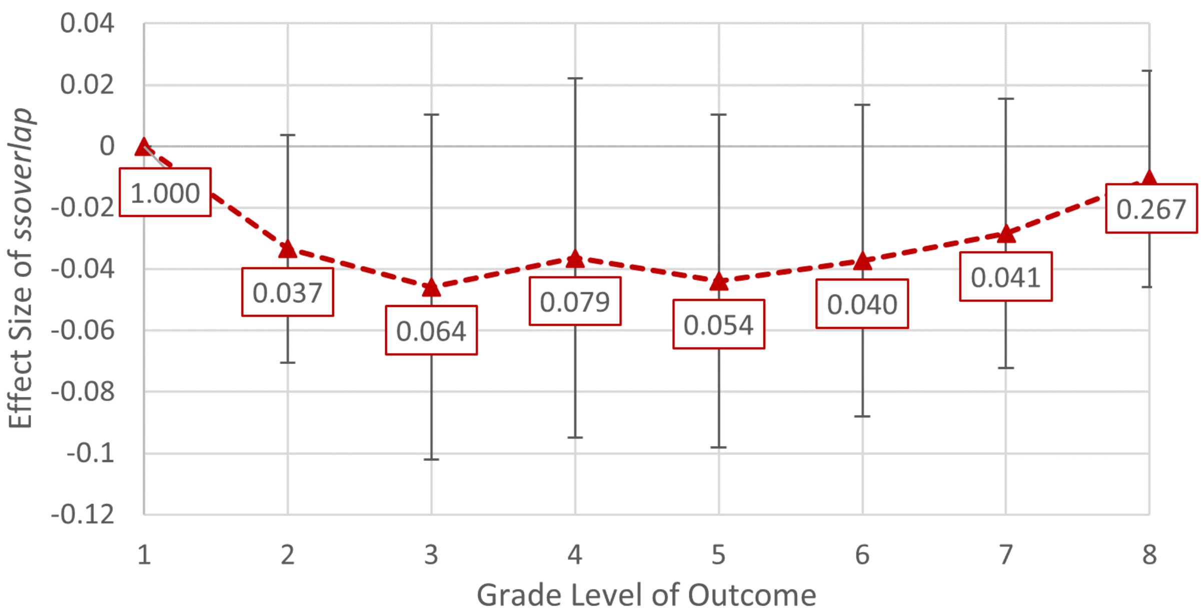
Notes: Table presents results of the main specification but with a cascading set of outcome variables to analyze possible channels for overlap's negative effect on schooling. Dependent variables and sub-samples defined as follows: Column (1) is an indicator equal to one if an individual started school by 2013, and zero otherwise. Conditional on starting school by 2013, column (2) is an indicator equal to one if an individual was enrolled in school in 2013, and zero otherwise. Conditional on enrollment in 2013, column (3) is an indicator equal to one if an individual is reported by the household respondent to miss more than two consecutive weeks of school in the last year, and zero otherwise. Finally, conditional on enrollment in school in 2009 and 2013 and not missing more than two consecutive weeks of school, column (4) is a measure of highest completed grade level ruling out possible effects on the extensive margins. $ssoverlap_\ell$ and included controls defined in Table 1.3. Conventional robust standard errors in parentheses. Randomization inference p-values in square brackets.

Figure 1.4: Schooling Impacts by Primary School Grade Level

(a) Grade Level Completion by Grade: Dummy if Grade Completed by 2013



(b) Survival Rate by Grade: Dummy if Grade Reached by 2013 after Starting School



Notes: Figure plots the $\hat{\beta}$ coefficient on $ssoverlap_\ell$ (on the vertical axis) from regressing the main specification on a vector of indicators corresponding to grades 1 through 8 (on the horizontal axis). Panel (a) plots the effect of $ssoverlap_\ell$ on a vector of indicators for completing grades 1 through 8 by 2013. Panel (b) plots the effect of $ssoverlap_\ell$ on a vector of indicators for attending grades 1 through 8 by 2013 conditional on having started school (often called "survival" to each grade level). Each point estimate is accompanied by error bars of the 95% confidence interval from conventional standard errors and RI p-values in adjacent boxes.

Table 1.9: Adjustments to Other Child Labor

VARIABLES	Farmed: Dry Season (1)	Tended Livestock (2)	Worked: Day Labor (3)	Worked: Unpaid (4)
$ssoverlap_{\ell}$	0.011	0.005	-0.014	-0.020
	(0.016)	(0.017)	(0.022)	(0.017)
	[0.443]	[0.739]	[0.289]	[0.199]
Observations	2,142	2,142	2,142	2,142
R-squared	0.082	0.036	0.083	0.022
Base DV Mean	0.01	0.04	0.03	0.01
Δ DV Mean	0.04	0.03	0.07	0.02

Notes: Table presents regressions of indicators for other types of child labor on the main specification. All dependent variables are equal to one if the individual worked any amount of the listed activity in the past 12 months, and are zero otherwise: column (1) is household-farm work during the most recent dry-season sowing and harvest; column (2) is tending to household livestock; column (3) is working as a day laborer (i.e., ganyu); and column (4) is working unpaid. Incidence of each activity falls below 5% at baseline and at or below 10% post-policy. Formal work is excluded as it employs less than 0.15% of the sample. $ssoverlap_{\ell}$ and included controls defined in Table 1.3. Conventional robust standard errors in parentheses. Randomization inference p-values in square brackets.

Table 1.10: Adjustments to the Household Farm

VARIABLES	Acres Farmed (1)	Cost: Hired Labor (2)	Cost: Seeds (3)	Cost: Other Inputs (4)	Farm Costs (5)	Farm Revenues (6)	Farm Profits (7)
ssoverlap_ℓ	0.003	835.0	394.8	-307.8	2451.0	4874.9	2211.8
	(0.112)	(413.6)	(188.3)	(1560.8)	(2253.7)	(3784.3)	(3244.3)
	[0.969]	[0.028]	[0.016]	[0.862]	[0.295]	[0.303]	[0.630]
Observations	1,174	1,174	1,174	1,174	1,174	1,174	1,174
R-squared	0.442	0.176	0.135	0.332	0.332	0.329	0.192
Base DV Mean	1.68	1175.0	542.3	4686.7	8051.2	9168.4	2030.6
Δ DV Mean	0.03	1076.4	617.7	5539.0	9107.6	13717	4773.1

Notes: Table presents regressions of household-level farm outcomes on a household-averaged version the main specification for all households containing sample individuals. Column (1) is the number of acres cultivated on the household farm throughout the year. Columns (2)-(4) measure different farm costs: labor hired from outside the household in column (2), expenses on seeds in column (3), and expenses on fertilizer, pesticide and herbicide in column (4). Column (5) is total farm costs including other expenses like land rent, coupon purchases, and transportation. Column (6) is total farm revenue including crop sales, land rent or coupon sales. Column (7) is total farm profits: revenue minus costs. All dependent variables are winsorized at their 95th percentile, and farm profits is also winsorized at its 5th percentile due to negative outliers. $ssoverlap_\ell$ and included controls defined in Table 1.3. Conventional robust standard errors in parentheses. Randomization inference p-values in square brackets.

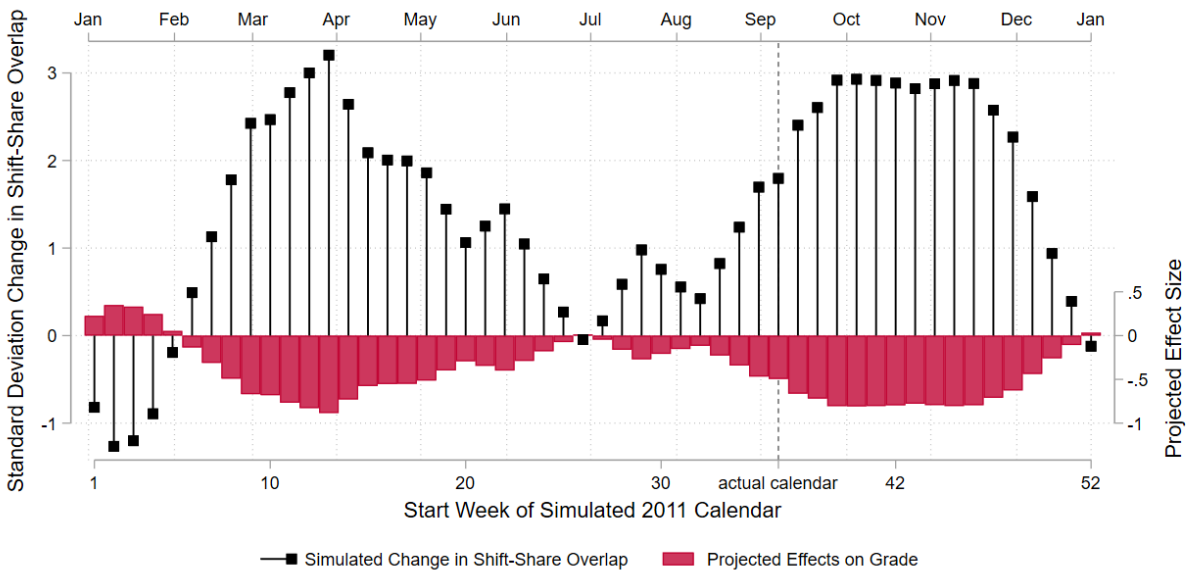
Table 1.11: Long-Run Impacts on Outcomes

	Grade		Farmed	
VARIABLES	2016	2019	2016	2019
	(1)	(2)	(3)	(4)
ssoverlap_ℓ	-0.375	-0.128	-0.102	-0.158
	(0.168)	(0.231)	(0.046)	(0.050)
	[0.014]	[0.537]	[0.022]	[0.010]
Observations	1,918	1,714	1,918	1,714
R-squared	0.727	0.604	0.291	0.164
Base DV Mean	0.65	0.62	0.17	0.17
Δ DV Mean	2.87	4.50	0.33	0.41

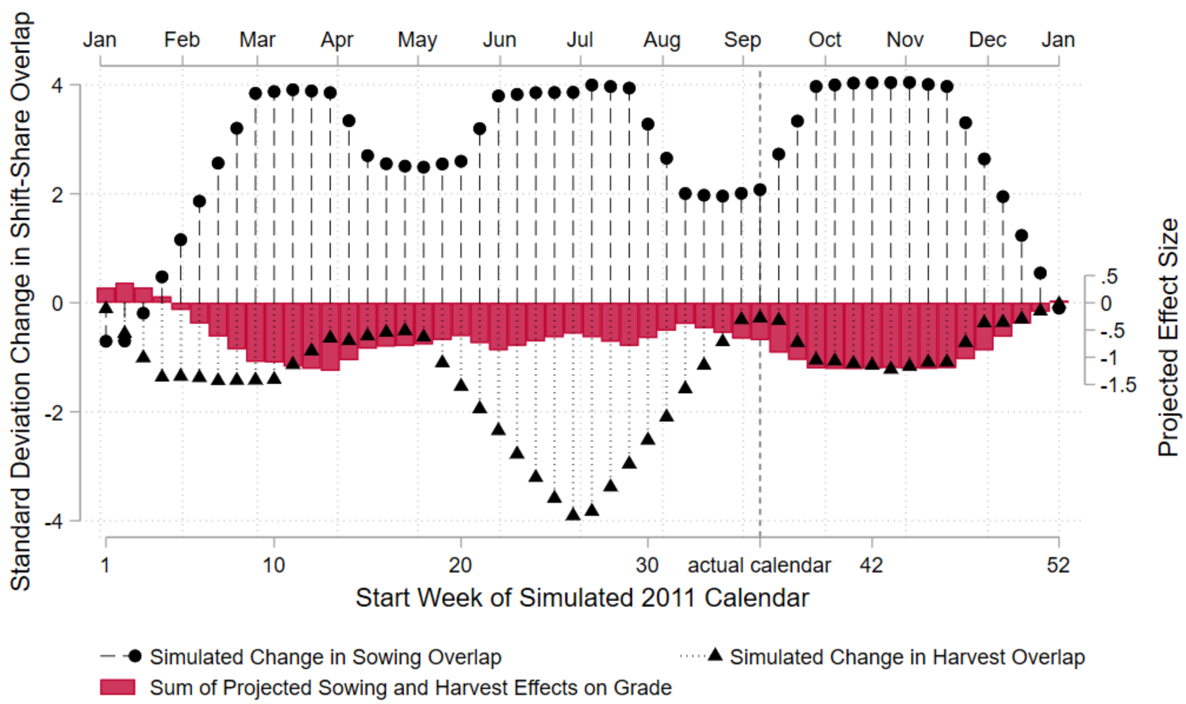
Notes: Regressions run main specification but on later panel samples of individuals ages 0-13 pre-policy. $ssoverlap_\ell$ and included controls defined in Table 1.3. Dependent variables were measured for a subset of individuals in follow-up panel surveys in either 2016 or 2019 and are defined as follows: $Grade_i$ and $Farmed_i$ are defined in Table 1.3 but instead refer to the listed year 2016 or 2019. Conventional robust standard errors in parentheses. Randomization inference p-values in square brackets. Results in Appendix A.6 breakdown results by age group.

Figure 1.5: Simulated Policy Impacts of Alternative 2011 School Calendars

(a) Overall Impact



(b) Sowing versus Harvest Impact



Notes: Figure depicts simulated policy impacts of alternative 2011 school calendars relative to the actual 2009 school calendar. Panel (a) plots the expected change in shift-share overlap for the average sample community using drop lines and projected effects on $Grade_i$ using crimson bars. Panel (b) plots these outcomes separately for both the sowing and harvest periods. Additionally, the top horizontal axis estimates the start of each month, and the vertical line denotes the actual 2011 school calendar that started on September 6, 2010.

CHAPTER II

Teaching and Incentives: Substitutes or Complements?

Co-authored with Arlete Mahumane, James Riddell IV, Tanya Rosenblat, Dean Yang, and Hang Yu

2.1 Introduction

Societies devote substantial resources to helping people acquire knowledge. These efforts often take place in educational institutions. In addition, outside of school settings, there are many efforts to promote learning about financial decision-making (raising "financial literacy"), public health (promoting "health literacy"), and many other areas. Efforts to promote learning commonly take one of two approaches. First, one can *teach*, via classroom instruction, broadcast media, advertising, social media, or other means. Second, one can improve learners' *incentives* to acquire knowledge, such as by informing them about the returns to education, or providing incentives for good performance on learning assessments (e.g., merit scholarships or other rewards based on test scores). These two broad approaches are often described as operating on the "supply" and "demand" sides of education, respectively (Banerjee and Duflo, 2011; Glewwe, 2014). Supply interventions provide educational inputs (e.g., teaching and instruction), reducing the marginal cost of learning. Demand interventions seek to raise learners' perceived marginal benefit of learning.

Supply and demand educational interventions often operate at the same time. Existing research, however, says little about *interactions* between such interventions. Crucially, are supply and demand interventions *substitutes* or *complements*? Understanding complementarities between interventions is key for cost-effectiveness analyses, and thus decision-making on optimal combinations of policies (Twinam, 2017). If two interventions are complements, the gains from implementing both exceed the sum of the gains of implementing each one singly. The greater the complementarity, the more attractive it could be to implement both policies together, rather than either one alone. If they are substitutes, by contrast, the gains from implementing both are less than the sum of the gains of implementing each one singly. In this case, it becomes more likely that the optimal course would be to implement just one or the other of the policies, not both together.

We implemented a randomized controlled trial of a supply and a demand intervention to promote learning, estimating the degree to which the two are substitutes or complements. We study learning about COVID-19 among adults in Mozambique, and implement treatments that are representative examples of supply and demand interventions to promote learning. Our supply treatment *teaches* about COVID-19. It provides information targeted at individuals' specific knowledge gaps, a pillar of the "teaching at the right level" (TaRL) pedagogical approach (Banerjee et al., 2007; Duflo et al., 2011). We view this feedback as an important component of teaching; however, we do not attempt to teach principles (e.g., of immunology) which would allow respondents to answer new questions correctly ("in-depth" teaching). The demand-side treatment offers individuals financial *incentives* for correct responses on a later COVID-19 knowledge test. This treatment is analogous to educational testing with non-zero stakes for test-takers.

Abiding by COVID-19 health protocols, we interacted with our 2,117 Mozambican study respondents solely by phone. We registered a pre-analysis plan prior to implementation. We assessed respondents' COVID-19 knowledge in a baseline survey, and then implemented the teaching and incentive treatments in a 2x2 cross-randomized design. The design created a control group and three treatment groups: "Incentive" only, "Teaching" only, and "Incentive plus Teaching" (or "Joint"). We measure impacts on a COVID-19 knowledge test several weeks later.

To theoretically examine interactions between teaching and incentives, we write down a simple model of knowledge acquisition. Individuals can exert effort to search for knowledge on their own, and can also learn from teaching. In the model, the Incentive and Teaching treatments can be either substitutes or complements, depending on the magnitudes of two countervailing effects. The Incentive treatment has a *motivation* effect, potentially enhancing the impact of Teaching. But Teaching can have a *crowding-out* effect, by reducing the need to search for knowledge, thus lowering the effectiveness of the Incentive treatment. We define a parameter λ, representing the degree of complementarity. If motivation effects dominate crowding-out effects, then Incentive and Teaching are complements ($\lambda > 0$). Otherwise, they are substitutes ($\lambda < 0$).

In advance of sharing our results publicly, we determined a reasonable "benchmark" λ by collecting expert predictions of our treatment effects. The vast majority of surveyed experts expected the two treatments to be substitutes, predicting that the effect on test scores of the combination of both treatments would be less than the sum of the effects of each treatment implemented singly. In the context of the theoretical model, expert predictors believed that when offering the Incentive and Teaching treatments together, the crowding-out effect would dominate the motivation effect.

We find substantially more complementarity than experts predicted: actual estimated λ is positive, and highly significantly different from experts' negative prediction of λ. The Incentive treatment raises COVID-19 knowledge test scores (fraction of questions answered correctly) by 1.56 percentage points, while Teaching does so by 2.88 percentage points. By contrast, the Joint treatment raises test scores by 5.81 percentage points, 31% larger than the sum (4.44 percentage points) of the effects of each treatment provided separately. Actual estimated λ is also marginally statistically significantly different from zero, another benchmark of interest. These results are consistent with the theoretical case in which the motivation effect dominates the crowding-out effect when providing both treatments together. The effect of the Joint treatment is large in magnitude: 0.5 test score standard deviations. Additionally, the Joint treatment's significant positive effect and complementarity pertain to newly asked questions (not just questions previously asked) and persist over nine months after the intervention.

We provide a simple illustration of the importance of the estimate of λ for cost-effectiveness comparisons. We use our actual treatment effect estimates and implementation costs to calculate cost-effectiveness of the individual Incentive and Teaching treatments, as well as the cost-effectiveness of the Joint treatment for different values of λ. Our estimated λ is below the threshold at which the Joint treatment would be the most cost-effective of our three treatments. That said, governments or NGOs implementing our treatments in different contexts may come to different cost-effectiveness rankings given their specific implementation costs.

This research contributes to economics research on education and learning. There is a substantial literature examining the impacts of supply- and demand-side educational interventions (Glewwe, 2014; Evans and Popova, 2015; Le, 2015; McEwan, 2015; Conn, 2017; Muralidharan, 2017).

On the supply side, studies have examined provision of educational supplies (Glewwe et al., 2000, 2009), school facilities (Duflo, 2001), new teaching technologies (Muralidharan et al., 2019b), and "teaching at the right level" (TaRL) (Banerjee and Duflo, 2011; Duflo et al., 2011). Angrist et al. (2020b) show that teaching via cellphone can offset learning loss during the COVID-19 pandemic. Mbiti et al. (2019) show complementarity between two supply-side interventions (increased school resources and teacher incentives). Outside of school settings, supply-side efforts are made to provide health education to promote "health literacy" (Batterham et al., 2016), financial education to promote "financial literacy" (see

Kaiser and Menkhoff (2017) for a review), and agricultural "extension" to improve farming knowledge (Anderson and Feder, 2007; Fabregas et al., 2019).[1] Our Teaching treatment implements a targeted approach to promote COVID-19 health literacy.

Demand-side educational interventions seek to increase the perceived returns to learning. In school settings, studies have examined impacts of providing information on the wage returns to schooling (Jensen, 2010), merit scholarships based on test performance (Kremer et al., 2009b; Berry et al., 2019), or incentives for test performance (Angrist and Lavy, 2009; Levitt et al., 2011; Fryer, 2011; Behrman et al., 2015; Burgess et al., October 2016; Fryer, 2016; Hirshleifer, 2017). Outside of school settings, studies have evaluated incentive-based strategies such as cash payments, deposit contracts, lotteries and non-cash rewards to promote healthy behaviors (Finkelstein et al., 2019), but do not target learning outcomes. Our Incentive treatment is analogous to policies providing financial incentives for test performance, making it a rare example of a demand-side policy to promote learning among non-students.[2]

The most novel feature of our work is that we explicitly highlight and measure the complementarity between a supply-side and a demand-side educational intervention. Behrman et al. (2015) and List et al. (2018) study the interactions between test-score incentives for teachers (supply-side) and students (demand-side), but do not estimate a complementarity parameter, as we do.[3] In addition to being of policy interest, we view this interaction as of particular theoretical interest due to the countervailing motivation and crowding-out effects of combining supply- and demand-side educational interventions.

Our study also contributes to understanding adult education in health crises. Broader research suggests that adults have higher economic and physiological barriers to learning (Aker and Sawyer, 2021), and that successful health informational interventions are comprehensive but not overly complex (Dupas et al., 2011). Additional challenges in health crises often include underlying institutional mistrust and misinformation (Vinck et al., 2019) and logistical obstacles to needs assessments and outreach with vulnerable populations (Checchi et al., 2017). In this context, we demonstrate simple interventions that can complement phone data collection during epidemics (Angrist et al., 2020a; Maffioli, 2020; Magaço et al., 2021). In particular, our Teaching intervention shows that providing feedback on knowledge-based

[1]There are also efforts to improve knowledge of legal issues, often referred to as "legal awareness" or "public legal education" (American Bar Association, 2021).

[2]Carpena et al. (2017) find no effect of financial incentives on adult financial literacy test performance. Thornton (2008) studies incentives to learn about HIV status.

[3]Fryer et al. (2016) study a supply-side intervention (teacher incentives) jointly with a demand-side intervention (student incentives). They do not examine the supply- and demand-side treatments separately, so cannot measure their complementarity. Li et al. (2014) compare results across two different experiments, rather than measuring complementarity in one experiment, and argue that there is complementarity between a peer-effects intervention (supply-side) and providing test-score financial incentives (demand-side).

questions is a feasible and impactful add-on to health surveys—for example, on ”knowledge, attitudes, and practices (KAP)” surveys common in public health.[4]

Related studies seek to improve COVID-19-related knowledge among adults. Alsan et al. (2020) show that messaging tailored to minorities improves their COVID-19-related knowledge. Mistree et al. (2021) and Maude et al. (2021) find that randomly assigned teaching interventions improve COVID-19-related knowledge in India and Thailand, respectively, while Bahety et al. (2021) find no evidence that COVID-19 SMS-based information campaigns improve knowledge in rural India. Angrist et al. (2020b) and Banerjee et al. (2020) use phone-based interventions to address issues during the pandemic.

2.2 A Simple Model of Learning

There are N dimensions of *knowledge*. On each dimension there are two possible states $\{A, B\}$: a *correct* state A and a *incorrect* state B. For example, one dimension of knowledge might be “Hot tea helps to prevent Covid-19,” with the two states being “correct” and “incorrect”.

Initial Knowledge. Every agent has independent priors on each state which we model as follows. The agent initially believes that both states are equally likely to be correct. She then receives a binary signal that informs her about the correct state – that signal is correct with probability $\mu > \frac{1}{2}$. This implies that a share μ of population have a posterior that places weight μ on the correct state while a share $1 - \mu$ of the population has a posterior that places weight μ on the incorrect state.

Actions. For each knowledge dimension i, an agent takes an action $x_i \in \{a, b\}$: a (b) will provide utility 1 if the correct state is A (B) and 0 otherwise. The agent will therefore always choose the action that is appropriate for the state on which she places a greater subjective probability on being correct. For example, equipped with initial knowledge a share μ of the population will derive utility 1 by taking the correct action and a share $1 - \mu$ of the population will derive utility 0. The initial expected utility of agents is therefore μ. Let R be the benefits or returns that agents gain for knowing the correct state of a knowledge dimension.

Teaching. Now assume that the government or some other authority seeks to teach the agent the correct state (our Teaching treatment). The agent will adopt this recommendation with probability $p(R)$ which captures the credibility of the source (and hence the agent's propensity to follow the advice) as well as the attention she pays to the advice. Otherwise the agent ignores the recommendation.

[4] See for Puspitasari et al. (2020) for a review of COVID-19 KAP surveys.

Importantly, attention can depend on the return the agent receives for being correct: $p(R)$ is (weakly) increasing in R. This creates a positive interaction effect between the return to knowledge and the propensity to absorb what is taught.

Teaching generates 3 types of posteriors:

- A share p of the population places subjective probability 1 on the correct state. This group is made up of all agents who followed the advice.
- A share $(1-p)\mu$ of the population places subjective probability μ on the correct state.
- A share $(1-p)(1-\mu)$ of the population places subjective probability $1-\mu$ on the correct state.

When the perceived returns to knowledge are negligible (i.e., $R = 0$), the Teaching treatment increases the share of correct answers to $p(0) + (1 - p(0))\mu$.

Returns to Knowledge. Recall that agents gain benefits or returns R for knowing the correct state of a knowledge dimension. She can spend effort $e \geq 0$ on searching for knowledge at a cost of αe^2 – this will provide a correct signal with probability e. Then with probability $1-e$ she does not find the correct answer and follows her initial belief μ. Returns R may be manipulated by a **learning incentive** (our Incentive treatment), which increases the share of correct answers to $e^* + (1-e^*)\mu$.

- Agents who already experienced the Teaching treatment and paid attention to it expend effort $e = 0$ since their posterior is already placing probability 1 on the correct state. Knowledge depreciation is ignored as it is assumed to be the same, on average, for all agents.
- The other two groups of agents will in equilibrium spend the same amount e^* on searching behavior. Their expected utility equals:

  $$(e^* + (1-e^*)\mu)R - \alpha(e^*)^2$$

  The first two terms capture the utility from taking the correct action when she finds the correct signal, and the last term captures the cost of searching for correct knowledge.

  The optimal action therefore equals $e^* = \frac{R}{2\alpha}(1-\mu)$: she will search more if their initial knowledge is less precise (lower μ), if searching is less expensive (lower α) or if the reward R is higher.

To summarize, the Teaching and Incentive treatments give rise to three types of posterior beliefs:

- A share $p(R) + (1 - p(R))e^*$ of the population places subjective probability 1 on the correct state. This group is made up of all agents who followed the advice.

- A share $(1 - p(R))(1 - e^*)\mu$ of the population places subjective probability μ on the correct state.

- A share $(1 - p(R))(1 - e^*)(1 - \mu)$ of the population places subjective probability $1 - \mu$ on the correct state.

Learning. The share of the population with correct knowledge *prior* to the Teaching and Incentive treatments is μ.

After the Teaching and Incentive treatments, the share of correct answers increases to:

$$p(R) + (1 - p(R))e^* + (1 - p(R))(1 - e^*)\mu \tag{2.2.1}$$

We organize the share of correct answers by treatment in Table 2.1.

We can now compare the effect of the Incentive plus Teaching (Joint) treatment with the simple sum of each treatment implemented separately. Let this difference be defined as the complementarity parameter λ:

$$\lambda \equiv \text{Joint} - (\text{Teaching only} + \text{Incentive only}) = \underbrace{(p(R) - p(0))(1 - \mu)}_{\text{motivation}} - \underbrace{e^* p(1 - \mu)}_{\text{crowding out}} \tag{2.2.2}$$

There are two opposing effects. The *motivation* effect captures that Teaching has greater impact when the return to knowledge is higher (e.g., because agents are more motivated to learn, she pays more attention to teaching, or exert more knowledge-search effort). On the other hand, there is a *crowding out* effect because Teaching reduces the need to search for knowledge and hence the effectiveness of the Incentive treatment.

Lemma 1. *The Teaching and Incentive treatments are* ***complements*** *if the motivation effect dominates the crowding out effect. Otherwise, the Teaching and Incentive treatments are* ***substitutes****.*

When the Teaching and Incentive treatments are complements, the complementarity parameter will be positive: $\lambda > 0$. When they are substitutes, on the other hand, it will be negative: $\lambda < 0$. When $\lambda = 0$, we say the two treatments are *additive*.

In our empirical analyses, we provide an estimated complementarity parameter, $\hat{\lambda}$.

2.3 Sample and Data

2.3.1 Data

We implemented three rounds of surveys by phone in July-November 2020: a pre-baseline, baseline and endline survey (see Figure 2.1 for a study timeline). Respondents were from households with phones in the sample of a prior study (Yang et al., 2021).[5] We surveyed one adult per household. Participants received a small gift of 50 meticais (approx. US$0.70) after completing each survey, as explained at study enrollment, which was transferred via MPesa over 93% of the time and phone credit recharge otherwise. Appendix B.1 provides details on the COVID-19 context and study communities.

Between a pre-baseline survey and baseline survey, we randomly assigned households to treatments and registered a pre-analysis plan (PAP). The baseline survey was immediately followed by over-the-phone treatment implementation. There was a minimum of 3.0 weeks and average of 6.3 weeks between baseline and endline surveys for all respondents. Baseline and endline surveys occurred when COVID-19 cases were rising rapidly.

The endline sample size is 2,117 respondents, following a sample size of 2,226 at baseline. The retention rate between baseline and endline is 95.1% overall, at least 94.4% in each of the seven districts surveyed, and balanced across treatment conditions.

We measured respondents' COVID-19 knowledge in three categories: 1) general knowledge (risk factors, transmission, and symptoms); 2) preventive actions (preventing spread to yourself and others); and 3) government policies (official actions taken by the national government of Mozambique). Pre-baseline, we tested numerous pilot questions. Then, at baseline and endline, we administered a pre-specified set of knowledge questions and their correct responses in our analysis plan submitted to the AEA RCT Registry. At baseline, we asked respondents knowledge questions randomly selected within each category, and respondents randomly assigned to the Teaching treatment were given feedback on incorrect and correct responses. At endline, respondents were asked a full set of knowledge questions to estimate treatment effects. Poor internet access and low ownership of electronic devices make it very unlikely that respondents looked up correct answers during the questionnaire. See Appendix B.2 for details on question selection and the list of questions.[6]

[5] AEA RCT Registry for Yang et al. (2021): https://doi.org/10.1257/rct.3990-5.1

[6] Examples of questions (correct responses in parentheses) include the following. General knowledge: "How is coronavirus spread? Mosquito bites (No)". Preventive actions: "Will this action prevent spreading coronavirus to yourself and others? Shop in crowded areas like informal markets (No)". Government policy: "Is the government currently... Asking households to not visit patients infected by COVID-19 at hospitals (Yes)".

2.3.2 Outcomes

Outcomes are COVID-19 knowledge test scores: the share of knowledge questions answered correctly. Responses are considered "correct" if they match the pre-specified correct answer and are "incorrect" otherwise. At baseline, each respondent was assigned a randomized subset of 20 out of 40 questions, distributed as follows across categories: 6 (out of 12) general knowledge, 8 (out of 16) preventive action, and 6 (out of 12) government policy questions.

We pre-specified two primary outcomes: First, the Overall test score is the share of correct answers to all 40 knowledge questions asked at endline: 12 on general knowledge, 16 on preventive actions, and 12 on government actions. In the control group (N=847), this outcome has a mean of 0.781 and a standard deviation of 0.108. Second, the Teaching-Eligible test score is the share of correct answers to the 20 knowledge questions that were also asked at baseline—that is, those that were eligible for feedback via the Teaching intervention: 6 on general knowledge, 8 on preventive actions, and 6 on government actions. In the control group, this outcome has a mean of 0.784 and a standard deviation of 0.123.

Secondary outcomes include test scores for Teaching-Ineligible questions, the remainder 20 questions NOT asked of the respondent at baseline, and newly asked questions, those questions randomly not asked of the respondent at either pre-baseline or baseline.[7] We also analyze test scores for knowledge categories: general knowledge, preventive actions, and government policies.

2.4 Empirical Approach

2.4.1 Treatments

To improve COVID-19 knowledge, we designed two interventions to be implemented at the end of the baseline survey following all baseline questions: 1) "Incentive" and 2) "Teaching". Respondents were randomly assigned to one of four groups (probabilities in parentheses): Incentive alone (20%), Teaching alone (20%), both treatments ("Incentive plus Teaching" or "Joint") (20%), or a control group (40%). Randomization was stratified within 76 communities. We describe the treatments briefly below. Complete implementation protocols can be found in Appendix B.3.

Incentive treatment: We informed respondents that they would earn 5 Mozambican meticais (approx. US$0.07) for every correct response to previously-asked and newly-asked

[7] In the control group, the Teaching-Ineligible test score has a mean of 0.778 (sd=0.125) and the newly-asked test score has a mean of 0.777 (sd=0.144). The number of Newly-asked questions at endline varies randomly based on the random selection of questions in the pre-baseline survey and has these summary statistics: mean=14.4; sd=1.8; min=7; max=20.

COVID-19 knowledge questions on the endline survey. They were also told that this would allow them to earn 200 meticais (approx. US$2.80), if they answered all 40 questions correctly, in addition to their 50 meticais survey completion gift. 250 meticais is equivalent to half of the sample median pre-pandemic (February 2020) weekly household income. After endline questioning, the number of correct answers and resulting payment were automatically calculated in SurveyCTO, displayed for enumerators, read to respondents, and added to the 50 meticais survey completion gift.

Teaching treatment: We provided respondents feedback on 80% of their incorrect answers and 20% of their correct answers, on average, to COVID-19 knowledge questions from the baseline survey. Feedback consisted of reminding respondents of their answer, telling them if they were correct or incorrect, and then telling them the correct answer.[8]

Joint treatment: We informed respondents of the Incentive treatment first, then implemented the Teaching treatment.

Sample sizes by treatment condition were as follows: Incentive (N=414, 19.6% of sample), Teaching (N=418, 19.7%), Joint (N=438, 20.7%) and control group (N=847, 40.0%). In Appendix C.6, we show that attrition between baseline and endline is low (4.9%) and balanced across treatment conditions. We also show that chance imbalance between the baseline outcome and the standalone Incentive treatment is heavily concentrated in only one district, and that our results are robust excluding it. Finally, we show that baseline measure of household income, food insecurity, and presence of an older adult in the household are balanced across treatment conditions.

Randomization of the Incentive, Teaching, and Joint treatments was also stratified by two cross-randomly assigned treatments to improve social distancing as part of a separate study (Allen IV et al., 2021): 1) misperceptions correction, which updated beliefs upwards or confirmed beliefs about high rates of community support for social distancing, and 2) leader endorsement, which reported to respondents previously collected social distancing endorsements by community opinion leaders. In Appendix C.8, we present regression results showing no meaningful interactions between the social distancing treatments and this paper's treatments. We also verify that our primary treatment effect estimates are very similar when the Test Score outcome measure excludes social distancing knowledge questions, which are most susceptible to being affected by the social distancing treatments.

[8]For example, one question asks respondents whether "drinking hot tea" helps prevent COVID-19 (which it does not). If respondents correctly responded "no" to this question, they are told "For 'drinking hot tea', you chose NO. Your answer is CORRECT. The correct answer is NO. This action will NOT prevent spreading coronavirus to yourself and others." If respondents incorrectly responded "yes", responded "don't know", or refused to answer, they were told "For 'drinking hot tea', you chose YES / DON'T KNOW / REFUSE TO ANSWER. Your answer is INCORRECT. The correct answer is NO. This action will NOT prevent spreading coronavirus to yourself and others."

2.4.2 Regression

As pre-specified, we estimate the following OLS regression equation:

$$Y_{i,j,t=3} = \beta_0 + \beta_1 Incentive_{ij} + \beta_2 Teaching_{ij} + \beta_3 Joint_{ij} + \eta \mathbf{B}_{ijt} + \gamma_i + \varepsilon_{ij} \tag{2.4.1}$$

where $Y_{i,j,t=3}$ is the COVID-19 knowledge test score for respondent i in community j. $Incentive_{ij}$, $Teaching_{ij}$, and $Joint_{ij}$ are indicator variables for inclusion in each treatment group. $\mathbf{B}_{ijt}$ is a vector representing the share of correct answers to questions asked at pre-baseline and baseline, respectively.[9] γ_i are community fixed effects, and ε_{ij} is a mean-zero error term. We report robust standard errors.

