Economics of Education Review 91 (2022) 102317

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Teaching and incentives: Substitutes or complements?

James Allen IV[a,b,f,*], Arlete Mahumane[c], James Riddell IV[d], Tanya Rosenblat[e], Dean Yang[a,b,f], Hang Yu[g,h]

[a] Department of Economics, University of Michigan
[b] Ford School of Public Policy, University of Michigan
[c] Beira Operational Research Center, National Institute of Health, Mozambique
[d] Division of Infectious Diseases, University of Michigan Medical School
[e] School of Information and Department of Economics, University of Michigan
[f] Population Studies Center, University of Michigan
[g] National School of Development, Peking University
[h] Institute of South-South Cooperation and Development, Peking University

ARTICLE INFO

JEL Classification:
I10
I21
D90

Keywords:
COVID-19
Teaching
Education
Learning
Cost-effectiveness
Mozambique
Africa

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 via targeted feedback, 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, though the standalone teaching treatment is the most cost-effective. The complementarity between teaching and incentives persists in the longer run, over nine months post-treatment.

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 & 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

* Corresponding author.
E-mail address: alleniv@umich.edu (J. Allen).

https://doi.org/10.1016/j.econedurev.2022.102317
Received 15 April 2022; Received in revised form 4 August 2022; Accepted 21 September 2022
Available online 10 October 2022

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

Abiding by COVID-19 health protocols, we interacted with our 2117 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 2×2 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 (Conn, 2017; Evans & Popova, 2015; Glewwe, 2014; Le, 2015; McEwan, 2015; Muralidharan, 2017).

On the supply side, studies have examined provision of educational supplies (Glewwe, Kremer, Moulin & Zitzewitz, 2000, 2009), school facilities (Duflo, 2001), new teaching technologies (Muralidharan, Romero & Wuthrich, 2019), and "teaching at the right level" (TaRL) (Banerjee & Duflo, 2011; Duflo et al., 2011). Angrist et al., (2020c) 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, Hawkins, Collins, Burchbinder & Osborne, 2016), financial education to promote "financial literacy" (see Kaiser and Menkhoff (2017) for a review), and agricultural "extension" to improve farming knowledge (Anderson & Feder, 2007; Fabregas, Kremer, Lowes, On & Zane, 2019).[1] Our Teaching treatment implements a targeted approach to promote COVID-19 health literacy.

Demand-side educational interventions seek to increase the perceived returns to learning. In school settings, studies have examined impacts of providing information on the wage returns to schooling (Jensen, 2010), merit scholarships based on test performance (Berry, Kim & Son, 2019; Kremer, Miguel & Thornton, 2009), or incentives for test performance (Angrist & Lavy, 2009; Behrman, Parker, Todd & Wolpin, 2015; Fryer, 2011; Levitt, List, Neckermann & Sadoff, 2011; 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, Bilger & Baid, 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, Livingston and Neckermann (2018) study the interactions between test-score incentives for teachers (supply-side) and students (demand-side), but do not estimate a complementarity parameter, as we do.[3] In addition to being of policy interest, we view this interaction as of particular theoretical interest due to the countervailing motivation and crowding-out effects of combining supply- and demand-side educational interventions.

Our study also contributes to understanding adult education in

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

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

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

health crises. Broader research suggests that adults have higher economic and physiological barriers to learning (Aker & Sawyer, 2021), and that successful health informational interventions are comprehensive but not overly complex (Dupas et al., 2011). Additional challenges in health crises often include underlying institutional mistrust and misinformation (Vinck, Pham, Bindu, Bedford & Nilles, 2019) and logistical obstacles to needs assessments and outreach with vulnerable populations (Checchi et al., 2017). In this context, we demonstrate simple interventions that can complement phone data collection during epidemics (Angrist et al., 2020; Maffioli, 2020; Magaço et al., 2021). In particular, our Teaching intervention shows that providing feedback on knowledge-based questions is a feasible and impactful add-on to health surveys—for example, on "knowledge, attitudes, and practices (KAP)" surveys common in public health.[4]

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

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 an *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 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

We 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:

