The Efficacy of Crowdsourced Nudges: Experimental Evidence

Nicholas G. Otis*
UC Berkeley

Abstract

In a series of large-scale experiments, I test the efficacy of a novel procedure for crowdsourcing behavioral science interventions. Participants ($n = 1,822$) generated nudges to encourage recipients to opt-in to receive notifications about the spread of COVID-19. Separate participants ($n = 1,360$) provided 324,160 incentive-compatible predictions of the effects of these nudges, and I selected the ten interventions with the largest anticipated effects for further testing. In two randomized field experiments I separately evaluate the effects of (1) all 1,822 nudges ($n = 40,911$) and (2) the ten *crowd-selected* nudges ($n = 34,844$). While the average nudge reduced uptake of the COVID-19 notification service by 19.6% relative to control (0.21 pp; p =0.33), the ten crowd-selected nudges increased adoption by 41.0% (0.34 pp; p=0.01), or 35.0% (0.23 pp; p=0.07) compared to benchmark nudges. These results demonstrate the efficacy of crowdsourced behavioral science interventions: local participants can produce effective nudges, and incentivized forecasts provide a mechanism to harness the wisdom of local crowds to identify these interventions.

*notis@berkeley.edu. I gratefully acknowledge financial support from Innovations for Poverty Action, the Fetzer Franklin Fund, GiveWell, and the Global Priorities Institute. I thank Ned Augenblick, Stefano DellaVigna, Lee Fleming, David Holtz, Supreet Kaur, and Dmitry Taubinsky for helpful comments. Chazel Hakim, Zimai Lan, and many others provided excellent research assistance, and Chaning Jang, Irene Ngina, and Pauline Wanjeri provided outstanding management of field operations. This project was approved by the University of California, Berkeley Office for Protection of Human Subjects.

1 Introduction

Researchers and policymakers increasingly rely on light-touch nudge interventions to change behavior (DellaVigna and Linos, 2022). The growing importance of these interventions is highlighted by their widespread use in attempts to control the COVID-19 pandemic (Bavel et al., 2020; Campos-Mercade et al., 2021; Dai et al., 2021; Rabb et al., 2022). Selecting interventions to evaluate or scale, however, remains difficult. For example, decision makers must navigate a literature that overstates the true effect of interventions due to publication bias (DellaVigna and Linos, 2022; Maier et al., 2022), and discern the extent to which interventions tested in different settings and with different outcomes generalize to their context (Bryan et al., 2021; Szaszi et al., 2022).

An alternative to this top-down approach, which relies on expert judgment, is to harness local knowledge about which interventions will be effective. Across many domains research has found that bottom-up information can be remarkably informative (e.g., in targeting funding (Iyer et al., 2016; Mollick and Nanda, 2016; Hussam et al., 2022), developing scientific innovations (Sobel, 2005; Bennett et al., 2007), and predicting research results (Dreber et al., 2015; DellaVigna and Pope, 2018)). If the efficacy of nudges is context-dependent, insights from individuals with local contextual knowledge may be especially informative.

In this paper I propose and test a novel procedure for *crowdsourcing* nudge interventions. The intuition behind this procedure is that individuals similar to nudge recipients possess rich information about the types of interventions that will be effective in their context, but that this decentralized information is noisy and requires the "wisdom of crowds" to extract promising candidate interventions.

Concretely, I crowdsourced interventions to increase opt-in to a text-message based COVID-19 notification service that provides updates from the Kenyan Ministry of Health on the positivity rate, deaths, and the distribution of cases across the country. Variations in opt-in rates between messages are a result of simple changes to communication materials, which is one of the most common classes of nudges (DellaVigna and Linos, 2022). Like many low-income countries, Kenya lacks a strong behavioral science evidence base. This means that the crowdsourcing framework will be tested in a setting where the need for effective interventions is especially high.

I evaluate the efficacy of crowdsourced nudges in a series of large-scale field experiments with 75,755 participants, which took place in four stages (Figure 1 provides an overview). First, I recruited 1,822 Kenyan participants using Facebook advertisements.