Due to treatment random assignment, coefficients β_1, β_2, and β_3 represent causal effects of the respective treatments on test scores. We estimate the complementarity parameter as a linear combination of regression coefficients: $\hat{\lambda} = \beta_3 - (\beta_1 + \beta_2)$.

2.4.3 Hypotheses

We hypothesize that each treatment has a positive effect on test scores. Specifically, as pre-specified, we hypothesize that the coefficient β_1 in a regression of the Overall test score, and the coefficients β_2 and β_3 in a regression of the Teaching-Eligible test score will be positive. We adjust p-values for multiple hypothesis testing across these three coefficients.[10]

Additionally, using our estimated $\hat{\lambda}$, we test the following null hypotheses: $\lambda = -0.0265$ (the mean of expert predictions, $\tilde{\lambda}$), and $\lambda = 0$.

2.4.4 Pre-Specification

Prior to baseline data collection, we uploaded our pre-analysis plan (PAP) to the AEA RCT Registry.[11] In this paper, we report on a subset of analyses pre-specified in the PAP. In Appendix C.8, we present the "Populated PAP" for our pre-specified primary analysis. These results are substantively duplicative of and yield very similar conclusions to the primary analyses we present here in the main text.

Hypotheses related to the complementarity parameter λ were not pre-specified in the PAP. The motivations for testing them are the theoretical model's ambiguous prediction as to whether λ should be positive or negative, and the fact that the vast majority of experts

[9]The average respondent correctly answered 72.1% and 77.3% of the 20 knowledge questions at pre-baseline and baseline, respectively.

[10]We use the method of List et al. (2019), as implemented by Barsbai et al. (2020) to allow inclusion of control variables in the regression.

[11]ID Number AEARCTR-0005862 (https://doi.org/10.1257/rct.5862-1.0).

predicted that $\lambda < 0$.

2.4.5 Expert Predictions

In advance of presenting our results publicly, we surveyed subject-matter experts on their expectations of our treatment effects.[12] The expert prediction survey provided respondents with an overview of the project, specifics of each intervention, and definitions of the primary outcomes (summarizing information available in the pre-analysis plan) as well as the control group mean and standard deviation for those outcomes. The survey then asked respondents to report their prediction of each treatment effect as a percentage point difference with respect to the control group mean (positive values representing positive treatment effects, and negative values representing negative treatment effects).

Experts were asked to predict the treatment effect on test scores (fraction of questions answered correctly). For the Incentive treatment, experts were asked to predict the treatment effect on the endline test score for all 40 questions asked. For the Teaching and Joint treatments, experts were asked to to predict the treatment effect on the endline test score for the 20 knowledge questions randomly selected at baseline that were eligible the Teaching treatment.

We received expert predictions from 67 survey respondents before the survey closed on January 2, 2021. Of these, 73% of respondents were in the field of economics, 45% were faculty members (most others were graduate students), and 57% had experience working on a randomized controlled trial.

Table 2.2 summarizes the expert predictions. To be consistent with the figures and tables in this paper, we display the predictions as fractions (bounded by 0 and 1) rather than percentage points. On average, respondents expected that Incentive would increase the test scores by 0.040, Teaching would increase test scores by 0.046, and Joint would increase test scores by 0.059.

For each expert who provided predictions, we calculate the complementarity parameter implied by their predictions: Predicted Joint Effect $-$ (Predicted Incentive Effect $+$ Predicted Teaching effect). [13] We refer to the average of expert-predicted complementarity parameters as $\tilde{\lambda}$. This average is negative ($\tilde{\lambda} = -0.0265$). The vast majority of experts

[12]We released an English version of the survey on the Social Science Prediction Platform (see https://socialscienceprediction.org/ for more information) and circulated an identical Portuguese version of the survey in Mozambique that we designed and distributed on Qualtrics.

[13]This requires us to assume that the expert-predicted effect of the Incentive treatment on the test score based on all 40 questions is the same as the experts-predicted effect on the test score based on the 20 Teaching-Eligible questions. Due to random selection of the subset of 20 questions in the latter case, we view this as a reasonable assumption—experts should not have predicted a different treatment effect on a randomly selected subset of 20 questions than on the full set of 40 questions.

(80.6%) expect the interventions to be substitutes, predicting that the joint treatment effect would be less than the sum of the standalone treatment effects. There is no significant difference in predicting that the interventions are substitutes across respondents who are or are not in the field of economics, faculty members, or have worked on a randomized controlled trial.

Figure 2.2 displays probability density functions (PDFs) of the predictions. For each treatment, the vast majority of experts predicted positive effects. The mean Incentive treatment effect (β_1) is 0.040, while for Teaching (β_2) it is 0.046. Notably, the mean predicted effect for the Joint treatment (β_2) is 0.059, lower than the sum of the mean predictions for the separate Incentive and Teaching treatments (0.086): experts expect the treatments to be substitutes rather than complements.

Graphically, the expectation of substitutability can be seen in the fact that the PDF of the Joint treatment has considerable overlap with the PDFs of Incentive and Teaching. Relatedly, in the figure we also display the complementarity parameter implied by each expert's predictions. For each expert, we take their predicted Joint treatment effect and subtract the sum of their predictions for the separate Incentive and Teaching treatments. The distribution of experts' λ estimates is the gray dotted line. Most of the mass of λ estimates lies to the left of zero: 81% of experts predicted negative λ. The mean of experts' λ estimates is -0.0265. We refer to this mean as $\tilde{\lambda}$, and will test the null that our estimated $\hat{\lambda}$ equals $\tilde{\lambda}$.

2.5 Results

2.5.1 Primary Analysis

Table 2.3 presents the results from testing this paper's primary hypotheses. In Column 1, we test our first pre-specified primary hypothesis regarding the effect of the Incentive treatment on the overall test score.[14] The Incentive treatment has a positive effect, and is statistically significantly different from zero (p-val=0.0003) after multiple hypothesis testing (MHT) adjustment. The point estimate indicates a 0.020 increase, relative to the 0.781 mean control group test score. This effect is substantial in magnitude, amounting to 0.19 standard deviations of the outcome variable.

In Column 2, we test our remaining pre-specified primary hypotheses on the effect of the Teaching treatment and Joint treatment on the Teaching-Eligible test score.[15] Coefficient estimates in Column 2 indicate that the Teaching and Joint treatments each also have positive

[14]Recall the Overall test score is the share of correct answers to all 40 knowledge questions asked at endline.

[15]Recall that the Teaching-Eligible test score is the share of correct answers to the 20 knowledge questions that were also asked at baseline and hence eligible for all interventions.

effects. The point estimate on Teaching indicates a 0.0288 increase (0.23 standard deviations of the outcome variable), while the Joint treatment causes a 0.0581 increase (0.47 standard deviations). Each of these coefficient estimates is statistically significantly different from zero (p-val=0.0003 for each) after MHT adjustment.

In Column 3, we also estimate treatment effects on the Teaching-Ineligible test score.[16] The Incentive intervention, which applied to newly-asked questions, indeed maintains a significantly positive effect; however, the Teaching treatment does not, suggesting that the intervention is effective in teaching specific facts but not related information on a topic. Finally, the Joint intervention maintains a significant but smaller positive effect.

For our analysis of treatment complementarity, we choose to use results on the Teaching-Eligible test score in Column 2, which contains two of our three pre-specified treatment effects. Also, as its outcome is based on questions that were eligible for all interventions, it maximizes the comparability of treatment effects across our treatment conditions.[17] The fourth row of the table displays the estimate, $\hat{\lambda}$, of the complementarity parameter, and its standard error. In Column 2, $\hat{\lambda} = 0.0137$, indicating that the Teaching and Incentive treatments are complements, rather than substitutes. The key benchmark is the mean of the expert predictions, $\tilde{\lambda} = -0.0265$. We reject the null that $\lambda = -0.0265$ (p-val$<$0.0001).

We also display the p-value of the test that $\lambda = 0$, which is 0.1460 in Column 2. Given the standard error on $\hat{\lambda}$, we can reject at the 95% confidence level that $\lambda < -0.0048$ (in other words, we can reject all but a very small amount of substitutability between the two treatments).

We also present these results on the Teaching-Eligible test score in Column 2 graphically. In Figure 2.3, we display the estimates of the three treatment effects, Joint treatment effects implied if λ took on the values of 0 or -0.0265, and p-values of relevant tests of pairwise differences. In Figure 2.4, we present cumulative distribution functions of test scores by treatment group, showing that the Joint treatment leads to the largest rightward shift of the test score distribution.

In sum, our estimates of the complementarity parameter indicate that the Incentive and Teaching treatments exhibit much more complementarity than experts predicted. We strongly reject the high degree of substitutability predicted by experts. In addition, we reject at a

[16]Recall that the Teaching-Ineligible test score is the share of correct answers to the other 20 questions NOT asked at baseline and hence NOT eligible for the Teaching intervention. For a given respondent, the Overall test score is the average of the Teaching-Eligible and Teaching-Ineligible test scores.

[17]The Teaching treatment effect can be made arbitrarily small simply by adding larger numbers of new questions to the knowledge-measurement test that were not asked before and that therefore would not have been eligible to be taught.

marginal level of statistical significance that $\lambda = 0$.

This complementarity is also present when evaluating treatment effects on newly asked questions, building confidence that results are driven by actual learning and not merely rote memorization or experimenter demand effects. In Column 4 of Table 2.3, we run regression 3.4.2 pre-specified in our PAP as of secondary interest that replaces the outcome with the share of correct answers to endline knowledge questions that were NOT randomly asked of the respondent at either pre-baseline or baseline. Thus respondents were not previously told the answers to these questions as part of the Teaching intervention, making it less obvious what the experimenters "wanted to hear". Both the Incentive and Joint treatments have a positive effect on the newly-asked test score (statistically significant at 1% level). Additionally, we continue to reject that $\lambda = -0.0265$ (the expert prediction) at the 1% level and $\lambda = 0$ at a marginal level of statistical significance.

2.5.2 Cost-Effectiveness

We now illustrate how the relative cost-effectiveness of the treatments we study depends on λ. We describe the analysis briefly here, providing details in Appendix B.6. The key inputs are:

- Treatment effect estimates for the Incentive and Teaching treatments (β_1 and β_2). The effect of the joint treatment is then $\beta_1 + \beta_2 + \lambda$.
- Implementation costs of each treatment, per treated beneficiary (derived from actual implementation costs in this study).

We consider cost-effectiveness of each treatment, the cost per unit (1-percentage-point) increase in the test score (lower numbers are better). For a range of values of λ we display the cost-effectiveness of each treatment in Figure B.4. The cost-effectiveness of the Incentive and Teaching treatments are horizontal, because they do not depend on λ. The cost-effectiveness of the Joint treatment is a decreasing function of λ: the greater the complementarity of the two treatments, the more cost-effective is the Joint treatment.

The intersection of the Joint treatment line with the horizontal lines indicates the "breakeven" λs, above which the Joint treatment is more cost effective than the respective single treatment. Breakeven λ is -0.0250 for the Incentive treatment, and 0.0290 for Teaching. The latter number is more important overall, since the Teaching treatment is the more cost-effective of the two individual treatments. λ must be above 0.0290 for the joint treatment to be the most cost-effective of the three treatment combinations.

For reference, we also show the mean expert prediction, $\tilde{\lambda} = -0.0265$, and our empirical estimate, $\hat{\lambda} = 0.0137$. At $\hat{\lambda}$, Joint is more cost-effective than Incentive, but not as cost-effective as Teaching. Actual costs in a scaled-up program may be different from those of our study, and could yield different cost-effectiveness rankings across treatments. In Appendix B.6 we provide an example of alternative relative implementation costs that would lead Joint to be the most cost-effective at $\hat{\lambda}$.

2.5.3 Knowledge Categories

We also estimate impacts of the treatments on Teaching-Eligible and Teaching-Ineligible test scores across the knowledge categories: general knowledge, preventive actions, and government policies. Results in Table 2.4 are broadly similar to the estimates in Table 2.3 Columns 2 and 3, though treatment effects for the Incentive and Teaching interventions are heterogeneous along different dimensions.

Results for the Incentive treatment vary across knowledge category. The results suggest that the Incentive treatment was least effective at increasing general knowledge (e.g., risk factors, transmission and symptoms) and most effective at increasing knowledge on government policy. As the government's COVID-19 policies changed just prior to and during the baseline and endline surveys, one possible interpretation is that the Incentive intervention was most effective at promoting learning of relatively new or updating information.

Results for the Teaching and Joint treatment vary less across knowledge category and more so between Teaching-Eligible and Teaching-Ineligible test scores. The Teaching treatment has a significantly positive effect on all knowledge categories for Teaching-Eligible questions, but insignificant effects otherwise. The Joint treatment remains significantly positive across all regressions. The estimated complementarity parameter $\hat{\lambda}$ appears largest (most positive) for the preventive actions subcategory (Columns 2 and 5).

2.5.4 Long-Run Analysis

We further estimate the longer-run effects of the treatments over nine months later, using COVID-19 knowledge questions included in a post-endline survey that had other primary aims. This analysis was not pre-specified, so results should be considered exploratory. We briefly summarize here, providing details in Appendix B.7.

In a post-endline phone survey from July-August 2021, we asked 1,886 respondents (89.1% retention from endline, balanced across treatment conditions) 20 pre-specified questions on general knowledge and preventive actions. We excluded government policy questions because many pre-specified questions/answers were no longer true or applicable. Respondents received the standard 50 meticais survey completion gift but were offered no other incentives. We compare endline and post-endline treatment effects on two modified Test Scores of questions assessing general knowledge and preventive actions: 1) Test Score for all relevant questions asked in each round, and 2) Test Score for the same set of relevant questions across baseline, endline, and post-endline. For robustness, we analyze both outcomes, noting that each deviate from our pre-specified primary outcome due to the exclusion of government policy questions, and only draw conclusions supported by all regression specifications.

Results are in Table B.13. The Joint treatment has positive effects on long-run COVID-19 knowledge (Columns 2 and 4, statistically significant at 1% level) in both post-endline regressions. In addition, the complementarity parameter remains positive over this longer run. We continue to reject that $\lambda = -0.0265$ (the expert prediction) at the 1% level, and in addition also reject that $\lambda = 0$ (at the 5% level or better) in all specifications. These results indicate that the Joint intervention's impact, and the complementarity between Incentives and Teaching, were not merely short-run phenomena.

2.6 Conclusion

When governments and educational institutions seek to promote knowledge acquisition, two approaches are common. First, they can *teach* the knowledge in question (a "supply" educational intervention). Second, they can provide *incentives* for learners to acquire the knowledge (an educational intervention on the "demand" side). This paper is among the first to examine the *interaction* between a supply-side and a demand-side intervention to promote knowledge gains, estimating a complementarity parameter (λ).

We implemented a randomized study among Mozambican adults studying whether a teaching and an incentive treatment are substitutes or complements in promoting learning about COVID-19. Most experts surveyed in advance expected the two treatments to be substitutes ($\lambda < 0$). In reality, the two treatments exhibit much more complementarity than experts predicted: we estimate λ to be positive and statistically significantly larger than the expert prediction.

Our findings provide a key input for policy-making. We use our empirical estimates combined with actual implementation costs to rank potential treatment combinations for different values of the complementarity parameter (λ) in terms of their cost-effectiveness (cost per unit gain in knowledge). We identify a threshold value of λ, above which it makes sense to implement both the Incentive and Teaching treatments, rather than just one or the other. Our actual estimate of λ does not exceed this threshold, implying that the Joint treatment is not the most cost-effective policy; rather, the Teaching treatment is. This conclusion about relative cost-effectiveness may vary in other contexts with different implementation costs.

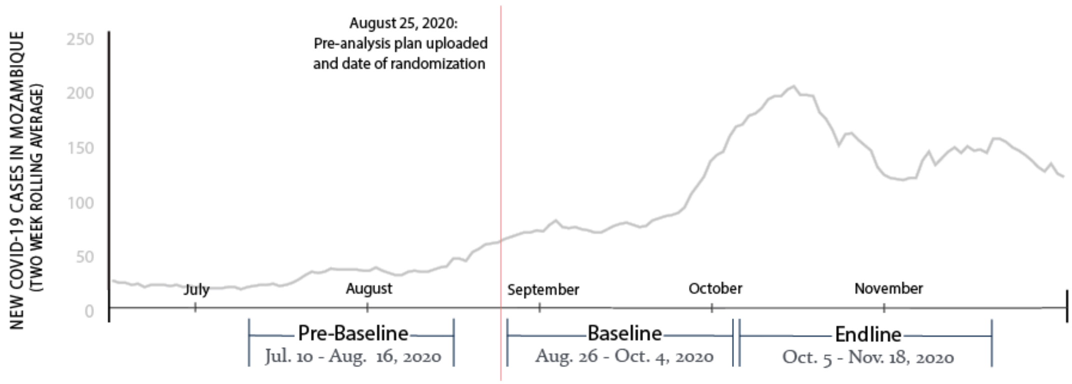
Future studies should gauge the generality of these findings. For example, they should measure the complementarity between teaching and incentive treatments in stimulating learning about other topic areas (for example, personal finance, legal rights, or agricultural techniques); motivating behavior change;[18] and in other study populations (e.g., students). It would also be valuable to examine the complementarity between other types of "demand" and "supply" interventions, particularly demand interventions that are more readily scalable than monetary payments,[19] or supply interventions that involve more actors (e.g., teachers) than our standardized enumerator-led phone-based interventions. We view these as promising directions for future research.

[18]In Appendix C.8, we find mixed and inconclusive effects on self-reported COVID-19 preventive behaviors. While disappointing, self-reported outcomes and relatively low case counts during surveying are just two reasons we are uncertain of the null results.

[19]For example, lottery tickets have been shown to promote safe sexual behavior (Bjorkman Nyqvist et al., 2018) and food vouchers have been shown to increase HIV testing (Nglazi et al., 2012).

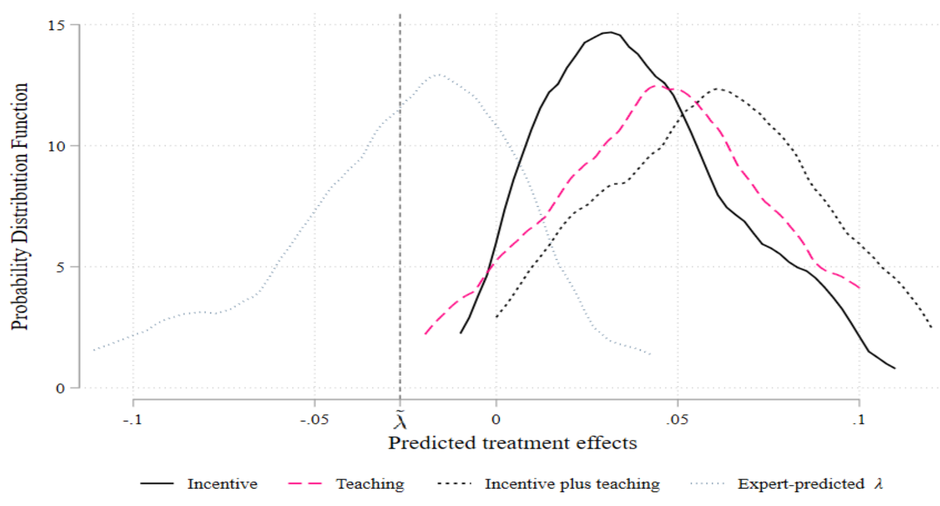
2.7 Tables and Figures

Figure 2.1: Study Timeline



Notes: Pre-analysis plan uploaded and treatments randomly assigned immediately prior to start of baseline survey, on Aug. 25, 2020. Treatments implemented immediately following baseline survey on same phone call. There was at least a three week gap between baseline and endline survey for any given study participant. Not depicted is the post-endline survey implemented between June 30 and August 30, 2021 that we use in the long-run analysis described in Section 2.5.4.

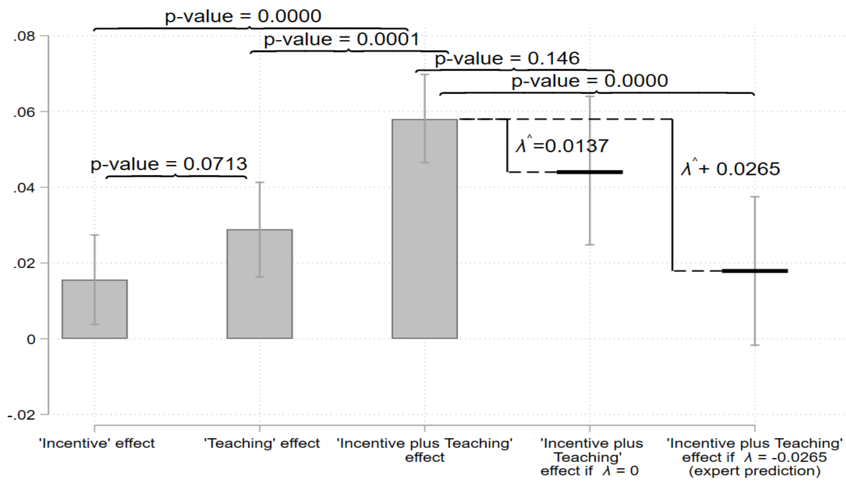
Figure 2.2: Distributions of Expert Predictions of Treatment Effects and Complementarity Parameter



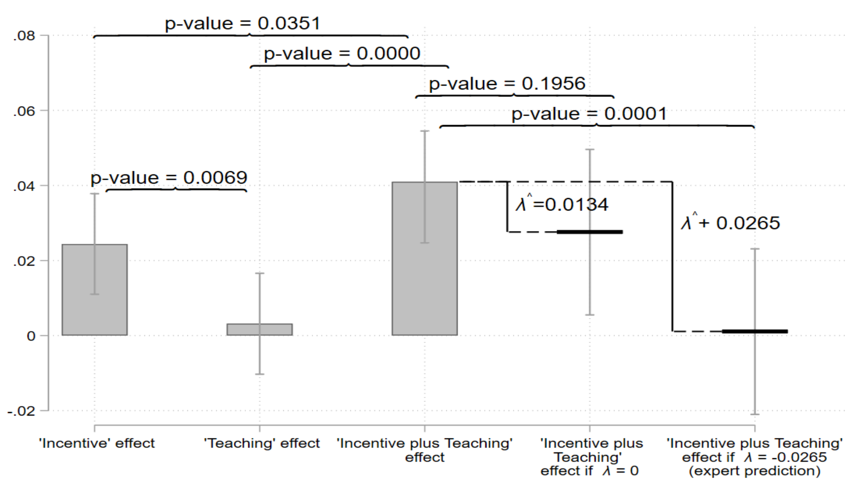
Notes: Probability density functions of predicted treatment effects of 67 experts surveyed prior to results being publicized (survey closing date Jan. 2, 2021). Experts predicted effects of "Incentive", "Teaching", and "Incentive plus Teaching" ("Joint") treatments on COVID-19 knowledge test score (fraction of questions answered correctly). Expert-predicted λ values are calculated from each expert's predictions. Mean of expert-predicted λ values is $\tilde{\lambda} = -0.0265$. Smoothing uses Epanechnikov kernel with bandwidth 0.9924.

Figure 2.3: Treatment Effects and Test of Complementarity Parameter λ Against Benchmark Values

(a) Teaching-Eligible Test Score



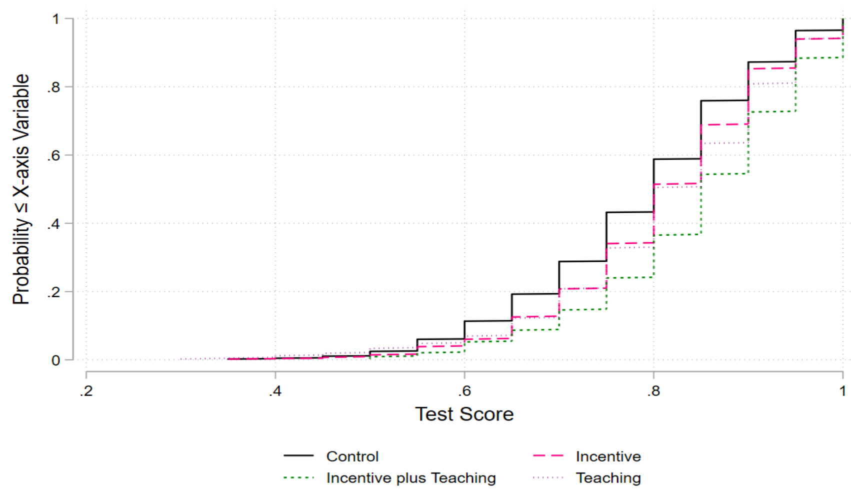
(b) Teaching-Ineligible Test Score



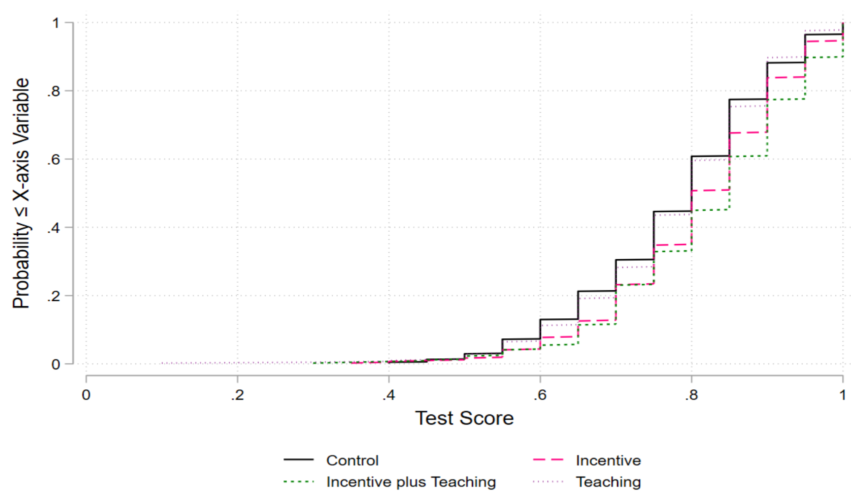
Notes: Panel (a) dependent variable on y-axis is the Teaching-Eligible test score (share of correct answers to knowledge questions asked at baseline and hence eligible for all treatments). Panel (b) dependent variable is Teaching-Ineligible test score (share of correct answers to knowledge questions NOT asked at baseline and hence NOT eligible for the Teaching intervention). Bars in first three columns display regression coefficients representing treatment effects (and 95% confidence intervals) for "Incentive", "Teaching", and "Incentive plus Teaching" ("Joint") treatments. Floating solid horizontal lines in fourth and fifth columns display "Incentive plus Teaching" ("Joint") treatment effects that would be implied by different benchmark values of complementarity parameter λ. Difference between values in 3rd and 4th columns is actual estimated complementarity parameter, $\hat{\lambda}$; the test that this difference is equal to zero tests the null that $\lambda = 0$. Difference between values in 3rd and 5th columns is difference between $\hat{\lambda}$ and mean expert prediction, $\tilde{\lambda} = -0.0265$; the test that this difference is equal to zero tests the null that $\lambda = -0.0265$.

Figure 2.4: Cumulative Distribution Functions of Test Score by Treatment Group

(a) Teaching-Eligible Test Score



(b) Teaching-Ineligible Test Score



Notes: Panel (a) dependent variable on y-axis is the Teaching-Eligible test score (share of correct answers to knowledge questions asked at baseline and hence eligible for all treatments). Panel (b) dependent variable is Teaching-Ineligible test score (share of correct answers to knowledge questions NOT asked at baseline and hence NOT eligible for the Teaching intervention). Figure displays cumulative distribution functions (CDFs) of test scores in "Control", "Incentive", "Teaching", and "Incentive plus Teaching" ("Joint") treatment groups.

Table 2.1: Test Scores and Treatment Effects Implied by Theoretical Model

Treatment	Share of Correct Answers	Boost (Versus Control)
Control	μ	0
Teaching Only	$p(0) + (1 - p(0))\mu$	$p(0)(1 - \mu)$
Incentives Only	$e^* + (1 - e^*)\mu$	$e^*(1 - \mu)$
Incentive plus Teaching (Joint)	$p(R) + (1 - p(R))e^*$ $+(1 - p(R))(1 - e^*)\mu$	$p(R)(1 - \mu) + e^*(1 - \mu)$ $-e^*p(1 - \mu)$

Table 2.2: Expert Predictions

Expert Prediction	Mean	Std. Dev.	Min	Max
Incentive Treatment Effect	0.0399	0.0256	0.0000	0.1000
Teaching Treatment Effect	0.0455	0.0307	-0.0196	0.1007
Joint Treatment Effect	0.0589	0.0296	0.0000	0.1200
Complementarity parameter (λ)	-0.0265	0.0333	-0.1108	0.0426
Indicator: Incentive and Teaching treatments are substitutes (λ¡0)	0.8060	0.3984	0.0000	1.0000

Notes: 67 experts provided predictions on the Social Science Prediction Platform (socialscienceprediction.org) prior to knowing results. Survey closing date January 2, 2021.

Table 2.3: Treatment Effects on COVID-19 Knowledge Test Scores

VARIABLES	Overall (1)	Teaching-Eligible (2)	Teaching-Ineligible (3)	Newly-asked (4)
Incentive	0.0200***	0.0156***	0.0244***	0.0209***
	(0.0054)	(0.0060)	(0.0069)	(0.0081)
	[0.0003]			
Teaching	0.0160***	0.0288***	0.0032	0.0017
	(0.0055)	(0.0064)	(0.0069)	(0.0078)
		[0.0003]		
Incentive plus Teaching	0.0496***	0.0581***	0.0410***	0.0416***
	(0.0055)	(0.0060)	(0.0069)	(0.0080)
		[0.0003]		
$\hat{\lambda}$	0.0136	0.0137	0.0134	0.0189
	(0.0084)	(0.0095)	(0.0104)	(0.0120)
Observations	2,117	2,117	2,117	2,117
R-squared	0.319	0.333	0.201	0.150
Control Mean DV	0.781	0.784	0.778	0.777
Control SD DV	0.108	0.123	0.125	0.144
p-value: $\lambda = 0$	0.1048	0.1462	0.1956	0.1145
p-value: λ = -0.0265	0.0000	0.0000	0.0001	0.0002
p-value: Incentive = Teaching	0.5292	0.0713	0.0069	0.0332
p-value: Incentive = Joint	0.0000	0.0000	0.0351	0.0235
p-value: Teaching = Joint	0.0000	0.0001	0.0000	0.0000

Notes: Column 1: COVID-19 Knowledge Overall test score, the share of correct answers to 40 knowledge questions asked at endline that were also randomly selected for the respondent to answer at baseline. Column 2: Teaching-Eligible test score, the share of correct answers to 20 knowledge questions asked at baseline and hence eligible for all treatments. Column 3: Teaching-Ineligible test score, the share of correct answers to 20 knowledge questions NOT asked at baseline and hence NOT eligible for the Teaching intervention. Column 4: Newly-asked test score, the share of correct answers to the 20 or fewer endline knowledge questions that were NOT randomly asked of the respondent at either pre-baseline or baseline. λ is the complementarity parameter (see Section 3.2). "$\hat{\lambda}$" is coefficient on "Incentive plus Teaching" ("Joint") minus sum of coefficients on "Incentive" and "Teaching". All regressions include community fixed effects and controls for pre-treatment (pre-baseline and baseline) test scores. Robust standard errors in parentheses. Significance levels in Columns 1 and 2 adjusted for multiple hypothesis testing across the three coefficients estimated (on Incentive, Teaching, and Joint treatments); p-values adjusted for multiple hypothesis testing in square brackets. *** p<0.01, ** p<0.05, * p<0.1.

Table 2.4: Regression of Test Score (TS) Categories on Treatments

	Teaching-Eligible Test Scores			Teaching-Ineligible Test Scores		
VARIABLES	General	Preventive	Government	General	Preventive	Government
	(1)	(2)	(3)	(4)	(5)	(6)
Incentive	0.0018	0.0118	0.0419***	0.0174	0.0249***	0.0422***
	(0.0099)	(0.0088)	(0.0099)	(0.0112)	(0.0090)	(0.0110)
Teaching	0.0265***	0.0234**	0.0299***	0.0044	0.0016	0.0146
	(0.0102)	(0.0092)	(0.0109)	(0.0111)	(0.0095)	(0.0111)
Incentive plus Teaching	0.0415***	0.0535***	0.0749***	0.0336***	0.0439***	0.0538***
	(0.0103)	(0.0087)	(0.0100)	(0.0106)	(0.0094)	(0.0108)
$\hat{\lambda}$	0.0133	0.0183	0.0031	0.0118	0.0173	-0.0030
	(0.0157)	(0.0136)	(0.0154)	(0.0166)	(0.0141)	(0.0165)
Observations	2,117	2,117	2,117	2,117	2,117	2,117
R-squared	0.206	0.257	0.189	0.117	0.080	0.139
Control Mean DV	0.797	0.827	0.789	0.782	0.710	0.790
Control SD DV	0.189	0.170	0.202	0.191	0.157	0.202
p-value: Incentive = Teaching	0.0354	0.276	0.309	0.313	0.0289	0.0268
p-value: Incentive = Joint	0.000845	3.64e-05	0.00254	0.193	0.0732	0.344
p-value: Teaching = Joint	0.213	0.00365	0.000135	0.0182	0.000110	0.00143

Notes: Columns 1-3: the Teaching-Eligible test scores for knowledge categories, the share of correct answers at endline to the 6 questions on general knowledge, 8 questions on preventive actions, and 6 questions on government policy, respectively. Columns 4-6: the Teaching-Eligible test scores for knowledge categories, the share of correct answers at endline to the 6 questions on general knowledge, 8 questions on preventive actions, and 6 questions on government policy, respectively. λ is the complementarity parameter (see Section 3.2 of main text). $\hat{\lambda}$ is coefficient on "Incentive plus Teaching" (Joint) minus sum of coefficients on "Incentive" and "Teaching". All regressions also include community fixed effects and controls for pre-treatment (Rounds 1 and 2) Test Scores. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

CHAPTER III

Correcting Misperceptions about Support for Social Distancing to Combat COVID-19

Co-authored with Arlete Mahumane, James Riddell IV, Tanya Rosenblat, Dean Yang, and Hang Yu

3.1 Introduction

Attitudes toward social distancing have changed rapidly during the pandemic (Janzwood, April 27, 2020). During such rapid change, people often underestimate support for social distancing in their communities. Early in the pandemic, 98% of our Mozambican sample thought that people should be social distancing, but estimated that only 69% of others in the community felt similarly. This gap motivates a public health policy: simply inform people of high rates of community support for social distancing. What impact would such messaging have on social distancing behavior?

In theory, the impact of such a "misperceptions correction" intervention on social distancing is ambiguous: on the one hand, informing people that more neighbors support social distancing than expected encourages *free-riding* and *lowers* the perceived benefits from social distancing. On the other hand, people should revise their belief about the seriousness of COVID-19 upwards in order to rationalize the observed number of infections in their neighborhood despite the higher than expected social distancing support. This *perceived infectiousness* effect *increases* the perceived benefits from social distancing and dominates free-riding in communities with high levels of infections.[1] Finally, the *norm adherence* effect should induce people to follow whatever local social norm is set by their neighbors - in our case this effect should always increase social distancing.

[1]Our model is related to the literature on decision-making under misspecified subjective models (Spiegler, 2020). Agents hold incorrect assumptions on one model parameter (e.g., share of population social distancing), leading them to incorrect conclusions about other parameters (e.g., disease infectiousness).

We implemented a randomized controlled trial testing the impact of informing people about high local support for social distancing. The treatment either updated beliefs upwards or confirmed beliefs about high rates of support for social distancing. Abiding by COVID-19 protocols, we conducted all treatments and surveys by phone among 2,117 Mozambican households.

Our outcome variable is the extent to which a household engages in social distancing. Measuring this behavior is challenging due to experimenter demand effects.[2] Yet most prior studies ask for self-reports about general social distancing compliance. When we do so, 95% claim to observe government social distancing recommendations. We therefore construct a novel measure of social distancing. First, we ask respondents to self-report several social distancing actions. Second, we ask *others* in the community to report on the respondent's social distancing. We are aware of no prior study that makes use of other-reports on a respondent's social distancing behavior. Incorporating self-reported actions and others' reports drops social distancing to a more discerning 8% (see Figure 3.1 and Section 3.3.3). Improved measurement leads the social distancing rate to fall by an order of magnitude.