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

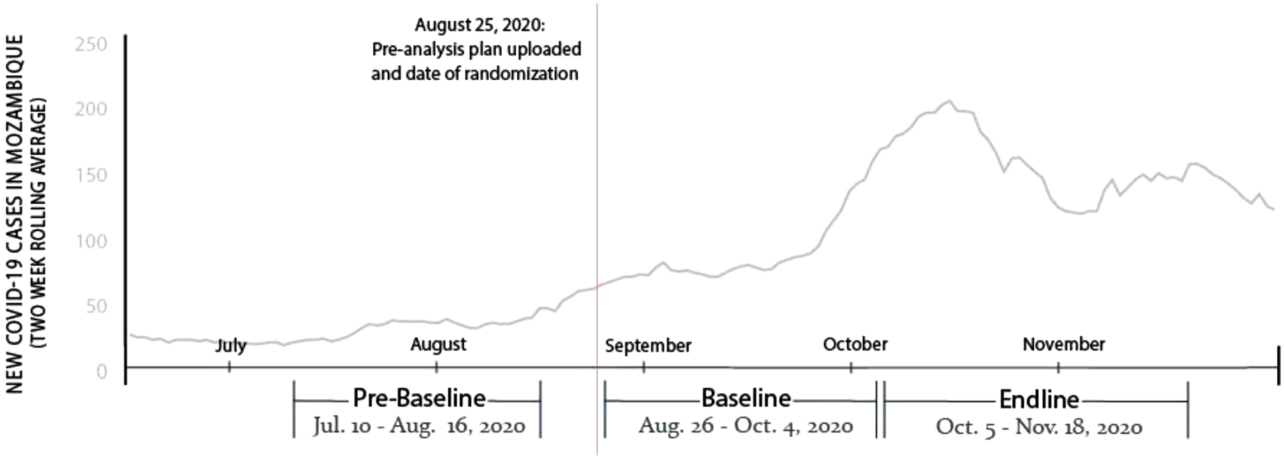

Fig. 1. Study Timeline
Notes: Pre-analysis plan uploaded and treatments randomly assigned immediately prior to start of baseline survey, on Aug. 25, 2020. Treatments implemented immediately following baseline survey on same phone call. There was at least a three week gap between baseline and endline survey for any given study participant. Not depicted is the post-endline survey implemented between June 30 and August 30, 2021 that we use in the long-run analysis described in Section 5.4.

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

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

3. Sample and data

3.1. Data

We implemented three rounds of surveys by phone in July-November 2020: a pre-baseline, baseline and endline survey (see Fig. 1 for a study timeline). Respondents were from households with phones in the sample of a prior study (Yang, Allen, Mahumane, Riddell, & Yu, 2023).[5] We surveyed one adult per household. Participants received a small gift of 50 meticais (approx. US$0.70) after completing each survey, as explained at study enrollment, which was transferred via MPesa over 93% of the time and phone credit recharge otherwise. Appendix A provides details on the COVID-19 context and study communities.

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

The endline sample size is 2117 respondents, following a sample size of 2226 at baseline. The retention rate between baseline and endline is 95.1% overall, at least 94.4% in each of the seven districts surveyed, and balanced across treatment conditions.

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

3.2. Outcomes

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

We pre-specified two primary outcomes: First, the Overall test score is the share of correct answers to all 40 knowledge questions asked at endline: 12 on general knowledge, 16 on preventive actions, and 12 on

[5] AEA RCT Registry for Yang, Allen, Mahumane, Riddell, & Yu, 2023: https://doi.org/10.1257/rct.3990-5.1

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

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

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

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%). Randomization was stratified within 76 communities. 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$2.80), if they answered all 40 questions correctly, in addition to their 50 meticais survey completion gift. 250 meticais is equivalent to half of the sample median pre-pandemic (February 2020) weekly household income. After endline questioning, the number of correct answers and resulting payment were automatically calculated in SurveyCTO, displayed for enumerators, read to respondents, and added to the 50 meticais survey completion gift.

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

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

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

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

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

4.3. Hypotheses

We hypothesize that each treatment has a positive effect on test scores. Specifically, as prespecified, we hypothesize that the coefficient β_1 in a regression of the Overall test score, and the coefficients β_2 and β_3 in a regression of the Teaching-Eligible test score will be positive. We adjust p-values for multiple hypothesis testing across these three coefficients.[10] Additionally, using our estimated $\widehat{\lambda}$, we test the following null hypotheses: $\lambda = -0.0265$ (the mean of expert predictions, $\widetilde{\lambda}$), 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 E, we present the "Populated PAP" for our pre-specified primary analysis. These results are substantively duplicative of and yield very similar conclusions to the primary analyses we present here in the main text.

Hypotheses related to the complementarity parameter λ were not pre-specified in the PAP. The motivations for testing them are the theoretical model's ambiguous prediction as to whether λ should be positive or negative, and the fact that the vast majority of experts predicted that $\lambda < 0$.

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

[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."

[9] The average respondent correctly answered 72.1% and 77.3% of the 20 knowledge questions at prebaseline 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).

