Teaching and Incentives: Substitutes or Complements?

James Allen IV, Arlete Mahumane, James Riddell IV, Tanya Rosenblat, Dean Yang, and Hang Yu

NBER Working Paper No. 28976
July 2021, Revised February 2022
JEL No. D90,I12

ABSTRACT

Interventions to promote learning are often categorized into supply- and demand-side approaches. In a randomized experiment to promote learning about COVID-19 among Mozambican adults, we study the interaction between a supply and a demand intervention, respectively: teaching, and providing financial incentives to learners. In theory, teaching and learner-incentives may be substitutes (crowding out one another) or complements (enhancing one another). Experts surveyed in advance predicted a high degree of substitutability between the two treatments. In contrast, we find substantially more complementarity than experts predicted. Combining teaching and incentive treatments raises COVID-19 knowledge test scores by 0.5 standard deviations. The complementarity between teaching and incentives persists in the longer run, over nine months post-treatment.

James Allen IV
Department of Economics and
Gerald R. Ford School of Public Policy
University of Michigan
735 S. State Street
Ann Arbor, MI 48109
alleniv@umich.edu

Arlete Mahumane
Beira Operational Research Center (CIOB)
Rua Correia de Brito 1323
Beira City
Sofala Province
Mozambique
cynthiwea@gmail.com

James Riddell IV
Department of Internal Medicine
University of Michigan Medical School
Ann Arbor, MI 48109
jriddell@umich.edu

Tanya Rosenblat
School of Information and
Department of Economics
University of Michigan
4332 N Quad
Ann Arbor, MI 48109
trosenbl@umich.edu

Dean Yang
University of Michigan
Department of Economics and
Gerald R. Ford School of Public Policy
735 S. State Street, Room 3316
Ann Arbor, MI 48109
and NBER
deanyang@umich.edu

Hang Yu
National School of Development
Peking University
hangyu@umich.edu

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, taking a "teaching at the right level" (TaRL) pedagogical approach (Banerjee et al., 2007; Duflo et al., 2011). 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., 2019), and "teaching at the right level" (TaRL) (Banerjee and Duflo, 2011; Duflo et al., 2011). Angrist et al. (2020) 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 TaRL 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., 2009; 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., 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

[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

particular theoretical interest due to the countervailing motivation and crowding-out effects of combining supply- and demand-side educational interventions.

Related studies seek to improve COVID-19-related knowledge. 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.

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.

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 peer-effects intervention (supply-side) and providing test-score financial incentives (demand-side).

- 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 correct 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.
- 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:

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

We can organize the share of correct answers by treatment, in Table 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:

$$\lambda \equiv \text{Joint} - (\text{Teaching only} + \text{Incentive only}) = \underbrace{(p(R) - p(0))\,(1 - \mu)}_{\text{motivation}} - \underbrace{e^* p(1 - \mu)}_{\text{crowding out}} \tag{2}$$

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

3 Sample and Data

3.1 Data

We implemented surveys by phone in July-November 2020. Respondents were from households with phones in the sample of a prior study (Yang et al., 2021).[4] We surveyed one adult per household. Appendix A provides details on the COVID-19 context, study communities and the study timeline.

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

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. See Appendix B for details on question selection and the list of questions.[5]

3.2 Primary Outcomes

Our primary outcome is a COVID-19 knowledge test score: the share of knowledge questions answered correctly. Responses are considered "correct" if they match the pre-specified correct answer and are "incorrect" otherwise. We construct this outcome as the share of correct answers to 20 knowledge questions asked at endline that were also randomly selected for the respondent to answer at baseline.[6] This allows us to track, within respondent, changes in test scores on exactly the same questions between the baseline and endline.[7]

In the control group (N=847), this outcome has a mean of 0.784 and a standard deviation of 0.123.

[5] 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)".

[6] At baseline, each respondent was assigned a randomized subset of 20 out of 40 questions, distributed as follows across subcategories: 6 (out of 12) general knowledge, 8 (out of 16) preventive action, and 6 (out of 12) government policy questions.

[7] This COVID-19 test score based on the 20 questions asked of the respondent at both baseline and endline is one of two primary outcomes pre-specified in our PAP. In this paper, we focus on only this outcome, for brevity. Results and conclusions (Appendix F) are very similar when examining the other pre-specified primary outcome, the COVID-19 knowledge test score constructed from an expanded set of 40 questions, including questions that the respondent was asked for the first time at endline (respondents would not have been asked these questions at baseline).

4 Empirical Approach

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%). We describe the treatments briefly below. Complete implementation protocols can be found in Appendix C.

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 COVID-19 knowledge questions on the endline survey. They were also told that this would allow them to earn 200 meticais (approx. US$2.71), if they answered all 40 questions correctly, in addition to their 50 meticais participation fee on the endline survey. 250 meticais is equivalent to half of the sample median pre-pandemic (February 2020) weekly household income.