The average effect of the misperceptions correction treatment in the full sample is small and not statistically significantly different from zero. However, as theory predicts, there is substantial treatment effect heterogeneity: the treatment effect is statistically significantly more positive when local COVID-19 cases (per 100,000 population) are higher. In districts with few cases, the treatment effect is negative. In the district with the most COVID-19 cases, the treatment increases social distancing by 9.2 percentage points (statistically significant at the 5% level), a 70% increase over that district's control-group mean.

This pattern is consistent with the theoretical prediction that as infection rates rise, the perceived-infectiousness effect should increasingly dominate the free-riding effect of the misperceptions correction treatment, leading the treatment effect to become more positive. We also test a further implication of the model: expectations of future infection rates should show similar treatment effect heterogeneity. Empirical analyses confirm this prediction, providing additional support for the theoretical model.

This paper contributes to understanding the impact of providing information about others' beliefs and attitudes (Benabou and Tirole, 2011; Bicchieri and Dimant, 2019). In health settings, Yu (2020) and Yang et al. (2021) find (in an overlapping Mozambican sample) that correcting overestimates of stigmatizing attitudes promoted HIV testing, though Banerjee et al. (2019b) find that informing Nigerian young adults of peers' attitudes on healthy sexual relationships did not change respondents' own attitude.[3] Regarding social distancing,

[2]Jakubowski et al. (2021) find that self-reported mask wearing is overstated relative to measures based on observations of others.

[3]In other contexts, correcting misperceptions of community support or approval (i.e., the injunctive norm)

Martinez et al. (2021) show that respondents are influenced by others' social distancing actions in hypothetical vignettes; however, no prior study has tested the impact of providing information on community support of social distancing on respondent behavior.

Our emphasis on interactions between free-riding and perceived-infectiousness effects is novel, but each effect has been studied separately. Free-riding has been studied in the context of vaccination decisions (Hershey et al., 1994; Lau et al., 2019) and social distancing (Cato et al., 2020) and in similar Mozambican settings Fafchamps et al. (2020). Perceived COVID-19 infection risk (e.g., due to vaccine anticipation, Andersson et al. (2021)) has been shown to lower social distancing intentions.

3.2 Theory

Our model focuses on the interaction between the free-riding and perceived infectiousness effects for communities with low and high overall infection rates. We view norm-adherence as a uniform effect that should always increase social distancing.

We consider a community where people have random pairwise meetings. People believe that a share x of the population supports social distancing and that the probability of becoming infected from unprotected meetings is α. People treat x as given, but infer the infectiousness α from the current infection rate R in the community which they can observe (we describe this inference below). The true infectiousness of the disease is $\hat{\alpha}$.

Importantly, people in the community have *miscalibrated beliefs*: the true share of the population supporting social distancing is $\hat{x}$ (we are interested in the case $\hat{x} > x$). People infer the true infectiousness $\hat{\alpha}$ of the disease only if they are correctly calibrated ($\hat{x} = x$).

Individual Effort A supporter engages in preventative effort e and assumes that other supporters choose effort e^* (in equilibrium we have $e = e^*$). Non-supporters choose effort $e = 0$.

When someone supporting social distancing meets another person, she escapes exposure with probability:

$$\begin{aligned} A(e, e_{other}) &= \sqrt{e + e_{other}} \\ &= \begin{cases} \sqrt{e + e^*} & \text{if other person is supporter} \\ \sqrt{e} & \text{if other person is non-supporter} \end{cases} \end{aligned} \tag{3.2.1}$$

has also been shown to change energy consumption (Schultz et al., 2007), female labor force participation (Bursztyn et al., 2020), donations to charities addressing climate change (Andre et al., 2021), and recycling program participation (Fuhrmann-Riebel et al., 2023).

Hence, the marginal benefit of effort decreases both with own effort e as well as the other person's effort e^*.[4]

The expected probability of escaping exposure is therefore:

$$\overline{A}(e, e^*) = (1 - x)\sqrt{e} + x\sqrt{e + e^*} \tag{3.2.2}$$

An agent becomes exposed with probability $1 - \overline{A}(e, e_{other})$. If exposed she gets infected with probability α and suffers disutility $-C$ from infection.[5] If she is not exposed then she does not get infected. Her baseline utility from no infection equals $\overline{U}$. The cost of preventative effort is e. Hence, her total utility equals:

$$\overline{U} - \alpha(1 - \overline{A}(e, e_{other}))C - e \tag{3.2.3}$$

The agent chooses e to maximize her utility, giving us the following first-order condition:

$$\frac{\alpha C}{2\sqrt{e}}\left[1 - x(1 - \frac{1}{\sqrt{1 + \frac{e^*}{e}}})\right] = 1 \tag{3.2.4}$$

In equilibrium it has to be the case that the population effort e^* equals e. Hence, we can characterize equilibrium effort as:

$$e = \left(\frac{\alpha C}{2}\left[1 - x(1 - \frac{1}{\sqrt{2}})\right]\right)^2 \tag{3.2.5}$$

This demonstrates the *free-riding effect*: increasing the share x of supporters *decreases* effort because the marginal utility from own effort decreases. Also, effort increases if the disease is more infectious (higher α) and if illness is costlier (higher C).

Infection Rate People observe the current infection rate in the community. Infections come from two sources: non-supporters become sick at rate $\alpha\left(1 - x\sqrt{e}\right)$ while supporters become sick at rate $\alpha(1 - \overline{A}(e, e))$. Hence, people in the community assume that the current

[4] We assume the other person's effort is unobservable. This is consistent with our finding that respondents underestimate the extent of social distancing.

[5] For simplicity, we assume that infectiousness does not vary with the agent's type (supporter or non-supporter). Otherwise, we would need to keep track of two levels of infectiousness. The qualitative results would not change.

infection rate is generated by the following process:

$$\begin{aligned} R &= \alpha \left[\underbrace{(1-x)(1-x\sqrt{e})}_{\text{non-supporters}} + \underbrace{x\left(1-\sqrt{e}(1+(\sqrt{2}-1)x)\right)}_{\text{supporters}} \right] \\ &= \alpha \left[1 - \sqrt{e}\, \underbrace{2x\left(1-x\left(1-\frac{1}{\sqrt{2}}\right)\right)}_{=G(x)} \right] \end{aligned} \tag{3.2.6}$$

However, the true process determining current infections is actually:

$$R = \hat{\alpha}\left[1 - \sqrt{e}G(\hat{x})\right] \tag{3.2.7}$$

In other words, the true infection process is driven by the same social distancing effort of supporters but different infectiousness $\hat{\alpha}$ and different $\hat{x}$.

3.2.1 Basic Equilibrium

Supporters initially assume that the disease has low infectiousness and they adjust their estimate of α upwards until the current infection rate R stabilizes.

Proposition 2. *In equilibrium, effort level e, the current infection rate R, and the assumed infectiousness α satisfy Equations 3.2.5, 3.2.6 and 3.2.7. Moreover, $\hat{\alpha} > \alpha$ if $\hat{x} > x$.*

In equilibrium, both the assumed infection process (Equation 3.2.6) and the real infection rate (Equation 3.2.7) must produce observed infection rate R. For the second part, note that $G(x)$ is increasing in $x \in [0, 1]$: hence, $\hat{x} > x$ implies $\hat{\alpha} > \alpha$ to generate the same infection rate R.

3.2.2 Treatment Effect

We now consider the effect of our treatment informing people that the population share supporting social distancing is really $\hat{x} > x$.

Proposition 2 implies that if supporters are informed that the true population share supporting social distancing is $\hat{x} > x$, they must infer higher disease infectiousness than they initially assumed (because their estimated disease infectiousness immediately jumps from α to true $\hat{\alpha}$). This is the *perceived-infectiousness effect.*

Supporters of social distancing will adjust their effort level to a new level $\hat{e}$, but there are two countervailing effects:

1. Holding assumed infectiousness α constant, the free-riding effect *decreases* effort.

2. The perceived-infectiousness effect *increases* effort, because the agent now believes the disease is more infectious than initially thought (perceived α increases), increasing the gain from social distancing.

Intuitively, the perceived-infectiousness effect varies monotonically with R: when infections are low, supporters' effort is low, and both supporters and non-supporters get infected at similar rates. Hence, agents revise the estimate of infectiousness α only slightly upwards in response to the treatment. On the other hand, when infections are high, supporters' effort is high and the upward revision will be larger.

The following theorem makes this intuition precise. Instead of doing comparative statics on R (which is determined in equilibrium) we state the comparative statics results in terms of the infectiousness $\hat{\alpha}$ (for given x and $\hat{x}$). Note that R increases with $\hat{\alpha}$.

Theorem 2. *Assume an agent is informed that a share $\tilde{x} > x$ of the population supports social distancing. Then there is a threshold $\hat{\alpha}^*$ such that for any $\hat{\alpha} < \hat{\alpha}^*$ the free-riding effect dominates and equilibrium effort decreases, and for $\hat{\alpha} > \hat{\alpha}^*$ the perceived-infectiousness effect dominates and the equilibrium effort increases.*

See Appendix C.1 for the proof.

The interplay between free-riding and perceived-infectiousness effects also yields analogous predictions about a central belief about COVID-19: the future infection rate. In the endline survey, we ask respondents to estimate this. The expected future rate differs from the current infection rate R, because this study occurs at a point when infection rates are clearly evolving. The misperceptions correction treatment changes respondent beliefs about social distancing support and about infectiousness, and therefore should change expected future infection rates. Recall that non-supporters are always infected with higher probability than supporters. The higher the infectiousness parameter $\hat{\alpha}$, the higher should be future infection rates for both groups. When $\hat{\alpha}$ is currently small, the perceived-infectiousness effect is small. Simultaneously, the treatment corrects beliefs about the share of social-distancing supporters upwards, which should *reduce* estimates of future infection rates because supporters have lower infection rates. Thus, the expected future infection rate *decreases* when $\hat{\alpha}$ is currently small. In contrast, when $\hat{\alpha}$ is currently large, the treatment leads to a large increase in perceived infectiousness, implying that the disease will infect higher shares of both supporters and non-supporters. This will tend to *increase* expected future infection rates.

To summarize, the misperceptions correction treatment effect on the expected future infection rate should show heterogeneity similar to that described in Theorem 2. The treatment effect on the expected future infection rate is strictly negative if the current local infection rate (R) (which moves monotonically with $\hat{\alpha}$) is small enough. The treatment effect on the expected future infection rate increases with the current infection rate, and can become positive if current infection rates are sufficiently high.

In our empirical analyses, we test these predictions regarding heterogeneity in the misperceptions correction treatment effect.

3.3 Sample and Data

3.3.1 Data

We implemented three rounds of surveys by phone in July-November 2020: a pre-baseline, baseline and endline survey (see Figure C.2 for a study timeline). Respondents were drawn across 76 communities in Central Mozambique from a sample of a prior study (Yang et al., 2021) that focused on HIV-vulnerable households—a policy-relevant sample especially vulnerable to COVID-19.[6] To avoid risk of spreading COVID-19 via in-person interaction with study participants, we also limited the sample to those households with phones. Thus both HIV-vulnerability and phone ownership are two relevant factors to bear in mind when considering the external validity of the results. We surveyed one adult per household. Appendix C.2 provides details on the COVID-19 context, study communities and the study timeline.

Between a pre-baseline survey and baseline survey, we randomly assigned households to treatments and registered a pre-analysis plan (PAP). The baseline survey was immediately followed by over-the-phone treatment implementation. There was a minimum of 3.0 weeks and average of 6.3 weeks between baseline and endline surveys for all respondents. Baseline and endline surveys occurred when COVID-19 cases were rising rapidly.

The endline sample size is 2,117 respondents, following a sample size of 2,226 at baseline. The retention rate between baseline and endline is 95.1% overall, at least 94.4% in each of the seven districts surveyed, and balanced across treatment conditions. We also surveyed 145 community opinion leaders over the 76 study communities—at least one, an average of

[6]AEA RCT Registry for Yang et al. (2021): https://doi.org/10.1257/rct.3990-5.1. In that prior study, we run a randomized evaluation of a bundled community-level HIV/AIDS program whose main component was home visits by case care workers to promote HIV testing to HIV-vulnerable households, such as those with HIV-positive or other chronically ill members, orphaned children, or a grandparent as the household head. In this study, we use community-stratified randomization and regress with community fixed effects to rule out the influence of this prior intervention on our results.

2.09, and at most 4 per community—for inputs to the primary outcome and treatments as described below.

3.3.2 Measuring Misperceptions

We measure both true and perceived support for social distancing as follows. First, to measure actual community support for social distancing, we asked respondents *"Do you support the practice of social distancing to prevent the spread of coronavirus? (Yes, No, Don't know, Refuse to Answer)"*, which captures the respondent's first-order belief of the injunctive norm for social distancing. We then calculated the fraction of "Yes" responses across the sample and within each community. Directly after, to measure perceived community support, we asked respondents *"For every 10 households in your community, how many do you think support the practice of social distancing to prevent the spread of coronavirus? (integer 0-10)"*, capturing the respondent's second-order beliefs of the injunctive norm for social distancing within their community. The difference between the true and perceived community support for social distancing is the respondent's misperception of the social norm.

Three possible concerns with our measure of perceived support for social distancing include the role of uncertainty, the restricted scale, and bias from experimenter demand effects. First, a possible concern is that unawareness and uncertainty around new social norms and others' beliefs—plausible at the start of the pandemic—may lead respondents away from the extreme points of the answer scale. However, in Appendix C.3, we present a cumulative distribution of our perceived community support measure across survey rounds that shows that respondents readily utilized the extreme ends of the scale, with 8% and 35% of the sample at pre-baseline reporting perceived community support of 0% and 100%, respectively, and 51% of the sample at baseline reporting 100%. Second, despite more common use of a 0-100 scale when measuring perceived norms (e.g., Andre et al. (2021); Fuhrmann-Riebel et al. (2023)), we simplified our scale to an 11-point 0-10 scale due to past difficulties eliciting "percentage" measures in this context, repeated feedback from our field team and local partners that a 0-100 scale was too complex, and the inability to use a "slider" mechanism over the phone. Given the high concentration of perceived community support at 100% at baseline, the restricted scale may attenuate the treatment effect of the misperceptions correction on perceived community support given that there is "little room to improve" for many respondents in the sample. Third, experimenter demand effects may have led respondents to report higher shares of perceived support for social distancing in order to make their communities look favorable. Such action would lead to an upward-biased estimate of true perceptions of community support and, in turn, an underestimate of the misperception of the social norm, which would also lead to an attenuation of the treatment effect for the misperceptions correction intervention.[7] We ask the reader to bear in mind these possible limitations when interpreting the results.

3.3.3 Primary Outcome

The primary outcome is an indicator that the respondent practiced social distancing, as pre-specified in our PAP. It is constructed from self-reports of social distancing as well as others' reports of the respondent's social distancing. The outcome is equal to one if the respondent is practicing social distancing according to both self-reports and other-reports, and zero otherwise.

Respondents are social distancing according to their self-report if both of the following are true: 1) they answer "yes" to "In the past 14 days, have you observed the government's recommendations on social distancing?", and 2) they report doing at least seven out of eight "social distancing actions" in the past seven days (higher than the sample median number, six).[8] A list of the social distancing actions and their corresponding summary statistics are presented in Appendix C.4.

To collect others' reports on a respondent's social distancing, study participants were asked about their social interactions with ten other community study participants. These ten others were identified from social network data and geographic proximity. Additionally, community leaders were also asked about social interactions with all study participants in their respective community.[9] At baseline, the average respondent household was known by 0.98 community leaders and 3.21 neighboring survey respondents. Other-reports were collected at baseline and endline.

In collecting other-reports, we asked others whether they had seen anyone from the respondent household in the last 14 days.[10] If so, we then asked: 1) Did he/she come closer than 1.5 meters to you or others not of his/her household at any point in the last 14 days?; 2)

[7]See Section 3.4.1 for description of how the misperceptions correction treatment. If upward-biased estimates of perceived support remain less than or equal to true community support, then the misperceptions correction is implemented and may still boost respondents' true perception of community support in a way not captured by our measure; however, if the bias leads to overestimating true community support, then respondents will become ineligible for the misperceptions correction treatment thereby attenuating the treatment effect (but not biasing upwards).

[8]While this threshold was pre-specified, regression results (in Appendix C.7.3) are robust to alternate definitions of this component, such as a threshold of six, or dropping social distancing actions 4 and 6 for which respondents might misinterpret and answer "no" if not showing symptoms.

[9]The average community leader was asked about 33.90 households (std. dev.=22.10, minimum=2, second-highest=99, maximum=228—a special case where one individual was the traditional leader across multiple communities). To mitigate survey fatigue, leaders were told upfront of the number and offered a stepwise incentive that increased for each additional set of 25 study households.

[10]As is common in this context, households were identified by the name of the household head and a list of other known household members.

Did he/she shake hands, try to shake hands, or touch you or others not of his/her household in the last 14 days?; and 3) In general, did he/she appear to be observing the government's recommendations on social distancing (avoid large gatherings and keep at least 1.5 meters distance from people not of his/her household)? Respondents are considered to be social distancing according to others if all others responded "no", "no", and "yes" (respectively) to these three questions, reported having not seen the respondent in the past 14 days, or reported not knowing the respondent.[11]

Figure 3.1 displays how these questions lead to the social distancing outcome. First, 95% of respondents say "yes" to the self-report on general social distancing. When considering self-reports of doing at least seven out of eight social distancing actions, the social distancing rate falls to 36%. Finally, incorporating others' reports reduces the rate further to 8%. Limited overlap between self-reports and others' reports of social distancing suggests that each is providing different sets of information. We suspect that self-reports likely over-report social distancing due to experimenter demand bias, whereas others' reports are likely less biased by experimenter demand and rather over-report due to recall bias or lack of observation (as respondents not known or not seen in the past 14 days were not assumed to violate social distancing behavior).[12] Together, we believe the combined measure is a novel improvement from simple self-reports, though we leave comparison of both measurement methods to observed behavior as an avenue for future work. Incorporating additional information into the social distancing measure—using self-reports of specific social distancing behaviors as well as other-reports—leads to substantially lower social distancing rates.

3.4 Research Design

3.4.1 Treatments

We implemented a randomized controlled trial estimating impacts on social distancing of two treatments: 1) misperceptions correction, and 2) leader endorsement. Before the baseline survey, we randomly assigned 30% of households completing the pre-baseline survey each to one of two treatments and the remaining 40% to a control group. Sample sizes by treatment condition were as follows: misperceptions correction (N=628, 29.7% of sample), leader endorsement (N=637, 30.1%), and control group (N=852, 40.3%). Treatment scripts are located in Appendix C.5.

For the misperceptions correction treatment, we used the following data: 1) respondents'

[11] At baseline, 90.55% of respondent households were known by some other respondent or community leader.

[12] For example, complete lack of observation by others was true for 9% of the sample (see footnote 11).

own support for social distancing from the pre-baseline survey, from which we estimated the true share of community support for social distancing (as the fraction of respondents expressing support within the community), and 2) respondents' perceived share of community support for social distancing at baseline (reported as an integer out of 10). Immediately after completing the baseline survey, treated individuals underestimating the share were told the true share supporting social distancing, rounded to an integer out of 10.[13] Treated individuals correctly estimating the share were also told that they were correct. In practice, 92.4% of treated respondents received this treatment, 53.2% of whom underestimated community support for social distancing and 46.8% of whom correctly estimated it. The small minority overestimating the share were not provided additional information.[14]

For the leader endorsement treatment, we identified and surveyed community opinion leaders prior to the baseline survey and requested their permission to tell others in their community that they "support social distancing, are practicing social distancing, and encourage others to do the same". Then, in this treatment, we reported this endorsement to respondents, mentioning the community leader(s) by name.[15]

Attrition between baseline and endline is low (4.9%). In Appendix C.6, we show that attrition and key baseline variables are balanced across treatment conditions. Further, at endline, 97.9% recall receiving the baseline survey and, of those, 99.4% report trusting the COVID-19 information we provided.[16]

3.4.2 Regressions

A pre-specified ordinary-least-squares regression equation provide treatment effect estimates:[17]

$$Y_{ijd} = \beta_0 + \beta_1 T1_{ijd} + \beta_2 T2_{ijd} + \eta B_{ijd} + \delta_{ijd}^{others} + \delta_{ijd}^{leaders} + \gamma_{jd} + \varepsilon_{ijd} \qquad (3.4.1)$$

where Y_{ijd} is the social distancing indicator for respondent i in community j and district d; $T1_{ijd}$ and $T2_{ijd}$ are indicator variables for the misperceptions correction and leader endorsement treatment groups, respectively; B_{ijd} is the baseline value of the dependent variable; δ_{ijd}^{others} is a vector of dummy variables for the number of other respondents who

[13]In 63 out of 76 communities (82.9%) the number we convey to respondents is 10 out of 10, and in 13 communities (17.1%) the number is 9 out of 10.

[14]While respondents were not incentivized to truthfully guess community support (for scalability), true beliefs can still be updated for all except those who overestimated true community support with an upward biased guess; however, the latter case should only attenuate our treatment effect and not bias it upward.

[15]Communities had at least one and an average of 2.09 endorsements from community leaders (std. dev.=0.94, maximum=4).

[16]Trust may have arisen from multiple in-person household surveys since 2017 (see Yang et al. (2021))

[17]Appendix C.7.1 shows that all conclusions are robust to logit and probit specifications.

report knowing the respondent's household from 0 to 8; $\delta_{ijd}^{leaders}$ is a vector of indicators for the number of community leaders who report knowing the respondent's household from 0 to 4;[18] γ_{jd} are community fixed effects; and ε_{ijd} is a mean-zero error term. We report robust standard errors.[19]

Coefficients β_1 and β_2 represent the intent-to-treat impacts of the misperceptions correction and leader endorsement treatments (respectively) on social distancing.

We modify Equation 3.4.1 to estimate heterogeneity in treatment effects with respect to local COVID-19 case loads:

$$\begin{aligned} Y_{ijd} = \beta_0 + \beta_1 T1_{ijd} + \beta_2 T2_{ijd} + \beta_3(T1_{ijd} * Covid_d) + \beta_4(T2_{ijd} * Covid_d) \\ + \eta B_{ijd} + \delta_{ijd}^{others} + \delta_{ijd}^{leaders} + \gamma_{jd} + \varepsilon_{ijd} \end{aligned} \tag{3.4.2}$$

Equation 3.4.2 adds interactions between treatment indicators and the cumulative number of district-level COVID-19 cases per 100,000 population at the start of the endline survey.[20] Coefficients β_1 and β_2 in Equation 3.4.2 now represent the impacts of the treatments in districts where COVID-19 cases are zero (slightly out of sample); β_3 and β_4 represent the change in the respective treatment effect for a one-unit increase in district-level COVID-19 cases per 100,000 population.

3.4.3 Hypotheses

We pre-specified the hypothesis that each treatment (β_1 and β_2 in Equation 3.4.1) would have positive effects. Subject-matter experts (surveyed without knowing results) concurred with this expectation.[21] The mean expert predictions were that the misperceptions correction and leader endorsement treatments would increase social distancing by 5.23 and 5.56 percentage points, respectively.

We also test the hypotheses that the impact of the misperceptions correction treatment on social distancing and on the expected future infection rate will be greater in areas with a higher current COVID-19 infection rate (β_3 in Equation 3.4.2 will be positive). We did not pre-specify these hypotheses, but advance them on the basis of our theoretical model.

[18] As pre-specified, we cap δ_{ijd}^{others} at the first integer that covers over 90% of the sample, and $\delta_{ijd}^{leaders}$ at the maximum number of leaders found in any community.

[19] Appendix C.7.2 shows that clustering standard errors by the 76 communities or 7 districts has minimal impact on standard errors and does not affect whether any coefficients are statistically significant at conventional levels.

[20] The main effect of $Covid_d$ is absorbed by γ_{jd}.

[21] Predictions by 71 individuals provided at https://socialscienceprediction.org/ (survey closing date January 2, 2021).

3.5 Results

3.5.1 Pre-Treatment Descriptives

Table B.4 presents pre-treatment summary statistics of social distancing support, perceptions and behavior in the first six months of the COVID-19 pandemic. First, we document a large and statistically significant gap between actual and perceived support for social distancing: at both pre-baseline and baseline, over 97% of respondents support social distancing; however, respondents underestimate the community share expressing such support, on average estimating 69% in a pre-baseline survey and 80% at baseline. Second, we observe a large and statistically significant 11 percentage point increase in the perceived share of community support between pre-treatment survey rounds, consistent with the idea that misperceptions for new public health behaviors are most prevalent at the start of the public health crisis and then diminish over time as social networks share information. Third, despite increases in reported and perceived support for social distancing, we see small decreases in self-reported social distancing behavior; in the theoretical model, this behavior is predicted where the current local infection rate is low, as was indeed the case for all study communities prior to the endline survey.[22]

3.5.2 Average Treatment Effects

In Table 3.2 Column (1), we present regression estimates for our primary outcome.[23] Both treatment coefficients are small in magnitude and neither is statistically significantly different from zero. These findings diverge from expert predictions of treatment effects. We strongly reject the null that our T1 and T2 treatment effect estimates are equal to the positive mean expert predictions (p-value$<$0.001 in each case).

However, we find the misperceptions correction has a positive effect on measures of perceived community support for social distancing. Analyses presented in Appendix C.3 (not pre-specified) shows that the treatment effect is concentrated on the lower end of the distribution, having a significant positive effect on a respondent perceiving that at least 50% of households in their community support social distancing.

3.5.3 Treatment Effect Heterogeneity

In Table 3.2 Column (2), we present regression estimates of treatment effect heterogeneity (Equation 3.4.2) with respect to the local infection rate, measured as COVID-19 cases per

[22] See Figure C.2 to see relatively low levels of new COVID-19 cases in Mozambique during the pre-baseline and baseline relative to the endline survey.

[23] The complete set pre-specified analyses are presented in Appendix C.8.

100,000 population in the respondent's district.

The misperceptions correction treatment effect is heterogeneous with respect to local COVID-19 cases. The coefficient on the interaction term with $T1_{ijd}$ is positive and statistically significant at the 1% level. The coefficient on the $T1_{ijd}$ main effect is the predicted effect of misperceptions correction in a district with zero cases (slightly out of sample), and suggests that the misperceptions correction would reduce social distancing by 3.4 percentage points in such a location (statistically significant at the 5% level).

Figure 3.2 displays this treatment effect heterogeneity. We plot district-specific treatment effects (estimating Equation 3.4.1 separately in each of seven districts) on the y-axis (with 95% confidence intervals) against district case counts on the x-axis. In the six districts with the lowest case counts, coefficients are negative. By contrast, in Chimoio, the district with the most cases (39.08/100,000) that also accounts for one-quarter of the sample, we estimate a large positive effect: 9.2 percentage points—a 70% increase over that district's control group (statistically significant at the 5% level).

This heterogeneous treatment effect holds up to various robustness checks (presented in Appendix C.7). First, we run logit and probit specifications of the primary results. Second, we cluster standard errors by community and district. Third, we vary the threshold by which self-reported "social distancing actions" were incorporated in the social distancing indicator. Fourth, we test four alternative measures of the local COVID-19 infection rate, including the simple case count (not per capita) and high-case-count indicators, to show that the treatment effect heterogeneity is not unique to our preferred measure. Fifth, we exclude the top-COVID-19 and largest-sampled district of Chimoio to verify that it alone is not driving our results. In all cases, we find that our primary results are very similar.

By contrast, the leader endorsement treatment effect is not heterogeneous with respect to local case loads. The coefficient on the corresponding interaction term in Column (2) is small in magnitude and not statistically significantly different from zero. Some reasons why this treatment may not be effective, even when COVID-19 cases are high, include limited familiarity of leaders among all community members or limited confidence that the leader's endorsement reflected true beliefs rather than political "lip service". Coupled with findings from Banerjee et al. (2019a) on gossips spreading information, the result suggests that network-central individuals may be effective at transmitting information but not necessarily because their opinions have a dominating influence on community members' beliefs.

The interplay between the free-riding and perceived-infectiousness effects is the distinctive feature of our theoretical model. When the perceived-infectiousness effect is large enough, it overcomes the countervailing free-riding effect, and the misperceptions correction treatment leads to more social distancing. An additional implication of the theory is that the treatment should have similar heterogeneous effects on the expected future infection rate.

We conduct this additional test of the theory, examining treatment effects on the expected future infection rate.[24] In Columns (3) and (4) of Table 3.2, the outcome is the share of the community the respondent thinks will get sick from COVID-19 (responses were integers out of 10; we divide by 10 to yield a 0-1 scale). In Column (3), we estimate average treatment effects. Each coefficient is small in magnitude and not statistically significantly different from zero.

In Column (4), we estimate heterogeneity in treatment effects with respect to local cases, and find the same pattern as in Column (2). The misperceptions correction decreases the expected future infection rate in districts with no cases, and this impact becomes more positive as current cases rise (the $T1_{ijd}$ main effect and interaction term coefficients are both statistically significant at the 5% level).

These treatment effect heterogeneity findings in Table 3.2 Columns (2) and (4) jointly support the theoretical model. When current infection rates are low, the misperceptions correction treatment does not change perceived infectiousness much, but leads to realizations that social distancing support is higher than previously thought. People therefore reduce estimates of the future infection rate, and also reduce their own social distancing (choosing to free-ride). By contrast, when current infection rates are high, the treatment causes larger increases in perceived infectiousness. Notwithstanding an increase in the share of social distancing supporters, people increase their estimate of the future infection rate and increase their social distancing.

3.6 Conclusion

Support for social distancing increased rapidly during the COVID-19 pandemic. If people are unaware of the extent to which others' beliefs on social distancing have changed, would revealing true high rates of such support lead to more social distancing? In theory, the impact of providing such information is ambiguous: it could reduce social distancing if free-riding effects dominate, but could have a positive effect on social distancing if perceived-infectiousness effects dominate. Perceived-infectiousness effects are more likely to dominate when the current local infection rate is higher.

We implemented a randomized controlled trial testing the impact of a "misperceptions correction" treatment revealing high community support for social distancing. The treatment

[24]The question is "For every 10 people in your community, how many do you think would get sick from coronavirus?" Sample sizes in these regressions are smaller. We implemented this question midway through the endline survey, after finding preliminary evidence suggesting the need to explore mechanisms behind treatment effect heterogeneity.

effect on social distancing exhibits the spatial heterogeneity predicted by theory: negative in areas with low infection rates (reflecting the dominance of free-riding effects), and more positive in areas with higher rates (as perceived-infectiousness effects become increasingly prominent). In the area with the most cases, amounting to one-quarter of our sample, the treatment effect is positive and large in magnitude. The treatment effect on the expected future infection rate shows similar heterogeneity, confirming an additional theoretical prediction.

Our results suggest that when local infection rates are high, health policies shifting perceptions of community support for social distancing upwards could help promote social distancing behavior. Future research is needed to confirm the external validity of these findings and determine how the results translate to other contexts. For example, in cities, looser social networks among neighbors might lead to larger misperceptions of community support while population-dense housing might further activate the perceived-infectiousness effect; alternatively, in communities with lower baseline support for social distancing, a misperceptions correction treatment may be less motivating but may also potentially "gain more ground" among those with the lowest support who also underestimate the social norm. These findings may also help predict the impacts of analogous public health messaging revealing community support for preventive measures against other infectious diseases.

3.7 Tables and Figures

Table 3.1: Summary Statistics of Pre-Treatment Social Distancing Measures

	VARIABLES	Pre-Baseline			Baseline			T-test
		N	Mean	SD	N	Mean	SD	p-value
(1)	Respondent supports social distancing (SD)	2,117	0.976^a	0.153	2,117	0.989^b	0.104	0.001
(2)	Perceived share of community supporting SD	2,109	0.689^a	0.313	2,114	0.800^b	0.262	0.000
(3)	Primary SD indicator: if (4) & (8)				2,117	0.078	0.269	
(4)	→ Self-report SD indicator: if (5) & (6)	2,117	0.383	0.486	2,117	0.355	0.479	0.045
(5)	→ Self-report: Followed govt rules last 14 days	2,117	0.952	0.214	2,117	0.949	0.219	0.692
(6)	→ Self-report: SD behaviors above median	2,117	0.396	0.489	2,117	0.361	0.481	0.012
(7)	→ Others SD indicator: if (9) & (10)				2,117	0.232	0.422	
(8)	→ Other households' report of SD				2,117	0.378	0.485	
(9)	→ Leaders' report of SD				2,117	0.519	0.500	

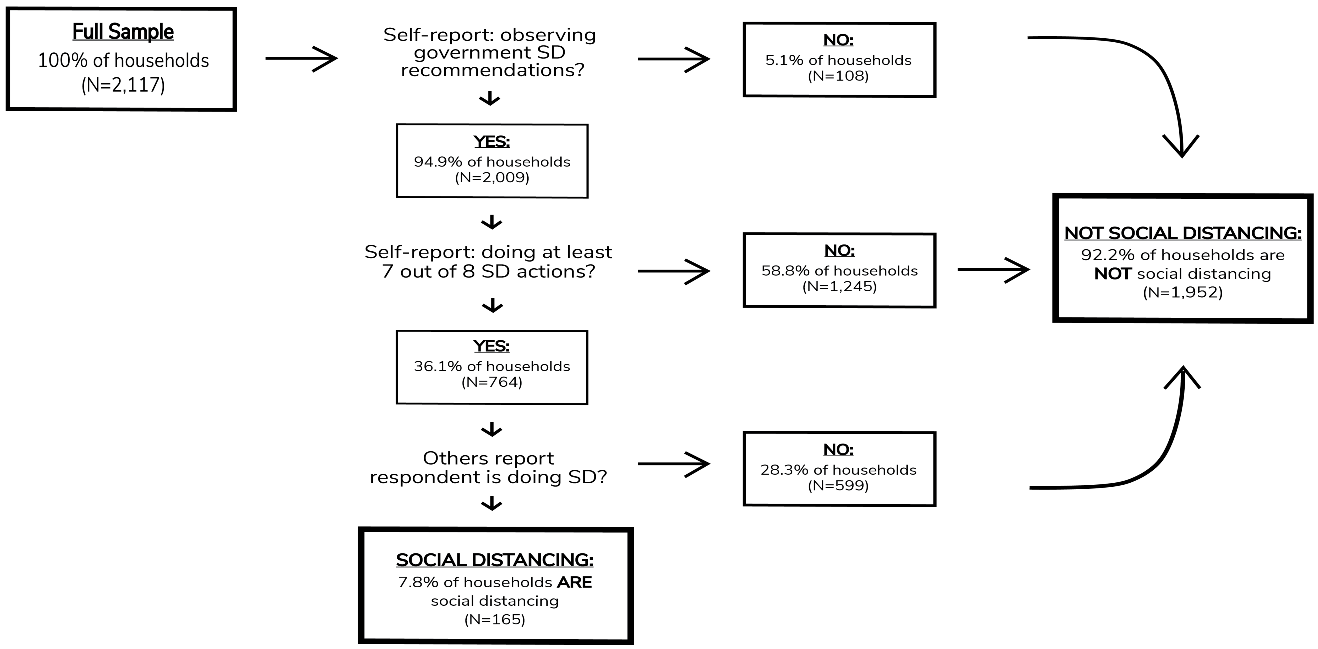
Notes: Pre-baseline data collected from July 10 to August 16, 2020, and baseline data collected from August 26 to October 4, 2020. Variables are as follows. Variables are as follows. Row 1: indicator equal to one if respondent answers "yes" to supporting "the practice of social distancing to prevent the spread of coronavirus" and zero otherwise. Row 2: perceived share of households (asked as "for every 10 households") in community that support social distancing (SD). Row 3: indicator for SD equal to one if respondent is SD according to self (Row 4) and others' reports (Row 8), and zero otherwise. Row 4: indicator for SD according to self if respondent answered "yes" to observing the government's recommendations on SD in the last 14 days (Row 5) and report following four out of four (above both sample's median of three) social distancing behaviors (Row 6), and zero otherwise. Row 7: indicator for SD according to others if all other respondents (Row 8) and community leaders (Row 9) reported not knowing the respondent household, not seeing the respondent household in the past 14 days, or—if seen—that the respondent household 1) did NOT come closer than 1.5 meters to others outside their household; 2) did NOT shake hands, try to shake hands, or touch others outside their household; and 3) appeared to be observing the government's recommendations on SD, and zero otherwise. All variables have a minimum of 0 and a maximum of 1. Last column displays the p-value of a paired t-test on the difference between pre-baseline and baseline measure (where pre-baseline data are available). Superscripts ^{a} and ^{b} indicate paired t-tests comparing reported and perceived support for social distancing at pre-baseline and baseline, respectively, which are significantly different (p-value=0.000).