Table 2
Expert Predictions.

Expert Prediction	Mean	Std. Dev.	Min	Max
Incentive Treatment Effect	.0399	.0256	0	.1
Teaching Treatment Effect	.0455	.0307	−0.0196	.1
Joint Treatment Effect	.0589	.0296	0	.1200
Complementarity parameter (λ)	−0.0265	.0333	−0.111	.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.

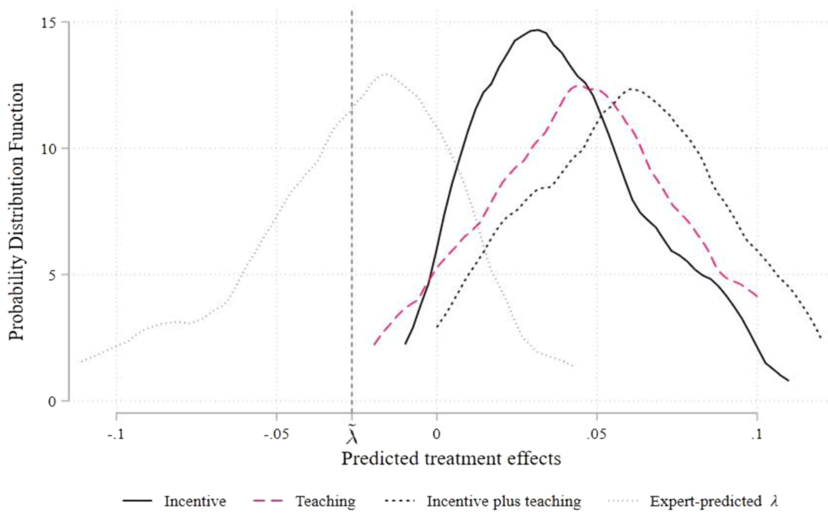

Fig. 2. Distributions of Expert Predictions of Treatment Effects and Complementarity Parameter
Notes: Probability density functions of predicted treatment effects of 67 experts surveyed prior to results being publicized (survey closing date Jan. 2, 2021). Experts predicted effects of "Incentive", "Teaching",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

4.5. *Expert Predictions*

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

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

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

Table 2 summarizes the expert predictions. To be consistent with the figures and tables in this paper, we display the predictions as fractions (bounded by 0 and 1) rather than percentage points. On average, respondents expected that Incentive would increase the test scores by 0.040, Teaching would increase test scores by 0.046, and Joint would increase test scores by 0.059.

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

Fig. 2 displays probability density functions (PDFs) of the predictions. For each treatment, the vast majority of experts predicted positive effects. The mean Incentive treatment effect (β_1) is 0.040, while for Teaching (β_2) it is 0.046. Notably, the mean predicted effect for the Joint treatment (β_2) is 0.059, lower than the sum of the mean predictions for the separate Incentive and Teaching treatments (0.086): experts expect the treatments to be substitutes rather than complements.

Graphically, the expectation of substitutability can be seen in the fact that the PDF of the Joint treatment has considerable overlap with the PDFs of Incentive and Teaching. Relatedly, in the figure we also display the complementarity parameter implied by each expert's predictions. For each expert, we take their predicted Joint treatment effect and

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

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

subtract the sum of their predictions for the separate Incentive and Teaching treatments. The distribution of experts' λ estimates is the gray dotted line. Most of the mass of λ estimates lies to the left of zero: 81% of experts predicted negative λ. The mean of experts' λ estimates is -0.0265. We refer to this mean as $\tilde{\lambda}$, and will test the null that our estimated $\hat{\lambda}$ equals $\tilde{\lambda}$.

5. Results

5.1. Primary Analysis

Table 3 presents the results from testing this paper's primary hypotheses. In Column 1, we test our first pre-specified primary hypothesis regarding the effect of the Incentive treatment on the overall test score.[14] The Incentive treatment has a positive effect, and is statistically significantly different from zero (p-val=0.0003) after multiple hypothesis testing (MHT) adjustment. The point estimate indicates a 0.020 increase, relative to the 0.781 mean control group test score. This effect is substantial in magnitude, amounting to 0.19 standard deviations of the outcome variable.

In Column 2, we test our remaining pre-specified primary hypotheses on the effect of the Teaching treatment and Joint treatment on the Teaching-Eligible test score.[15] Coefficient estimates in Column 2 indicate that the Teaching and Joint treatments each also have positive effects. The point estimate on Teaching indicates a 0.0288 increase (0.23 standard deviations of the outcome variable), while the Joint treatment causes a 0.0581 increase (0.47 standard deviations). Each of these coefficient estimates is statistically significantly different from zero (p-val=0.0003 for each) after MHT adjustment.