Each participant created an intervention to increase adoption of the COVID-19 notification service. I randomly assigned these participants to one of three different incentive schemes that provided bonus payments based on the efficacy of their nudges (Gibbs et al., 2017; Charness and Grieco, 2019). The nudges created by these local participants are diverse, spanning topics ranging from collective goals (e.g., "*Come join us in the fight against coronavirus.*"), to highlighting the risks of COVID-19 (e.g., "*Corona is real and it kills.*"), and all messages were accompanied by the same invitation to join the COVID-19 notification service. Panel A of Table A1 provides a random sample of example nudges developed by participants.

Second, I develop a *crowd-selection* procedure which leverages the wisdom of crowds to reduce noise in the crowdsourcing process. Motivated by recent literature demonstrating that the average predicted effects of interventions contain substantial information on their relative efficacy (DellaVigna and Pope, 2018; Thomas et al., 2020; Otis, 2022), I collect forecasts of the causal effects of crowdsourced nudges from Study 1. 1,360 new Kenyan participants provided 324,160 forecasts, and following my pre-registration, I selected the ten nudge interventions with the largest predicted effects for additional testing.

Finally, I ran two large field experiments to estimate the causal effects of these crowdsourced nudges on recipients' willingness to opt into the COVID-19 notification service. The first experiment ($n = 40,911$) evaluates (a) the average impact of all 1,822 crowdsourced nudges relative to a control condition, and (b) whether the efficacy of these nudges varies based on the incentives faced by the nudge producers. In the second experiment (n =34,844) I estimate the causal effect of the 10 nudges with the highest crowd-predicted effects relative to the same control and three benchmark nudges from the literature (Legate et al., 2022) (2 message) and based on communications from the Kenyan Ministry of Health (1 message).

I present three main results. First, the average treatment effect across all 1,822 nudges is negative, reducing opt-in rates by 19.6% (0.21 percentage points (pp); p=0.33) relative to a control message. Second, this result is not driven by a lack of incentives for participants to create persuasive content (Gibbs et al., 2017; Charness and Grieco, 2019); those facing higher incentives for nudge efficacy produce messages that are no more effective. Third, in a separate experiment I evaluate the effects of the ten *crowd-selected* nudges that local participants predicted would have the largest causal effect. This procedure allows me to leverage the wisdom of crowds by aggregating the beliefs of many participants, which reduces noise in the crowdsourcing process. These crowd-selected nudges increase opt-in to the COVID-19 notification service by 41.0% (0.34 pp; p=0.01) compared to

the control group or by 35.0% (0.23 pp; p=0.07) relative to three benchmark conditions. Together, these results provide encouraging evidence on the effectiveness of crowdsourced and crowd-selected nudges: local participants can create effective nudges, and incentivized crowd predictions identify effective nudges from a larger menu of interventions.

The remainder of the paper is structured as follows. Section 2 provides an overview of the experiments and results. Section 3 concludes. Details on the design of the four studies and the empirical strategy are provided in Section 4 and in the appendix.

2 Results

Study 1: Crowdsourcing nudges. Study 1 is designed to generate a large set of crowdsourced nudge interventions to increase opt-in to an SMS-based notification service that provides updates on the spread of COVID-19 in Kenya. Using Facebook ads, I recruited 1,822 Kenyan participants who completed a screening survey and passed pre-registered exclusion criteria described in Appendix B. Panel B of Figure 1 provides an overview of the message design task, the message that nudge recipients would see, and an example of the type of information provided by the notification service. Participants creating nudge interventions were randomly assigned to one of three different incentive contracts that paid either 0, 4, or 10 Kenyan Shillings for each randomly assigned recipient that opted in to receive the notification. This crowdsourcing procedure produced a diverse set of nudge interventions (Panel A of Table A1 provides a randomly selected list of ten example messages).

Study 2: Crowd selection of nudges. The previous study crowdsourced a large menu of messages, but some of these messages are going to be from producers of low ability or who exert minimal effort. In other words, while crowdsourcing produces many nudges, they are going to be of variable quality. Study 2 extends the crowdsourcing process to crowd *selection* of nudges. Building on recent work (DellaVigna and Pope, 2018; Thomas et al., 2020; Otis, 2021, 2022), I collected 324,160 forecasts of the causal effects of 1,496 messages in Study 1 from a new sample of 1,360 forecasters. Following my pre-registration, I selected the ten nudges with the largest predicted effects for experimental evaluation in Study 4. See Appendix B for details on the set of predicted nudges and for pre-registered exclusion criteria.