Table 3.2: Treatment Effects on Social Distancing and Expected COVID-19 Illnesses

VARIABLES	(1) Primary SD Indicator	(2) Primary SD Indicator	(3) Perceived share of people in community that will get sick from COVID-19	(4) Perceived share of people in community that will get sick from COVID-19
T1: Misperceptions Correction	0.0042	-0.0466**	0.0418	-0.1936**
	(0.0140)	(0.0191)	(0.0322)	(0.0944)
T2: Leader Endorsement	-0.0054	-0.0258	-0.0209	-0.0598
	(0.0137)	(0.0198)	(0.0308)	(0.0944)
T1 × District COVID-19 Cases		0.0030***		0.0073**
		(0.0011)		(0.0029)
T2 × District COVID-19 Cases		0.0012		0.0013
		(0.0010)		(0.0029)
Observations	2,117	2,117	812	812
R-squared	0.158	0.163	0.146	0.152
Control Mean DV	0.0857	0.0857	0.3590	0.3590
Control SD DV	0.2801	0.2801	0.3685	0.3685

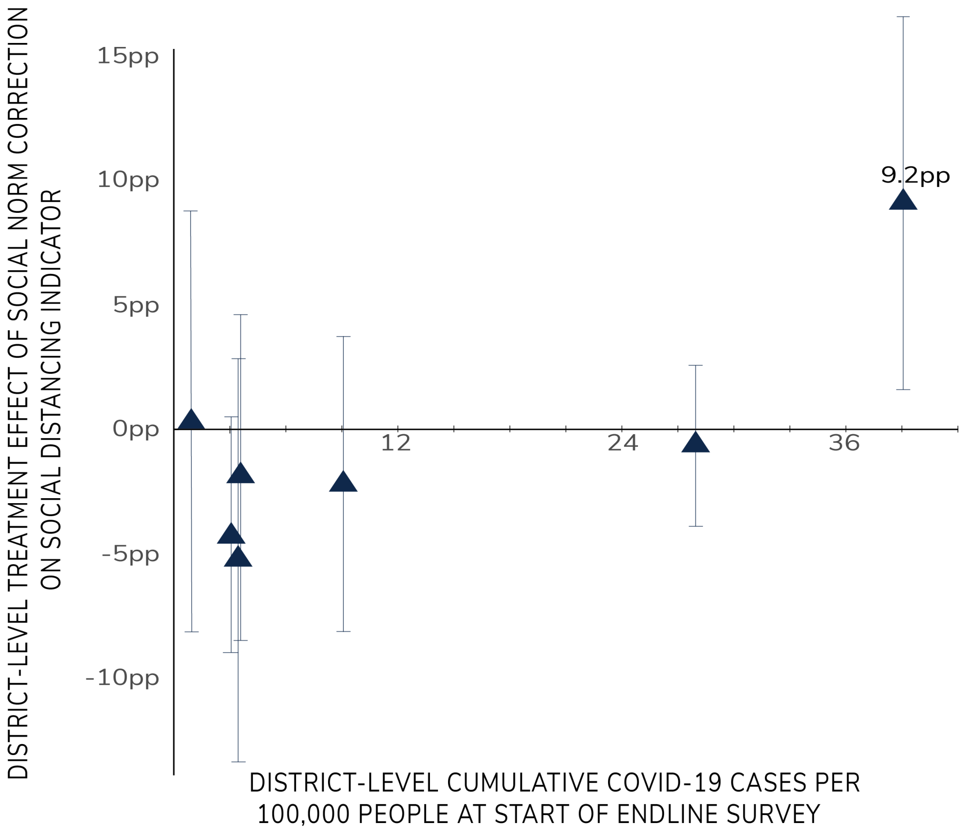
Notes: Dependent variable in Columns 1-2 defined in Table B.4. Dependent variable in Columns 3-4 is the expected future infection rate: "For every 10 people in your community, how many do you think would get sick from coronavirus?" (converted to share from 0 to 1). "T1: Misperceptions Correction" is equal to one if respondent was randomly assigned to the misperceptions correction treatment, and zero otherwise. "T2: Leader Endorsement" is equal to one if respondent was randomly assigned to the leader endorsement treatment, and zero otherwise. "T1 x District COVID-19 Cases" & "T2 x District COVID-19 Cases" are the respective treatment indicators interacted with district-level cumulative COVID-19 cases per 100,000 population at the start of the endline survey (see Appendix C.2.3, Table C.1, Column 2). All regressions control for a baseline measure of the dependent variable, a vector of indicators for number of community leaders knowing the respondent at baseline (0 through 4), and a vector of indicators for number of other respondents knowing the respondent at baseline (0 through 8). All regressions also include community fixed effects. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Figure 3.1: The Social Distancing Measure



Notes: As pre-specified, respondents considered social distancing (SD) if: 1) self-report they are following government SD recommendations, 2) self-report they are doing at least seven out of eight SD actions, and 3) be reported by others in community to be SD. Percentages reported are all shares of full sample (N=2,117). See Table B.4 and Section 3.3.3 and Appendix C.4 of main text for social distancing question definitions.

Figure 3.2: District-Level Misperceptions Correction Treatment Effects by COVID-19 Cases



Notes: Misperceptions correction treatment effects (triangles) estimated separately for each of seven districts (with 95% confidence intervals). District-level treatment effects plotted on vertical axis against district-level cumulative COVID-19 case loads at start of endline survey (per 100,000 population) on horizontal axis.

APPENDICES

APPENDIX A

Appendix to Chapter 1

An outline of this Appendix: Section A.1 details the sub-Saharan Africa cross-sectional analysis used to create Figure 1.1. Section A.2 provides additional details of the data cleaning process used for primary outcomes and crop calendars. Section A.3 provides additional details on the paper's simulations. Section A.4 details alternative inference procedures used in Table 1.4. Section A.5 provides details of "back-of-the-envelope" calculations made in the Section 1.5.1.1 of the main text. Section A.6 further details the long-run secondary analysis with description of the data and additional results by age group.

A.1 Sub-Saharan Africa Country-Level Analysis

This appendix provides data description and robustness checks for Figure 1.1, which shows across sub-Saharan African (SSA) countries[1] a negative correlation between overlap in the school and farming calendars and primary school survival rates. Specifically, it shows that as the overlapping percent of total school days plus total sowing/harvest days increases, the rate at which children who start school complete primary school decreases. Table A.1 summarizes the sources of the country-level data on school calendars, farming calendars, crop production and the outcome used for this analysis.

[1]My target sample was 46 countries belonging to SSA according to the United National Development Programme, listed here: https://www.africa.undp.org/content/rba/en/home/regioninfo.html. Due to data limitations explained in this appendix, Guinea-Bissau, Seychelles, and South Sudan were unable to be included in the final sample.

A.1.1 Data

School Calendars: To estimate the day-level overlap between school and farming calendars, I start by assembling a novel dataset of the daily primary and secondary school calendars for 82% (37) of SSA countries.[2] Table A.1 lists sources for school calendars. All school calendars start in either 2018 or 2019 and hence were announced well before and are not affected by the COVID-19 pandemic. Most school calendar data are taken from official announcements on government websites, Facebook pages, or correspondence with government officials. Detailed school calendar announcements include dates of all school holidays, breaks between terms, differences between primary and secondary schedules, and occasionally exam schedules. I also include school calendars from other reputable sources that provided start and end dates of major terms. As an example, Figure A.1 shows an excerpt of the official 2019/2020 school calendar for Malawi. For the few countries in which I was unable to locate a recent school calendar (after various attempts to contact government officials via email, phone, or third-parities), I use UNESCO UIS data for school calendar start month and end month in 2019 to estimate days in the school calendar. To do so, I estimate the average number of school days in each month as 5/7 of the number of days in the month in 2019 (i.e., not during a leap year) and assume that school operates for the start month, the end month, and each month in between.[3] I translate these school calendars into a vector $\{s_1, s_2, \ldots s_{365}\}$ of 365 indicator variables (each representing one day of the year) equal to one if school was scheduled on day d and zero otherwise. Further, I define country-level school requirements as $S = \sum_{d=1}^{365} s_d$, the total number of days during which school is scheduled.

Farming calendars: Farming calendars are conceptualized as country-level crop calendars weighted by country-level crop production. Table A.1 lists sources for crop calendars. Crop calendars are mostly taken from "Country Briefs" written by the UN's Food and Agriculture Organization (FAO) Global Information and Early Warning System (GIEWS).[4] Each FAO/GIEWS crop calendar depicts the sowing, growing, and harvesting periods in roughly within-month 10-day increments for between three to seven locally important crops.

[2]This dataset will soon be available on my website https://www.jamesalleniv.com. I thank my amazing research assistants Danielle-Andree Atangana and Noelle Seward for their efforts in locating and digitizing school and farming calendars as well as Max Diaz, Flavia Lorenzon and Laston Manja for finding additional school calendars.

[3]As this school calendar estimated with UIS data likely overestimates the total number of school days, I include an indicator for UIS-derived school calendars in regressions below to show the main finding is not driven by differences in data sources.

[4]For three countries in which "Country Briefs" were not available, I use data on planting and harvest periods from the FAO's Crop Calendar Tool. Finally, the crop calendar for Comoros came from a World Food Programme (WFP) report: https://documents.wfp.org/stellent/groups/public/documents/ena/wfp085419.pdf.

Most crop calendars also highlight "major foodcrops", such as maize or cassava, specific to each country. As an example, Figure A.2 depicts the FAO/GIEWS crop calendar for Malawi. I translate these into vector $\{f_{c,1}, f_{c,2}, \ldots f_{c,365}\}$ of 365 indicator variables (each representing one day of the year) equal to one if either sowing and harvesting of crop calendar c occurs on day d and zero otherwise. Further, I define farming time requirements for crop calendar c as $F_c = \sum\limits_{d=1}^{365} f_{c,d}$, the sum of the total number of days during which sowing or harvesting of the crop occurs. A crop calendar was not found for Seychelles where farming is not common, so Seychelles had to be excluded from the sample.

Production Data: Crop production and land allocation data come from FAO Statistics Division (FAOSTAT). Production is measured in metric tons and land allocation in hectares. I use data only for crops with a crop calendar. In a few countries where some crops were assigned different calendars based on season or region, I assumed that land allocation and production were split 75% and 25% across primary and secondary growing seasons (respectively) and split evenly across regions.

Calculating Overlap: Using these data, I construct my measure of country-level overlap as follows. First, I calculate overlap for crop c as the product of the school and crop calendar indicators on day d, summed across all days to get the total number of days during which both school is scheduled and sowing/harvesting occurs – i.e., $overlap_c = \sum_{d=1}^{365}(s_d * f_{c,d})$. Second, I aggregate to the country-level by weighting crop-level farming time requirements F_c and $overlap_c$ by crop c's share of country-level production for listed crops. Finally, I divide the sum of country-level school and farming time requirements by country-level overlap to get Overlap Percent, the fraction of total school and farming days that overlap. Estimating overlap as a share of total schooling and farming time requirements, as opposed to just total number of days, effectively controls for possible correlations between length of school and/or farming calendar and the primary outcome.

Outcome Data: The primary outcome is the survival rate to grade 5 for both sexes from UNESCO's Institute for Statistics (UIS), which measures the fraction of a cohort of students enrolled in first grade who are expected to reach grade 5 of primary school.[5] Thus lower survival rates suggest lower level of retention and higher incidence of dropout within school. Other outcomes I will test include survival rate to grade 4 and survival rate to the last grade of primary school, both from UIS, and also primary school complete rate from the World Development Indicators (WDI). All outcomes are important indicators for

[5]UIS calculates the outcome by using two consecutive years of enrollment data at each grade level to "reconstruct" a cohort's progression through primary school, and then divides the number of students expected to reach the last grade by the total number of the students in the cohort (i.e., those who originally enrolled in the first grade). See the UIS Glossary for more information.

monitoring universal primary education. However, since the number of grades in primary school varies across SSA countries, I select survival rate to grade 5 as the primary outcome as the furthest-along consistent measure of primary school progress. I use the most recent data point available; data come from as recently as 2018, as far back as 2002, and from 2015 on average across countries. These data are not available for Guinea Bissau and South Sudan, which are hence excluded from the analysis.

Summary Statistics: Table A.2 presents summary statistics for the 43 countries for which the primary outcome data are available. Across these countries, the average survival rate to grade 5 is 71.6% (standard deviation of 17.8). The Overlap Percent of total school and farming days has a mean of 22.8% (standard deviation of 4.5 percentage points), which is estimated by dividing the sum of total farm days and total school days by total overlap days. I use daily school calendars for 81% of sample countries. Other country-level variables show that 65% of countries are least developed countries, 35% are landlocked, and 9% are small island developing states. Descriptive statistics on last colonial power and region are also provided.

A.1.2 Analysis

Regression: Figure 1.1 visually depicts β in the following regression:

$$Y_j = \alpha + \beta \text{ Overlap Percent}_j + \epsilon_j \tag{A.1}$$

where Y_j is the survival rate to grade 5 in country j; Overlap Percent$_j$ is the percent of country j's total school and farming time requirements that overlap as measured in days; and ϵ_j is an error term. I test the robustness of this result by also regressing Equation A.1 with those country-level controls summarized in Table A.2 as well as year fixed effects. In other specifications, I Y_j with a country's survival rate to grade 4, survival rate to the last grade of primary school, and primary school completion rate.

Results: Regression results are presented in Table A.3. Column (1) shows the results visually depicted in Figure 1.1, finding that a one percentage point increase in the percent of total schooling and farming requirements that overlap is correlated with a 2.39 percentage point decline in an SSA country's survival rate to grade 5. With Overlap Percent$_j$ ranging from 14.6 to 31.8, the coefficient maps to a more than 40 percentage point gap across SSA countries in the survival rate to grade 5.

Remaining columns show that the significant negative result is robust to other specifications. Column (2) adds country-level controls. Column (3) adds year fixed effects. Column (4) define crop shares by land allocation (measured in hectares) instead of production.

Column (5) conceptualizes country-level farming calendars by only those crops listed by FAO/GIEWS as "major food crops". Column (6) excludes countries defined by UIS-derived school calendars to assess the subsample of SSA countries with only verified daily school calendars. In Column (7), I analyze using only the UIS-derived school calendars; however, because these data are available for all countries in the sample annually from 1997 to 2018, I analyze the data as a panel including the same country-level controls and year fixed effects. These robustness checks build confidence in the relationship depicted in Figure 1.1.

Additionally, regression results on different but related outcomes are presented in Table A.4. Outcomes include survival rate to grade 4 in columns (1)-(3), survival rate to the last grade of primary school in columns (4)-(6), and primary school completion rate in columns (7)-(9). In each set, the first column has no controls, the second adds country-level controls, and the third adds year fixed effects. Results are statistically significant in all but one regression, giving some assurance that the main finding is not specific to its outcome measure.

Table A.1: Data Sources for Sub-Saharan Africa (SSA) Country-Level Analysis

COUNTRIES	School Calendar	Crop Calendar	Production Data	Outcome Data	Included in Sample
Angola	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Benin	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Botswana	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Burkina Faso	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Burundi	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Cameroon	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Cape Verde	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Central African Republic	UIS Monthly	FAO/GIEWS	FAOSTAT	UIS	Yes
Chad	UIS Monthly	FAO/GIEWS	FAOSTAT	UIS	Yes
Comoros	UIS Monthly	WFP Report	FAOSTAT	UIS	Yes
Côte d'Ivoire	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Demc Repub of the Congo	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Equatorial Guinea	UIS Monthly	FAO CCT	FAOSTAT	UIS	Yes
Eritrea	UIS Monthly	FAO/GIEWS	FAOSTAT	UIS	Yes
Ethiopia	Reliable Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Gabon	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Gambia	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Ghana	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Guinea	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Guinea-Bissau	Official Daily	FAO/GIEWS	FAOSTAT	X	No
Kenya	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Lesotho	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Liberia	Reliable Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Madagascar	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Malawi	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Mali	Reliable Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Mauritania	Reliable Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Mauritius	Official Daily	FAO CCT	FAOSTAT	UIS	Yes
Mozambique	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Namibia	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Niger	UIS Monthly	FAO/GIEWS	FAOSTAT	UIS	Yes
Nigeria	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Republic of the Congo	UIS Monthly	FAO/GIEWS	FAOSTAT	UIS	Yes
Rwanda	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Sao Tome and Principe	UIS Monthly	FAO CCT	FAOSTAT	UIS	Yes
Senegal	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Seychelles	Official Daily	X	FAOSTAT	UIS	No
Sierra Leone	Reliable Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
South Africa	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
South Sudan	Official Daily	FAO/GIEWS	FAOSTAT	X	No
Swaziland / Eswatini	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Tanzania	Reliable Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Togo	Reliable Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Uganda	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Zambia	Official Daily	FAO/GIEWS	FAOSTAT	UIS	Yes
Zimbabwe	Reliable Daily	FAO/GIEWS	FAOSTAT	UIS	Yes

Notes: Target sample was 46 countries belonging to SSA according to the United Nations Development Programme. School Calendar: "Official Daily" is a daily calendar from a government announcement, "Reliable Daily" is a daily calendar from a reputable website or correspondence (e.g., in-country international schools), and "UIS Monthly" refers to monthly calendars from the UNESCO Institute for Statistics (UIS). Crop Calendar: "FAO/GIEWS" refers to country briefs written by the UN's Food and Agriculture Organization (FAO) Global Information and Early Warning System (GIEWS), "FAO CCT" refers to the FAO's Crop Calendar Tool, and the calendar for Comoros came from a World Food Programme (WFP) report. No crop calendar was found for Seychelles where farming is not common. All production data came from the FAO Statistics Division (FAOSTAT). All outcome data came from the UNESCO Institute for Statistics (UIS). Due to indicated data limitations, Guinea-Bissau, Seychelles, and South Sudan are not in the final sample.

Figure A.1: Example of an Official School Calendar: Malawi

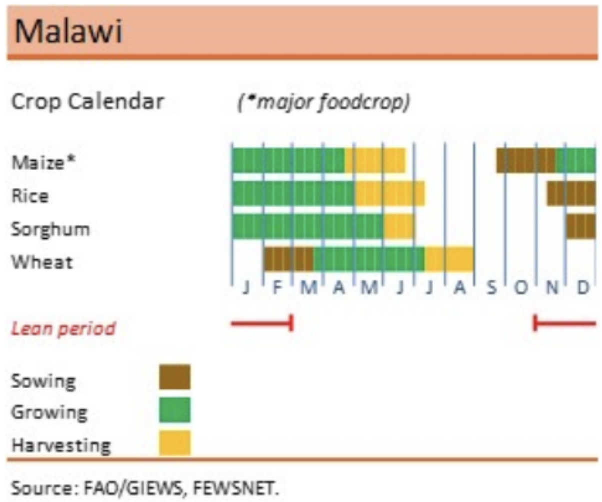
The Primary, Secondary School and Teacher Training College academic year will comprise 41 weeks as follows:

2019/2020 ACADEMIC CALENDAR

TERM	OPENING	CLOSING	TOTAL WEEKS	HOLIDAY
1	16th September, 2019	20th December, 2019	14 weeks	2 weeks
2	6th January, 2020	3rd April, 2020	13 weeks	2 weeks
3	20th April, 2020	24th July, 2020	14 weeks	7 weeks end of year

Notes: The figure shows an excerpt of the official 2019/2020 school calendar for Malawi, which shows the daily schedule including breaks between terms. Not shown is the description of other school holidays. The dataset of sub-Saharan African school calendars and source documentation will soon be available on my website: https://www.jamesalleniv.com.

Figure A.2: Example of FAO/GIEWS Crop Calendars: Malawi



Notes: The figure depicts the most common source of crop calendars in the analysis. Crop calendars are mostly taken from "Country Briefs" written by the UN's Food and Agriculture Organization (FAO) Global Information and Early Warning System (GIEWS). Each FAO/GIEWS crop calendar depicts the sowing, growing, and harvesting periods in roughly within-month 10-day increments for between three to seven locally important crops. Most crop calendars also highlight "major foodcrops", such as maize or cassava, specific to each country.

Table A.2: Summary Statistics for Sub-Saharan Africa Country-Level Analysis

VARIABLE	N	Mean	SD	Min	Max
Survival Rate to Grade 4	43	77.73	15.15	37.55	98.50
Survival Rate to Grade 5	43	71.64	17.77	27.92	97.21
Survival Rate to Last Grade of Primary School	43	64.34	19.52	24.16	94.61
Primary Completion Rate	43	72.52	17.43	40.56	100.66
Overlap Percent of School and Farm Days	43	22.75	4.51	14.57	31.82
Total Overlap Days between School and Farming	43	81.71	25.98	42.97	142.86
Total Farm Days	43	156.63	45.14	60.16	244.41
Total School Days	43	195.45	15.33	155.00	228.00
Indicator if Daily School Calendar was Found	43	0.81	0.39	0.00	1.00
Indicator if Least Developed Country	43	0.65	0.48	0.00	1.00
Indicator if Landlocked	43	0.35	0.48	0.00	1.00
Indicator is Small Island Developing State	43	0.09	0.29	0.00	1.00
Last Colonial Power: France	43	0.35	0.48	0.00	1.00
Last Colonial Power: Britain	43	0.35	0.48	0.00	1.00
Last Colonial Power: Other	43	0.30	0.46	0.00	1.00
Region: Eastern Africa	43	0.33	0.47	0.00	1.00
Region: Middle Africa	43	0.21	0.41	0.00	1.00
Region: Southern Africa	43	0.12	0.32	0.00	1.00
Region: Western Africa	43	0.35	0.48	0.00	1.00

Notes: Sample size of 43 (out of a possible 46) sub-Saharan African countries due to data availability. Summary statistics presented for outcomes, overlap measures, and other country-level variables. In the primary specification, the outcome of the survival rate to grade 5 of primary school is regressed on Overlap Percent, the sum of Total Farm Days and Total School Days then divided by Total Overlap Days.

Table A.3: Results for Sub-Saharan Africa Country-Level Analysis

VARIABLES	Survival Rate to the Last Grade of Primary School						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Overlap Percent	-2.457***	-2.172***	-2.210***	-1.756**	-2.021***	-2.590***	-1.200**
	(0.560)	(0.613)	(0.715)	(0.719)	(0.648)	(0.568)	(0.463)
Observations	43	43	43	43	43	35	464
R-squared	0.323	0.721	0.763	0.731	0.773	0.869	0.541
Country-level Controls	N	Y	Y	Y	Y	Y	Y
Year FE	N	N	Y	Y	Y	Y	Y
Crop Shares Defined by	Production	Production	Production	Land	Production	Production	Production
Crop Selection	All Given	All Given	All Given	All Given	Major Crops	All Given	All Given
School Calendars Used	Best Given	Best Given	Best Given	Best Given	Best Given	Daily	UIS monthly
Standard Errors	Robust	Robust	Robust	Robust	Robust	Robust	Cluster
DV Mean	64.34	64.34	64.34	64.34	64.34	64.34	63.33

Notes: The table presents results of a regression of survival rate to grade 5 on Overlap Percent, the fraction of total school days and total farm days that overlap with each other, for 43 countries in sub-Saharan Africa. Column (1) shows the results visually depicted in Figure 1.1. Remaining columns show that the significant negative result is robust to other specifications: column (2) adds country-level controls, column (3) adds year fixed effects, column (4) define crop shares by land allocation (measured in hectares) instead of production, column (5) conceptualizes country-level farming calendars by only those crops listed by FAO/GIEWS as “major food crops”, column (6) excludes countries defined by UIS-derived school calendars to assess the subsample of SSA countries with only verified daily school calendars, and column (7) analyze using only the UIS-derived school calendars as a panel from 1997 to 2018 with the same country-level controls and year fixed effects.

Table A.4: Additional Results for Sub-Saharan Africa Country-Level Analysis

VARIABLES	Survival Rate to Grade 4			Survival Rate to Last Grade			Primary Completion Rate		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Overlap Percent	-2.017***	-1.922***	-2.018***	-2.457***	-2.172***	-2.210***	-1.173**	-0.955*	-0.661
	(0.466)	(0.569)	(0.634)	(0.560)	(0.613)	(0.715)	(0.530)	(0.470)	(0.595)
Observations	43	43	43	43	43	43	44	44	44
R-squared	0.361	0.695	0.759	0.323	0.721	0.763	0.092	0.635	0.723
Country Controls	N	Y	Y	N	Y	Y	N	Y	Y
Year FE	N	N	Y	N	N	Y	N	N	Y
DV Mean	77.73	77.73	77.73	64.34	64.34	64.34	72.35	72.35	72.35

Notes: The table presents results of different outcomes on Overlap Percent, the fraction of total school days and total farm days that overlap with each other, for 43 countries in sub-Saharan Africa. Outcomes include survival rate to grade 4 in columns (1)-(3), survival rate to the last grade of primary school in columns (4)-(6), and primary school completion rate in columns (7)-(9). In each set, the first column has no controls, the second adds country-level controls, and the third adds year fixed effects.

A.2 Data Cleaning Details

The appendix builds on Sections 1.3.3 and 1.3.4 to provide additional details on the data cleaning procedures.

A.2.1 Primary Outcomes

Outcome data come from Malawi's Integrated Household Panel Survey (IHPS) 2010-2013, which was implemented as part of the World Bank Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) initiative.[6] While the IHPS 2010 was implemented during the school calendar change "transition year", key measures of schooling and agricultural production were recalled from the pre-policy period, and I use these whenever possible. At baseline, the IHPS sample was selected to be representative at the national and regional levels, covers 26 of Malawi's 28 districts and 4 urban areas, and surveys households in 204 16-household enumeration areas (Government of Malawi, 2012). The IHPS defined a "community" as the village or urban location surrounding an enumeration area; similarly, I define a "community" as each 16-household enumeration area. Outcome data include highest grade level completed and hours spent working on the household farm.

First, $Grade_i$ is the highest grade level completed for individual i for the reference academic year. These data come from the Education module of the Household Questionnaire. Highest grade for the current academic year at the time of surveying equals the integer reported for "What class are you in or what was the highest class level you ever attended?" if the individual reported NOT attending school in the current academic year; one less the integer reported for "What class are you in or what was the highest class level you ever attended?" if the individual reported WAS currently attending school; or zero if responded "No" to "Have you ever attended school?" or if both a reason for never attending school was given and class information was missing. Highest grade for the previous academic year was constructed similarly but using other questions that referred to the "last completed academic year". Then, I used the daily school calendar and the recorded date of the visit during which the Education module was administered to determine to which academic year each measured referred. The baseline value of $Grade_i$ refers to the pre-policy 2009 academic year, while the outcome $Grade_i$ refers to the 2013 academic year. A similar process was used to generate a dummy if an individual started school or enrolled in the referenced academic year, both of which are used in secondary analyses.

To clean $Grade_i$, I set $Grade_i$ equal to zero if missing for individuals under five years at the time of surveying (as they were not eligible for the Education module), though this is

[6] Documentation can be found at: https://microdata.worldbank.org/index.php/catalog/2248/study-description.

only relevant for the long-run analysis. Otherwise I perform no additional cleaning but check that the results are not driven by erroneous or improbable panel data in Table A.5. First, in column (1), I analyze a dummy variable equal to one if the difference in an individual's 2013 $Grade_i$ and 2009 $Grade_i$ is either less than zero or greater than five (i.e., improbable changes in highest completed grade level), and zero otherwise, which equals one for only 3.6% of the sample. I find that this dummy is not significantly correlated with shift-share overlap. Further, Table A.5 columns (2)-(8) show that the main result is robust to alternative cleaning procedures for $Grade_i$. Columns (2)-(4) assume the 2013 value is the "true" reference: column (2) prevents "grade regression" by setting the maximum 2009 value equal to the 2013 value; column (3) prevents "unrealistic grade progression" by setting the minimum 2009 value equal to the 2013 value minus 5 (preventing six or more grades completed in four years); column (4) does both. Columns (5)-(7) assume the 2009 value is the "true" reference: column (5) prevents "grade regression" by setting the minimum 2013 value equal to the 2009 value; column (6) prevents "unrealistic grade progression" by setting the maximum 2013 value equal to the 2009 value plus 5; column (7) does both. Column (8) drops the 3.6% of the sample with such inconsistent observations. The regressions show that the significantly negative effect on $Grade_i$ is robust to these alternative cleaning procedures.

Second, $Farmed_i$ is an indicator if individual i was reported to work any hours on the household farm during the rainy-season sowing and harvest periods. These data come from the Household Labor section of the Rainy Season Module of the Agriculture Questionnaire, which reports for each agricultural plot the number of weeks, days per week, and hours per day of work during the land preparation and planting (i.e., sowing) period and harvesting period for up to four household members. For each individual i, these time-use variables are multiplied together to generate plot-level total hours worked and then summed across plots to construct $Farm\ Hours_i$ as the total of all hours worked on household plots during the sowing and harvest periods in the rainy season. $Farmed_i$ is equal to one if $Farm\ Hours_i > 0$ and zero otherwise. The baseline value refers to the 2009/10 rainy season, while the outcome itself refers to the 2012/2013 rainy season.

Third, $Farm\ Hours_i$, defined in the previous paragraph, is cleaned to address outliers. In the primary specification, I winsorize it by replacing any value beyond the 95th percentile with the value at the 95th percentile. Then in Table A.6, I present balance tests and main effects for alternative cleaning procedures. First, columns (1)-(4) follow the balance test specification described in Table 1.1 to regress $Farm\ Hours_i$ not winsorized in column (1), winsorized at the 95th percentile in column (2) and the 90th percentile in column (3), and with an inverse hyperbolic sine transformation in column (4). All four regressions reveal a *positive* correlation between shift-share overlap and these measures at varying degrees of significance. Second, columns (5)-(8) follow the main specification described in Table 1.3 to test for the effect of shift-share overlap on the 2013 post-policy measure of $Farm\ Hours_i$ similarly adjusted for extreme outliers. These regressions reveal significant *negative* effects that, relative to each estimate's corresponding baseline imbalance, are larger in columns where outliers are addressed.

A.2.2 Crop Calendars

Data on community crop production are matched to crop calendars from the Food and Agriculture Organization (FAO) Crop Calendar Tool, which provides start and end months for the sowing and harvest periods for 45 major crops in Malawi. The FAO Crop Calendar Tool defines a crop c as a unique combination of its altitude zone (high, medium, or low),[7] season of production (rainy, dry, or permanent), and basic crop type (e.g., maize, soybean, etc.). The most common crops, such as maize, have different calendars for different altitudes and seasons; in general, crop calendars are longer at lower altitudes and in the rainy versus the dry season.

Crop calendars from the FAO Crop Calendar Tool match to 83% of pre-policy cultivated acres in the IHPS data. For the remaining 17% of cultivated acres, I use the modal sowing month and harvest month reported by IHPS households in 2010 (the earliest available). In using IHPS-based crop calendars from 2010, I assume that the modal crop calendars were unaffected by the one-month school calendar change between 2009 and 2010, or at least that any endogeneity has minuscule effect on my analysis after these crops are weighted by their relatively smaller share of total cultivated acres.

[7]High altitude is defined as greater than 1300 meters, low altitude is defined as less than 600 meters, and medium altitude is in between these cutoffs.

Table A.5: Grade Level in 2013 following Different Cleaning Procedures

VARIABLES	Dummy: '13-'09 <0 or >5	Assume 2009 value is "true" reference: Fix Max '09 if '09>'13	Fix Min '09 if '13>'09+5	Both (2) & (3)	Assume 2013 value is "true" reference: Fix Min '13 if '09>'13	Fix Max '13 if '13>'09+5	Both (5) & (6)	Drop Obs if '09>'13
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ssoverlap_ℓ	-0.010	-0.274	-0.259	-0.256	-0.275	-0.265	-0.264	-0.259
	(0.013)	(0.101)	(0.101)	(0.091)	(0.103)	(0.103)	(0.095)	(0.092)
	[0.268]	[0.004]	[0.015]	[0.007]	[0.007]	[0.015]	[0.008]	[0.011]
Observations	2,142	2,142	2,142	2,142	2,142	2,142	2,142	2,064
R-squared	0.042	0.703	0.710	0.775	0.685	0.682	0.734	0.759
Base DV Mean	0.04	1.38	1.47	1.42	1.43	1.43	1.43	1.39
Δ DV Mean		2.35	2.26	2.31	2.35	2.26	2.31	2.33

Notes: Dependent variables relate to $Grade_i$, which receives minimal cleaning the primary analysis. Column (1) is a dummy variable equal to one if the difference in an individual's 2013 $Grade_i$ and 2009 $Grade_i$ is either less than zero or greater than five and is zero otherwise. Columns (2)-(8) tests robustness of the primary results after implementing different reasonable data cleaning procedures for $Grade_i$. Columns (2)-(4) assume the 2013 value is the "true" reference: column (2) sets the maximum 2009 value equal to the 2013 value; column (3) sets the minimum 2009 value equal to the 2013 value minus 5; column (4) does both. Columns (5)-(7) assume the 2009 value is the "true" reference: column (5) sets the minimum 2013 value equal to the 2009 value; column (6) sets the maximum 2013 value equal to the 2009 value plus 5; column (7) does both. Column (8) drops inconsistent observations from the sample. $ssoverlap_\ell$ and included controls defined in Table 1.3. Conventional robust standard errors in parentheses. Randomization inference p-values in square brackets.

Table A.6: Household-Farm Hours: Balance, Effects, Robustness

VARIABLES	Balance Test on Baseline Values:				Main Test on Post-Policy Values:			
	No Winz	Winz 95	Winz 90	IHS	No Winz	Winz 95	Winz 90	IHS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ssoverlap_{ℓ}	9.058	3.165	1.384	0.148	-6.289	-5.170	-5.354	-0.364
	(3.711)	(1.576)	(1.066)	(0.135)	(5.455)	(3.704)	(2.860)	(0.163)
	[0.004]	[0.031]	[0.107]	[0.122]	[0.090]	[0.087]	[0.042]	[0.043]
Observations	2,142	2,142	2,142	2,142	2,142	2,142	2,142	2,142
R-squared	0.071	0.125	0.138	0.150	0.131	0.203	0.226	0.249
Base DV Mean	1.88	1.88	1.88	1.88	13.96	9.82	7.21	1.03
Δ DV Mean	12.08	7.94	5.33	-0.85	23.65	21.49	19.59	1.06

Notes: Dependent variables relate to the continuous measure of household-farm hours $Farm\ Hours_i$. Columns (1)-(4) follow the balance test specification described in Table 1.1 to regress $Farm\ Hours_i$ not winsorized in column (1), winsorized at the 95th percentile in column (2) and the 90th percentile in column (3), and with an inverse hyperbolic sine transformation in column (4). Columns (5)-(8) follow the main specification described in Table 1.3 to test for the effect of shift-share overlap on the 2013 post-policy measure of $Farm\ Hours_i$ similarly adjusted for extreme outliers. $ssoverlap_{\ell}$ and included controls defined in Table 1.3. Conventional robust standard errors in parentheses. Randomization inference p-values in square brackets.

A.3 Simulation Details

A.3.1 Overlap Comparison

With its shift-share construction and normalization, changes in the overlap measure $ssoverlap_\ell$ can be hard to interpret in "real" terms. To put $ssoverlap_\ell$ into perspective, I ran a simulation and estimate that a 1.21 standard deviation increase in $ssoverlap_\ell$ is roughly equivalent to adding 10 days of overlap during sowing and harvest of rainy-season maize in the average sample community.

To determine this, I averaged crop shares across the 135 sample communities, added new school days to the 2009 pre-policy school calendar, and re-calculated $ssoverlap_\ell$ to simulate the effect of increasing overlap on the $ssoverlap_\ell$ measure. I chose to add school days during periods of sowing (late November to late December) and harvest (early May to early June) of rainy-season maize, as it alone accounts for over half of all planted acres in the sample (when summed across the three altitude zones) and thus likely represents periods of peak labor demand for most farming communities.

I find that adding five days spaced out during the sowing of rainy-season maize and another five days spaced out during the harvest of rainy-season maize increases normalized $ssoverlap_\ell$ by 1.21 standard deviations. By comparison, adding ten days spaced out during rainy-season maize sowing alone increases normalized $ssoverlap_\ell$ by 1.35, and adding ten days spaced out during rainy-season maize harvest alone increases normalized $ssoverlap_\ell$ by 1.07—the difference caused by the fact that sowing periods across all crops are more concentrated from mid-November through December, whereas harvest periods across all crops vary from February through July depending on the length of the growing season.