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

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

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

Table 3
Treatment Effects on COVID-19 Knowledge Test Scores.

VARIABLES	Overall (1)	Teaching-Eligible (2)	Teaching-Ineligible (3)	Newly-asked (4)
Incentive	0.0200***	0.0156***	0.0244***	0.0209***
	(0.0054)	(0.0060)	(0.0069)	(0.0081)
	[0.0003]			
Teaching	0.0160***	0.0288***	0.0032	0.0017
	(0.0055)	(0.0064)	(0.0069)	(0.0078)
		[0.0003]		
Incentive plus Teaching	0.0496***	0.0581***	0.0410***	0.0416***
	(0.0055)	(0.0060)	(0.0069)	(0.0080)
		[0.0003]		
$\hat{\lambda}$	0.0136	0.0137	0.0134	0.0189
	(0.0084)	(0.0095)	(0.0104)	(0.0120)
Observations	2117	2117	2117	2117
R-squared	0.319	0.333	0.201	0.150
Control Mean DV	0.781	0.784	0.778	0.777
Control SD DV	0.108	0.123	0.125	0.144
p-value: $\lambda = 0$	0.1048	0.1462	0.1956	0.1145
p-value: $\lambda = -0.0265$	0.0000	0.0000	0.0001	0.0002
p-value: Incentive = Teaching	0.5290	0.0713	0.0069	0.0332
p-value: Incentive = Joint	0.0000	0.0000	0.0351	0.0235
p-value: Teaching = Joint	0.0000	0.0001	0.0000	0.0000

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

We also present these results on the Teaching-Eligible test score in Column 2 graphically. In Fig. 3, we display the estimates of the three treatment effects, Joint treatment effects implied if λ took on the values of 0 or -0.0265, and p-values of relevant tests of pairwise differences. In Fig. 4, we present cumulative distribution functions of test scores by treatment group, showing that the Joint treatment leads to the largest rightward shift of the test score distribution.

In sum, our estimates of the complementarity parameter indicate that the Incentive and Teaching treatments exhibit much more complementarity than experts predicted. We strongly reject the high degree of substitutability predicted by experts. In addition, we reject at a marginal level of statistical significance that $\lambda = 0$.

This complementarity is also present when evaluating treatment effects on newly asked questions, building confidence that results are driven by actual learning and not merely rote memorization or experimenter demand effects. In Column 4 of Table 3, we run regression 3 pre-specified in our PAP as of secondary interest that replaces the outcome with the share of correct answers to endline knowledge questions that were NOT randomly asked of the respondent at either pre-baseline or baseline. Thus respondents were not previously told the answers to these questions as part of the Teaching intervention, making it less obvious

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

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

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

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

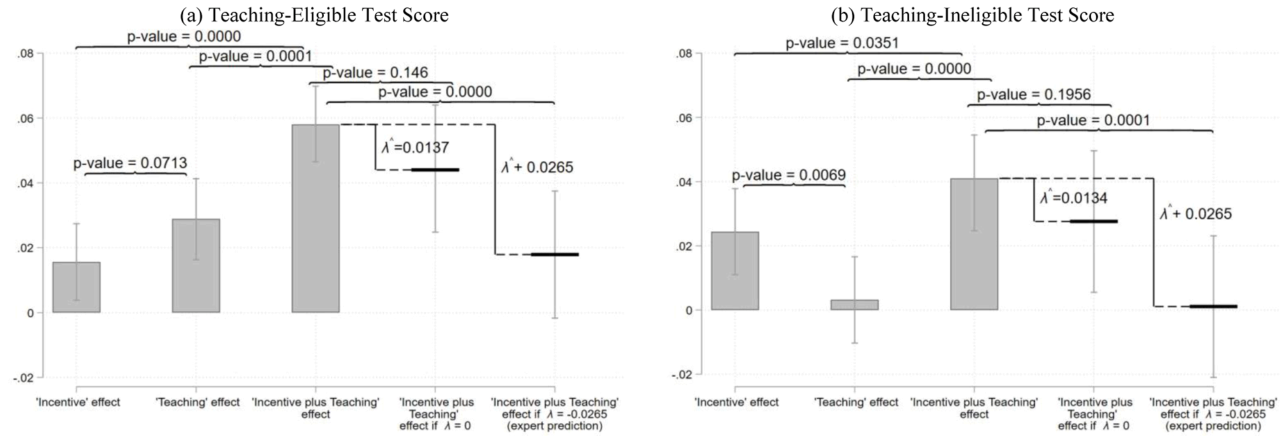

Fig. 3. Treatment Effects and Test of Complementarity Parameter λ Against Benchmark Values
Notes: Panel (a) dependent variable on y-axis is the Teaching-Eligible test score (share of correct answers to knowledge questions asked at baseline and hence eligible for all treatments). Panel (b) dependent variable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