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.

We also randomly assigned treatments to improve social distancing (Allen IV et al., 2021). Randomization of the Incentive, Teaching, and Joint treatments were stratified within 76 communities and within the separate social distancing treatment conditions. Further details are in Appendix F, where we also present regression results showing that there are no interactions between the social distancing treatments and this paper's treatments.

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%). Attrition between baseline and endline is low (4.9%). In Appendix D, we show that attrition between baseline and endline as well as key baseline variables are balanced across treatment conditions.

[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."

4.2 Regression

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

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

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: $\hat{\lambda} = \beta_3 - (\beta_1 + \beta_2)$.

We also analyze impacts on test scores for question subcategories: general knowledge, preventive actions, and government policies.

4.3 Hypotheses

We hypothesize that each treatment has a positive effect on test scores: the coefficients β_1, β_2 and β_3 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 $\lambda = 0$.

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 F, 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.

In the analyses presented in this paper, we depart from the PAP in minor ways. First, we present results for only one out of two primary outcomes, the test score based on the 20 questions (randomly selected) that a respondent was asked in both the baseline and endline surveys. Results shown in Appendix F for the other pre-specified primary outcome (a test score based on a larger set of 40 questions, including those not asked in the baseline and

[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).

therefore not eligible for the Teaching treatment) are very similar and yield the same substantive conclusions. We present just one of the two outcomes in this paper for brevity, and chose the outcome based on questions asked in both rounds to maximize the comparability of treatment effects across our treatment conditions.[12]

Second, when adjusting coefficient p-values for multiple hypothesis testing (MHT), we adjust across the three coefficients in Regression 3. By contrast, in the PAP, we pre-specified that we would adjust p-values across three coefficients, but where one coefficient (on Incentive) was from one regression (where the outcome was the test score based on 40 questions) and the other two coefficients (on Teaching and Incentive plus Teaching) were from another regression (where the outcome was the test score based on 20 previously-asked questions). We depart from the PAP for simplicity, because we chose not to show the regression for the test score based on 40 questions (see previous paragraph). In Appendix F we adjust for MHT across the three coefficients from the two regressions as pre-specified, and p-values are very similar (all are $<$0.001).

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 predicted that $\lambda < 0$.

4.5 Expert Predictions

In advance of presenting our results publicly, we surveyed subject-matter experts on their expectations of our treatment effects.[13] In Appendix E we describe the expert prediction survey in greater detail, and provide summary statistics of the predictions.

Figure 1 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

[12] 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.

[13] Predictions were provided at https://socialscienceprediction.org/ (survey closing date January 2, 2021).

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 $\tilde{\lambda}$.

5 Results

5.1 Primary Analysis

Table 2 presents the results from testing this paper's primary hypotheses. Our first pre-specified primary hypothesis refers to the effect of the incentive treatment on the COVID-19 knowledge test score. This is the coefficient β_1 on the incentive treatment in Equation 3, for which we present the coefficient estimate in Column 1 of the table. The Incentive treatment has a positive effect on test scores, and is statistically significantly different from zero (p-val=0.0133) after multiple hypothesis testing (MHT) adjustment. The point estimate indicates a 0.0156 increase, relative to the 0.784 mean control group test score. This effect is substantial in magnitude, amounting to 0.13 standard deviations of the outcome variable.

We also pre-specified primary hypotheses on the effect of the Teaching treatment and Joint treatment on COVID-19 knowledge test scores (β_2 and β_3 in Equation 3, respectively). Coefficient estimates (also in Column 1) indicate that the Teaching and Joint treatments each also have positive 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.

The fourth row of the table displays the estimate, $\hat{\lambda}$, of the complementarity parameter, and its standard error. In Column 1, $\hat{\lambda} = 0.0137$, 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. 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 graphically. In Figure 2, 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 3, 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 marginal level of statistical significance that $\lambda = 0$.

This complementarity is also present when evaluating treatment effects on newly asked questions, alleviating concern that results are driven by rote memorization. In Column 2 of Table 2, we run regression 3 replacing 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.[14] This analysis was pre-specified in our PAP as of secondary interest. 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.

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 G. 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 A.5. 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 G we provide an example of alternative relative implementation costs that would lead Joint to be the most cost-effective at $\hat{\lambda}$.

[14] Summary statistics for the number of new questions at endline: mean=14.4; sd=1.8; min=7; max=20.

5.3 Long-Run Analysis

We also 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 H.

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. 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 A.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.

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); among students; and in other study populations. 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.[15] We view these as promising directions for future research.

[15]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).

References

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* (28651).

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.

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. Kares, M. C. H. Jukes, and T. Lestsomo (2020). Practical Lessons for Phone-Based Assessments of Learning. *BMJ Global Health 5*.

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. 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* (27346).

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.

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.

Berry, J., H. Kim, and H. Son (2019). When Student Incentives Don't Work: Evidence from a Field Experiment in Malawi. pp. 1–54.