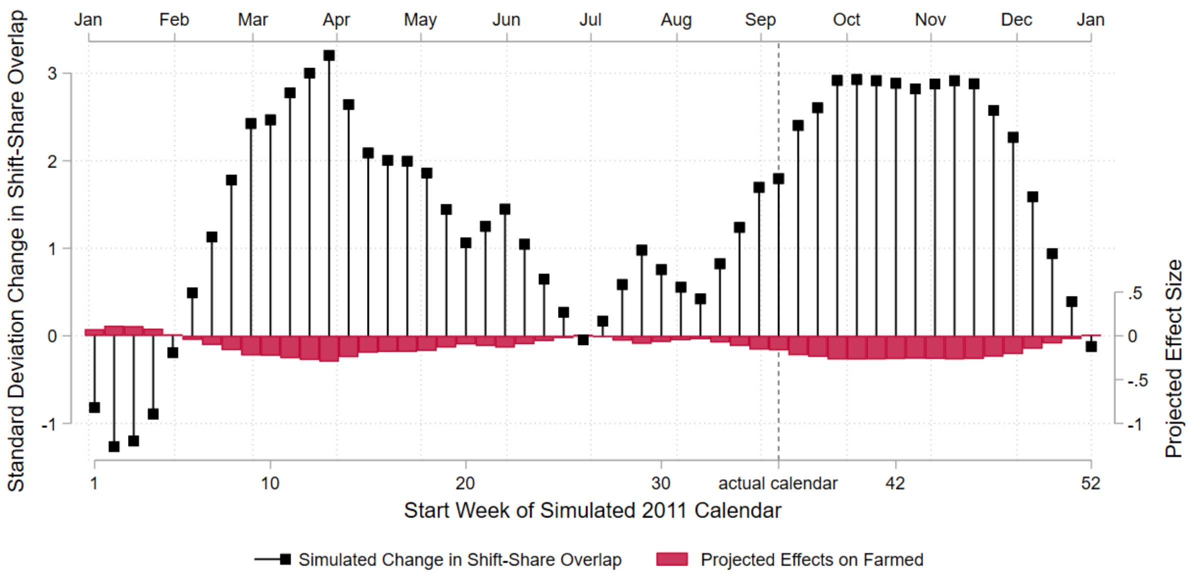
A.3.2 Policy Simulation

To identify Malawi's overlap-minimizing school calendar, I simulate 52 other potential school calendars that could have been used for the 2011 school year, each starting on a Monday and maintaining the structure and length of the original 2011 school calendar. This is done by effectively shifting the school calendar backward to previous Mondays in the year or forward to future Mondays using July 1st as the cutoff for the year. Then, for each of 52 simulated school calendars in 2011, I estimate the counterfactual change in shift-share overlap for each community relative to the original 2009 school calendar. Next, I multiply the counterfactual change in shift-share overlap by the coefficient in Table 1.3 column (1) to approximate the calendar's potential effect on $Grade_i$. A similar process simulates 52 other potential school calendars that could have been used for the original 2009 school year.

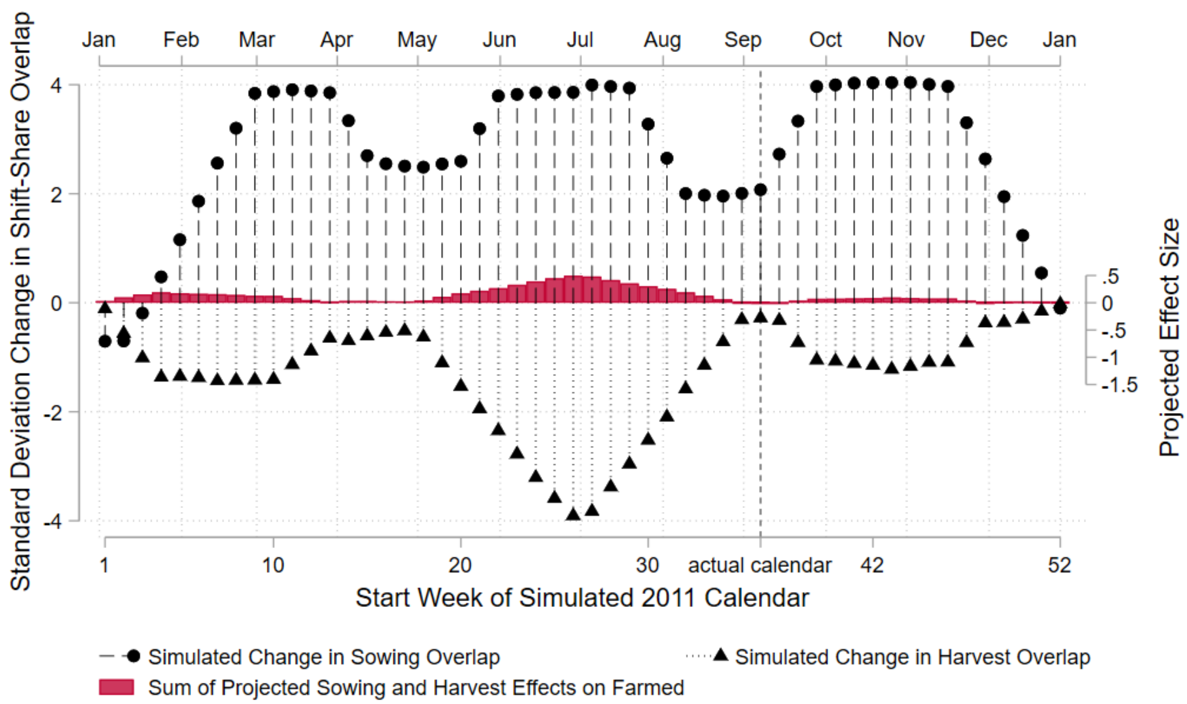
Additionally I use the policy simulation to estimate potential effects on $Farmed_i$, which are presented in Figure A.3 below. Projected overall effects on $Farmed_i$ in panel (a) follow a similar pattern to projected effects on $Grade_i$ in Figure 1.5 panel (a). However, different estimates in sowing and harvest periods from Table 1.7 creates divergent patterns in panel (b) across the two figures. Here, $Farmed_i$ is actually projected to increase under many alternative school calendars, especially when harvest-specific overlap is minimized given that harvest overlap appears to drive overlap's overall negative effect on $Farmed_i$.

Figure A.3: Simulated Impacts on Household-Farm Labor

(a) Overall Impact



(b) Sowing versus Harvest Impact



Notes: Figure depicts simulated policy impacts of alternative 2011 school calendars relative to the actual 2009 school calendar. Panel (a) plots the expected change in shift-share overlap for the average sample community using drop lines and projected effects on $Farmed_i$ using crimson bars. Panel (b) plots these outcomes separately for sowing and harvest periods. Additionally, the top horizontal axis estimates the start of each month, and the vertical line denotes the actual 2011 school calendar starting on Sept. 6, 2010.

A.4 Alternative Inference Procedures

In this appendix, I provide additional details for the alternative inference procedures listed in Table 1.4. This includes conventional ordinary least squares (OLS), several randomization inference (RI) procedures, and Borusyak et al. (2022)'s share-weighted shock-level regression technique. The fact that these alternative procedures produce similar p-values builds confidence in the primary results.

Table 1.4 starts in row (a) with conventional ordinary least squares (OLS) estimates. OLS produces valid inference under certain asymptotic assumptions including the absence of omitted variable bias. However, OLS standard errors may be invalid in shift-shares settings due to unobserved correlation between observations with similar exposure shares (Adão et al., 2019; Borusyak et al., 2022). In my setting, this could be unobserved correlation between individuals living in different communities that have similar sets of crop shares. Of most concern, when residuals are positively correlated, conventional OLS estimation will likely overreject. Hence there is motivation for turning to other inference procedures.

Two types of inference that can account for this issue in shift-share settings are RI approaches (e.g., Borusyak and Hull (2021)) and shift-share asymptotic approaches (e.g., Borusyak et al. (2022)). Each type has its strengths and weaknesses. As described by Borusyak and Hull (2021), RI requires specifying the full shock assignment process used to generate shock counterfactuals, whereas asymptotic approximation only requires specifying its first moment. Yet, RI is valid even when asymptotic assumptions of homoskedasticity or distribution symmetry are violated in the data. One additional assumption for Borusyak et al. (2022)'s share-weighted shock-level regression approach is that of having many shocks such that the largest share in the regression converges to zero as the sample size increases, which ensures a large effective sample size for the shock-level regression. However, the prevalence of medium-altitude rainy-season maize in my data, which accounts for 41.3% of the sum of crop shares, threatens to violate this assumption. Indeed, I only estimate an effective sample size of 15.5, although Borusyak et al. (2022)'s simulations conclude that an effective sample size of 20 "may be considered satisfactory" with a rejection rate near 7% instead of 5%. Therefore, I pursued a RI approach ex ante, although it turns out that both procedures produce similar results.

A.4.1 Alternative RI procedures

In this section, I specify the different *shock assignment processes* used in alternative RI procedures, which are each estimated as follows. First, I use the shock assignment process to generate a set of crop-level shock counterfactuals. Second, I weight the shocks counterfactuals by the existing crop shares $share_{c,\ell}$, and sum across crops to estimate a counterfactual shift-share overlap measure for each location (as in Equation 1.4.3). Third, I run the regression specified in Equation 1.4.4, replacing only the actualy shift-share overlap measure with the counterfactual one, and collect the counterfactual β, called $\tilde{\beta}$. Then, I repeat these three steps 1000 times. Finally, I calculate Fisher exact p-values as the fraction of $\tilde{\beta}$ for which $|\tilde{\beta}| \geq |\hat{\beta}|$.

The various shock assignment processes described below are labeled according to their row in Table 1.4. Row (b) randomly re-draws with replacement the crop-level shock from the actual distribution of shocks, which is the RI approach used in the primary analysis (and hence corresponds Table 1.3). This approach avoids putting additional structure on the data and assumes that 1) shock assignment is uncorrelated with crop characteristics, and 2) the actual distribution is representative of true distribution. Row (c) relaxes the first assumption by imposing that shocks are correlated by season by randomly re-drawing with replacement shocks from the actual distribution of shocks for crops within the same season (i.e., rainy, dry, permanent). Row (d) relaxes the second assumption by randomly re-drawing with replacement from a normal distribution defined by the actual distribution's first and second moments. Row (e) combines both features to randomly re-draws with replacement from a normal distribution defined by same-season crops' first and second moments.

Rows (f) and (g) take a different approach in specifying the shock assignment process. Rather than rely on information from the existing distribution, this approach recognizes Malawi's school calendar change as the underlying source of variation in shift-share overlap and thus generates shock counterfactuals from simulations of all possible school calendar changes that could have occurred between 2009 and 2011. First, for both 2009 and 2011, I construct 52 hypothetical school calendars, each starting a Monday and maintaining the structure and length of the year's original school calendar. This is done by effectively shifting the school calendar backward to previous Mondays in the year or forward to future Mondays. I use July 1st as the cutoff for the year (rather than January 1st) so that the simulated 2009, 2010, and 2011 school years do not intersect. Second, for each of 52 simulated school calendars for 2009 and 2011, I estimate the overlap between it and crop-level farming calendars as in Equation 1.4.1. Finally, for the 2,704 possible school calendar changes between 2009 and 2011 (i.e., 52 x 52), I estimate the shock counterfactual $\Delta overlap_c$ for crop c. Together, the $s = 2,704$ simulated shock counterfactuals form the shock distribution for crop c. In row (f), I assume that shock assignment is correlated by crop season (i.e., rainy, dry, permanent), so for each iteration I randomly re-draw with replacement the shock counterfactuals for crop c from the same simulation s for same-season crops. In row (g), I more conservatively assume that shock assignment is correlated across all crops, and thus for each iteration I draw all crop-specific shock counterfactuals from the same simulation s. Further, in row (g), rather than randomly re-drawing with replacement for 1000 iterations, I cycle through all possible 2,704 simulations when constructing the Fisher exact p-values.

A.4.2 Share-weighted shock-level regression

For completion and robustness, I also use share-weighted shock-level regression technique pioneered by Borusyak et al. (2022) to estimate exposure-robust standard errors. First, I transform my location-level outcome and shift-share overlap data into a dataset of exposure-weighted "crop-level" aggregates using their *ssaggregate* command, partialling out $Y_{base,i,\ell}$, $farmshare_{\ell}$, and $w_{i,\ell}$ via the controls option. This aggregatation is equivalent to the "recentering" method described in Borusyak and Hull (2021) for eliminating omitted variable bias when regressing a shift-share variable. Second, following their Proposition 5, I estimate a crop-share-weighted shock-level regression of:

$$\bar{Y}_c^{\perp} = \alpha + \beta ss\bar{overlap}_c^{\perp} + season_c'\gamma + \bar{\epsilon}_c^{\perp} \tag{A.4.1}$$

where $ss\bar{overlap}_c^{\perp}$ is instrumented by crop-level shocks $\Delta overlap_{c,2011-2009}$, and $season_c$ is a vector of crop-level seasonal dummies (rainy, dry, permanent) and a dummy for crops classified as grains included as non-transformed crop-level controls. I cluster standard errors by season and further specify Stata's *ivreg2*'s small option, which requests small-sample statistics (F and t-statistics) and performs a finite sample adjustment. Estimates from Equation A.4.1 produce numerically equivalent estimates of $\hat{\beta}$ as well as exposure-robust standard errors reported in Table 1.4.

A.5 Perceived Value of School Calculations

The appendix provides details for the "back-of-the-envelope" calculations used to estimate the perceived value of time in school in Section 1.5.1.1.

In my reference estimate, I approximate the perceived value of one completed grade level for the average sample household at \$21 USD. I start by equating effect sizes in Table 1.3 columns (1) and (3): 0.34 grades with 3 days of household-farm work (1 day annually for three post-policy years). In Malawi's 2010 Integrated Household Survey (IHS3) sample, children under 15 who are hired to do farm labor are paid an average daily wage of 331 kwacha, or \$2.35 in 2009 USD (following an exchange rate of 1 USD to 141 Malawian kwacha in 2009 (St. Louis Federal Reserve Bank, FRED Economic Data, 2012)). Thus, 0.34 grades = 3 days of farm work = \$2.35 * 3 = \$7.05 $\implies$ 1 grade = \$7.05 USD / 0.34 = \$20.74.

In my upper-bound estimate, I approximate the perceived value of one completed grade level for the average sample household at \$33 USD. To be more generous, this estimate assumes 1) that the positive baseline imbalance is muting the same magnitude of negative effect in the main analysis, and 2) the average daily wage for hired *adult* farm labor. Baseline imbalance in Table 1.1 under $Farm\ Hours_i$ is 3.165 hours, which translates to about 0.59 days of additional lost household-farm hour when applying the conversation assumption discussed in Section 1.5.1: 3.165 effect * 1.21 to convert effect into 10-day effect / 6.5 hours per day = 0.59 days. Then, 0.34 grades = 3.59 days of farm work. In the IHS3 sample, adult hired farm labor is paid an average daily wage of 442 kwacha, or \$3.13 in 2009 USD. Thus, 0.34 grades = 3.59 days of farm work = \$3.13 * 3.59 = \$11.24 $\implies$ 1 grade = \$11.24 / 0.34 = \$33.05.

To approximate the present discounted value to an individual of an additional year of schooling from the Montenegro and Patrinos (2014)'s "Mincerian" estimates, I use income data from the IHS3 survey and reasonable discounting assumptions. First, I consider a range of possible returns to an additional year of schooling corresponding with Montenegro and Patrinos (2014) smaller estimate of 5.2% in 2004 (SD of 3.7) and larger estimate of 9.8% in 2010 (SD of 4.5). In the IHS3 sample, the average annual income across wage and ganyu work for working-age adults with no schooling is 10,165 kwacha or \$72.15 in 2009 USD. Multiplying this by the returns estimates, the annual return for another year of schooling for an non-educated worker is between \$3.75–\$7.07 USD. Treating annual income as an annuity, its present discounted value PDV = Annual return $*\ [1 - (1/1 + r)^n]/r$ is then between \$94.76–\$178.66, where I assume interest rate $r = 0.03$ and number of time periods $n = 48$ as the number of working-age years from ages 18–65.

A.6 Long-Run Analysis

A.6.1 Data Details

Long-run outcome data come from Malawi's Integrated Household Panel Survey (IHPS) 2010-2016 and 2010-2019, both of which were also implemented as part of the World Bank Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) initiative.[8] Citing an increasing number of households and budget/resource constraints, the IHPS tracked households from only half of baseline enumeration areas (EAs) starting in 2016. EA selection was stratified by region and urban/rural designation, and selection oversampled urban areas in order to secure reliable national estimates for both urban and rural areas (Government of Malawi, 2017).

Unfortunately, due to the reduced target sample and additional participant attrition, 2016 and 2019 data are only available for 43.5% and 38.6% of my sample. Additionally, retention into these surveys rounds from the IHPS 2010 is positively correlated with shift-share overlap. Regressing retention on shift-share overlap and controls (akin to Table 1.1's test of retention in the 2013 sample) produces a coefficient of 0.154 [RI p-value = 0.079] for 2016 retention and a coefficient of 0.124 [RI p-value = 0.123] for 2019 retention, suggesting slight over-sampling from locations that experienced greater overlap due to the school calendar change. To address both issues, I broaden my sample to include those ages 0-5 pre-policy who become school-aged in later years (ages 7-12 in 2016 and 10-15 in 2019). This sample includes 1,918 and 1,714 individuals in 2016 and 2016, respectively, and their retention is slightly less imbalanced. In the broader age 0-13 sample, shift-share overlap's correlation has a coefficient of 0.145 [RI p-value = 0.084] for 2016 retention and a coefficient of 0.116 [RI p-value = 0.136] for 2019 retention. Still, I interpret long-run results with some caution.

A.6.2 Additional Results

Table A.7 and Table A.8 break down results of Table 1.11 by age groups: ages 0-5 in 2009 (pre-policy) who become ages 7-12 in 2016 and 10-15 in 2019; ages 6-9 in 2009 who become ages 13-16 in 2016 and 16-19 in 2019; and 10-13 in 2009 who become ages 17-20 in 2016 and 20-23 in 2019. See brief description and summary of results at the end of Section 1.5.2.5.

[8] Documentation for 2010-2016 can be found at: https://microdata.worldbank.org/index.php/catalog/2939, and documentation for 2010-2019 can be found at: https://microdata.worldbank.org/index.php/catalog/3819.

Table A.7: Long-Run Impacts on Grade Level

VARIABLE: Grade in	Pre-Policy Age 0-13 2016	2019	Pre-Policy Age 0-5 2016	2019	Pre-Policy Age 6-9 2016	2019	Pre-Policy Age 10-13 2016	2019
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ssoverlap_{ℓ}	-0.375	-0.128	-0.259	-0.212	-0.264	-0.042	-0.426	0.556
	(0.168)	(0.231)	(0.160)	(0.233)	(0.319)	(0.445)	(0.528)	(0.671)
	[0.014]	[0.537]	[0.094]	[0.220]	[0.249]	[0.913]	[0.061]	[0.195]
Observations	1,918	1,714	929	834	562	515	427	365
R-squared	0.727	0.604	0.483	0.484	0.452	0.414	0.540	0.508
Age at Outcome	Age 7-20	Age 10-23	Age 7-12	Age 10-15	Age 13-16	Age 16-19	Age 17-20	Age 20-23
Base DV Mean	0.65	0.62	0.01	0.00	0.48	0.46	2.26	2.24
Δ DV Mean	2.87	4.50	1.50	3.25	3.93	5.68	4.47	5.68

Notes: Specification and variables are as defined in Table 1.3. Dependent variable $Grade_i$ was measured for a subset of individuals in follow-up panel surveys in either 2016 or 2019. Conventional robust standard errors in parentheses. Randomization inference p-values in square brackets.

Table A.8: Long-Run Impacts on Farm Work

	Pre-Policy Age 0-13		Pre-Policy Age 0-5		Pre-Policy Age 6-9		Pre-Policy Age 10-13	
VARIABLE: Farmed in	2016	2019	2016	2019	2016	2019	2016	2019
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ssoverlap_ℓ	-0.102	-0.158	-0.132	-0.218	-0.053	-0.120	-0.036	-0.048
	(0.046)	(0.050)	(0.065)	(0.072)	(0.081)	(0.083)	(0.100)	(0.112)
	[0.022]	[0.010]	[0.023]	[0.014]	[0.345]	[0.013]	[0.493]	[0.674]
Observations	1,918	1,714	929	834	562	515	427	365
R-squared	0.291	0.164	0.188	0.178	0.263	0.215	0.209	0.213
Age at Outcome	Age 7-20	Age 10-23	Age 7-12	Age 10-15	Age 13-16	Age 16-19	Age 17-20	Age 20-23
Base DV Mean	0.65	0.62	0.01	0.00	0.48	0.46	2.26	2.24
Δ DV Mean	-0.15	-0.03	0.30	0.50	0.19	0.20	-1.55	-1.57

Notes: Specification and variables are as defined in Table 1.3. Dependent variable $Farmed_i$ was measured for a subset of individuals in follow-up panel surveys in either 2016 or 2019. Conventional robust standard errors in parentheses. Randomization inference p-values in square brackets.

APPENDIX B

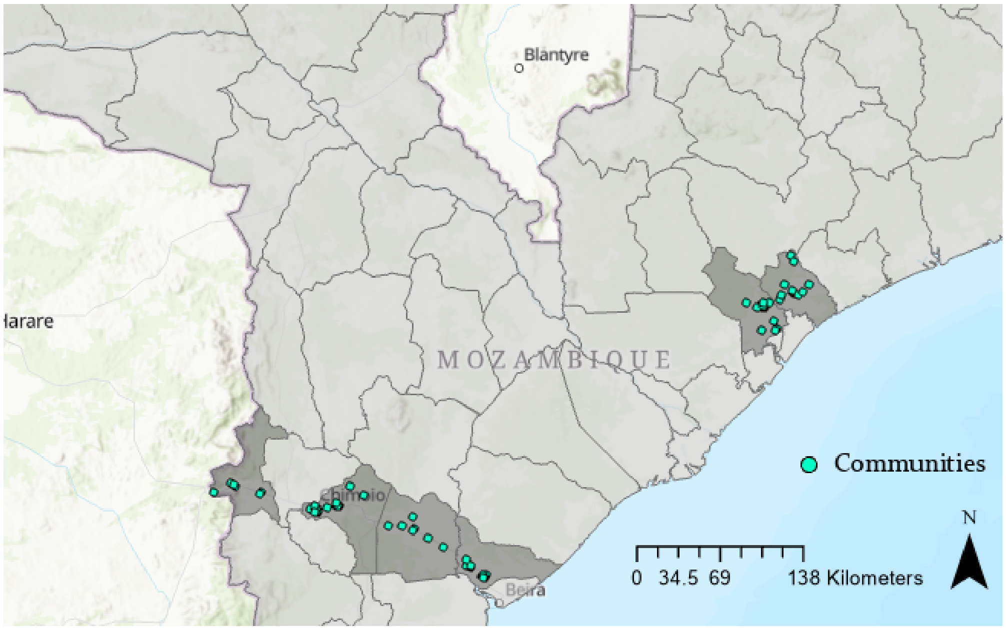
Appendix to Chapter 2

In this Appendix, we often refer to survey by its round number: Pre-baseline is Round 1, baseline is Round 2, endline is Round 3, and post-endline is Round 4.

B.1 Study Area

The Mozambican government declared a State of Emergency due to the COVID-19 pandemic on March 31, 2020 (Republic of Mozambique, 3/31/2020). The government recommended social distancing (at least 1.5 meters) and required it at public and private institutions and gatherings. The government also suspended schools, required masks at funerals and markets, banned gatherings of 20 or more, and closed bars, cinemas and gymnasiums (Republic of Mozambique, 4/1/2020). The government stopped short of implementing a full economic "lockdown" due to its economic costs (Siuta and Sambo, April 1, 2020; Jones et al., 2020). On August 5, 2020, the government renewed the State of Emergency (Republic of Mozambique, 8/5/2020), called for improved mask-wearing, and announced a schedule for loosening restrictions (Nyusi, 8/5/2020). In September 2020, the government loosened some restrictions, including resuming religious services at 50% capacity (Nyusi, 9/5/2020; U.S Embassy in Mozambique).

Figure B.1: Study Area



Notes: The country of Mozambique is shaded in light gray. District borders are defined by a black line. Districts within this sample are shaded in dark gray. The geographic center for the 76 communities encompassed in this sample are highlighted as cyan points on the map.

Study participants come from 76 communities in central Mozambique. The study communities are in seven districts of three provinces: Dondo and Nhamatanda in Sofala province; Gondola, Chimoio and Manica in Manica province; and Namacurra and Nicoadala in Zambezia province. These 76 communities are mapped in Figure C.1. Compared to other communities in Mozambique, the study areas are relatively accessible to transport corridors (highways and ports) and are thus important geographic conduits for infectious disease.

We collected survey data in three rounds between July 10 and November 18, 2020. Figure 2.1 depicts the study timeline below a rolling average of new Mozambican COVID-19 cases. We piloted surveys in Round 1. Immediately before the Round 2 survey, we randomly assigned households to treatments and submitted our pre-analysis plan to the AEA RCT Registry. The Round 2 survey served as a baseline, and was immediately followed (on the same phone call) by our treatment interventions. Round 3 was our endline survey. Surveys collected data on COVID-19 knowledge, beliefs, and behaviors. While data collection for Round 3 began only one day after completion of Round 2, there was a minimum of 3.0 weeks and average of 6.3 weeks between Rounds 2 and 3 surveys for any given respondent. While the Round 1 survey occurred when new COVID-19 cases remained relatively steady, both the Round 2 and Round 3 surveys occurred during a period of substantial growth in new COVID-19 cases.

Details on our Round 4 survey to test long-run impacts can be found in Appendix B.7.

B.2 COVID-19 Knowledge Questions

Survey questions measured COVID-19-related knowledge in the three main categories: 1) general knowledge, which included questions on risk factors, transmission, and symptoms; 2) preventive actions, which included questions on social distancing (i.e., how to prevent spreading COVID-19 to others), and household prevention (i.e., how to prevent spreading COVID-19 to yourself and your household); and 3) government policies (i.e., official actions taken by the national government of Mozambique to address COVID-19).

In Round 1, we piloted a set of 71 questions (larger than our eventual pre-specified set for Rounds 2 and 3). The Round 1 question pool had 71 possible knowledge questions: 21 on general knowledge (6 on risk factors, 8 on transmission, 7 on symptoms), 30 on preventive actions (14 on social distancing, 16 on household prevention), and 20 on government policy. For brevity, we do not list the full set of 71 questions in this appendix.[1]

In Round 1, we asked each respondent 20 knowledge questions randomly selected from within each question type: 6 on general knowledge (2 on risk factors, 2 on transmission and 2 on main symptoms), 8 on preventative actions (4 on social distancing actions and 4 on household prevention actions), and 6 on government policy. The Round 1 Test Score (used as a pre-specified control variable in regressions) is the share of these 20 knowledge questions answered correctly by a respondent.

Criteria for selecting questions from the Round 1 pilot for the final set of Round 2 and 3 questions included identifying Round 1 questions with larger shares of incorrect answers and wide variance in responses, each question's medical significance and relevance to COVID-19 prevention, as well as the diversity of the final question pool (e.g., a mix of "yes" and "no" correct responses). In total, 33 knowledge questions were taken from Round 1, six questions were slightly modified from Round 1 to clarify or update the wording to reflect current information, and one new question was added.

The final question pool used for Round 2 and Round 3 has 40 questions: 12 on general knowledge (4 on risk factors, 4 on transmission, 4 on symptoms), 16 on preventive actions (8 on social distancing, 8 on household prevention), and 12 on government policy. This question pool was pre-specified.[2]. The questions are listed in Tables B.1, B.2, and B.3. Details on questions included in our Round 4 survey can be found in Appendix B.7.

In Round 2, respondents were asked 20 knowledge questions from the pre-specified question pool, randomly selected from within each question subcategory: 6 on general knowledge (2 on risk factors, 2 on transmission and 2 on main symptoms), 8 on preventative actions (4 on

[1] The list of 71 Round 1 pilot questions can be found on our project website [URL here].

[2] See American Economic Association's RCT Registry, ID number AEARCTR-0005862: [URL here]

social distancing actions and 4 on household prevention actions), and 6 on government policy. The Round 2 Test Score (used as a pre-specified control variable in regressions) is the share of these 20 knowledge questions answered correctly by a respondent.

In Round 3, we asked respondents all 40 knowledge questions from the pre-specified question pool: 12 on general knowledge, 16 on preventive action, and 12 on government policy. The Overall Test Score (one of two pre-specified primary outcome variables) is the share of these 40 knowledge questions answered correctly by a respondent. Of these 40 knowledge questions, survey respondents will have been asked 20 of these knowledge questions in Round 2, immediately prior to treatment implementation. The Teaching-Eligible Test Score (the other one of two pre-specified primary outcome variables) is the share of these 20 knowledge questions (also asked in Round 2) answered correctly by a respondent in Round 3. The other 20 knowledge questions asked in Round 3 would not have been asked in Round 2 (but could have been asked in Round 1).

Table B.4 presents summary statistics in the control group (N=847) of the Overall Test Score and the Teaching-Eligible Test Score, as well as the Rounds 1 and 2 Test Scores. In Rounds 1 and 2, respondents answered 71.6% and 76.9% of questions correctly. We observe a small increase in COVID-19 knowledge over time, with knowledge in both Round 3 indices increasing to over 78%. We also observe in Round 3 that the Overall Test Score and the Teaching-Eligible Test Score are remarkably similar, suggesting that the small increase in knowledge over time is not likely to be driven by repeated exposure to the same questions.

Table B.1: Pre-specified "General Knowledge" Questions and Correct Answers

Risk Factors: Who do you think is more likely to die from a coronavirus infection?	
(1)	An adult who does not smoke or an adult who does smoke (Second)
(2)	A 60-year-old man with diabetes and hypertension and 60-year-old man with blindness and hearing loss (First)
(3)	A grandparent or their grandchild (First)
(4)	A healthy 30-year-old adult or a healthy 60-year-old adult (Second)
Transmission: How is coronavirus spread?	
(5)	Droplets from the cough of an infected person (Yes)
(6)	Drinking unclean water (No)
(7)	Sexually transmitted (No)
(8)	Mosquito bites (No)
Symptoms: What are the main symptoms of coronavirus?	
(9)	Fever (Yes)
(10)	Cough and breathing difficulties (Yes)
(11)	Pain with urination (No)
(12)	New loss of taste or smell (Yes)

Notes: Correct answers in parentheses. In Round 2, two questions were randomly selected to be asked of the respondent from each sub category. In Round 3 all questions were asked of each respondent.

Table B.2: Pre-specified "Preventive Actions" Questions and Correct Answers

Social Distancing Actions: Will this action prevent spreading coronavirus to yourself and others?	
(1)	Shop in crowded areas like informal markets (No)
(2)	Gather with several friends (No)
(3)	Help the elderly avoid close contact with other people, including children (Yes)
(4)	If show symptoms of coronavirus, immediately inform my household and avoid people (Yes)
(5)	Drinking alcohol in bars (No)
(6)	Wear a face mask if showing symptoms of coronavirus (Yes)
(7)	Instead of meeting in person, call on the phone or send text message (Yes)
(8)	Allow children to build immunity by playing with children from other households (No)
Household Prevention Actions: Will this action prevent spreading coronavirus to yourself and others?	
(9)	Drinking hot tea (No)
(10)	Open the windows to increase air circulation (Yes)
(11)	Wear a face mask in public when you are healthy (Yes)
(12)	Eat foods with lemons or garlic or pepper (No)
(13)	Drink only treated water (No)
(14)	Spray alcohol and chlorine all over your body (No)
(15)	Avoid close contact with anyone who has a fever and cough (Yes)
(16)	Avoid taking taxi-bicycle or taxi-mota to go out (Yes)

Notes: Correct answers in parentheses. In Round 2, four questions were randomly selected to be asked of the respondent from each sub category. In Round 3 all questions were asked of each respondent.

Table B.3: Pre-specified "Government Policy (Actions)" Questions and Correct Answers

Government Actions: is the government of Mozambique currently taking this action to address coronavirus?	
(1)	Order a 14 day home quarantine for all persons who have had direct contact with confirmed cases of COVID-19 (Yes)
(2)	Close all airports (No)
(3)	Suspend religious services and celebrations (Yes)
(4)	Allow a maximum of 50 participants in funeral ceremonies where COVID-19 is NOT the cause of death (Yes)
(5)	Banning personal travel between provinces (No)
(6)	Prohibit use of minibuses for public transportation (No)
(7)	Ask household to not visit patients infected by COVID-19 at hospitals (Yes)
(8)	Close government offices not related to health (No)
(9)	Order all citizens to wear masks when going out of their homes (No)
(10)	Prohibit funerals for those with coronavirus or COVID-19 (No)
(11)	Declare a State of Emergency (Yes)
(12)	Plan to resume Grade 12 classes this year before other primary and secondary grades (Yes)

Notes: Correct answers in parentheses. In Round 2, six questions were randomly selected to be asked of the respondent. In Round 3 all questions were asked of each respondent.

Table B.4: Summary Statistics of Test Score (TS) in Control Group

Outcome	Round	Mean	Std. Dev.	Min	Max
Round 1 TS	Round 1	0.716	0.116	0.25	1
Round 2 TS	Round 2	0.769	0.121	0.35	1
Overall TS	Round 3	0.781	0.108	0.45	1
Teaching-Eligible TS	Round 3	0.784	0.123	0.35	1

Notes: Number of observations in control group is 847. Rounds 1 and 2 Test Scores pre-specified as control variables in regressions. Overall test score and Teaching-Eligible test score (Round 3) are the two pre-specified primary outcome variables in this study. They were referred to in the pre-analysis plan (PAP) as "Knowledge Index" and "Feedback-Eligible Knowledge Index", respectively.

B.3 Treatment Details

We randomized respondents to one of four treatment arms: 1) Incentive, 2) Teaching, 3) Incentive plus Teaching (Joint), and 4) a control group. Table B.5 shows the distribution of respondents across treatment arms in the Round 2 and Round 3 samples. Retention in the sample is balanced across treatment arms.

All treatments were initiated by enumerators directly following the Round 2 (baseline) survey as part of the same phone call. If a respondent was randomly assigned to a treatment, the corresponding intervention text would appear on the enumerator's computer tablet. Enumerators read a script aloud exactly as shown below. Following the treatment, respondents were asked if they would like the information repeated. Of the N=832 receiving the incentive treatment and N=856 receiving the teaching treatment, only 6.0% and 6.7% asked for the script to be repeated, respectively.

Table B.5: Distribution of Respondents Across Treatment Groups

Treatment Arm	Round 2 Sample	Round 3 Sample	Probability of Random Assignment
Incentive	433 (19.5%)	414 (19.6%)	20%
Teaching	441 (19.8%)	418 (19.7%)	20%
Incentive plus Teaching (Joint)	464 (20.8%)	438 (20.7%)	20%
Control Group	888 (39.9%)	847 (40.0%)	40%
TOTAL	2,226	2,117	100%

Notes: Randomization of respondents to treatment groups occurred immediately prior to administration of Round 2 baseline survey and treatment.

Script for Incentive treatment. At baseline, after questioning: "We plan to call you for another follow-up phone survey in about two or three weeks. During this survey, we will ask you many of the same questions that we asked you today, and some new questions. This survey will also be confidential. For responding to this additional survey, you will receive 50Mts. Additionally, we will offer you 5Mts for every correct response you give us in our next phone survey to reward your knowledge of coronavirus! This reward will apply to the same questions that we asked you today and new questions about coronavirus symptoms, prevention, how it spreads, who is most at risk, and actions taken by the government of Mozambique. If you answer all of the questions correctly, you could earn up to 200Mts in addition to your 50Mts participation fee in our next survey!"

For the Incentive treatment, additional text was read to respondents at endline. First, at the start of the endline survey, enumerators reminded treated respondents that both previous and new knowledge questions were eligible for the Teaching incentive. Second, at the end of the endline survey, the number of correct answers and the resulting incentive were calculated in the SurveyCTO program (and not by enumerators). Then this information was presented in a final text in which enumerators told respondents how many questions they answered correctly and additional meticais consequently earned.

At endline, before questioning, just after consent: "As you were told in the previous survey, we will offer you 5Mts for every correct response you give us today to reward your knowledge of coronavirus! This reward will apply to the same questions that we asked you in the previous survey and new questions about coronavirus symptoms, prevention, how it spreads, who is most at risk, and actions taken by the government of Mozambique. If you answer all of the questions correctly, you could earn up to 200Mts in addition to your 50Mts participation fee!"