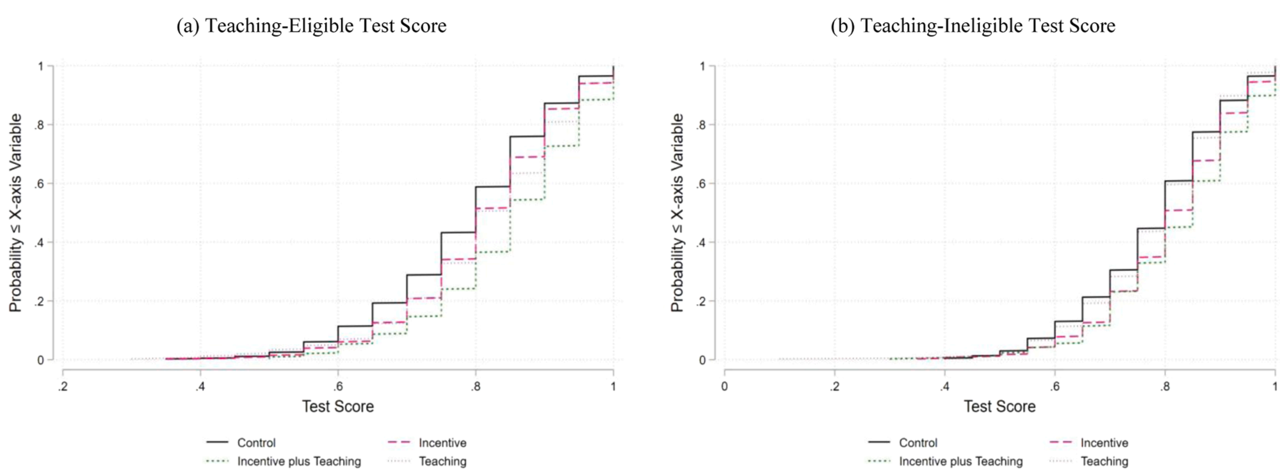

Fig. 4. Cumulative Distribution Functions of Test Score by Treatment Group
Notes: Panel (a) dependent variable on y-axis is the Teaching-Eligible test score (share of correct answers to knowledge questions asked at baseline and hence eligible for all treatments). Panel (b) dependent variable is Teaching-Ineligible test score (share of correct answers to knowledge questions NOT asked at baseline and hence NOT eligible for the Teaching intervention). Figure displays cumulative distribution functions (CDFs) of test scores in "Control", "Incentive", "Teaching", and "Incentive plus Teaching" ("Joint") treatment groups.

what the experimenters "wanted to hear". Both the Incentive and Joint treatments have a positive effect on the newly-asked test score (statistically significant at 1% level). Additionally, we continue to reject that $\lambda = -0.0265$ (the expert prediction) at the 1% level and $\lambda = 0$ at a marginal level of statistical significance.

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 F. 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 Fig. A.4. The cost-effectiveness of the Incentive and Teaching treatments are horizontal, because they do not depend on λ. The cost-effectiveness of the Joint treatment is a decreasing function of λ: the greater the complementarity of the two treatments, the more cost-effective is the Joint treatment.

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

0.0290 for the joint treatment to be the most cost-effective of the three treatment combinations.

For reference, we also show the mean expert prediction, $\widetilde{\lambda} = -0.0265$, and our empirical estimate, $\widehat{\lambda} = 0.0137$. At $\widehat{\lambda}$, Joint is more cost-effective than Incentive, but not as costeffective 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 F we provide an example of alternative relative implementation costs that would lead Joint to be the most cost-effective at $\widehat{\lambda}$.

5.3. Knowledge Categories

We also estimate impacts of the treatments on Teaching-Eligible and Teaching-Ineligible test scores across the knowledge categories: general knowledge, preventive actions, and government policies. Results in Table 4 are broadly similar to the estimates in Table 3 Columns 2 and 3, though treatment effects for the Incentive and Teaching interventions are heterogeneous along different dimensions.