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.

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*.

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.

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.

Duflo, E. (2001). Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. *American Economic Journal 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.

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*.

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.

Fryer, R. G. (2011). Financial Incentives and Student Achievement: Evidence from Randomized Trails. *The Quarterly Journal of Economics 126*, 6755–1798.

Fryer, R. G. (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* (17752).

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.

Hirshleifer, S. (2017). Incentives for Effort or Outputs? A Field Experiment to Improve Student Performance. *Abdul Latif Jameel Poverty Action Lab (J-PAL)*.

Jensen, R. (2010). The (Perceived) Returns to Education and the Demand for Schooling. *The Quarterly Journal of Economics 125*, 515–548.

Jones, S., E. Egger, and R. Santos (2020). Is Mozambique Prepared for a Lockdown During the COVID-19 Pandemic? *UNU-WIDER Blog*.

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.

Kremer, M., E. Miguel, and R. Thornton (2009). Incentives to Learn. *The Review of Economics and Statistics 91*, 437–456.

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.

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.

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.

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 (2019). Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments. *NBER Working Paper*.

Muralidharan, K., A. Singh, and A. J. Ganimian (2019). Disrupting education? Experimental Evidence on Technology-Aided Instruction in India. *American Economic Review 109*, 1426–60.

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.

Nyusi, F. J. (August 5, 2020a). *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.

Nyusi, F. J. (September 5, 2020b). *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.

Republic of Mozambique (April 2, 2020c). *"Bulletin of the Republic", I Series, No. 64*. Maputo, Mozambique.

Republic of Mozambique (August 5, 2020a). *"Bulletin of the Republic", I Series, No. 149*. Maputo, Mozambique.

Republic of Mozambique (March 31, 2020b). *"Bulletin of the Republic", I Series, No. 62*. Maputo, Mozambique.

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.

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

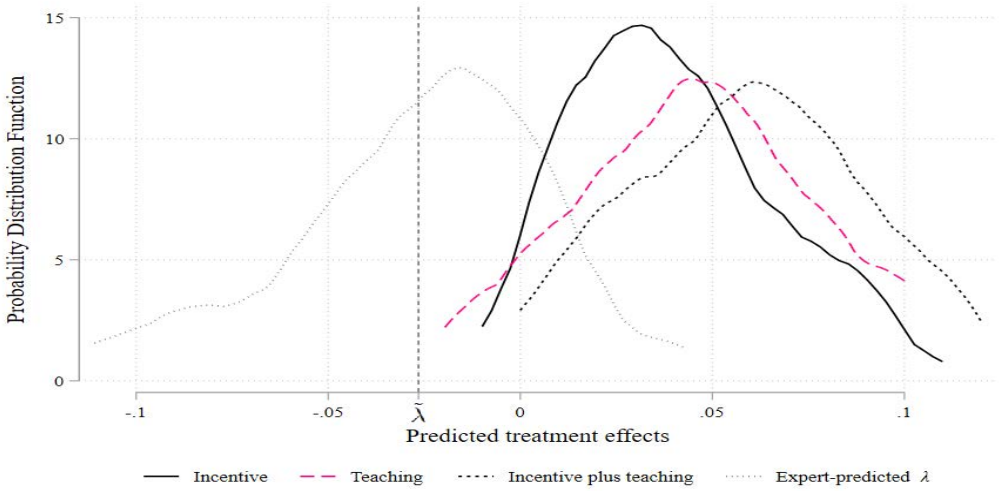
Twinam, T. (2017). Complementarity and Identification. *Econometric Theory 33*, 1154–1185.

U.S Embassy in Mozambique (2020). COVID-19 Information.

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. *NBER Working Paper* (28716).