At endline, after questioning, just prior to payment: "In our previous survey, we offered you 5Mts for every correct response you gave us today to reward your knowledge of coronavirus. Today you correctly answered XX out of 40 coronavirus knowledge questions. Therefore, today you will receive an additional XX Mts in addition to your 50 Mts participation gift!" The additional amount was then added to the respondent's MPesa transfer or phone credit recharge.

Script for Teaching treatment. "Now, I want to provide you some feedback on your responses from today's survey on questions about actions that prevent the spread of coronavirus.

- Respondents are randomly given tailored feedback to their response to COVID-19 **prevention questions**. We inform them of a subset of their correct responses and correct a subset of their incorrect responses. The script for each action is as follows: For " *[insert action]*", you chose *[insert respondents choice]* . Your answer is *[insert respondents choice]* . The correct answer is *[insert pre-specified correct choice: YES or NO]* . This action *[insert pre-specified correct choice: WILL or WILL NOT]* prevent spreading coronavirus to yourself and others."

- Respondents are randomly given tailored feedback to their response to COVID-19 **general knowledge questions**. We inform them of a subset of their correct responses and correct a subset of their incorrect responses. The script for each question is as follows: "For " *[insert question]*", you chose *[insert respondents choice]* but the correct answer is *[insert pre-specified correct answer]* . *[insert pre-specified correct answer statement]*."

For the 6 general knowledge and 6 government action questions asked in Round 2, feedback was given for all incorrect answers. For the 8 preventive action questions asked in Round 2, feedback was given for roughly half of all correct answers and half of all incorrect answers. This was done to test the efficacy of positive feedback versus negative feedback, which is currently under analysis and not discussed in this paper.

Script for Incentive plus Teaching (Joint) treatment. This is a combination of the Incentive and Teaching treatments. Both scripts are read to the respondent. The Incentive script is always read first, before the Teaching script.

B.4 Attrition and Balance

Table B.6 checks that attrition and baseline variables are balanced with respect to treatment assignment.

Attrition between Round 2 (baseline) and Round 3 (endline) is low, at only 4.6% overall, and is less than 5.6% in each of the seven districts surveyed. Balance in attrition is confirmed in column 1, which starts with the Round 2 (baseline) sample and regresses treatments on an indicator equal to one if the respondent was not reached for the Round 3 (endline) survey. None of the treatments have a large or statistically significant effect on attrition. Achieving balance in attrition was not obvious *a priori* since respondents offered the knowledge incentive treatment had a higher expected payoff for participation in the Round 3 survey, though empirically this has no effect.

We examine balance in baseline household characteristics in columns 2-4, which examine the final Round 3 sample and regresses treatments on Round 1 measures of household income, an index of food insecurity, and an indicator for presence of an older adult over 60 years. Treatments are balanced at the 95% confidence level across all three outcomes. In column 5, we test for balance in the baseline Round 2 Test Score, the primary outcome at baseline.[3] We unfortunately find chance imbalance: a statistically significantly positive correlation between the baseline outcome and the standalone Incentive treatment, but not in other treatment arms. Further analysis revealed that this imbalance is heavily concentrated in Nhamatanda, one of the seven districts surveyed, and that the imbalance is no longer statistically significant when Nhamatanda is excluded from the sample: results shown in columns 6 and 7.

Note that our pre-specified primary regression equations include controls for Round 1 and Round 2 test scores, including this Round 2 Test Score for which we are finding baseline imbalance. To further verify that baseline imbalance in Nhamatanda is not driving our primary results, we re-run our primary analysis as described in the 3.4.2 section but excluding observations from Nhamatanda district from the sample. Columns 8 and 9 present this robustness check, showing that the results are not qualitatively different from the ones presented in Table B.8. Indeed, when excluding Nhamatanda, the p-values on the tests that $\lambda = 0$ are even smaller than in our main analyses. We conclude that our primary results are not driven by the chance imbalance in the Round 2 (baseline) values of the outcome variables.

We further test for baseline balance in educational attainment in Table B.7. As this was not measured in the pre-baseline or baseline surveys, we link respondents to their individual-level data from a prior household survey (Yang et al., 2021) and obtain a measure

[3] In Round 2 there is only one Test Score, based on a randomly-selected 20 questions, as described previously.

of years of schooling for 74.8% of the endline sample. In Column 1, we test for balance in the availability of these data and find balance. In Column 2, we test for balance in educational attainment if these data are available and find a positive correlation between a respondent's years of schooling and the Incentive plus Teaching (Joint) treatment, significant at the 95% confidence level. Due to this imbalance, in Columns 3-4, we test and confirm the robustness of the two pre-specified primary analyses to controlling for years of schooling in the regression.[4] Statistical significance for all treatment coefficients remains unchanged from the main results in Table 2.3 Columns 1-2, and adjusted point estimates differ by less than 0.001. Further, in Columns 5-6, we test for heterogeneous treatment effects by years of schooling and find no significant interaction. We conclude that despite chance imbalance in educational attainment for the Joint treatment, the main conclusions of the paper remain valid.

[4] We "dummy out" missing observations by setting missing values of years of schooling to zero and including a dummy variable for data availability as a control in the regression as well.

Table B.6: Attrition and Baseline Balance

VARIABLES	(1) Dummy if attrited between R2 & R3	(2) R1: Household income last week	(3) R1: Food insecurity index	(4) R1: Older adult (60+) in Household	(5) Baseline test score (TS)	(6) Baseline TS (Nhamatanda)	(7) Baseline TS (Not Nhamatanda)	(8) Overall TS (Not Nhamatanda)	(9) Teaching-Eligible TS (Not Nhamatanda)
Incentive	-0.0031	-14.91	0.0844	0.0149	0.0145	0.0673	0.0083	0.0193	0.0141
	(0.0121)	(180.50)	(0.0904)	(0.0289)	(0.0066)	(0.0218)	(0.0069)	(0.0057)	(0.0064)
Teaching	0.0065	209.90	0.0262	0.0185	0.0023	0.0235	-0.0002	0.0153	0.0274
	(0.0128)	(210.70)	(0.0911)	(0.0283)	(0.0070)	(0.0240)	(0.0073)	(0.0056)	(0.0065)
Incentive plus Teaching (Joint)	0.0120	206.30	0.0724	0.0367	0.0055	0.0016	0.0053	0.0494	0.0573
	(0.0130)	(211.70)	(0.0930)	(0.0282)	(0.0068)	(0.0255)	(0.0070)	(0.0058)	(0.0063)
$\hat{\lambda}$								0.0149	0.0158
								(0.0087)	(0.0099)
Observations	2,226	1,873	2,117	2,096	2,117	214	1,903	1,903	1,903
R-squared	0.030	0.043	0.125	0.058	0.114	0.061	0.114	0.312	0.321
Districts	All	All	All	All	All	Nhamatanda	NOT Nhamatanda	NOT Nhamatanda	NOT Nhamatanda
Control Mean DV	0.0462	1049	2.407	0.335	0.769	0.719	0.775	0.787	0.790
Control SD DV								0.107	0.123
p-value: $\lambda = 0$								0.0871	0.1113
p-value: Incentive = Teaching								0.5381	0.0794
p-value: Incentive = Joint								0.0000	0.0000
p-value: Teaching = Joint								0.0000	0.0001

Notes: Column 1: For Round 2 sample, dummy if attrited between Round 2 baseline (post-intervention) and Round 3 endline. Columns 2-4: Round 1 baseline variables—Household income last week is the specific amount reported, if given, or otherwise is imputed from the selected income range. The food insecurity index is the total of five indicator variables: 1) lack of food in last seven days; unable to buy usual amount of food due to 2) market shortages, 3) high prices, 4) drop in income; and reduction in number of meals/portions. Older adult in household is a dummy variable indicating if the respondent reports that anyone in the household is aged 60 years or over. Column 5: Round 2 baseline Test Score (TS). Column 6: Baseline TS for sample in Nhamatanda district. Column 7: Baseline TS for sample not in Nhamatanda. Column 8-9: Endline outcomes as described Table 2.3 Columns 1-2 for sample not in Nhamatanda. $\hat{\lambda}$ is coefficient on "Incentive plus Teaching" (Joint) minus sum of coefficients on "Incentive" and "Teaching". All regressions also include community fixed effects. Robust standard errors in parentheses.

Table B.7: Years of Schooling: Baseline Balance and Treatment Heterogeneity

VARIABLES	(1) Dummy if have schooling data	(2) Years of schooling	(3) Overall test score	(4) Feedback-eligible test score	(5) Overall test score	(6) Feedback-eligible test score
Incentive	-0.0317	0.289	0.0195	0.0151	0.0290	0.0204
	(0.0263)	(0.256)	(0.00537)	(0.00598)	(0.00832)	(0.00924)
Teaching	-0.0138	0.160	0.0157	0.0285	0.0157	0.0303
	(0.0259)	(0.254)	(0.00547)	(0.00636)	(0.00926)	(0.0107)
Incentive plus Teaching	0.00404	0.564	0.0487	0.0572	0.0464	0.0550
	(0.0257)	(0.244)	(0.00551)	(0.00596)	(0.00926)	(0.00992)
Years of schooling			0.00237	0.00247	0.00261	0.00263
			(0.000706)	(0.000774)	(0.000881)	(0.000987)
Incentive x Years of schooling					-0.00159	-0.000887
					(0.00107)	(0.00119)
Teaching x Years of schooling					3.17e-06	-0.000299
					(0.00117)	(0.00134)
Incentive plus Teaching x Years of schooling					0.000330	0.000316
					(0.00114)	(0.00122)
Dummy if have schooling data			-0.0280	-0.0269	-0.0278	-0.0268
			(0.00777)	(0.00847)	(0.00776)	(0.00847)
Observations	2,117	1,584	2,117	2,117	2,117	2,117
R-squared	0.046	0.149	0.325	0.337	0.326	0.337
Control Mean DV	0.753	7.853	0.784	0.784	0.784	0.784
Control SD DV	0.431	3.776	0.123	0.123	0.123	0.123

Notes: Column 1: Dummy if respondent's years of schooling is known from a prior household survey (Yang et al., 2021). Column 2: Respondent's years of schooling (if known). Columns 3-4: The two pre-specified analyses described in Table 2.3 with additional controls for respondent's years of school and a dummy for data availability. Column 5-6: Testing for treatment effect heterogeneity by years of schooling in regressions of the two pre-specified outcomes. All regressions also include community fixed effects. Columns 3-6 including controls for pre-treatment test scores. Robust standard errors in parentheses.

B.5 Populated Pre-analysis Plan

On August 25, 2020, prior to baseline data collection, we uploaded our pre-analysis plan (PAP) "Learning about COVID-19: Improving Knowledge via Incentives and Feedback" to the American Economic Association's RCT Registry, registration ID number AEARCTR-0005862: https://doi.org/10.1257/rct.5862-1.0.

We follow Duflo et al. (2020), assembling the full set of pre-specified analyses in a Populated PAP document. The full Populated PAP can be accessed at our research website: https://fordschool.umich.edu/mozambique-research/combating-covid-19. Additionally, in this appendix, we present results from the Populated PAP for the pre-specified primary analysis. These results are substantively duplicative of and yield very similar conclusions to the primary analyses we present in the main text.

Note that we adhere to the nomenclature we used in the main text to refer to outcomes and treatment conditions that differ from some nomenclature used in a Pre-Analysis Plan (PAP). Therefore, we refer to the treatments as "Incentive" and "Teaching", whereas in the PAP these are referred to as "Knowledge Incentive" and "Tailored Feedback", respectively. Additionally, we refer to the two primary outcome variables as 1) "Overall Test Score" and 2) "Teaching-Eligible Test Score", whereas in the PAP these are referred to 1) the Knowledge Index, and 2) the Feedback-Eligible Knowledge Index, respectively.

B.5.1 Primary Analyses

We estimate intent-to-treat (ITT) effects using the following ordinary-least-squares (OLS) regression specifications. To estimate the causal effect of the Incentive treatment, we run:

$$Y_{i,j,t=3}^{all} = \alpha_0 + \alpha_1 Incentive_{ij} + \alpha_2 Teaching_{ij} + \alpha_3 Joint_{ij} + \eta \mathbf{B}_{ijt} + \gamma_i + \varepsilon_{ij} \tag{B.5.1}$$

where $Y_{i,j,t=3}^{all}$ is the Overall Test Score for respondent i in community j, measured in Round 3 survey; $Incentive_{ij}$, $Teaching_{ij}$, and $Joint_{ij}$ are indicators for inclusion in the respective treatment groups; $\mathbf{B}_{ijt}$ is a vector representing the share of correct answers to questions asked in Round 1 and Round 2, respectively [5]; γ_i are community fixed effects; and ε_{ij} is a mean-zero error term. We report robust standard errors.

To estimate the causal effect of the Teaching and Joint treatments, we run:

$$Y_{i,j,t=3}^{teaching} = \beta_0 + \beta_1 Incentive_{ij} + \beta_2 Teaching_{ij} + \beta_3 Joint_{ij} + \eta \mathbf{B}_{ijt} + \gamma_i + \varepsilon_{ij} \tag{B.5.2}$$

[5] The average respondent correctly answered 72.1% and 77.3% of the 20 knowledge questions in Rounds 1 and 2, respectively.

where $Y_{i,j,t=3}^{teaching}$ is the Teaching-Eligible Test Score for respondent i in community j, measured in Round 3 (endline survey), and other right-hand side variables are as specified in Equation B.5.1.

Results from estimating these equations are in Table B.8. Overall, the coefficient signs, magnitudes, and statistical significance levels are very similar in Column 1 (for the Overall Test Score) and Column 2 (for the Teaching-Eligible Test Score). Each of the treatments has positive effects on the outcomes that are statistically significant at conventional levels even after pre-specified multiple hypothesis testing adjustment across three coefficients in the two regressions (p-values in square brackets, <0.001 in each case). The estimate, $\hat{\lambda}$, of the complementarity parameter is nearly identical across the two regressions.

Table B.8: Regression of Test Score (TS) on Treatments

VARIABLES	(1) Overall Test Score (TS)	(2) Teaching-Eligible TS
Incentive	0.0200	0.0156
	(0.0054)	(0.0060)
	[0.0003]	
Teaching	0.0160	0.0288
	(0.0055)	(0.0064)
		[0.0003]
Incentive plus Teaching (Joint)	0.0496	0.0581
	(0.0055)	(0.0060)
		[0.0003]
$\hat{\lambda}$	0.0136	0.0137
	(0.0084)	(0.0095)
Observations	2,117	2,117
R-squared	0.319	0.333
Control Mean DV	0.781	0.784
Control SD DV	0.108	0.123
p-value: $\lambda = 0$	0.1048	0.1462
p-value: λ = -0.0265	0.0000	0.0000
p-value: Incentive = Teaching	0.5292	0.0713
p-value: Incentive = Joint	0.0000	0.0000
p-value: Teaching = Joint	0.0000	0.0001

Notes: The Overall Test Score (TS) is the share of correct answers to all 40 knowledge questions in Round 3: 12 on general knowledge, 16 on preventive actions, and 12 on government policy. The Teaching-Eligible TS is the share of correct answers to the 20 knowledge questions in Round 3 that were eligible for the Teaching treatment (i.e., also asked in Round 2): 6 on general knowledge, 8 on preventive actions, and 6 on government policy. λ is the complementarity parameter (see Section 2 of main text). $\hat{\lambda}$ is coefficient on "Incentive plus Teaching" (Joint) minus sum of coefficients on "Incentive" and "Teaching". P-values adjusted for pre-specified multiple hypothesis testing are in square brackets. All regressions also include community fixed effects and controls for pre-treatment (Rounds 1 and 2) Test Scores. Robust standard errors in parentheses.

We also pre-specified other secondary analyses. First, we pool the Incentive, Teaching, and Joint treatments together to examine the effect of any treatment on the primary outcomes. Results in Table B.9 for the coefficient on the indicator for receiving any treatment, "Pooled Treatment", is statistically significantly positive at conventional levels in each regression.

Second, we analyze impacts of the treatments on test scores based on topical categories:

general knowledge, preventive actions, and government policies. Regressions are as described above but replacing the respective test scores with corresponding outcomes for the indicated categories. Results in Table B.10 are broadly similar to the estimates in Table B.8. The estimated complementarity parameter $\hat{\lambda}$ appears largest (most positive) for the preventive actions subcategory (Columns 2 and 5).

Table B.9: Regression of Test Score (TS) on Pooled Treatment

	(1)	(2)
VARIABLES	Overall Test Score (TS)	Teaching-Eligible TS
Pooled Treatments	0.0289	0.0346
	(0.0041)	(0.0045)
Observations	2,117	2,117
R-squared	0.308	0.320
Control Mean DV	0.781	0.784
Control SD DV	0.108	0.123

Notes: Column 1: the Overall Test Score (TS) is the share of correct answers to all 40 knowledge questions in Round 3: 12 on general knowledge, 16 on preventive actions, and 12 on government policy. Column 2: the Teaching-Eligible TS is the share of correct answers to the 20 knowledge questions in Round 3 that were eligible for the Teaching treatment (i.e., also asked in Round 2): 6 on general knowledge, 8 on preventive actions, and 6 on government policy. All regressions also include community fixed effects and controls for pre-treatment (Rounds 1 and 2) Test Scores. Robust standard errors in parentheses.

Third, we analyze impacts of the treatments on self-reported COVID-19 preventive behaviors. Outcomes include respondents' stated support for social distancing, self-report of following government social distancing recommendations, and the number of preventive actions taken by the household to prevent the spread of COVID-19. All outcomes are socially desirable and advocated by the government, so positive coefficients would be considered "good". Results in Table B.11 are mixed and inconclusive. Six out of nine coefficients in the table are positive, and three are negative. Two out of nine coefficients are statistically significantly different from zero at conventional levels: the negative coefficient on Teaching in Column 1, and the positive coefficient on Incentive in Column 2.

Table B.10: Regression of Test Score (TS) Categories on Treatments

VARIABLES	(1) General TS	(2) Preventive TS	(3) Government TS	(4) Teaching-Eligible General TS	(5) Teaching-Eligible Preventive TS	(6) Teaching-Eligible Government TS
Incentive	0.0094	0.0184	0.0421	0.0018	0.0118	0.0419
	(0.0084)	(0.0065)	(0.0083)	(0.0099)	(0.0088)	(0.0099)
Teaching	0.0154	0.0125	0.0223	0.0265	0.0234	0.0299
	(0.0085)	(0.0067)	(0.0087)	(0.0102)	(0.0092)	(0.0109)
Incentive plus Teaching (Joint)	0.0374	0.0487	0.0644	0.0415	0.0535	0.0749
	(0.0087)	(0.0065)	(0.0084)	(0.0103)	(0.0087)	(0.0100)
$\hat{\lambda}$	0.0126	0.0178	0.0001	0.0133	0.0183	0.0031
	(0.0131)	(0.0100)	(0.0127)	(0.0157)	(0.0136)	(0.0154)
Observations	2,117	2,117	2,117	2,117	2,117	2,117
R-squared	0.199	0.204	0.211	0.206	0.257	0.189
Control Mean DV	0.790	0.768	0.790	0.797	0.827	0.789
Control SD DV	0.159	0.116	0.165	0.189	0.170	0.202
p-value: $\lambda = 0$	0.3333	0.0759	0.9955	0.3985	0.1774	0.8410
p-value: Incentive = Teaching	0.5361	0.4486	0.0410	0.0354	0.2756	0.3090
p-value: Incentive = Joint	0.0048	0.0001	0.0170	0.0008	0.0000	0.0025
p-value: Teaching = Joint	0.0278	0.0000	0.0000	0.2130	0.0036	0.0001

Notes: The Overall Test Score (TS) categories (Columns 1-3) are the share of correct answers in Round 3 to the 12 questions on general knowledge, 16 questions on preventive actions, and 12 questions on government policy, respectively. The Teaching-Eligible TS categories (Columns 4-6) are the share of correct answers to the questions in Round 3 that were eligible for the Teaching treatment (i.e., also asked in Round 2): 6 on general knowledge, 8 on preventive actions, and 6 on government policy, respectively. λ is the complementarity parameter (see Section 2 of main text). $\hat{\lambda}$ is coefficient on "Incentive plus Teaching" (Joint) minus sum of coefficients on "Incentive" and "Teaching". All regressions also include community fixed effects and controls for pre-treatment (Rounds 1 and 2) Test Scores. Robust standard errors in parentheses.

Table B.11: Regressions of Behavior on Treatments

VARIABLES	(1) Supports Social Distancing	(2) Followed Government Recommendation in past 14 days	(3) Preventive Action Practice in Past 14 Days
Incentive	0.0067	0.0278	0.0130
	(0.0040)	(0.0110)	(0.0072)
Teaching	-0.0175	0.0121	-0.0007
	(0.0085)	(0.0123)	(0.0075)
Incentive plus Teaching (Joint)	-0.0017	0.0104	0.0076
	(0.0058)	(0.0127)	(0.0072)
Observations	2,117	2,117	2,117
R-squared	0.067	0.065	0.278
Control Mean DV	0.992	0.945	0.764
Control SD DV	0.0906	0.229	0.138
p-value: Incentive = Teaching	0.0051	0.2019	0.1122
p-value: Incentive = Joint	0.1398	0.1697	0.5232
p-value: Teaching = Joint	0.1053	0.9049	0.3361

Notes: Column 1: indicator equal to one if respondent answers “yes” to supporting “the practice of social distancing (SD) to prevent the spread of coronavirus” and zero otherwise. Column 2: indicator for SD according to self if respondent answered “yes” to observing the government’s recommendations on SD in the last 14 days, and zero otherwise. Column 3: share of eight social distancing behaviors (Column 4) and five household prevention behaviors (Column 5) that the respondents report doing in the last 14 days. All regressions also include community fixed effects and controls for pre-treatment (Rounds 1 and 2) Test Scores. Robust standard errors in parentheses.

Table B.12: Regressions of Interactions of Knowledge Treatments and Social Distancing Treatments

VARIABLES	(1) Overall Test Score (TS)	(2) Teaching-eligible TS	(3) Overall TS without SD Index	(4) Teaching-eligible TS without SD Index
Incentive	0.0159	0.00236	0.0205	0.0169
	(0.00862)	(0.00977)	(0.00619)	(0.00694)
Teaching	0.00318	0.0120	0.0199	0.0350
	(0.00882)	(0.0102)	(0.00620)	(0.00727)
Incentive plus Teaching	0.0477	0.0528	0.0581	0.0688
	(0.00842)	(0.00895)	(0.00636)	(0.00704)
Social Norm Correction (SNC)	-0.0101	-0.0151		
	(0.00764)	(0.00833)		
Leader Endorsement (LE)	-0.00797	-0.0169		
	(0.00728)	(0.00790)		
Incentive × SNC	0.00654	0.0159		
	(0.0128)	(0.0143)		
Incentive × LE	0.00677	0.0279		
	(0.0133)	(0.0147)		
Teaching × SNC	0.0181	0.0229		
	(0.0134)	(0.0152)		
Teaching × LE	0.0242	0.0323		
	(0.0136)	(0.0157)		
Incentive plus Teaching × SNC	-0.00304	0.000286		
	(0.0138)	(0.0151)		
Incentive plus Teaching × LE	0.00840	0.0161		
	(0.0130)	(0.0138)		
Observations	2,117	2,117	2,117	2,117
R-squared	0.322	0.336	0.291	0.311
Control Mean DV	0.781	0.784	0.748	0.751
Control SD DV	0.108	0.123	0.121	0.141

Notes: Dependent variable in Columns 1 and 2 defined in Table B.8. Dependent variable in Column 3: Overall TS calculated without the 8 knowledge questions on social distancing actions – that is, the share of correct answers to 32 knowledge questions in Round 3: 12 on general knowledge, 8 on household preventive actions, and 12 on government policy. Dependent variable in Column 4: Teaching-Eligible TS calculated without the 4 Teaching-Eligible knowledge questions on social distancing actions. All regressions also include community fixed effects and controls for pre-treatment (Rounds 1 and 2) Test Scores. Robust standard errors in parentheses.

Fourth, we run a regression with indicators for knowledge treatments, the cross-randomized social distancing treatments and their interaction terms to test for significant interactions between the treatments implemented for two separate experiments in the same population. Results are in Table B.12, Columns 1 and 2. There are six interaction terms in each regression. In Column 1, one coefficient (Teaching x LE) is statistically significant at the 10% level. In Column 2, that same coefficient is statistically significant at the 5% level, and another in that column (Incentive x LE) is significant at the 10% level. Looking at the patterns of coefficients overall, these appear to be chance occurrences. There is no corresponding effect of the LE (leader endorsement) treatment on the "Incentive plus Teaching" (Joint) treatment, which we should expect to also appear if the LE treatment truly interacted with the knowledge treatments. In Columns 3 and 4, we also verify that our primary treatment effect estimates are very similar when the Test Score outcome measure excludes social distancing knowledge questions, which are most susceptible to being affected by the social distancing treatments. Overall, there does not appear to be substantial evidence of interactions between the set of knowledge treatments and the set of social distancing treatments.[6]

B.5.2 Additional Figures

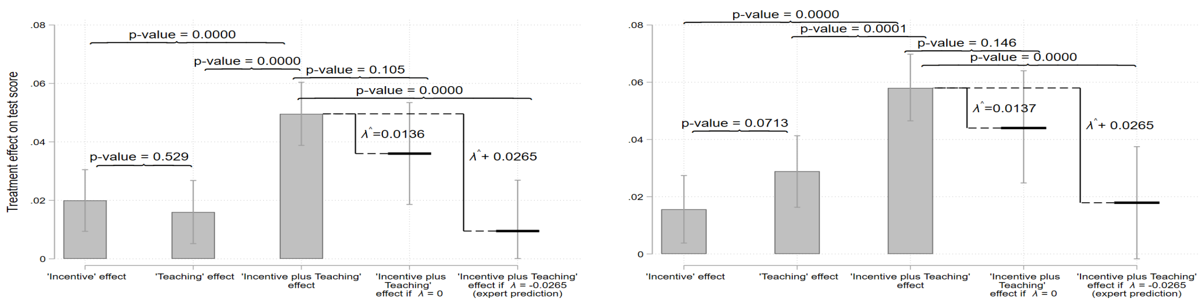
We show here additional figures that correspond to those in the main text, but that relate to the other pre-specified primary outcome (the Overall Test Score based on 40 COVID-19 knowledge questions). We show these to emphasize that key findings and conclusions are robust to examination of either of the two pre-specified primary outcome variables.

In Figure B.2, we display in Panel (a) treatment effects and the complementarity parameter from analyses of the Overall Test Score based on 40 COVID-19 knowledge questions. The corresponding main text Figure 2.3 Panel (a) is replicated here in Panel (b) for comparison. The key conclusion is stable across the two figures: the test that $\lambda = 0$ is rejected at marginal levels of statistical significance (in fact, in Panel (a) the p-value is a bit closer to conventional levels of statistical significance, at 0.105).

[6]Note these are separate experiments with different pre-specified outcomes of interest. As our primary interest was never to examine interactions between these treatments sets, we do not believe it would be accurate to characterize our results as focusing on the "short model" (a weighted average of effects across different cross-randomized treatment groups), along the lines of Muralidharan et al. (2019a)

Figure B.2: Treatment Effects and Test of Complementarity Parameter λ

(a) Overall Test Score (b) Teaching-Eligible Test Score

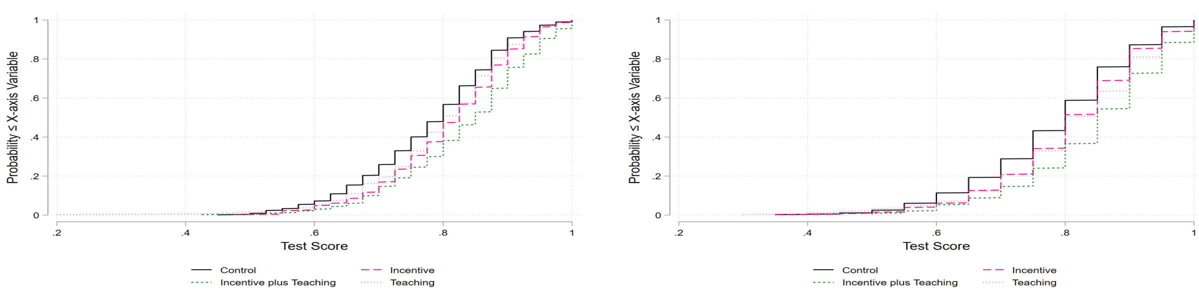


Notes: Overall Test Score is fraction of correct responses on COVID-19 knowledge out of 40 questions. Teaching-Eligible Test Score is a fraction of correct responses on COVID-19 knowledge out of 20 questions previously asked in the Round 2 (baseline) survey. In each panel of figure, bars in first three columns display regression coefficients representing treatment effects (and 95% confidence intervals) for "Incentive", "Teaching", and "Incentive plus Teaching" ("Joint") treatments. Floating solid horizontal lines in fourth and fifth columns display "Incentive plus Teaching" ("Joint") treatment effects that would be implied by different benchmark values of complementarity parameter λ. Difference between values in 3rd and 4th columns is actual estimated complementarity parameter, $\hat{\lambda}$; the test that this difference is equal to zero tests the null that $\lambda = 0$. Difference between values in 3rd and 5th columns is difference between $\hat{\lambda}$ and mean expert prediction, -0.0265; the test that this difference is equal to zero tests the null that $\lambda = -0.0265$.

In Figure B.3, we display in Panel (a) CDFs of the Overall Test Score based on 40 COVID-19 knowledge questions. The corresponding main text Figure 2.4 is replicated in Panel (b) for comparison. Both figures show that the Joint treatment is the most effective, shifting the CDFs of test scores furthest to the right.

Figure B.3: Cumulative Distribution Functions of Test Score by Treatment Group

(a) Overall Test Score (b) Teaching-Eligible Test Score



Notes: Overall Test Score is fraction of correct responses on COVID-19 knowledge out of 40 questions. Teaching-Eligible Test Score is a fraction of correct responses on COVID-19 knowledge out of 20 questions previously asked in the Round 2 (baseline) survey. Figure depicts the cumulative distribution function of this variable for the "Control" group, the "Incentive" treatment arm, the "Teaching" treatment arm, and the "Incentive plus Teaching" ("Joint") treatment arm.

B.6 Cost-Effectiveness

The estimate of the complementarity parameter λ is a key input into policy-making, because it determines the relative cost-effectiveness of the different combinations of treatments (Incentive, Teaching, or Joint). The decision as to which of the three possibilities to implement in practice is highly influenced by their relative cost-effectiveness. The treatment that is the most cost-effective among the three would be a strong candidate to prioritize for implementation from an economic standpoint.

We now illustrate how the relative cost-effectiveness of the treatments we study depends on λ. Cost-effectiveness in our context is the cost of achieving a unit (1-percentage-point, or 0.01) increase in the COVID-19 knowledge test score. The key inputs in the calculation of cost-effectiveness are:

- Treatment effect estimates for the Incentive and Teaching treatments (β_1 and β_2) taken from estimates of Table 2.3 Column 2 in main text. The effect of the joint treatment is then $\beta_1 + \beta_2 + \lambda$.

- Implementation costs of each treatment, per treated beneficiary, derived from actual implementation costs in this study. For the Incentive, Teaching, and Joint treatments we denote the implementation cost per beneficiary as, respectively, c_I, c_T, and c_J. Specifically, we use $c_I = 5.80$, $c_T = 2.83$, and $c_J = 7.21$ (c_J is less than the sum of c_I and c_T because there are some economies of scale from providing both treatments together.)[7]

For each treatment i, cost-effectiveness e_i (cost per 0.01 increase in test scores) is:

- Incentive treatment: $e_I = 100 * c_I/\beta_1$

- Teaching treatment: $e_T = 100 * c_T/\beta_2$

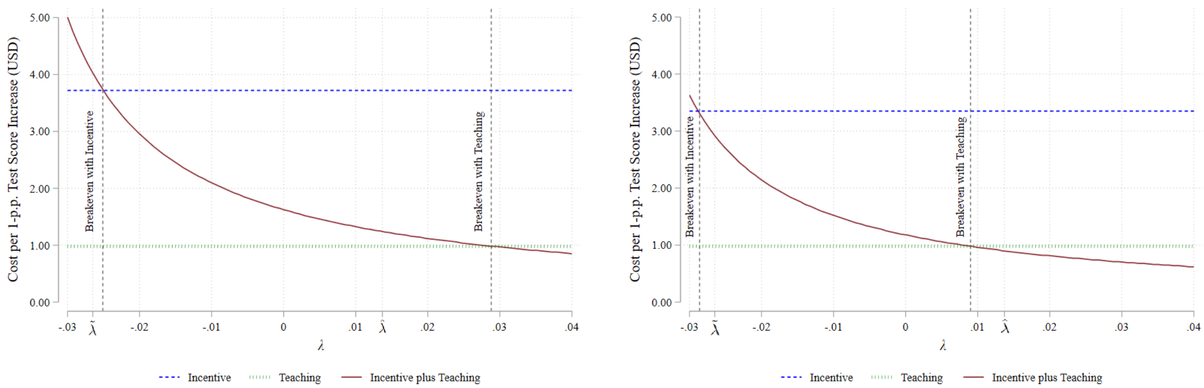
- Joint treatment: $e_J = 100 * c_J/(\beta_1 + \beta_2 + \lambda)$

In Figure B.4 panel (a), we display the cost-effectiveness of each treatment, using actual treatment effects for the Incentive and Teaching treatments (β_1 and β_2) and Joint treatment effects implied by a range of values of λ. The cost-effectiveness of the Incentive and Teaching treatments are horizontal, because they do not depend on λ. The cost-effectiveness of the Joint treatment is a decreasing function of λ: the greater the complementarity of the two treatments, the more cost-effective is the Joint treatment.

[7]These are marginal costs (project staff wages and study participant incentives) of adding one additional treatment beneficiary, estimated based on our own study cost data. We use marginal costs, presuming that fixed costs per beneficiary will be negligible in a sufficiently scaled-up program. Costs expressed in USD using the nominal exchange rate of 70.74 Mozambican meticais per USD as of August 26, 2020.

Figure B.4: Cost-Effectiveness of Treatments as Functions of λ

(a) Estimated Treatment Costs per Beneficiary (b) Alternative Treatment Costs per Beneficiary



Notes: Cost per unit (0.01, or 1-percentage-point) increase in COVID-19 Knowledge Test Score as a function of complementarity parameter λ, for Incentive treatment (horizontal dashed blue line), Teaching treatment (horizontal dotted green line), and Incentive plus Teaching (Joint) treatment (downward-sloping solid red line). In the left panel, implementation cost per beneficiary for Incentive, Teaching, and Joint treatments are, respectively, $c_I = 5.80$, $c_T = 2.83$, and $c_J = 7.21$. In the right panel, alternative implementation costs per beneficiary for Incentive, Teaching, and Joint treatments are, respectively, $c_I = 5.23$, $c_T = 2.83$, and $c_J = 5.23$. Impact of Incentive and Teaching treatments on test scores (β_1 and β_2) taken from estimates of Table 2.3 Column 2 in main text. Impact of Joint treatment is $\beta_1 + \beta_2 + \lambda$. Vertical lines indicate "breakeven" values of λ, at which Joint treatment is as cost-effective as the respective individual treatment: leftmost vertical line is breakeven with Incentive treatment, and rightmost vertical line is breakeven with Teaching treatment. Expert-predicted $\tilde{\lambda}$ (-0.0265) and actual estimated $\hat{\lambda}$ (0.0137) are also indicated on horizontal axis.

The intersection of the Joint treatment line with the horizontal lines indicates the "breakeven" λs, above which the Joint treatment is more cost effective than the respective single treatment. Break-even λ is -0.0250 for the Incentive treatment, and 0.0290 for Teaching. The latter number is the more relevant for policy decision-making, since the Teaching treatment is the more cost-effective of the two individual treatments. For the Joint treatment to be the most cost-effective of the three treatment combinations, λ must be above 0.0290.

For reference, we also show the mean expert prediction, $\tilde{\lambda}$, -0.0265, and our estimated $\hat{\lambda}$. At $\hat{\lambda} = 0.0137$, the Joint treatment is more cost-effective ($e_J = 1.24$) than the Incentive treatment ($e_I = 3.72$), but not as cost-effective as Teaching ($e_T = 0.98$). Actual costs in a scaled-up program may be different from those of our study, and could yield different cost-effectiveness rankings across treatments.