Evaluating crowdsourced nudges. I ran two large experiments ($n_{\text{study3}} = 40,911$ and

$n_{\text{study4}} = 34,844$) to evaluate the effects of the crowdsourced nudges. All message recipients were sent an invitation to opt in to the COVID-19 notification service, which was accompanied by a randomly assigned nudge intervention in the treatment conditions (see Figure 1 for details). The main outcome is the percent of participants that opt into the notification service, which they were only able to do through the text message invitation. While the focus of this study is on the effects of crowdsourced *nudges*, in Panel C of Table A3 and Table A6 I provide evidence on the effectiveness of financial incentives for signing up for the information service.

Study 3. The average effect of crowdsourced nudges. How effective are the 1,822 crowdsourced nudges at increasing adoption of the COVID-19 notification service? I randomly assigned these nudges to a sample of 36,517 participants and compare the pooled effect of these interventions to a control condition ($n = 4,394$) that received the invitation depicted in Panel B of Figure 1 absent any additional motivating message. Panel A of Figure 3 depicts the average effects of the crowdsourced nudges. Compared to the control condition which opted in on average 1.07% of the time, the 1,822 nudges on average *decrease* opt-in rates by 19.6% (0.21 pp; p =0.33).

Effect of incentives on nudge efficacy. Are participants creating ineffective nudges due to lack of incentives? In Study 1, I randomized the sample of 1,822 participants to one of three outcome contingent contracts that paid 0, 4, or 10 Kenyan Shillings for each message recipient that signed up for the COVID-19 notification service. Panel B of Figure 3 shows that incentives did not improve average nudge effectiveness: the opt-in rate for the highest-paid incentive condition is 0.74%, compared to 0.94% for the no-pay paid condition (p=0.09).

Study 4. The effect of crowd-selected nudges. Next, I test the effects of the ten messages from Study 3 that participants predicted would be most effective. Each of these *crowd-selected* nudges was sent to an average of 1,846 new recipients (18,457 participants total). I benchmark these crowd-selected nudges against (1) the control message from Study 3 ($n = 10,867$), and against three benchmark messages (sent to a total of 5,520 participants) based on recent experimental literature (Legate et al., 2022) and a social media campaign from the Kenyan government (see Appendix D for details). Panel A of Figure 4 depicts the effects of the crowd-selected and benchmark messages relative to the pure control, and Panel B displays the effects of the ten crowd-selected messages and

the three benchmark messages. The crowd-selected nudges lead to an average increase in opt-in rates of 41.0% over the pure control (0.34 pp; p=0.01), and a 35.0% (0.23 pp; p=0.07) improvement over the three benchmark messages.

3 Discussion

Despite widespread adoption of light-touch nudges, the process of developing and selecting behavioral science interventions remains opaque, relying on expert judgment to evaluate the literature and assess contextual fit. In this paper I test a new method for producing behavioral science interventions, which involves (1) crowdsourcing a large set of interventions from individuals with local contextual knowledge, and (2) using incentive compatible forecasts of the causal effects of these nudges to prioritize interventions for testing. This "bottom-up" strategy for designing nudges is appealing because it provides a mechanism for interventions to be tailored to the local setting.

My results highlight the value of *crowd-selection* in the crowdsourcing process: the average crowdsourced nudge is ineffective and if anything reduces participation in the COVID-19 notification service. In contrast, the average prediction from crowds of incentivized forecasters identifies a set of nudges that increase opt-in rates by 41.0% (0.34 pp; p=0.01). An implication of this result is that, even if the mean effect from a menu of crowdsourced nudges is negative, the effect of crowd-selected nudges can still be positive if the crowd can identify effective interventions from this menu.