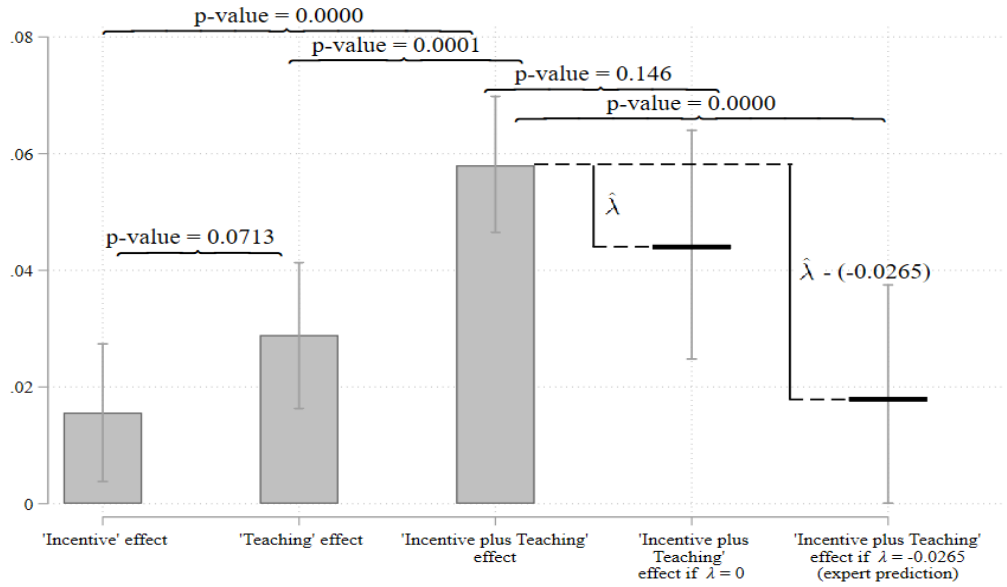
7

Figure 1: **Distributions of Expert Predictions of Treatment Effects and Complementarity Parameter**



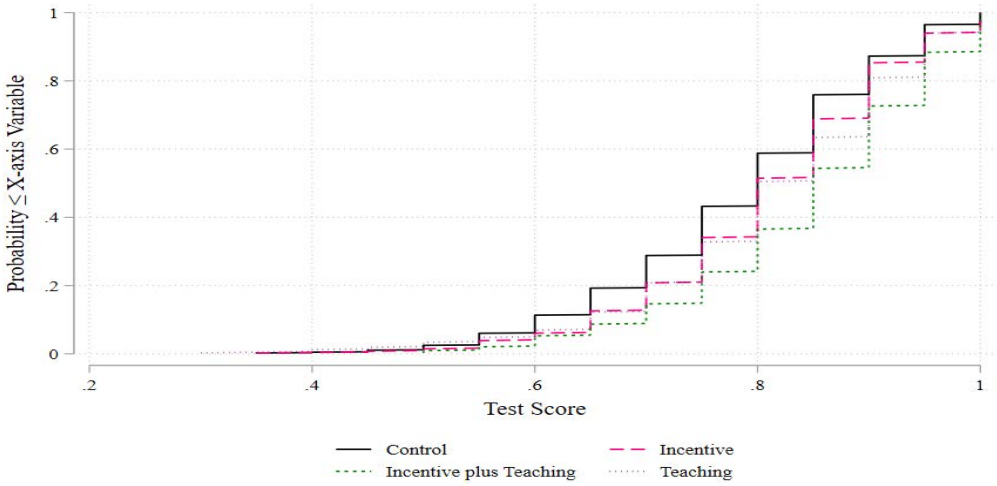
Notes: Probability density functions of predicted treatment effects of 67 experts surveyed prior to results being publicized (survey closing date Jan. 2, 2021). Experts predicted effects of "Incentive", "Teaching", and "Incentive plus Teaching" ("Joint") treatments on COVID-19 knowledge test score (fraction 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: **Treatment Effects and Test of Complementary Parameter λ Against Benchmark Values**



Notes: Dependent variable on y-axis: COVID-19 Knowledge Test Score (fraction of questions answered correctly). 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 3: **Cumulative Distribution Functions of Test Score by Treatment Group**



Notes: Dependent variable is COVID-19 Knowledge Test Score (fraction of questions answered correctly). Figure displays cumulative distribution functions (CDFs) of test scores in "Control", "Incentive", "Teaching", and "Incentive plus Teaching" ("Joint") treatment groups.

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

VARIABLES	(1) COVID-19 knowledge test score	(2) Test score: newly asked
Incentive	0.0156	0.0209
	(0.0060)	(0.0081)
	[0.0133]	
Teaching	0.0288	0.0017
	(0.0064)	(0.0078)
	[0.0003]	
Incentive plus Teaching	0.0581	0.0416
	(0.0060)	(0.0080)
	[0.0003]	
$\hat{\lambda}$	0.0137	0.0189
	(0.0095)	(0.0120)
Observations	2,117	2,117
R-squared	0.333	0.150
Control Mean DV	0.784	0.777
Control SD DV	0.123	0.144
p-value: $\lambda = 0$	0.1460	0.1140
p-value: λ = -0.0265	0.000	0.000
p-value: Incentive = Teaching	0.0713	0.0332
p-value: Incentive = Joint	0.0000	0.0235
p-value: Teaching = Joint	0.0001	0.0000

Notes: Dependent variable in Column 1 is COVID-19 Knowledge Test Score, the share of correct answers to 20 knowledge questions asked at endline that were also randomly selected for the respondent to answer at baseline. Dependent variable in Column 2 is 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 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 column 1 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.

Online Appendix

In this Online Appendix, we adhere to the nomenclature we used in the main text to refer to the treatment conditions. In the main text of this paper, we refer to these treatments, respectively, as "Incentive" and "Teaching". (In the PAP, we refer to the two individual treatments as "Knowledge Incentive" and "Tailored Feedback".)