Governments or NGOs implementing our treatments in different contexts may come to different cost-effectiveness rankings given their specific implementation costs. We provide an example of alternative relative implementation costs that would lead the Joint treatment to be the most cost-effective at $\hat{\lambda} = 0.0137$. We use the same implementation cost per beneficiary for the Teaching treatment ($c_T = 2.83$), but assume that the implementation cost of the Incentive treatment can be somewhat lower ($c_I = 5.23$). We also assume substantial economies of scale in implementing both treatments together, so that the cost per beneficiary of the Joint treatment is not the sum but just the maximum of the individual treatments: $c_J = 5.23$ (equal to the cost of the Incentive treatment).

Panel (b) of Figure B.4 displays the cost-effectiveness of each treatment in this alternative case. It is identical to panel (a) except we have changed the assumptions regarding the cost per beneficiary of the Incentive and Joint treatments. In this case, breakeven levels of λ are lower: -0.0288 for the Incentive treatment, and 0.0088 for Teaching. At $\hat{\lambda} = 0.0137$, the Joint treatment is the most cost-effective of the three treatments, with $e_J = 0.90$, compared with $e_I = 0.98$ and $e_I = 3.35$.

B.7 Long-Run Analysis

We collected a fourth round of survey data over the phone between June 30 and August 30, 2021. We refer to this as the post-endline or Round 4 survey. For any given respondent, the Round 4 survey came at least 41 weeks and average of 45.8 weeks after treatment implementation and at least 36 weeks and an average of 39.5 weeks after Round 3 (endline). Reported COVID-19 cases during the Round 4 survey were significantly higher than previous survey rounds, with Mozambique's 7-day average jumping from 78 and 144 at the start of Rounds 2 and 3, respectively, to 456 at the start of Round 4, a trend we confirmed with district-level data available in 3 of our 7 districts. In total, Round 4 surveyed 1,886 of the 2,117 respondents surveyed in Round 3, achieving a retention rate of 89.1% overall that is balanced across treatment conditions.

In Round 4, we measured COVID-19-related knowledge in two main categories: 1) general knowledge and 2) preventive action, drawing from the same question pool used at baseline and endline. We did not survey questions on government policy, as many policies had changed since Round 3 making many questions irrelevant. Specifically, we asked respondents 20 knowledge questions from the pre-specified question pool detailed in Appendix B.2: 12 on general knowledge (6 of which were asked in Rounds 2 and 3, and 6 of which were only asked in Round 3 but not Round 2), and 8 on preventive action (all of which were asked in Rounds 2 and 3).

Using these data, we calculated two modified Test Scores that resemble our pre-specified primary outcomes less the inclusion of questions on government policy:

1. Test Score of all general knowledge and preventive action questions asked of respondents in each round:
   - In Round 4 (post-endline), this includes 12 general knowledge and 8 preventive action questions;
   - In Round 3 (endline), this includes 12 general knowledge and 16 preventive action questions.
2. Test Score of general knowledge and preventive action questions that were eligible for the Teaching intervention (i.e., randomly selected to be asked of the respondent at baseline in Round 2). For a given respondent, this includes the same set of 6 general knowledge and 8 preventive action questions asked in Rounds 2, 3, and 4.

As this analysis was not pre-specifed, we evaluate long-term impacts by regressing on both Round 4 (post-endline) Test Scores outcomes above, running regressions on the equivalent Round 3 (endline) modified Test Scores for comparison, and only draw conclusions supported by both outcomes. Specifically, we estimate regression Equation 3.4.2 in four specifications where:

- Outcomes are the Test Scores (described above) in Round 4 and, for direct comparison, Round 3.
- $\mathbf{B}_{ijt}$ is modified to be a vector representing the share of correct answers to general knowledge and preventive action questions in Rounds 1 and 2, respectively (i.e., excluding government policy questions).

We present results in Table B.13 and discuss their relevance to verifying the robustness of the Joint intervention's positive effect and complementarity over time in Section 2.5.4.

Table B.13: Treatment Effects on Long-Run COVID-19 Knowledge Test Scores

VARIABLES	(1) Post-endline Overall TS	(2) Endline equivalent Overall TS	(3) Post-endline Teaching-Eligible TS	(4) Endline equivalent Teaching-Eligible TS
Incentive	-0.0104	0.0068	-0.0155	-0.0001
	(0.0065)	(0.0058)	(0.0071)	(0.0073)
Teaching	0.0124	0.0113	0.0149	0.0236
	(0.0067)	(0.0063)	(0.0073)	(0.0080)
Incentive plus Teaching	0.0342	0.0407	0.0368	0.0462
	(0.0066)	(0.0062)	(0.0071)	(0.0073)
$\hat{\lambda}$	0.0321	0.0226	0.0374	0.0227
	(0.0101)	(0.0094)	(0.0110)	(0.0116)
Observations	1,886	1,886	1,886	1,886
R-squared	0.203	0.275	0.195	0.282
Control Mean DV	0.797	0.783	0.794	0.819
Control SD DV	0.116	0.108	0.123	0.137
p-value: $\lambda = 0$	0.0014	0.0162	0.0007	0.0505
p-value: $\lambda =$ -0.0265	0.0000	0.0000	0.0000	0.0000
p-value: Incentive = Teaching	0.0026	0.5275	0.0002	0.0089
p-value: Incentive = Joint	0.0000	0.0000	0.0000	0.0000
p-value: Teaching = Joint	0.0043	0.0001	0.0086	0.0119

Notes: Column 1-2: fraction of general knowledge and preventive action questions answered correctly in Rounds 4 and 3, respectively. Columns 3-4: fraction of general knowledge and preventive action questions answered correctly in Rounds 4 and 3, respectively, that were eligible for the Teaching intervention (i.e., asked in Round 2). λ is the complementarity parameter (see Section 3.2 of main text). "$\hat{\lambda}$" is coefficient on "Incentive plus Teaching" ("Joint") minus sum of coefficients on "Incentive" and "Teaching". All regressions include community fixed effects and controls for corresponding pre-treatment (pre-baseline and baseline) Test Scores. Robust standard errors in parentheses.

APPENDIX C

Appendix to Chapter 3

C.1 Proofs

C.1.1 Proof of Theorem 2

The agent will adjust her effort level in response to the treatment to $\sqrt{\hat{e}} = \frac{\hat{\alpha}C}{2}H(\hat{x})$ where $H(x) = 1 - x(1 - \frac{1}{\sqrt{2}})$. Hence, the prior and posterior effort levels satisfy:

$$\frac{\sqrt{\tilde{e}}}{\sqrt{e}} = \frac{\tilde{\alpha}}{\alpha}\frac{H(\tilde{x})}{H(x)} \tag{C.1.1}$$

We take the ratios of Equations 3.2.6 and 3.2.7:

$$\frac{\hat{\alpha}}{\alpha} = \frac{1 - \sqrt{e}G(x)}{1 - \sqrt{e}G(\hat{x})} \tag{C.1.2}$$

We therefore obtain:

$$\frac{\sqrt{\tilde{e}}}{\sqrt{e}} = \frac{H(\hat{x})(1 - \sqrt{e}G(x))}{H(x)(1 - \sqrt{e}G(\hat{x}))} \tag{C.1.3}$$

Effort increases iff $\frac{\sqrt{\tilde{e}}}{\sqrt{e}} > 1$:

$$\begin{aligned} \frac{H(\hat{x})(1 - \sqrt{e}G(x))}{H(x)(1 - \sqrt{e}G(\hat{x}))} &> 1 \\ \sqrt{e}\left[H(x)G(\hat{x}) - H(\hat{x})G(x)\right] &> H(x) - H(\hat{x}) \end{aligned}$$

Now note that $G(x) = 2xH(x)$ such that:

$$\begin{aligned} \sqrt{e}2H(x)H(\hat{x})(\hat{x}-x) &> (1-\frac{1}{\sqrt{2}})(\hat{x}-x) \\ \sqrt{e} &> \frac{1-\frac{1}{\sqrt{2}}}{H(x)H(\hat{x})} \end{aligned} \tag{C.1.4}$$

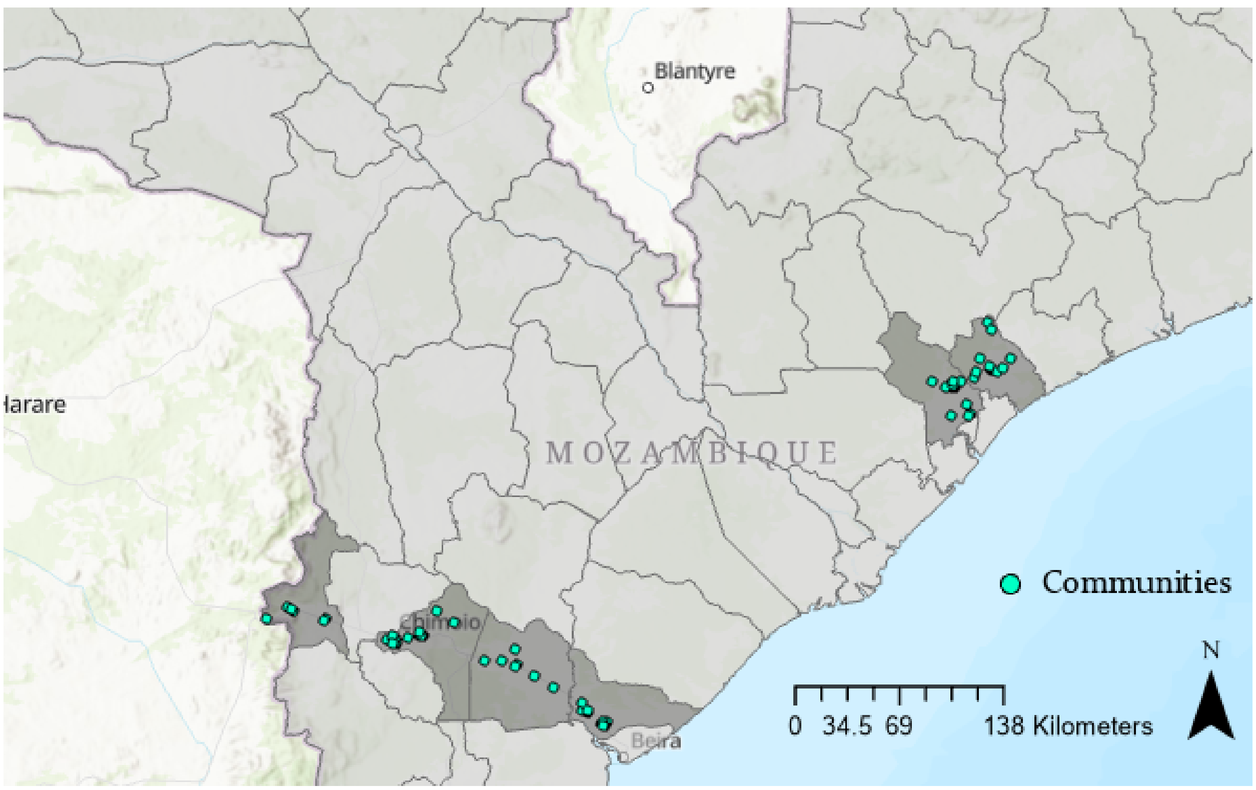
This shows that the perceived-infectiousness effect dominates if the initial effort level e is high enough. Effort is determined by Equation 3.2.5 and increases with α (which increases with $\hat{\alpha}$). Therefore, for sufficiently large $\hat{\alpha}$ the perceived-infectiousness effect dominates.

C.2 Study Context

C.2.1 Study Area

Study participants come from 76 communities in central Mozambique. The study communities are in seven districts of three provinces: Dondo and Nhamatanda in Sofala province; Gondola, Chimoio and Manica in Manica province; and Namacurra and Nicoadala in Zambezia province. These 76 communities are mapped in Figure C.1. Compared to other communities in Mozambique, the study areas are relatively accessible to main transport corridors (highways and ports), and are thus important geographic conduits for infectious disease.

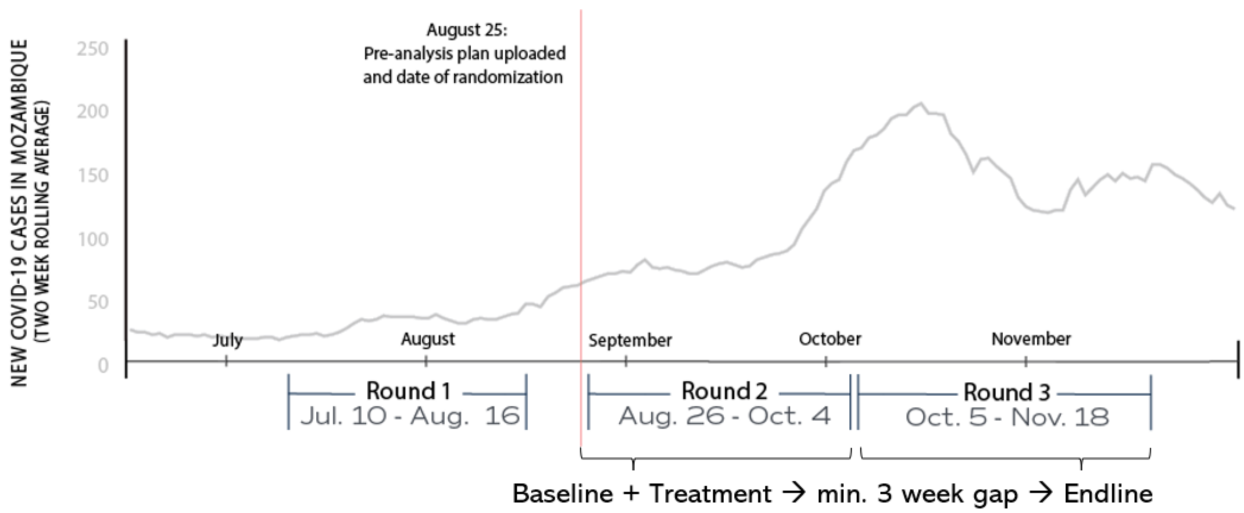
Figure C.1: Study Area



C.2.2 Study Timeline

We collected survey data in three rounds between July 10 and November 18, 2020. Figure C.2 depicts the study timeline below a rolling average of new Mozambican COVID-19 cases. We piloted surveys in Round 1 (pre-baseline). Immediately before the Round 2 survey, we randomly assigned households to treatments and submitted our pre-analysis plan to the AEA RCT Registry. The Round 2 survey served as a baseline and was immediately followed (on the same phone call) by our treatment interventions. Round 3 was our endline survey. Surveys collected data on COVID-19 knowledge, beliefs, and behaviors. While data collection for Round 3 began only one day after completion of Round 2, there was a minimum of 3.0 weeks and average of 6.3 weeks between Rounds 2 and 3 surveys for any given respondent. While the Round 1 survey occurred when new COVID-19 cases remained relatively steady, both the Round 2 and Round 3 surveys occurred during a period of substantial growth in new COVID-19 cases.

Figure C.2: Study Timeline



Notes: Round 1 is pre-baseline survey to collect social distancing support data, Round 2 is baseline survey, and Round 3 is endline survey. There is at least a three week gap between baseline and endline survey for any given study participant. Pre-analysis plan uploaded and treatments randomly assigned immediately prior to start of Round 2 baseline survey, on Aug. 25, 2021. Treatments implemented immediately following baseline survey on same phone call. Baseline measures reported in Table B.4 come from Round 2 surveys and endline measures come from Round 3 surveys.

C.2.3 COVID-19 Context

The Mozambican government declared a State of Emergency due to the COVID-19 pandemic on March 31, 2020, recommending social distancing (at least 1.5 meters) and requiring it at public and private institutions and gatherings. The government also suspended schools, required masks at funerals and markets, banned gatherings of 20 or more, and closed bars, cinemas and gymnasiums (Republic of Mozambique, 4/1/2020). The government stopped short of implementing a full economic "lockdown" due to its economic costs (Jones et al., 2020). On August 5, 2020, the government renewed the State of Emergency, called for improved mask-wearing, and announced a schedule for loosening restrictions (Nyusi, 8/5/2020). In September, the government loosened some restrictions including resuming religious gatherings at 50% capacity (U.S Embassy in Mozambique). Throughout this period, the government's social distancing recommendation remained constant.

COVID-19 cases by district at the start of the Round 3 (endline) survey are estimated as follows. Data on district-level population come from Mozambique's 2017 Census (National Institute of Statistics (INE), 2017). District COVID-19 case counts come from the government's COVID-19 Mozambique dashboard (Ministry of Health, 2020) and correspondence with provincial health offices. Each district's case count is from the start date of the endline survey in the district (ranging from October 5 to November 1, 2020). We also show the number of respondents in our study sample in each district.

Table C.1: COVID-19 Cases by District

DISTRICT	(1) Cumulative COVID-19 Cases	(2) Cases per 100,000 people	(3) Population	(4) Number of Study Respondents
Sofala Province				
Dondo	8	4.137	193,382	323
Nhamatanda	12	4.300	279,081	214
Manica Province				
Gondola	3	3.553	84,429	224
Chimoio	142	39.082	363,336	524
Manica	20	9.292	215,239	290
Zambezia Province				
Namacurra	4	1.652	242,126	244
Nicoadala	52	28.779	180,686	298

Notes: COVID-19 cases by district at the start of the Round 3 (endline) survey. Column 1: District COVID-19 case counts come from the government's COVID-19 Mozambique dashboard (Ministry of Health, 2020) and correspondence with provincial health offices, measured at the start date of the endline survey in the district (ranging from October 5 to November 1, 2020). Column 2: Calculated from Columns 1 and 3. Column 3: District-level population come from Mozambique's 2017 Census (National Institute of Statistics (INE), 2017). Column 4: Number of respondents in our study sample in each district.

C.3 Effect on Perceived Community Support

Table C.2 presents the cumulative distribution of this perceived community support measure in the final samples at pre-baseline, baseline and endline, and subdivided by treatment arm at endline. Even at pre-baseline, the distribution is skewed upwards with over 80% of the sample reporting that the majority (50% or greater) of households in their community support social distancing, though a sizable 8% use the extreme lower end of the scale to report that none (0%) of the households in their community support social distancing. At baseline, the distribution is further skewed upwards with over 90% of the sample reporting that the majority (50% or greater) of households in their community support social distancing and over half of the sample reporting that 100% of households do the same.

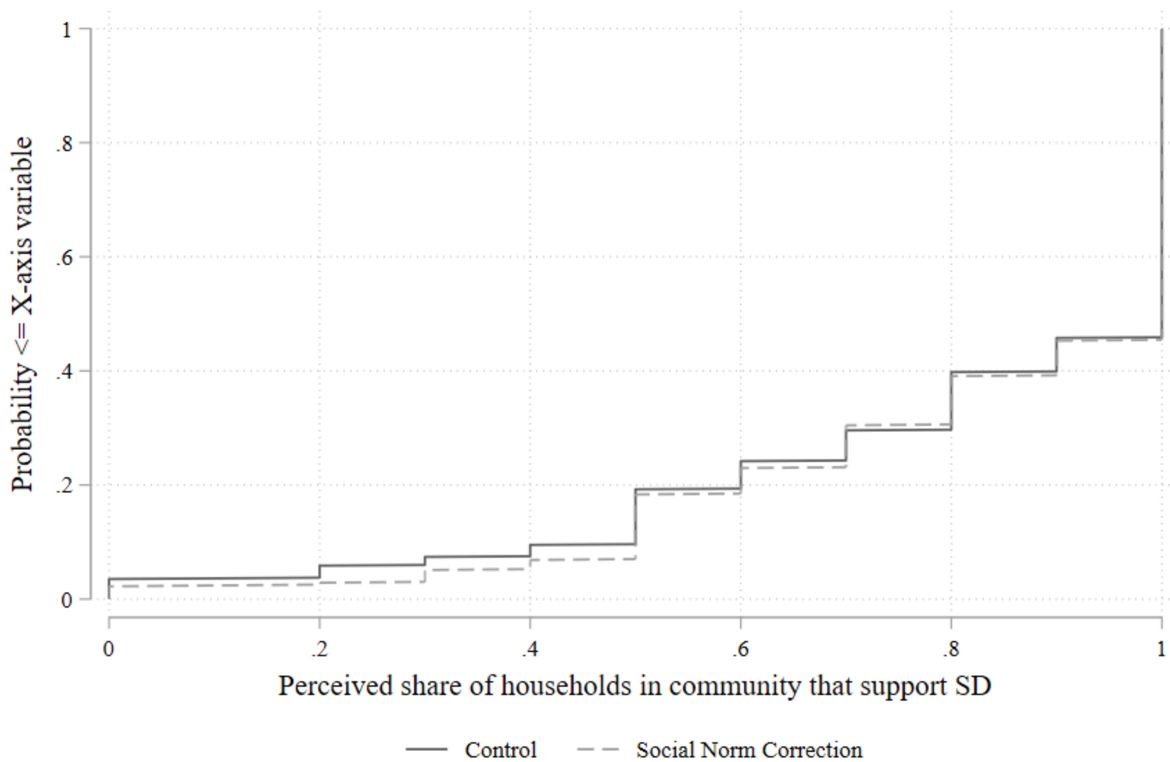
Table C.2: Sample Distribution (Cumulative %) by Perceived Community Support

	Pre-Baseline	Baseline	Endline			
Perceived Share	Total	Total	Total	Control	T1	T2
0%	8.0	2.7	2.8	3.5	2.2	2.5
10%	9.0	3.1	3.1	3.6	2.4	3.0
20%	10.7	4.4	4.4	5.9	2.9	3.9
30%	14.0	6.5	6.5	7.4	5.1	6.6
40%	19.1	9.6	8.8	9.5	6.9	9.9
50%	34.3	21.1	19.0	19.3	18.3	19.2
60%	41.8	27.1	23.9	24.2	23.0	24.5
70%	49.7	33.4	30.3	29.6	30.5	31.1
80%	59.7	43.4	40.8	39.8	39.1	44.0
90%	65.2	48.9	46.8	45.8	45.3	49.6
100%	100.0	100.0	100.0	100.0	100.0	100.0

Notes: Perceived share of households in the community who support social distancing is estimated by dividing responses to the question "For every 10 households in your community, how many support social distancing?" by 10, and hence has 11 categories from 0%, 10%... 90%, 100%. Cells report cumulative percentages from 0% up to the row in question. Pre-baseline "Total" refers to all responses from the final sample in Round 1 (N=2,109), and Baseline "Total" refers to all responses from the final sample in Round 2 (N=2,114). At endline, "Total" refers to Round 3 responses from the final sample (N=2,116), "Control" from the control group, "T1" from the misperceptions correction treatment group, and "T2" from the leader endorsement treatment group.

We find that the misperceptions correction treatment did increase respondents' perceived community support, particularly for those at the lower end of the distribution. Figure C.3 shows the cumulative distribution function for the perceived community support measure at endline. Relative to the control group, those receiving the misperceptions correction treatment were less likely to report that fewer than 50% of households in their community supported social distancing, instead reporting higher perceptions of community support. Further, Table C.3 presents three regressions estimating the treatment effects on the perceived community support. In Column (1), the dependent variable is the perceived share of households in the community who support social distancing. The coefficient is positive and marginally statistically significant (p-value=0.12). Regressions in Columns (2) and (3) find that the misperceptions correction treatment has a positive effect on an indicator for the respondent believing the majority (50% or more) of households in their community support social distancing, and an indicator that the respondent's perceived community support increased between baseline and endline (both coefficients are statistically significantly different from zero at the 5% level).

Figure C.3: Cumulative Distribution of Perceived Community Support by Treatment



Notes: Perceived share of households in the community who support social distancing is estimated by dividing responses to the question "For every 10 households in your community, how many support social distancing?" by 10, and hence has 11 categories from 0%, 10%... 90%, 100%. Figure depicts the cumulative distribution function of this variable for the "Control" group and "Misperceptions Correction" treatment arm. The leader endorsement treatment is excluded for clarity.

Table C.3: Treatment Effects on Perceived Community Support (PCS)

VARIABLES	(1) Continuous PCS	(2) Indicator if PCS $\geq$ 50%	(3) Indicator if PCS increased
T1: Misperceptions Correction	0.0196	0.0291**	0.0507**
	(0.0128)	(0.0138)	(0.0241)
T2: Leader Endorsement	0.0041	0.0036	0.0358
	(0.0128)	(0.0149)	(0.0236)
Observations	2,116	2,116	2,113
R-squared	0.164	0.118	0.043
Control Mean DV	0.8115	0.9049	0.2550
Control SD DV	0.2681	0.2935	0.4361

Notes: Dependent variables are defined as follows. Column 1 is the perceived share of households in community that support social distancing, which takes on the values shown in Table C.2. Column 2 is an indicator equal to one if respondent reports that majority (50% or more) of households in community support social distancing, and zero otherwise. Column 3 is an indicator equal to one if the respondent's perceived community support increased between the baseline (pre-treatment) and endline (post-treatment) surveys. "T1: Misperceptions Correction" "T2: Leader Endorsement" and controls are as defined in Table 3.2, except column 3 does not include a baseline value of the outcome as a control as it was used to calculate the outcome. All regressions also include community fixed effects. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

C.4 Social Distancing Index

The list of actions included in the Social Distancing Index and their corresponding summary statistics are presented below.

Social Distancing Actions: Is this something your household has been doing for the last seven days? (Answers indicating social distancing in parentheses.[1])

1. Shop in crowded areas like informal markets (No)
2. Gather with several friends (No)
3. Help the elderly avoid close contact with other people, including children (Yes)
4. If show symptoms of coronavirus, immediately inform my household and avoid people (Yes)
5. Drink alcohol in bars (No)
6. Wear a face mask if showing symptoms of coronavirus (Yes)
7. Instead of meeting in person, call on the phone or send text message (Yes)
8. Allow children to build immunity by playing with children from other households (No)

Below are the summary statistics for the questions that comprise the self-reported social distancing index at baseline and endline. Respondents were asked "Is this something your household has been doing for the last seven days?" about a randomly determined four social distancing actions at baseline and all eight social distancing actions at endline. Responses were coded as indicators equal to one if indicative of social distancing (answers that indicate social distancing shown in parentheses), and zero otherwise.

[1]For items 4 and 6 that are conditional on showing symptoms, survey staff instructed respondents to answer "Yes" (doing social distancing) if not showing symptoms.

Table C.4: Summary Statistics for Components of Social Distancing Index

	Baseline			Endline			T-test
VARIABLES	N	Mean	SD	N	Mean	SD	p-value
Shop in crowded areas like informal markets (No)	1,032	0.642	0.480	2,115	0.678	0.467	0.5011
Gather with several friends (No)	1,047	0.349	0.477	2,113	0.414	0.493	0.0357
Help the elderly avoid close contact with other people, including children (Yes)	1,094	0.877	0.329	2,114	0.923	0.266	0.0000
If show symptoms of coronavirus, immediately inform my household and avoid people (Yes)	1,050	0.836	0.370	2,113	0.860	0.347	0.0314
Drink alcohol in bars (No)	1,082	0.226	0.419	2,113	0.272	0.445	0.0152
Wear a face mask if showing symptoms of coronavirus (Yes)	1,034	0.902	0.297	2,114	0.885	0.319	0.3993
Instead of meeting in person, call on the phone or send text message (Yes)	1,039	0.935	0.247	2,112	0.930	0.255	0.5922
Allow children to build immunity by playing with children from other households (No)	1,070	0.439	0.497	2,113	0.456	0.498	0.0814

Notes: Variables are coded as indicators equal to one if indicative of social distancing (answers that indicate social distancing shown in parentheses), and zero otherwise. Respondents were asked "Is this something your household has been doing for the last seven days?" about a randomly determined four social distancing actions at baseline and all eight social distancing actions at endline. The baseline sample was asked a subset of these questions which explains the smaller number of observations at baseline. Last column displays the p-value of a paired t-test on the difference between baseline and endline measure (where baseline data are available).

C.5 Treatment Details and Scripts

Both the misperceptions correction and leader endorsement treatments were implemented directly following the baseline survey, on the same phone call. If a respondent was randomly assigned to a treatment, the corresponding intervention text would appear on the enumerator's tablet. Enumerators read a script aloud exactly as shown below. Following the treatment, respondents were asked if they would like the information repeated. Of the N=628 receiving the misperceptions correction and N=637 receiving the leader endorsement, only 8.6% and 9.4% asked for the script to be repeated, respectively.

Script for T1: Misperceptions Correction – "Now I want to give you some information about social distancing. In this survey, you indicated that you think *[insert respondent's answer here]* of every 10 households in your community support the practice of social distancing."

- *If response UNDERESTIMATES community support for social distancing:* "However, more households support social distancing than you think! Based on the results of our first COVID-19 survey, approximately *[insert actual community support for social distancing here]* of every 10 households in your community support social distancing to prevent the spread of the coronavirus."

- *If response CORRECTLY ESTIMATES community support for social distancing:* "You are correct! Based on the results of our first COVID-19 survey, approximately *[insert actual community support for social distancing here]* of every 10 household in your community support social distancing to prevent the spread of the coronavirus."

- *If response OVERESTIMATES community support for social distancing: (no information given)*

Script for T2: Leader Endorsement – "Our research team recently called and talked to your *[list leaders' titles and names here]*. They said that they support social distancing, are practicing social distancing themselves, and encourage others to do the same."

C.6 Attrition and Balance

Appendix Table C.5 presents regressions examining whether attrition and baseline variables are balanced with respect to treatment assignment.[2] Attrition between Round 2 (baseline) and Round 3 (endline) is only 4.9% and is less than 5.6% in each of the seven districts surveyed. Balance in attrition is confirmed in Column (1), which starts with the Round 2 (baseline) sample and regresses treatments on an indicator equal to one if the respondent was not reached for the Round 3 (endline) survey. Balance in baseline social distancing outcomes is confirmed in Columns (2)-(4), which examines the Round 2 social distancing outcomes. Balance in baseline household characteristics is confirmed in Columns (6)-(8), which examines the final Round 3 sample and regresses treatments on Round 1 measures of household income, an index of food insecurity, and an indicator for presence of an older adult over 60 years. In not a single regression in the table is a coefficient on a treatment indicator statistically significant at conventional levels.

[2]Figure C.2 shows the study timeline for the three survey rounds collected. Round 1 is a pre-baseline measure, Round 2 measures baseline values and Round 3 measures endline outcomes.

Table C.5: Treatment Effect on Attrition and Balance

VARIABLES	(1) Attrition	(2) Primary SD Indicator	(3) Others' Report of SD	(4) Self-Report of SD	(5) Perceived Social Norm	(6) Hh Income	(7) Food Insecurity	(8) Older Adult in Hh
T1: Misperceptions Correction	-0.0127	-0.0176	-0.0005	-0.0096	-0.0101	-159.46	0.0011	-0.0029
	(0.0111)	(0.0134)	(0.0203)	(0.0247)	(0.0138)	(181.66)	(0.0191)	(0.0250)
T2: Leader Endorsement	-0.0015	-0.0032	0.0090	0.0042	-0.0201	-39.95	-0.0240	0.0240
	(0.0113)	(0.0143)	(0.0206)	(0.0249)	(0.0137)	(181.76)	(0.0193)	(0.0252)
Observations	2,226	2,117	2,117	2,117	2,114	1,873	2,117	2,096
R-squared	0.030	0.096	0.199	0.076	0.047	0.043	0.090	0.058
Control Mean DV	0.0533	0.0833	0.2289	0.3556	0.8095	1176	0.8415	0.3424
Control SD DV	0.2248	0.2765	0.4204	0.4790	0.2618	4029	0.3654	0.4748

Notes: Dependent variables are as follows. Column 1: indicator if respondent attrited from the sample between baseline and endline. Columns 2-4: baseline SD outcomes defined in Table B.4. Column 5: baseline perceived share of community supporting SD, defined further in Table B.4. Column 6: at pre-baseline, self-reported total income for the previous week (in Mozambican meticais). Column 7: at pre-baseline, indicator if, in the last 7 days, household has 1) lacked food; 2) reduced number of meals/portions; or was unable to buy their usual amount of food due to 3) market shortages, 4) high prices, 5) reduced income. Column 8: at pre-baseline, indicator if adult age 60 or older is present in the household. Controls are as defined in Table 3.2. All regressions also include community fixed effects. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

C.7 Robustness of Treatment Effects Estimates

In this appendix, we show that our primary results are robust to 1) using a logit or probit specification, 2) clustering standard errors by community or district, 3) two alternative measures of the social distancing indicator as the primary outcome, 4) four alternative measures of COVID-19 case intensity used to test the interaction, and 5) excluding the district with the highest number of COVID-19 cases.

C.7.1 Logit and Probit Specifications

The primary social distancing indicator is a binary variable that is analyzed using an ordinary least-squares (OLS) regression, as pre-specified. As a robustness check, we adapt Equation 3.4.1 to be run using logit and probit regressions.

Table C.6 presents results from the logistic regression on the primary outcomes, while Table C.7 presents corresponding probit regression results. Regression coefficients are presented as marginal effects. Results in both tables are consistent with the results from OLS linear probability models presented in Table 3.2.

Table C.6: Treatment Effects Estimated Using Logistic Regression

VARIABLES	(1) Primary SD Indicator	(2) Primary SD Indicator	(3) Perceived share of households in community that will get sick from COVID-19	(4) Perceived share of households in community that will get sick from COVID-19
T1: Misperceptions Correction	0.0100	-0.0756**	0.0270	-0.4034***
	(0.0221)	(0.0376)	(0.0395)	(0.1376)
T2: Leader Endorsement	-0.0069	-0.0398	-0.0274	-0.3005**
	(0.0222)	(0.0349)	(0.0394)	(0.1354)
T1 × District COVID-19 Cases		0.0038***		0.0132***
		(0.0013)		(0.0040)
T2 × District COVID-19 Cases		0.0016		0.0084**
		(0.0013)		(0.0040)
Observations	1,285	1,285	806	806
Control Mean DV	0.1415	0.1415	0.3563	0.3563
Control SD DV	0.3488	0.3488	0.3680	0.3680

Notes: Dependent variables are defined in Tables B.4 and 3.2. Coefficients presented are marginal effects from logit regression. Social distancing treatments and controls are as defined in Table 3.2. All regressions also include community fixed effects. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table C.7: Treatment Effects Estimated Using Probit Regression

VARIABLES	(1) Primary SD Indicator	(2) Primary SD Indicator	(3) Perceived share of households in community that will get sick from COVID-19	(4) Perceived share of households in community that will get sick from COVID-19
T1: Misperceptions Correction	0.0089	-0.0709**	0.0288	-0.4015***
	(0.0212)	(0.0347)	(0.0390)	(0.1316)
T2: Leader Endorsement	-0.0084	-0.0356	-0.0298	-0.3057**
	(0.0214)	(0.0330)	(0.0392)	(0.1346)
T1 × District COVID-19 Cases		0.0037***		0.0132***
		(0.0013)		(0.0038)
T2 × District COVID-19 Cases		0.0014		0.0085**
		(0.0012)		(0.0040)
Observations	1,285	1,285	806	806
Control Mean DV	0.1415	0.1415	0.3563	0.3563
Control SD DV	0.3488	0.3488	0.3680	0.3680

Notes: Dependent variables are defined in Tables B.4 and 3.2. Coefficients presented are marginal effects from probit regression. Social distancing treatments and controls are as defined in Table 3.2. All regressions also include community fixed effects. Robust standard errors in parentheses. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

C.7.2 Clustering Standard Errors

In our primary analysis, we report robust standard errors, as pre-specified. As a robustness check, Table C.8 shows our main regressions clustering standard errors by the study's 76 communities or 7 districts. The results show that clustering has minimal impact on standard errors and does not affect whether any coefficients are statistically significant at conventional levels.

Table C.8: Treatment Effects Estimated with Clustered Standard Errors

	Primary SD Indicator			
	(1)	(2)	(3)	(4)
VARIABLES	Clustered at Community	Clustered at District	Clustered at Community	Clustered at District
T1: Misperceptions Correction	0.0042	0.0042	-0.0466***	-0.0466***
	(0.0133)	(0.0269)	(0.0150)	(0.0124)
T2: Leader Endorsement	-0.0054	-0.0054	-0.0258	-0.0258
	(0.0132)	(0.0133)	(0.0169)	(0.0163)
T1 × District COVID-19 Cases			0.0030***	0.0030***
			(0.0009)	(0.0005)
T2 × District COVID-19 Cases			0.0012	0.0012**
			(0.0010)	(0.0005)
Observations	2,117	2,117	2,117	2,117
R-squared	0.158	0.158	0.163	0.163
Control Mean DV	0.0857	0.0857	0.0857	0.0857
Control SD DV	0.2801	0.2801	0.2801	0.2801

Notes: Standard errors (in parentheses) are clustered at the level of 76 communities (Columns 1 and 3) or 7 districts (Columns 2 and 4). Dependent variable defined in Table B.4, and variables and suppressed controls (including community fixed effects) are defined in Table 3.2. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

C.7.3 Social Distancing Indicator Alternatives

For the social distancing indicator in the primary outcome, one condition is that respondents must report doing more than the sample median number of "social distancing actions" in the past seven days, as pre-specified, which worked out to being at least seven out of eight actions (as the sample median number was six). One concern might be that this relatively arbitrary threshold of the sample median may be driving the primary results.