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

Results for the Teaching and Joint treatment vary less across knowledge category and more so between Teaching-Eligible and Teaching-Ineligible test scores. The Teaching treatment has a significantly positive effect on all knowledge categories for Teaching-Eligible questions, but insignificant effects otherwise. The Joint treatment remains significantly positive across all regressions. The estimated complementarity parameter $\widehat{\lambda}$ appears largest (most positive) for the preventive actions subcategory (Table 4, Columns 2 and 5).

5.4. Long-Run Analysis

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

In a post-endline phone survey from July-August 2021, we asked 1886 respondents (89.1% retention from endline, balanced across treatment conditions) 20 pre-specified questions on general knowledge and preventive actions. We excluded government policy questions because many pre-specified questions/answers were no longer true or applicable. Respondents received the standard 50 meticais survey completion gift but were offered no other incentives. We compare endline and post-endline treatment effects on two modified Test Scores of questions assessing general knowledge and preventive actions: 1) Test Score for all relevant questions asked in each round, and 2) Test Score for the same set of relevant questions across baseline, endline, and post-endline. For robustness, we analyze both outcomes, noting that each deviate from our pre-specified primary outcome due to the exclusion of government policy questions, and only draw conclusions supported by all regression specifications.

Results are in Table 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

Table 4
Regression of Test Score (TS) Categories on Treatments.

VARIABLES	Teaching-Eligible Test Scores General (1)	Preventive (2)	Government (3)	Teaching-Ineligible Test Scores General (4)	Preventive (5)	Government (6)
Incentive	0.0018	0.0118	0.0419***	0.0174	0.0249***	0.0422***
	(0.0099)	(0.0088)	(0.0099)	(0.0112)	(0.0090)	(0.0110)
Teaching	0.0265***	0.0234**	0.0299***	0.0044	0.0016	0.0146
	(0.0102)	(0.0093)	(0.0109)	(0.0111)	(0.0095)	(0.0111)
Incentive plus Teaching	0.0415***	0.0535***	0.0749***	0.0336***	0.0439***	0.0538***
	(0.0103)	(0.0087)	(0.0010)	(0.0106)	(0.0094)	(0.0108)
$\widehat{\lambda}$	0.0133	0.0183	0.0031	0.0118	0.0173	−0.0030
	(0.0157)	(0.0136)	(0.0154)	(0.0166)	(0.0141)	(0.0165)
Observations	2117	2117	2117	2117	2117	2117
R-squared	0.206	0.257	0.189	0.117	0.080	0.139
Control Mean DV	0.797	0.827	0.789	0.782	0.710	0.790
Control SD DV	0.189	0.170	0.202	0.191	0.157	0.202
p-value: Incentive = Teaching	0.0354	0.276	0.309	0.313	0.0289	0.0268
p-value: Incentive = Joint	0.000845	3.64e-05	0.00254	0.193	0.0732	0.344
p-value: Teaching = Joint	0.213	0.00365	0.000135	0.0182	0.000110	0.00143

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

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

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

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

Author statement

We have no conflicts of interest to report.

Acknowledgements

Faustino Lessitala provided top-notch leadership and field management. Patricia Freitag, Ryan McWay, and Maggie Barnard provided excellent research assistance. Julie Esch, Laura Kaminski, and Lauren Tingwall's grant management was world-class. We appreciate feedback from Hoyt Bleakley, Brian Jacob, Laston Manja, Kwasi Tabiri, and participants in Michigan's Health, History, Development, and Demography (H2D2) Seminar. This work is supported by the Abdul Latif Jameel Poverty Action Lab (J-PAL) Innovation in Government Initiative through a grant from The Effective Altruism Global Health and Development Fund (award no. IGI-1366); the UK Foreign, Commonwealth & Development Office awarded through Innovations for Poverty Action (IPA) Peace & Recovery Program (award no. MIT0019-X9); the Michigan Institute for Teaching and Research in Economics via the Ulmer Fund (award no. G024289); Dubai Cares through the Evidence for Education in Emergencies (E-Cubed) Research Envelope; the United States Agency for International Development (USAID) awarded through the Feed the Future Innovation Lab for Markets, Risk and Resilience (MRR) Innovation Lab (award no. A20–1825-S007); and the National Institutes of Health Eunice Kennedy Shriver National Institute of Child Health & Human Development (award no. 1-R01-HD102382–01A1) and National Institute on Aging (award no. T32AG000221). The content is solely the responsibility of the authors and does not necessarily represent the official views of our funding organizations. Our study protocols were approved by Institutional Review Boards (IRBs) at the University of Michigan (Health Sciences and Social and Behavioral Sciences IRB, approval number HUM00113011) and the Mozambique Ministry of Health National Committee on Bioethics for Health (Portuguese acronym CNBS, reference number 302/CNBS/20). The study was submitted to the American Economic Association's RCT Registry on August 25, 2020, registration ID number AEARCTR-0,005,862: 10.1257/rct.5862–3.0.