It is also important to clarify how we refer to the primary outcome variables. In the PAP, we refer to the primary outcome variables as 1) the Knowledge Index (based on 40 questions), and 2) the Feedback-Eligible Knowledge Index (based on 20 questions). In the main text of this paper, we focus on the second of these two, and refer to it for simplicity as the "COVID-19 Knowledge Test Score". Throughout this Appendix, for clarity, we refer to the primary outcomes, respectively, as 1) the Overall Test Score, and 2) the Feedback-Eligible Test Score. In other words, the primary outcome of focus in the main text of the paper, the "COVID-19 Knowledge Test Score", is synonymous with the Feedback-Eligible Test Score.

A Study Area and Timeline

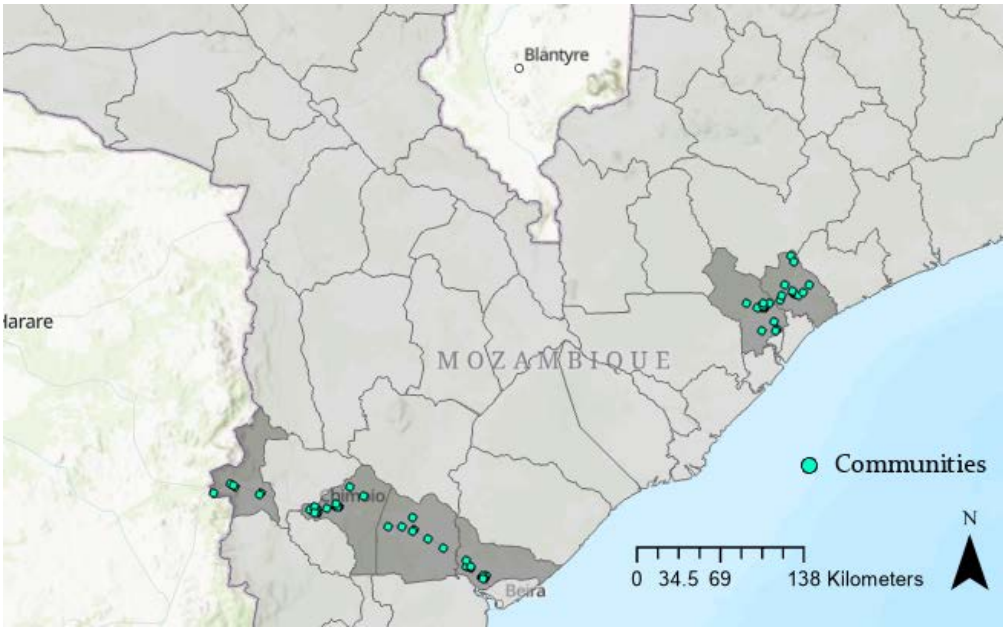
The Mozambican government declared a State of Emergency due to the COVID-19 pandemic on March 31, 2020 (Republic of Mozambique, 3/31/2020). The government recommended social distancing (at least 1.5 meters) and required it at public and private institutions and gatherings. The government also suspended 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, 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).

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 A.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. Appendix Figure A.2 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.

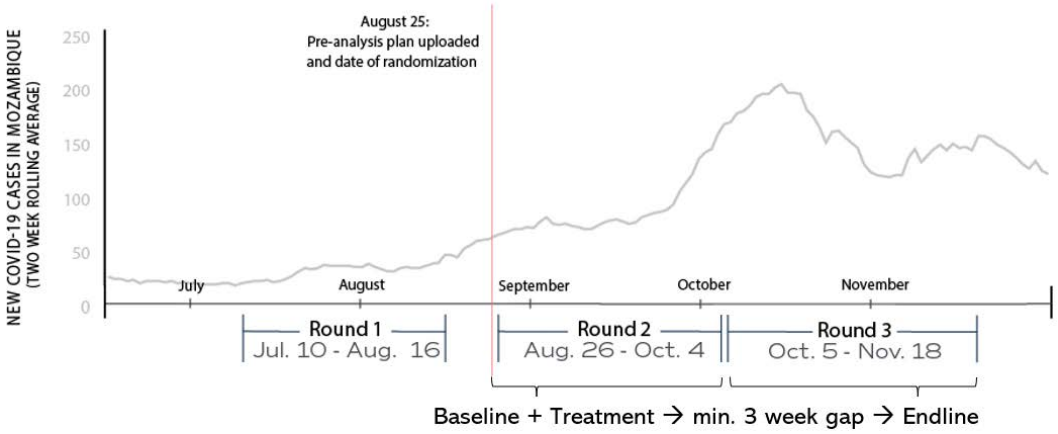
Details on our Round 4 survey to test long-run impacts can be found in Appendix H.

Figure A.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.

Figure A.2: **Study Timeline**



Notes: Round 1 is pre-baseline survey to pilot knowledge questions, 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 A.4 come from Round 2 surveys and endline measures come from Round 3 surveys.