Formalizing the crowdsourcing process into two stages—developing a choice set and selecting interventions from this choice set—clarifies several directions for future research. Policymakers may be interested in interventions that will induce a high-variance distribution of nudges if the wisdom of crowds can identify top performers in the right tail of this distribution. In this paper I test the effect of *linear* incentives for developing effective nudges that may have prevented participants from developing riskier messages that would produce a longer-tailed distribution of effects (Ederer and Manso, 2013). Refinements to the crowd-selection process, such as applying differential weights to forecasters based on past performance (Tetlock and Gardner, 2016), or allowing for communication between forecasters (Becker et al., 2017) may lead to even better crowd choices. Finally, in future work it will be crucial to understand the boundary conditions under which crowds can produce and identify effective interventions, and to consider other benchmarks for the production and selection of nudges; a mixture of local participants and academic experts may be able to discover nudges that leverage insights from the academic literature and

that are tailored to the local context.

4 Materials and methods

Study 1 design

1,822 participants were recruited over Facebook, completed a pre-treatment screening survey, and passed pre-registered exclusion criteria. For details on pre-registered exclusion criteria in all four studies see the Appendix.

Incentives. Participants were randomized to three different experimental conditions paying bonuses of $\{0, 4, 10\}$ Kenyan Shillings for each message recipient that was randomly assigned to receive their message and signed up for the COVID-10 notification service, and 99.23% of participants passed a multiple-choice comprehension check where they were asked about the magnitude of their incentives (they were given two tries to answer correctly). Table A2 shows that treatment groups were balanced on a range of covariates.

Study 2 design

1,360 participants were recruited over Facebook and passed pre-registered exclusion criteria. Forecasters provided predictions of 1,496 nudges from Study 1 (see Appendix C for details).

Forecast elicitation. Participants were provided with a benchmark opt-in rate for the control group of 1%, and predicted the conditional means on a slider scale, bounded at 0 and 3 to reduce noise. They received bonus payments based on the accuracy of their predictions.

Aggregation and policy choice. I pre-registered that I would calculate the average predicted effect for each message, and that I would experimentally evaluate the ten messages predicted to be most effective (see Appendix C).

Study 3 design

Study 3 participants were randomly selected from the sample pool of my implementing partner the Busara Center for Behavioral Economics. Messages were delivered to 40,911 participants. All messages were sent from the same SMS shortcode, a five-digit phone number used for large-scale communication. An example invitation is depicted in Appendix E. The control message was identical to the crowdsourced messages except that

it did not include the motivating text. Each participant was assigned to only one experimental message. The outcome is whether the message recipient opted into the COVID-19 notifications. If no response was received after 6 hours, a reminder was sent that was identical to the first message, but which started with "This is a reminder".

Study 4 design

Study 4 tested the effects of ten crowd-selected messages relative to the same control message used in Study 3 and three benchmark messages. Participants were from a new sample drawn from the Busara Center's pool. Messages were delivered to 34,844 participants and were balanced across experimental conditions on gender (see Table A5). Details on the benchmark messages and crowdsourced nudges can be found in Appendix E.

Statistical Methods

I evaluate the average effects of crowdsourced nudges using the following equation:

$$y_i = \alpha + \beta \mathbf{T}_i + \varepsilon_i \tag{1}$$

where $y_i \in \{0, 1\}$ is a variable taking a value of 1 if recipient i opts in to the COVID-19 notification service, and ε_i is the error term. $\mathbf{T}_i$ is a vector of dichotomous treatment variables that changes based on the experimental comparison being tested:

- Average effect of crowdsourced nudges (Study 2): $\mathbf{T}_i$ is a single dichotomous variable taking a value of 1 if recipient i receives a crowdsourced message.
- Effect of incentives (Study 3): $\mathbf{T}_i$ is a vector of three dichotomous variables each taking a value of 1 if recipient i is randomly assigned to a message from a producer facing bonus payments of 0, 4, or 10 Kenyan Shillings per recipient who opts in. As robustness checks I also exclude messages that were not included in the crowd-selection experiment (Study 2) and include pre-registered stratification variables.
- Effect of crowd-selected and benchmark messages (Study 4): $\mathbf{T}_i$ is a vector of two dichotomous variables. The first takes a value of 1 if recipient i received one of the ten crowdsourced messages. The second takes a value of 1 if the recipient receives one of the three benchmark messages. I also separate this set into a vector of thirteen message-level dichotomous variables. Alternative specifications control for recipient gender (observed in Study 4 but not 3).