Table C.9 shows that our primary results are robust to alternative definitions of the social distancing indicator based on the a respondent's self-reported number of social distancing actions. First, the dependent variable in Columns (1)-(2) is social distancing measure in which threshold number of self-reported actions to be considered social distancing is 6 out of 8 (i.e., at or above the sample median). Second, the dependent variable in Columns (3)-(4) is social distancing measure which excludes social distancing actions #4 and #6 from Section C.4 as these are conditional on experiencing symptoms and thus might be inadvertently misreported, thereby the threshold number of self-reported actions is changed to 5 out of 6 actions.

Under both alternative social distancing indicators, the main treatment effects in Columns (1) and (3) remain statistically insignificant while the coefficients relating to the misperceptions correction treatment in Columns (2) and (4) demonstrate similar treatment effect heterogeneity with respect to the local infection rate. Interestingly, Column (4) also shows treatment effect heterogeneity for the leader endorsement.

Table C.9: Treatment Effects with Alternative Social Distancing Measures

VARIABLES	(1) Alternative SD Indicator 1	(2) Alternative SD Indicator 1	(3) Alternative SD Indicator 2	(4) Alternative SD Indicator 2
T1: Misperceptions Correction	-0.0060	-0.0459**	-0.0013	-0.0452**
	(0.0159)	(0.0215)	(0.0144)	(0.0198)
T2: Leader Endorsement	-0.0073	-0.0313	-0.0018	-0.0372*
	(0.0159)	(0.0226)	(0.0144)	(0.0204)
T1 × District COVID-19 Cases		0.0024**		0.0026**
		(0.0012)		(0.0011)
T2 × District COVID-19 Cases		0.0014		0.0021**
		(0.0012)		(0.0011)
Observations	2,117	2,117	2,117	2,117
R-squared	0.197	0.199	0.163	0.167
Control Mean DV	0.1244	0.1244	0.0939	0.0939
Control SD DV	0.3302	0.3302	0.2919	0.2919

Notes: Dependent variable in Columns 1-2 is social distancing measure in which threshold number of self-reported actions to be considered social distancing is 6 out of 8 (instead of 7 out of 8). Dependent variable in Columns 3-4 is social distancing measure which excludes social distancing actions 4 and 6 that are conditional on experiencing symptoms (threshold number of self-reported actions is changed to 5 out of 6 actions). Variables and suppressed controls (including community fixed effects) are defined in Table 3.2. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

C.7.4 Local COVID-19 Infection Rate Alternatives

In the main paper, we measure the local COVID-19 infection rate as the district-level cumulative COVID-19 cases per 100,000 population at the start of the endline survey. One concern might be that we only observe treatment effect heterogeneity using this one measure of the local infection rate but not other potentially justifiable measures.

Table C.10 instead shows that the primary results are robust to various other measures of the local infection rate: 1) "District COVID-19 Case Count" is the total count of district-level cumulative COVID-19 cases at the start of the endline survey (i.e., Table C.1, Column 1); 2)

"$\mathbb{1}$(Cases Above District Median)" is an indicator for if the respondent's district has above median COVID-19 cases relative to the sample's seven districts, thus applying to the top three districts; 3) "$\mathbb{1}$(Cases Above Sample Median)" is an indicator for if the respondent's district has above median COVID-19 cases relative to all sample respondents, applying to the top two districts (due to the large sample in the top-COVID-19 district); and 4) "$\mathbb{1}$(Cases Above National Average)" is an indicator if the respondent's district-level cumulative COVID-19 cases per 100,000 population is estimated at above Mozambique's national average at the start of the endline survey, which applies to only the top-COVID-19 district in the sample.[3]

In Columns (1)-(4), we observe that the misperception correction intervention has statistically significant treatment effect heterogeneity with respect to the local infection rate, as before, except in Column (4) where the standalone treatment effect is marginally significant (p-value=0.101). We conclude then that the finding is not an exception driven by our specific measure of the local COVID-19 infection rate.

[3]The national average of COVID-19 cases per capita at the start of the endline survey is estimated by taking the 9,296 cumulative COVID-19 cases in Mozambique on October 5, 2020 (https://coronavirus.jhu.edu/region/mozambique) and dividing by the World Bank estimate of the 2021 population of Mozambique at 320.8 per 100,000 (https://data.worldbank.org/indicator/SP.POP.TOTL?locations=MZ). Thus, we estimate 28.98 cases per 100,000 as the national average at this time.

Table C.10: Treatment Effects with Alternative Local COVID-19 Infection Rates

VARIABLES	(1) Primary SD Indicator	(2) Primary SD Indicator	(3) Primary SD Indicator	(4) Primary SD Indicator
T1: Misperceptions Correction	-0.0375**	-0.0342*	-0.0312*	-0.0218
	(0.0164)	(0.0188)	(0.0159)	(0.0133)
T2: Leader Endorsement	-0.0208	-0.0212	-0.0203	-0.0140
	(0.0172)	(0.0199)	(0.0168)	(0.0141)
T1 × District COVID-19 Case Count	0.0009***			
	(0.0003)			
T2 × District COVID-19 Case Count	0.0003			
	(0.0003)			
T1 × $\mathbb{1}$(Cases Above District Median)		0.0727***		
		(0.0281)		
T2 × $\mathbb{1}$(Cases Above District Median)		0.0292		
		(0.0274)		
T1 × $\mathbb{1}$(Cases Above Sample Median)			0.0909***	
			(0.0305)	
T2 × $\mathbb{1}$(Cases Above Sample Median)			0.0378	
			(0.0288)	
T1 × $\mathbb{1}$(Cases Above National Average)				0.1034**
				(0.0410)
T2 × $\mathbb{1}$(Cases Above National Average)				0.0358
				(0.0375)
Observations	2,117	2,117	2,117	2,117
R-squared	0.163	0.161	0.163	0.163
Control Mean DV	0.0857	0.0857	0.0857	0.0857
Control SD DV	0.2801	0.2801	0.2801	0.2801

Notes: Dependent variable defined in Table B.4, and treatment indicators and suppressed controls (including community fixed effects) are defined in Table 3.2. Remaining displayed variables interact treatment indicators with ways to specify the local COVID-19 infection rate other than our preferred measure: district-level cumulative COVID-19 cases per 100,000 population. "District COVID-19 Case Count" is the total count of district-level cumulative COVID-19 cases at the start of the endline survey (i.e., Table C.1, Column 1). "$\mathbb{1}$(Cases Above District Median)" is equal to one if the respondent's district has above median COVID-19 cases relative to the sample's seven districts, and zero otherwise. "$\mathbb{1}$(Cases Above Sample Median)" is equal to one if the respondent's district has above median COVID-19 cases relative to all sample respondents, and zero otherwise. "$\mathbb{1}$(Cases Above National Average)" is equal to one if the respondent's district-level cumulative COVID-19 cases per 100,000 population is estimated at above Mozambique's national average at the start of the endline survey, and zero otherwise. Robust standard errors in parentheses. Significance levels: *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

C.7.5 Excluding Chimoio District

A central finding of the paper is the heterogeneity in the treatment effect of the misperceptions correction treatment with respect to local COVID-19 cases per 100,000 population (Table 3.2 Column 2). A question that arises is whether this heterogeneity is entirely driven by the Chimoio district, which has the highest COVID-19 case load in the sample by a fair margin (see Figure 3.2 and Appendix C.2.3). We therefore test the robustness of our findings to excluding from the sample the 524 respondents in Chimoio district (one-quarter of the sample), thereby only exploiting the more limited variation in district-level case loads across the remaining six districts.

Table C.11 below presents coefficient estimates from this restricted sample. First of all, Column (1) reveals that the coefficient on the misperceptions correction treatment is negative and statistically significant at the 10% level. Because this sample drops the district with the highest case loads, this result is consistent with theoretical predictions and previous findings that at lower case loads, the misperceptions correction treatment effect is more likely to be negative.

In Column (2), where we test for heterogeneity in the treatment effect, results are quite similar to the findings in Column (2) of Table 3.2 in the main text. The T1 main effect and interaction term coefficients are of similar magnitudes to those in Column (2) of Table 3.2, and maintain statistical significance at conventional levels (the T1 interaction term coefficient is now significant at the 10% instead of 5% level).

In sum, our central findings regarding heterogeneity in the treatment effect of the misperceptions correction treatment are robust to excluding from the sample respondents from the district (Chimoio) with the highest COVID-19 case loads.

Table C.11: Treatment Effects Excluding Chimoio District

VARIABLES	(1) Primary SD Indicator	(2) Primary SD Indicator
T1: Misperceptions Correction	-0.0237*	-0.0410**
	(0.0131)	(0.0194)
T2: Leader Endorsement	-0.0150	-0.0263
	(0.0141)	(0.0208)
T1 × District COVID-19 Cases		0.0019*
		(0.0010)
T2 × District COVID-19 Cases		0.0012
		(0.0010)
Observations	1,593	1,593
R-squared	0.141	0.142
Control Mean DV	0.0710	0.0710
Control SD DV	0.2570	0.2570

Notes: Regressions exclude 524 respondents from Chimoio district. Dependent variable defined in Table B.4, and variables and suppressed controls (including community fixed effects) are defined in Table 3.2. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

C.8 Populated Pre-analysis Plan

On August 25, 2020, prior to baseline data collection, we uploaded our pre-analysis plan (PAP) to the American Economic Association's RCT Registry, registration ID number AEARCTR-0005862: https://doi.org/10.1257/rct.5862-3.0.

In our PAP, we specify the following regression for our primary analysis, which is the same as Equation 3.4.1 in the main text:

$$Y_{ijd} = \beta_0 + \beta_1 T1_{ijd} + \beta_2 T2_{ijd} + \eta B_{ijd} + \delta_{ijd}^{others} + \delta_{ijd}^{leaders} + \gamma_{jd} + \varepsilon_{ijd} \tag{C.8.1}$$

where Y_{ijd} is the social distancing indicator for household i in community j and district d; $T1_{ijd}$ and $T2_{ijd}$ are indicator variables for the misperceptions correction and leader endorsement treatment groups, respectively; B_{ijd} is the baseline value of the dependent variable; γ_{jd} are community fixed effects; and ε_{ijd} is a mean-zero error term. We report robust standard errors. The regression also controls for the number of other survey respondents and community leaders who report knowing the survey respondent at baseline (in Round 2). Specifically, δ_i^{others} is a vector of dummy variables for the distinct number of other surveyed study respondents who report knowing the household (0, 1, 2..., 7, 8 or more; where 8 is the first integer where over 90% of the sample is represented by previous non-negative integers), and $\delta_i^{leaders}$ is a vector of dummy variables for the distinct number of community leaders who report knowing the household (0, 1, 2, 3, 4; where 4 is maximum number of leaders found within one of the 76 sample communities). Including this control variable helps reduce residual variance in the dependent variable, because respondents who are known by more others in the community will also have more reports of social interactions with others. These results are presented in the main paper in Table 3.2 Column (1) and are also replicated in Column (1) of Table C.12.

Additionally, we pre-specified the following secondary analyses. First, we analyze impacts of the social distancing treatments on the separate components of the social distancing index—the others' and self-report. These results are presented in Table C.12 Columns (2) and (3), respectively. Treatment effects on these outcomes are very similar to those in Column (1).

Second, we also pool SD1 and SD2 together to examine the effect of some endorsement of social distancing (whether by other community members or by community leaders) on the primary social distancing outcome. These coefficient in Table C.12 Column (4) is small in magnitude and not statistically significantly different from zero at conventional levels.

Table C.12: Additional Pre-specified Analyses

	(1)	(2)	(3)	(4)
VARIABLES	Primary SD Indicator	Others' Report of SD	Self-Report of SD	Primary SD Indicator
T1: Misperceptions Correction	0.0042	0.0010	0.0134	
	(0.0140)	(0.0181)	(0.0238)	
T2: Leader Endorsement	-0.0054	0.0145	-0.0189	
	(0.0137)	(0.0183)	(0.0234)	
Pooled SD Treatments				-0.0006
				(0.0116)
Observations	2,117	2,117	2,117	2,117
R-squared	0.158	0.333	0.211	0.158
Control Mean DV	0.0857	0.2113	0.4061	0.0857
Control SD DV	0.2801	0.4084	0.4914	0.2801

Notes: Dependent variables are defined in Table B.4. "T1: Misperceptions Correction" is an indicator equal to one if respondent was randomly assigned to the misperceptions correction treatment, and zero otherwise. "T2: Leader Endorsement" is an indicator equal to one if respondent was randomly assigned to the leader endorsement treatment, and zero otherwise. "Pooled SD Treatments" is an indicator equal to one if respondent was randomly assigned to the misperceptions correction treatment or leader endorsement treatment, and zero otherwise. Controls are as defined in Table 3.2. All regressions also include community fixed effects. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

We also randomly assigned a family of treatments to improve COVID-19 knowledge in the same study population.[4] Randomization of the misperceptions correction and leader endorsement treatments were stratified within 76 communities and within the separate

[4]The pre-analysis plan (PAP) for the knowledge study can be found here: https://fordschool.umich.edu/mozambique-research/combatting-COVID-19.

knowledge treatment conditions (i.e., the knowledge and social distancing treatments were cross-randomized). As pre-specified, we run a regression on the primary social distancing outcome with indicators for social distancing treatments, the cross-randomized knowledge treatments and their interaction terms. Results are presented in Table C.13, and show no large or statistically significant interaction effects between the social distancing and knowledge treatments.

Table C.13: Interactions between Social Distancing and Knowledge Treatments

	(1)
VARIABLES	Primary SD Indicator
T1: Misperceptions Correction	-0.0237
	(0.0214)
T2: Leader Endorsement	-0.0210
	(0.0222)
K1: Incentive	-0.0218
	(0.0241)
K2: Feedback	-0.0025
	(0.0251)
K3: Incentive & Feedback	-0.0144
	(0.0238)
T1 × K1	0.0545
	(0.0390)
T2 × K1	0.0249
	(0.0372)
T1 × K2	0.0467
	(0.0397)
T2 × K2	0.0139
	(0.0385)
T1 × K3	0.0404
	(0.0382)
T2 × K3	0.0374
	(0.0372)
Observations	2,117
R-squared	0.160
Control Mean DV	0.0857
Control SD DV	0.2801

Notes: Dependent variable is defined in Table B.4. Social distancing treatments are defined in Table 3.2. "K1 Incentive", "K2 Feedback", and "K3 Incentive & Feedback" are indicators equal to one if respondent was randomly assigned to one of these knowledge treatments, and zero otherwise. Remaining regressors represent interactions between social distancing treatments and the knowledge treatments. Controls are as defined in Table 3.2. Regression also includes community fixed effects. Robust standard errors in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.1.

BIBLIOGRAPHY

BIBLIOGRAPHY

ADHVARYU, A. R. AND A. NYSHADHAM (2012): "Schooling, child labor, and the returns to healthcare in Tanzania," *Journal of Human resources*, 47, 364–396.

ADMASSIE, A. (2003): "Child labour and schooling in the context of a subsistence rural economy: can they be compatible?" *International Journal of Educational Development*, 23, 167–185.

ADÃO, R., M. KOLESÁR, AND E. MORALES (2019): "Shift-share designs: Theory and inference," *The Quarterly Journal of Economics*, 134, 1949–2010.

AKER, J. C. AND M. SAWYER (2021): "Making Sense of the Signs: What Do We Know About Learning in Adulthood?" *Working Paper*.

ALLEN IV, J., A. MAHUMANE, J. RIDDELL IV, T. ROSENBLAT, D. YANG, AND H. YU (2021): "Correcting Perceived Social Distancing Norms to Combat COVID-19," *NBER Working Paper*.

ALSAN, M., F. CODY STANFORD, A. BANERJEE, E. BREZA, A. G. CHANDRASEKHAR, S. EICHMEYER, P. GOLDSMITH-PINKHAM, L. OGBU-NWOBODO, B. A. OLKEN, C. TORRES, A. SANKAR, P. VAUTREY, AND E. DUFLO (2020): "Comparison of Knowledge and Information-Seeking Behavior After General COVID-19 Public Health Messages and Messages Tailored for Black and Latinx Communities: A Randomized Controlled Trial," *Annals of Internal Medicine*, 174, 484–492.

AMERICAN BAR ASSOCIATION (2021): *Division of Public Education*, Washington D.C., USA https://www.americanbar.org/groups/public_education/.

ANDERSON, J. R. AND G. FEDER (2007): "Agricultural Extension," *Handbook of Agricultural Economics*, 3, 2343–2378.

ANDERSSON, O., P. CAMPOS-MERCADE, A. MEIER, AND E. WENGSTROM (2021): "Anticipation of COVID-19 Vaccines Reduces Social Distancing," *SSRN Working Paper*.

ANDRE, P., T. BONEVA, F. CHOPRA, AND A. FALK (2021): "Fighting climate change: The role of norms, preferences, and moral values," .

ANGRIST, J. AND V. LAVY (2009): "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial," *American Economic Review*, 99, 1384–1414.

Angrist, N., P. Bergman, D. K. Evans, S. Hares, M. C. H. Jukes, and T. Letsomo (2020a): "Practical Lessons for Phone-based Assessments of Learning," *BMJ Global Health*, 5.

Angrist, N., P. Bergman, and M. Matsheng (2020b): "School's Out: Experimental Evidence on Limiting Learning Loss Using "Low-Tech" in a Pandemic," *NBER Working Paper*.

Atkin, D. (2016): "Endogenous Skill Acquisition and Export Manufacturing in Mexico," *American Economic Review*, 106, 2046–85.

Bahety, G., S. Bauhoff, D. Patel, and J. Potter (2021): "Texts don't nudge: An adaptive trial to prevent the spread of COVID-19 in India," *Journal of Development Economics*, 153, 102747.

Baker, M., J. Gruber, and K. Milligan (2008): "Universal child care, maternal labor supply, and family well-being," *Journal of political Economy*, 116, 709–745.

Banerjee, A., M. Alsan, E. Breza, A. G. Chandrasekhar, A. Chowdhury, E. Duflo, P. Goldsmith-Pinkham, and B. A. Olken (2020): "Messages on Covid-19 Prevention in India Increased Symptoms Reporting and Adherence to Preventive Behaviors Among 25 Million Recipients with Similar Effects on Non-recipient Members of Their Communities," *NBER Working Paper*.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2019a): "Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials," *The Review of Economic Studies*, 86, 2453–2490.

Banerjee, A., S. Cole, E. Duflo, and L. Linden (2007): "Remedying education: Evidence from Two Randomized experiments in India," *Quarterly Journal of Economics*, 122, 1235–1264.

Banerjee, A., E. La Ferrara, and V. H. Orozco-Olvera (2019b): "The Entertaining Way to Behavioral Change: Fighting HIV with MTV," *NBER Working Paper*.

Banerjee, A. V. and E. Duflo (2011): *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*, New York, United States: Public Affairs.

Barsbai, T., V. Licuanan, A. Steinmayr, E. Tiongson, and D. Yang (2020): "Information and the Acquisition of Social Network Connections," *NBER Working Paper*.

Batterham, R. W., M. Hawkins, P. A. Collins, R. Burchbinder, and R. H. Osborne (2016): "Health Literacy: Applying Current Concepts to Improve Health Services and Reduce Health Inequalities," *Public Health*, 132, 3–12.

Becker, G. S. (1962): "Investment in Human Capital: A Theoretical Analysis," *Journal of Political Economy*, 70, 9–49.

Behrman, J. R., S. W. Parker, P. E. Todd, and K. I. Wolpin (2015): "Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools," *Journal of Political Economy*, 123, 325–364.

Benabou, R. and J. Tirole (2011): "Identity, Morals, and Taboos: Beliefs as Assets," *Quarterly Journal of Economics*, 126, 805–855.

Berry, J., H. Kim, and H. Son (2019): "When Student Incentives Don't Work: Evidence from a Field Experiment in Malawi," 1–54.

Bicchieri, C. and E. Dimant (2019): "Nudging with Care: The Risks and Benefits of Social Information," *Public choice*, 1–22.

Bjorkman Nyqvist, M., L. Corno, D. de Walque, and J. Svensson (2018): "Incentivizing Safer Sexual Behavior: Evidence from a Lottery Experiment on HIV Prevention," *American Economic Journal: Applied Economics*, 10, 287–314.

Borusyak, K. and P. Hull (2021): "Non-Random Exposure to Exogenous Shocks: Theory and Applications," .

Borusyak, K., P. Hull, and X. Jaravel (2022): "Quasi-experimental shift-share research designs," *The Review of Economic Studies*, 89, 181–213.

Burgess, S., R. Metcalfe, and S. Sadoff (October 2016): "Understanding the Response to Financial and Non-Financial Incentives in Education: Field Experimental Evidence Using High-Stakes Assessments," *IZA Institute of Labor Economics*.

Bursztyn, L., A. L. González, and D. Yanagizawa-Drott (2020): "Misperceived social norms: Women working outside the home in Saudi Arabia," *American economic review*, 110, 2997–3029.

Bustillo, I. (1989): *Female Educational Attainment in Latin America : A Survey*, Washington, DC: World Bank.

Carpena, F., S. Cole, J. Shapiro, and B. Zia (2017): "The ABCs of Financial Education: Experimental Evidence on Attitudes, Behavior, and Cognitive Biases," *Management Science*, 65, 346–369.

Cato, S., T. Iida, K. Ishida, A. Ito, K. McElwain, and M. Shoji (2020): "Social Distancing as a Public Good Under the COVID-19 Pandemic," *Public Health*, 188, 51–53.

Checchi, F., A. Warsame, V. Treacy-Wong, J. Polonsky, M. Van Ommeren, and C. Prudhon (2017): "Public health information in crisis-affected populations: a review of methods and their use for advocacy and action," *The Lancet*, 390, 2297–2313.

Conn, K. M. (2017): "Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations," *Review of Educational Research*, 87, 863–898.

Dillon, B. (2021): "Selling crops early to pay for school A large-scale natural experiment in Malawi," *Journal of Human Resources*, 56, 1296–1325.

Dizon-Ross, R. (2019): "Parents' beliefs about their children's academic ability: Implications for educational investments," *American Economic Review*, 109, 2728–65.

Duflo, E. (2001): "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 91, 795–813.

Duflo, E., A. Banerjee, A. Finkelstein, L. Katz, B. Olken, and A. Sautmann (2020): "In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plan for RCTs in Economics," *NBER Working Paper Series W26993*.

Duflo, E., P. Dupas, and M. Kremer (2011): "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, 101, 1739–1774.

Dupas, P. et al. (2011): "Health Behavior in Developing Countries," *Annual Review of Economics*, 3, 425–449.

Edmonds, E. V. (2007): "Child labor," *Handbook of Development Economics*, 4, 3607–3709.

Evans, D. K. and A. Popova (2015): "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews," *Oxford University Press on behalf of the World Bank*, 31, 242–70.

Fabregas, R., M. Kremer, M. Lowes, R. On, and G. Zane (2019): "SMS-extension and Farmer Behavior: Lessons from Six RCTs in East Africa," *ATAI Research Publications*.

Fafchamps, M., A. Vaz, and P. C. Vicente (2020): "Voting and peer effects: Experimental evidence from Mozambique," *Economic Development and Cultural Change*, 68, 567–605.

FEWS NET (2013): "Malawi: Food Security Outlook," Tech. rep., Famime Early Warning System Network.

Finkelstein, E. A., M. Bilger, and D. Baid (2019): "Effectiveness and Cost-effectiveness of Incentives as a Tool for Prevention of Non-communicable Diseases: A Systematic Review," *Social Science & Medicine*, 232, 340–350.

Fisher, R. A. (1935): "The logic of inductive inference," *Journal of the royal statistical society*, 98, 39–82.

Frye, M. (2011): "Malawi School Calendar Change 2009-2001, Memo for the Tsogolo La Thanzi Centre," Tech. rep.

Fryer, R. G. (2011): "Financial Incentives and Student Achievement: Evidence from Randomized Trails," *The Quarterly Journal of Economics*, 126, 6755–1798.

——— (2016): "Information, Non-financial Incentives, and Student Achievement: Evidence from a Text Messaging Experiment," *Journal of Public Economics*, 144, 109–121.

Fryer, R. G., T. Devi, and R. T. Holden (2016): "Vertical Versus Horizontal Incentives in Education: Evidence from Randomized Trails," *NBER Working Paper*.

Fuhrmann-Riebel, H., B. D'Exelle, K. López Vargas, S. Tonke, and A. Verschoor (2023): "Correcting Misperceptions About Trends and Norms to Address Weak Collective Action," *Available at SSRN 4311583*.

Gibson, M. and J. Shrader (2018): "Time use and labor productivity: The returns to sleep," *Review of Economics and Statistics*, 100, 783–798.

Glewwe, P. (2014): *'Overview of Education Issues in Developing Countries', in Education Policy in Developing Countries*, Chicago, USA: University of Chicago Press.

Glewwe, P., M. Kremer, and S. Moulin (2009): "Many Children Left Behind? Textbooks and Test Scores in Kenya," *American Economic Journal*, 1, 112–135.

Glewwe, P., M. Kremer, S. Moulin, and E. Zitzewitz (2000): "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya," *Journal of Development Economics*, 74, 251–268.

Government of Malawi (2012): "Malawi Integrated Household Panel Survey (IHPS) 2010-11 Basic Information Document," Tech. rep., Malawi Government, National Statistical Office.

——— (2017): "Malawi Integrated Household Panel Survey (IHPS) 2016 Basic Information Document," Tech. rep., Malawi Government, National Statistical Office.

Heckman, J. J. (2015): "Introduction to a Theory of the Allocation of Time by Gary Becker," *The Economic Journal*, 125, 403–409.

Hershey, J. C., D. A. Asch, T. Thumasathit, J. Meszaros, and V. W. Waters (1994): "The Roles of Altruism, Free Riding, and Bandwagoning in Vaccination Decisions," *Organizational Behavior and Human Decision Processes*, 59, 177–187.

Hirshleifer, S. (2017): "Incentives for Effort or Outputs? A Field Experiment to Improve Student Performance," *Abdul Latif Jameel Poverty Action Lab (J-PAL)*.

Ito, S. and A. Shonchoy (2020): "Seasonality, Academic Calendar and School Drop-outs in Developing Countries," Working Papers 2013, Florida International University, Department of Economics.

Jakubowski, A., D. Egger, C. Nekesa, L. Lowe, M. Walker, and E. Miguel (2021): "Self-Reported Mask Wearing Greatly Exceeds Directly Observed Use: Urgent Need for Policy Intervention in Kenya," *medRxiy Preprint*.

Janzwood, S. (April 27, 2020): *The Social Distancing Norm Cascade: The Role of Belief Systems in Accelerating Normative Change during the COVID-19 Pandemic*, Canada: Cascade Institute.

Jensen, R. (2010): "The (perceived) returns to education and the demand for schooling," *The Quarterly Journal of Economics*, 125, 515–548.

Jensen, R. and N. H. Miller (2017): "Keepin' 'em Down on the Farm: Migration and Strategic Investment in Children's Schooling," Working Paper 23122, National Bureau of Economic Research.

Jones, S., E. Egger, and R. Santos (2020): "Is Mozambique Prepared for a Lockdown During the COVID-19 Pandemic?" *UNU-WIDER Blog*.

Kadzamira, E. and P. Rose (2003): "Can free Primary Education meet the Needs of the Poor?: Evidence from Malawi," *International Journal of Educational Development*, 23, 501–516.

Kaiser, T. and L. Menkhoff (2017): "Does Financial Education Impact Financial Literacy and Financial Behavior, and if so, When?" *World Bank Economic Review*, 31, 611–630.

Kattan, R. B. and N. Burnett (2004): "User Fees in Primary Education." *World Bank Education Advisory Service*.

Kremer, M., E. Miguel, and R. Thornton (2009a): "Incentives to learn," *The Review of Economics and Statistics*, 91, 437–456.

——— (2009b): "Incentives to Learn," *The Review of Economics and Statistics*, 91, 437–456.

Lau, K., M. Miraldo, M. M. Galizzi, and K. Hauck (2019): "Social Norms and Free-riding in Influenza Vaccine Decisions in the UK: An Online Experiment," *The Lancet*, 394, 65.

Le, V. (2015): "Should Students be Paid for Achievement? A Review of the Impact of Monetary Incentives on Test Performance," *NORC at the University of Chicago*.

Levitt, S. D., J. A. List, S. Neckermann, and S. Sadoff (2011): "The Impact of Short-term Incentives on Student Performance," *University of Chicago*.

Li, T., L. Han, L. Zhang, and S. Rozelle (2014): "Encouraging Classroom Peer Interactions: Evidence from Chinese Migrant Schools," *Journal of Public Economics*, 111, 29–45.

List, J., A. Shaikh, and Y. Xu (2019): "Multiple Hypothesis Testing in Experimental Economics," *Experimental Economics*, 22, 773–793.

List, J. A., J. A. Livingston, and S. Neckermann (2018): "Do Financial Incentives Crowd Out Intrinsic Motivation to Perform on Standardized Tests?" *Economics of Education Review*, 66, 125–136.

Maffioli, E. (2020): "Collecting Data During an Epidemic: A Novel Mobile Phone Research Method," *Journal of International Development*, 32, 1231–1255.

Magaço, A., K. Munguambe, A. Nhacolo, C. Ambrósio, F. Nhacolo, S. Cossa, E. Macete, and I. Mandomando (2021): "Challenges and needs for social behavioural research and community engagement activities during the COVID-19 pandemic in rural Mozambique," *Global Public Health*, 16, 153–157.

Mankiw, N. G., D. Romer, and D. N. Weil (1992): "A contribution to the empirics of economic growth," *The quarterly journal of economics*, 107, 407–437.

Martinez, D., C. Parilli, C. Scartascini, and A. Simpser (2021): "Let's (not) get together! The role of social norms on social distancing during COVID-19," *PLoS ONE*, 16, 1–14.

Maude, R. R., M. Jongdeepaisal, S. Skuntaniyom, T. Muntajit, S. D. Blacksell, W. Khuenpetch, W. Pan-Ngum, K. Taleangkaphan, K. Malathum, and R. J. Maude (2021): "Improving Knowledge, Attitudes and Practices to Prevent COVID-19 Transmission in Healthcare Workers and the Public in Thailand," *BMC Public Health*, 21, 749.

Mbiti, I., K. Muralidharan, M. Romero, Y. Schipper, C. Manda, and R. Rajani (2019): "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania," *The Quarterly Journal of Economics*, 134, 1627–1673.

McEwan, P. J. (2015): "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments," *Review of Educational Research*, 85, 353–394.

Ministry of Health (2020): *COVID-19 Epidemiological Situation, Republic of Mozambique*, Maputo, Mozambique: Government of Mozambique.

Mistree, D., P. Loyalka, R. Fairlie, A. Bhuradia, M. Angrish, J. Lin, A. Karoshi, S. J. Yen, J. Mistri, and V. Bayat (2021): "Instructional Interventions for Improving COVID-19 Knowledge, Attitudes, Behaviors: Evidence from a Large-scale RCT in India," *Social Science & Medicine*, 276, 1–6.

Montenegro, C. E. and H. A. Patrinos (2014): "Comparable estimates of returns to schooling around the world," *World Bank policy research working paper*.

Montero, E. and D. Yang (2022): "Religious Festivals and Economic Development: Evidence from the Timing of Mexican Saint Day Festivals," *American Economic Review*.

Muralidharan, K. (2017): "Field Experiments in Education in Developing Countries," *Handbook of Economic Field Experiments*, 2, 323–385.

Muralidharan, K., M. Romero, and K. Wuthrich (2019a): "Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments," *NBER Working Paper*.

Muralidharan, K., A. Singh, and A. J. Ganimian (2019b): "Disrupting education? Experimental Evidence on Technology-Aided Instruction in India," *American Economic Review*, 109, 1426–60.

National Institute of Statistics (INE) (2017): *Mozambique Population and Housing Census 2017*, Maputo, Mozambique: Government of Mozambique.

Nglazi, M. D., N. Van Schaik, K. Kranzer, MRCP(UK), S. D. Lawn, R. Wood, and L. Bekker (2012): "An Incentivized HIV Counseling and Testing Program Targeting Hard-to-Reach Unemployed Men in Cape Town, South Africa," *Journal of Acquired Immune Deficiency Syndromes*, 59, 28–34.

Nguyen, T. (2008): "Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar," .

Nyusi, F. J. (August 5, 2020): *Communication to the Nation of His Excellency Philip Jacinto Nyusi, President of Republic of Mozambique, on the New State of Emergency, within the Scope of the Coronavirus Pandemic COVID-19*, Maputo, Mozambique: Maputo Mozambique.

——— (September 5, 2020): *Communication to the Nation of His Excellency Philip Jacinto Nyusi, President of Republic of Mozambique, on the New State of Emergency, within the Scope of the Coronavirus Pandemic COVID-19*, Maputo, Mozambique: Maputo Mozambique.

Psacharopoulos, G. and H. A. Patrinos (2018): "Returns to Investment in Education: a Decennial Review of the Global Literature," *Education Economics*, 26, 445–458.

Puspitasari, I. M., L. Yusuf, R. K. Sinuraya, R. Abdulah, and H. Koyama (2020): "Knowledge, attitude, and practice during the COVID-19 pandemic: a review," *Journal of multidisciplinary healthcare*, 13, 727.

Republic of Mozambique (April 2, 2020): *"Bulletin of the Republic", I Series, No. 64*, Maputo, Mozambique.

——— (August 5, 2020): *"Bulletin of the Republic", I Series, No. 149*, Maputo, Mozambique.

——— (March 31, 2020): *"Bulletin of the Republic", I Series, No. 62*, Maputo, Mozambique.

Santos, R. J. (2014): "Not All that Glitters is Gold: Gold Boom, Child Labor and Schooling in Colombia," *SSRN, Documento CEDE No. 2014-31.*

Schiefelbein, E. (1987): "Education Costs and Financing Policies in Latin America," *World Bank, Education and Training Department Discussion Paper EDT60.*

Schultz, W. P., J. M. Nolan, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius (2007): "The Constructive, Destructive, and Reconstructive Power of Social Norms," *Psychological Science*, 18, 429–434.

Shah, M. and B. M. Steinberg (2017): "Drought of Opportunities: Contemporaneous and Long-Term Impacts of Rainfall Shocks on Human Capital," *Journal of Political Economy*, 125, 527–561.

Siuta, M. and M. Sambo (April 1, 2020): *COVID-19 Em Mocambique: Dimensao e Possiveis Impactos. Boletim No. 124p*, Maputo, Mozambique: Instituto de Estudos Socias e Economicos.

Spiegler, R. (2020): "Behavioral Implications of Causal Misperceptions," *Annual Review of Economics*, 12, 81–106.

St. Louis Federal Reserve Bank, FRED Economic Data (2012): "Exchange rate to U.S. dollar for Malawi," .

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

Twinam, T. (2017): "Complementarity and Identification," *Econometric Theory*, 33, 1154–1185.

UIS and UNICEF (2015): "Fixing the Broken Promise of Education for All: Findings from the Global Initiative on Out-of-School Children." Tech. rep.

UNESCO, Institute for Statistics (2022): .

U.S Embassy in Mozambique (2020): "COVID-19 Information," .

Vinck, P., P. N. Pham, K. K. Bindu, J. Bedford, and E. J. Nilles (2019): "Institutional trust and misinformation in the response to the 2018–19 Ebola outbreak in North Kivu, DR Congo: a population-based survey," *The Lancet Infectious Diseases*, 19, 529–536.

Watkins, K. (2000): "The Oxfam Education Report," Tech. rep., Oxfam.

World Bank (2010): *The Education System in Malawi*, The World Bank.

——— (2017): *World Development Report 2018: Learning to Realize Education's Promise*, The World Bank.

——— (2021): "Employment in Agriculture (% of Total Employment)(modeled ILO estimate)," Tech. rep.

Yang, D., J. Allen IV, A. Mahumane, J. Riddell IV, and H. Yu (2021): "Knowledge, Stigma, and HIV Testing: An Analysis of a Widespread HIV/AIDS Program," Tech. rep., National Bureau of Economic Research.

Yu, H. (2020): *Three Essays in Development Economics. Dissertation*, Ann Arbor, MI: University of Michigan.