B COVID-19 Knowledge Questions

Survey questions measured COVID-19-related knowledge in the three main subcategories: 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 A.1, A.2, and A.3.

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 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 Feedback-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).

[1] The list of 71 Round 1 pilot questions can be found on our project website: https://fordschool.umich.edu/sites/default/files/2021-06/round1-questions-learing-covid-210614.pdf.

[2] See American Economic Association's RCT Registry, registration ID number AEARCTR-0005862: https://doi.org/10.1257/rct.5862-1.0

Table A.1: **Pre-specified "General Knowledge" Questions and Corresponding 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 A.2: **Pre-specified "Preventive Actions" Questions and Corresponding 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 A.3: **Pre-specified "Government Policy (Actions)" Questions and Corresponding 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 A.4 presents summary statistics in the control group (N=847) of the Overall Test Score and the Feedback-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 Feedback-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 A.4: **Summary Statistics of Test Score (TS) in Control Group**

Outcome	Round	Mean	Std. Dev.	Min	Max
Round 1 Test Score (TS)	Round 1	0.716	0.116	0.25	1
Round 2 Test Score (TS)	Round 2	0.769	0.121	0.35	1
Overall Test Score (TS)	Round 3	0.781	0.108	0.45	1
Feedback-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 Feedback-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.

Details on questions included in our Round 4 survey can be found in Appendix H.

C 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 A.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 three treatments were implemented directly following the Round 2 (baseline) survey, at the end 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 A.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. "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!"

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.

D Attrition and Balance

Table A.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 4.2 subsection 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 A.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.

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

Table A.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 (Excluding Nhamatanda)	(9) Feedback-eligible TS (Excluding 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)	(0.0065)
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)	(0.0074)	(0.0056)	(0.0065)
Incentive plus Teaching (Joint)	0.0120	206.30	0.0724	0.0367	0.0055	0.0016	0.0054	0.0494	0.0573
	(0.0130)	(211.70)	(0.0930)	(0.0282)	(0.0068)	(0.0255)	(0.0071)	(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	0.1110
p-value: Incentive = Teaching								0.5380	0.0794
p-value: Incentive = Joint								0.0000	0.0000
p-value: Teaching = Joint								0.0000	0.0000

Notes: Round 1 baseline variables are defined as follows—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. $\hat{\lambda}$ 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.

E Expert Predictions

Prior to releasing the results of our analyses, we implemented an expert prediction survey to elicit expectations of our treatment effects. We released an English version of the survey on the Social Science Prediction Platform,[4] and circulated an identical Portuguese version of the survey in Mozambique that we designed and distributed on Qualtrics. 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 our 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 Overall Test Score (based on the full set of 40 Round 2 and 3 knowledge questions). For the Teaching and "Incentive plus Teaching" (Joint) treatments, experts were asked to to predict the treatment effect on the "Feedback-Eligible Test Score" (based on the 20 knowledge questions randomly selected at the study respondent level from the full set of 40) that were the subject of the Teaching treatment.

We received expert predictions from 67 survey respondents. 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 A.7 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 Incentive plus Teaching (Joint) would increase test scores by 0.059.

Table A.7: **Expert Predictions**

Expert Prediction	Mean	Std. Dev.	Min	Max
Incentive Treatment Effect	.0399	.0256	0	.1
Teaching Treatment Effect	.0455	.0307	-.0196	.1
Teaching plus Incentive Treatment Effect	.0589	.0296	0	.012
Complementarity parameter (λ)	-.0265	.0333	-.111	.0426
Indicator: Incentive and teaching treatments are substitutes (λ<0)	0.81	0.40	0	1

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

For each expert who provided predictions, we calculate the complementarity parameter implied by their predictions: Predicted Joint Effect − (Predicted Incentive Effect + Predicted Teaching effect). This requires us to assume that the expert-predicted effect of the Incentive treatment on the Overall Test Score (based on all 40 questions) would have been the same if we had asked them to predict the Incentive treatment effect on the Feedback-Eligible Test Score (based on a randomly-selected 20 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

[4]See https://socialscienceprediction.org/ for more information.

of 40 questions.

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 (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 1 in the main text also presents the distributions of these predictions.

F 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 the treatment conditions (i.e., "Incentive" and "Teaching") rather than that used in the PAP (i.e., "Knowledge Incentive" and "Tailored Feedback", respectively). Additionally, while the PAP refers to the primary outcome variables as 1) the Knowledge Index (based on 40 questions), and 2) the Feedback-Eligible Knowledge Index (based on 20 questions), the main text of this paper only focuses on the second of these two, referring to it as the "COVID-19 Knowledge Test Score". Throughout this appendix, we refer to these outcomes, respectively, as 1) the Overall Test Score (or simply Test Score), and 2) the Feedback-Eligible Test Score.