Online Appendix

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

AStudy Area

The Mozambican government declared a State of Emergency due to the COVID-19 pandemic on March 31, 2020 (Republic of Mozambique 2020a). The government recommended social distancing (at least 1.5 m) 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 2020b). The government stopped short of implementing a full economic "lockdown" due to its economic costs (Jones, Egger & Santos, 2020; Siuta & Sambo, 2020). On August 5, 2020, the government renewed the State of Emergency (Republic of Mozambique 2020c), called for improved mask-wearing, and announced a schedule for loosening restrictions (Nyusi, 2020a). In September 2020, the government loosened some restrictions, including resuming religious services at 50% capacity (Nyusi, 2020b; U.S Embassy in Mozambique 2020).

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

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

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

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 (postendline) 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 Eq. (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 5.4.

References

Aker, J.C., .& Sawyer,.M. (2021). Making sense of the signs: What do we know about learning in adulthood? Working Paper.

Allen IV, J., Mahumane, A., Riddell IV, J., Rosenblat, T., Yang, D., & Yu, H. (2021). *Correcting perceived social distancing norms to combat COVID-19*. NBER Working Paper (28651).

Alsan, M., Cody Stanford, F., Banerjee, A., Breza, E., Chandrasekhar, A. G., Eichmeyer, S., et al. (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., & Feder, G. (2007). Agricultural Extension. *Handbook of Agricultural Economics, 3*, 2343–2378.

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

Angrist, N., Bergman, P., Evans, D. K., Hares, S., Jukes, M. C. H., & Letsomo, T. (2020). Practical lessons for phone-based assessments of learning. *BMJ Global Health, 5*.

Angrist, N., Bergman,.P., & Matsheng,.M. (2020). School's Out: Experimental Evidence on Limiting Learning Loss Using "Low-Tech" in a Pandemic. NBER Working Paper (28205).

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

Banerjee, A. V., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. New York, United States: Public Affairs.

Banerjee, A., Alsan,.M., Breza,.E., Chandrasekhar, A.G., Chowdhury,.A., Duflo,.E.… Olken, B.A. (2020). Messages on Covid-19 Prevention in India Increased Symptoms Reporting and Adherence to Preventive Behaviors Among 25 Million Recipients with Similar Effects on Non-recipient Members of Their Communities. NBER Working Paper (27496).

Banerjee, A., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from Two Randomized experiments in India. *Quarterly Journal of Economics, 122*, 1235–1264.

Barsbai, T., Licuanan,.V., Steinmayr,.A., Tiongson,.E., & Yang,.D. (2020). Information and the Acquisition of Social Network Connections. NBER Working Paper (27346).

Batterham, R. W., Hawkins, M., Collins, P. A., Burchbinder, R., & Osborne, R. H. (2016). Health Literacy: Applying Current Concepts to Improve Health Services and Reduce Health Inequalities. *Public health, 132*, 3–12.

Behrman, J. R., Parker, S. W., Todd, P. E., & Wolpin, K. I. (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., Kim, H., & Son, H. (2019). *When student incentives don't work: Evidence from a field experiment in malawi* (pp. 1–54).

Bjorkman Nyqvist, M., Corno, L., de Walque, D., & Svensson, J. (2018). Incentivizing Safer Sexual Behavior: Evidence from a Lottery Experiment on HIV Prevention. *American Economic Journal: Applied Economics, 10*, 287–314.

Burgess, S., Metcalfe, R., & Sadoff, S. (2016, October). *Understanding the response to financial and nonfinancial incentives in education: Field experimental evidence using high-stakes assessments*. IZA Institute of Labor Economics.

Carpena, F., Cole, S., Shapiro, J., & Zia, B. (2017). The ABCs of Financial Education: Experimental Evidence on Attitudes, Behavior, and Cognitive Biases. *Management Science, 65*, 346–369.

Checchi, F., Warsame, A., Treacy-Wong, V., Polonsky, J., Van Ommeren, M., & Prudhon, C. (2017). Public health information in crisis-affected populations: A review of methods and their use for advocacy and action. *The Lancet, 390*(10109), 2297–2313.

Conn, K. M. (2017). Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations. *Review of Educational Research, 87*, 863–898.

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., Banerjee,.A., Finkelstein,.A., Katz,.L., Olken,.B., & Sautmann,.A. (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., Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review, 101*, 1739–1774.