F.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:

$$Y^{all}_{i,j,t=3} = \alpha_0 + \alpha_1 Incentive_{ij} + \alpha_2 Teaching_{ij} + \alpha_3 Joint_{ij} + \eta \mathbf{B}_{ijt} + \gamma_i + \varepsilon_{ij} \tag{F.1}$$

where $Y^{all}_{i,j,t=3}$ 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:

$$Y^{feedback}_{i,j,t=3} = \beta_0 + \beta_1 Incentive_{ij} + \beta_2 Teaching_{ij} + \beta_3 Joint_{ij} + \eta \mathbf{B}_{ijt} + \gamma_i + \varepsilon_{ij} \tag{F.2}$$

where $Y^{feedback}_{i,j,t=3}$ is the Feedback-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 F.1.

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

Results from estimating these equations are in Table A.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 Feedback-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 A.8: **Regression of Test Score (TS) on Treatments**

	(1)	(2)
VARIABLES	Overall Test Score (TS)	Feedback-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)	(0.0059)
		[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: $\lambda = 0$	0.1050	0.1460
p-value: λ = -0.0265	0.0000	0.0000
p-value: Incentive = Teaching	0.5290	0.0713
p-value: Incentive = Joint	0.0000	0.0000
p-value: Teaching = Joint	0.0000	0.0000

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

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

	(1)	(2)
VARIABLES	Overall Test Score (TS)	Feedback-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 Feedback-Eligible TS is the share of correct answers to the 20 knowledge questions in Round 3 that were eligible for the tailored feedback 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 A.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 A.10: **Regression of Test Score (TS) Subcategories on Treatments**

!

VARIABLES	(1) General TS	(2) Preventive TS	(3) Government TS	(4) Feedback-eligible General TS	(5) Feedback-eligible Preventive TS	(6) Feedback-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)	(0.0093)	(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)	(0.0099)
$\hat{\lambda}$	0.0126	0.0178	0.0000	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: $\lambda = 0$	0.3330	0.0759	0.9950	0.3990	0.1707	0.8410
p-value: Incentive = Teaching	0.5360	0.4490	0.0410	0.0354	0.2760	0.3090
p-value: Incentive = Joint	0.0048	0.0000	0.0170	0.0008	0.0000	0.0025
p-value: Teaching = Joint	0.0278	0.0000	0.0000	0.2130	0.0037	0.0001

Notes: The Overall Test Score (TS) subcategories (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 Feedback-Eligible TS subcategories (Columns 4-6) are the share of correct answers to the questions in Round 3 that were eligible for the tailored feedback 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 A.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	0.0068	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	0.2020	0.1120
p-value: Incentive = Joint	0.1400	0.1700	0.5230
p-value: Teaching = Joint	0.1050	0.9050	0.3360

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 A.12: **Regressions of Interactions of Knowledge Treatments and Social Distancing Treatments**

VARIABLES	(1) Overall Test Score (TS)	(2) Feedback-eligible TS	(3) Overall Test Score (TS) without SD Index	(4) Feedback-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 A.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: Feedback-Eligible TS calculated without the 4 Feedback-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 A.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]

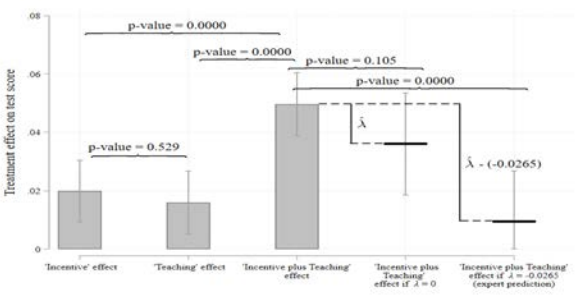
F.2 Additional Figures

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.

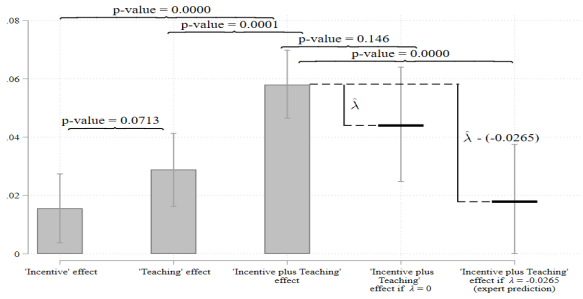
In Figure A.3, we display in Panel (a) treatment effects and the complementarity parameter from analyses of the Overall Test Score based on 40 COVID-19 knowledge questions. The corresponding main text Figure 2 is replicated in Panel (b) for comparison. The key conclusion is stable across the two figures: 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).

Figure A.3: **Treatment Effects and Test of Complementarity Parameter λ**

(a) Overall Test Score



(b) Feedback-Eligible Test Score



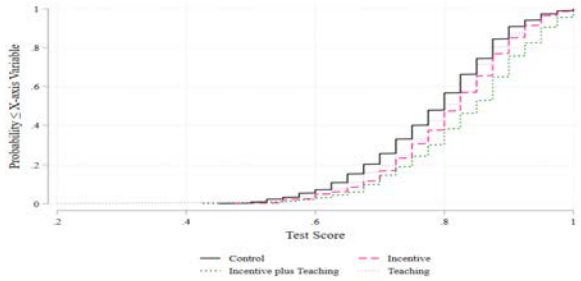
Notes: Overall Test Score is fraction of correct responses on COVID-19 knowledge out of 40 questions. Feedback-Eligible Test Score is a fraction of correct responses on COVID-19 knowledge out of 20 questions previously asked in the Round 2 (baseline) survey. In each panel of figure, bars in first three 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 A.4, we display in Panel (a) CDFs of the Overall Test Score based on 40 COVID-19 knowledge questions. The corresponding main text Figure 3 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.

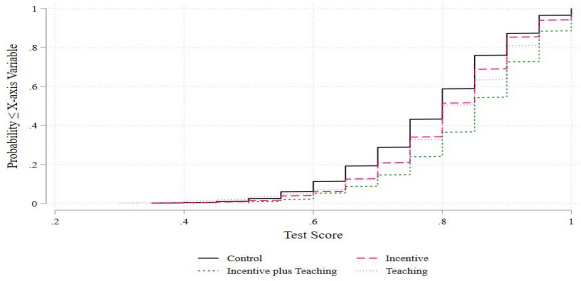
[6]Note these are separate experiments with different pre-specified outcomes of interest. As our primary interest was never to examine interactions between these treatments sets, we do not believe it would be accurate to characterize our results as focusing on the "short model" (a weighted average of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

Figure A.4: **Cumulative Distribution Functions of Test Score by Treatment Group**

(a) Overall Test Score



(b) Feedback-Eligible Test Score



Notes: Overall Test Score is fraction of correct responses on COVID-19 knowledge out of 40 questions. Feedback-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.

G 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). 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:

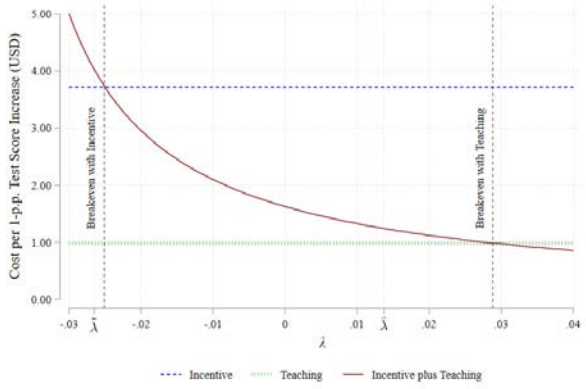
- Incentive treatment: $e_I = 100 * c_I/\beta_1$
- Teaching treatment: $e_T = 100 * c_T/\beta_2$
- Joint treatment: $e_J = 100 * c_J/(\beta_1 + \beta_2 + \lambda)$

[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.

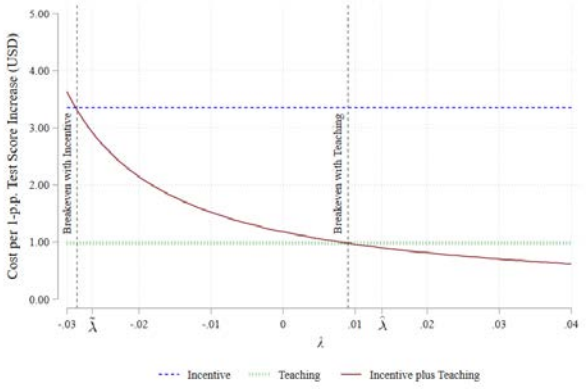
In Figure A.5 panel (a), we display the cost-effectiveness of each treatment, using actual treatment effects for the Incentive and Teaching treatments (β_1 and β_2) and Joint treatment effects implied by a range of values of λ. The cost-effectiveness of the Incentive and Teaching treatments are horizontal, 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.

Figure A.5: **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, Column 1 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 A.5 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 $e_I = 3.35$.

H 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 subcategories: 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: 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 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 A.13 and discuss their relevance to verifying the robustness of the Joint intervention's positive effect and complementarity over time in Section 5.3.

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

VARIABLES	(1) Post-endline test score (TS)	(2) Endline equivalent TS	(3) Post-endline Teaching-eligible TS	(4) Endline equivalent Teaching-eligible TS
Incentive	-0.0104	0.0068	-0.0155	-0.0000
	(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: λ = -0.0265	0.0000	0.0000	0.0000	0.0000
p-value: Incentive = Teaching	0.0026	0.5270	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: Dependent variable in column 1 is COVID-19 Knowledge Test Score in Round 4 (fraction of questions answered correctly). Dependent variable is column 2 if the Round 3 equivalent (fraction of all general knowledge and preventive action questions answered correctly). Dependent variables in columns 3-4 are the 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 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.