Dupas, P., et al. (2011). Health behavior in developing countries. *Annual Review of Economics, 3*(1), 425–449.

Evans, D. K., & Popova, A. (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*, 242–270.

Fabregas, R., Kremer, M., Lowes, M., On, R., & Zane, G. (2019). *SMS-extension and farmer behavior: Lessons from six RCTs in east africa*. ATAI Research Publications.

Finkelstein, E. A., Bilger, M., & Baid, D. (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., .Devi,.T., & Holden, R.T. (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. *Education policy in developing countries*. Chicago, USA: University of Chicago Press.

Glewwe, P., Kremer, M., & Moulin, S. (2009). Many Children Left Behind? Textbooks and Test Scores in Kenya. *American Economic Journal, 1*, 112–135.

Glewwe, P., Kremer, M., Moulin, S., & Zitzewitz, E. (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., Egger, E., & Santos, R. (2020). *Is mozambique prepared for a lockdown during the COVID-19 pandemic?* UNU-WIDER Blog.

Kaiser, T., & Menkhoff, L. (2017). Does Financial Education Impact Financial Literacy and Financial Behavior, and if so, When? *World Bank Economic Review, 31*, 611–630.

Kremer, M., Miguel, E., & Thornton, R. (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., List, J. A., Neckermann, S., & Sadoff, S. (2011). *The impact of short-term incentives on student performance*. University of Chicago.

Li, T., Han, L., Zhang, L., & Rozelle, S. (2014). Encouraging Classroom Peer Interactions: Evidence from Chinese Migrant Schools. *Journal of Public Economics, 111*, 29–45.

List, J. A., Livingston, J. A., & Neckermann, S. (2018). Do Financial Incentives Crowd Out Intrinsic Motivation to Perform on Standardized Tests? *Economics of Education Review, 66*, 125–136.

List, J., Shaikh, A., & Xu, Y. (2019). Multiple Hypothesis Testing in Experimental Economics. *Experimental Economics, 22*, 773–793.

Maffioli, E. (2020). Collecting data during an epidemic: A novel mobile phone research method. *Journal of International Development, 32*, 1231–1255.

Magaço, A., Munguambe, K., Nhacolo, A., Ambrósio, C., Nhacolo, F., Cossa, S., et al. (2021). Challenges and needs for social behavioural research and community engagement activities during the covid-19 pandemic in rural mozambique. *Global Public Health, 16*(1), 153–157.

Maude, R. R., Jongdeepaisal, M., Skuntaniyom, S., Muntajit, T., Blacksell, S. D., Khuenpetch, W., et al. (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., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (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., Loyalka, P., Fairlie, R., Bhuradia, A., Angrish, M., Lin, J., et al. (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., Romero,.M., & Wuthrich,.K. (2019). Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments. NBER Working Paper.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019b). Disrupting education? Experimental Evidence on Technology-Aided Instruction in India. *American Economic Review, 109*, 1426–1460.

Nglazi, M. D., Van Schaik, N., Kranzer, K., MRCP(UK), Lawn, S. D., Wood, R., & Bekker, L. (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. (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. (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*. September 5. Maputo, Mozambique: Maputo Mozambique.

Puspitasari, I. M., Yusuf, L., Sinuraya, R. K., Abdulah, R., & Koyama, H. (2020). Knowledge, attitude, and practice during the covid-19 pandemic: A review. *Journal of multidisciplinary healthcare, 13*, 727.

Republic of Mozambique. (2020a). *"Bulletin of the republic", i series, no. 149*. Maputo, Mozambique.

Republic of Mozambique. (2020b). *"Bulletin of the republic", i series, no. 62*. Maputo, Mozambique.

Republic of Mozambique. (2020c). *"Bulletin of the republic", i series, no. 64*. Maputo, Mozambique.

Siuta, M., & Sambo, M. (2020). *COVID-19 em mocambique: Dimensao e possiveis impactos. boletim no. 124p*. April 1. 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.

Vinck, P., Pham, P. N., Bindu, K. K., Bedford, J., & Nilles, E. J. (2019). Institutional trust and misinformation in the response to the 2018–19 ebola outbreak in north kivu, dr congo: A population-based survey. *The Lancet Infectious Diseases, 19*(5), 529–536.

Yang, D., Allen, . J., Mahumane, . A., Riddell, . J., & Yu, . H. (2023). Knowledge, Stigma and HIV Testing: An Analysis of a Widespread HIV/AIDS Program. *Journal of Development Economics, 160*, 102958.