For regressions that pool effects across messages I generate confidence intervals using the wild bootstrap. Otherwise, I use robust standard errors.

References

Bavel, J. J. V., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., et al. (2020). Using social and behavioural science to support covid-19 pandemic response. *Nature human behaviour*, 4(5):460–471.

Becker, J., Brackbill, D., and Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences*, 114(26):E5070–E5076.

Bennett, J., Lanning, S., et al. (2007). The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. Citeseer.

Bryan, C. J., Tipton, E., and Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature human behaviour*, 5(8):980–989.

Campos-Mercade, P., Meier, A. N., Schneider, F. H., Meier, S., Pope, D., and Wengström, E. (2021). Monetary incentives increase covid-19 vaccinations. *Science*, 374(6569):879–882.

Charness, G. and Grieco, D. (2019). Creativity and incentives. *Journal of the European Economic Association*, 17(2):454–496.

Dai, H., Saccardo, S., Han, M. A., Roh, L., Raja, N., Vangala, S., Modi, H., Pandya, S., Sloyan, M., and Croymans, D. M. (2021). Behavioural nudges increase covid-19 vaccinations. *Nature*, 597(7876):404–409.

DellaVigna, S. and Linos, E. (2022). Rcts to scale: Comprehensive evidence from two nudge units. *Econometrica*, 90(1):81–116.

DellaVigna, S. and Pope, D. (2018). Predicting experimental results: who knows what? *Journal of Political Economy*, 126(6):2410–2456.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., and Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50):15343–15347.

Ederer, F. and Manso, G. (2013). Is pay for performance detrimental to innovation? *Management Science*, 59(7):1496–1513.

Gibbs, M., Neckermann, S., and Siemroth, C. (2017). A field experiment in motivating employee ideas. *Review of Economics and Statistics*, 99(4):577–590.

Hussam, R., Rigol, N., and Roth, B. N. (2022). Targeting high ability entrepreneurs using community information: Mechanism design in the field. *American Economic Review*, 112(3):861–98.

Iyer, R., Khwaja, A. I., Luttmer, E. F., and Shue, K. (2016). Screening peers softly: Inferring the quality of small borrowers. *Management Science*, 62(6):1554–1577.

Legate, N., Ngyuen, T.-v., Weinstein, N., Moller, A., Legault, L., Vally, Z., Tajchman, Z., Zsido, A. N., Zrimsek, M., Chen, Z., et al. (2022). A global experiment on motivating social distancing during the covid-19 pandemic. *Proceedings of the National Academy of Sciences*, 119(22).

Maier, M., Bartoš, F., Stanley, T., Shanks, D. R., Harris, A. J., and Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, 119(31):e2200300119.

Mollick, E. and Nanda, R. (2016). Wisdom or madness? comparing crowds with expert evaluation in funding the arts. *Management science*, 62(6):1533–1553.

Otis, N. G. (2021). Forecasting in the field. *Working paper*. Retrieved from `https://nicholasotis.com/Research/Otis_ForecastingField.pdf`.

Otis, N. G. (2022). Policy choice and the wisdom of crowds. *Available at SSRN 4200841*.

Rabb, N., Swindal, M., Glick, D., Bowers, J., Tomasulo, A., Oyelami, Z., Wilson, K. H., and Yokum, D. (2022). Evidence from a statewide vaccination rct shows the limits of nudges. *Nature*, 604(7904):E1–E7.

Sobel, D. (2005). *Longitude: The true story of a lone genius who solved the greatest scientific problem of his time*. Macmillan.

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S., and Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences*, 119(31):e2200732119.

Tetlock, P. E. and Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Thomas, C. C., Otis, N. G., Abraham, J. R., Markus, H. R., and Walton, G. M. (2020). Toward a science of delivering aid with dignity: Experimental evidence and local forecasts from kenya. *Proceedings of the National Academy of Sciences*, 117(27):15546–15553.

Figure 1: Overview of studies

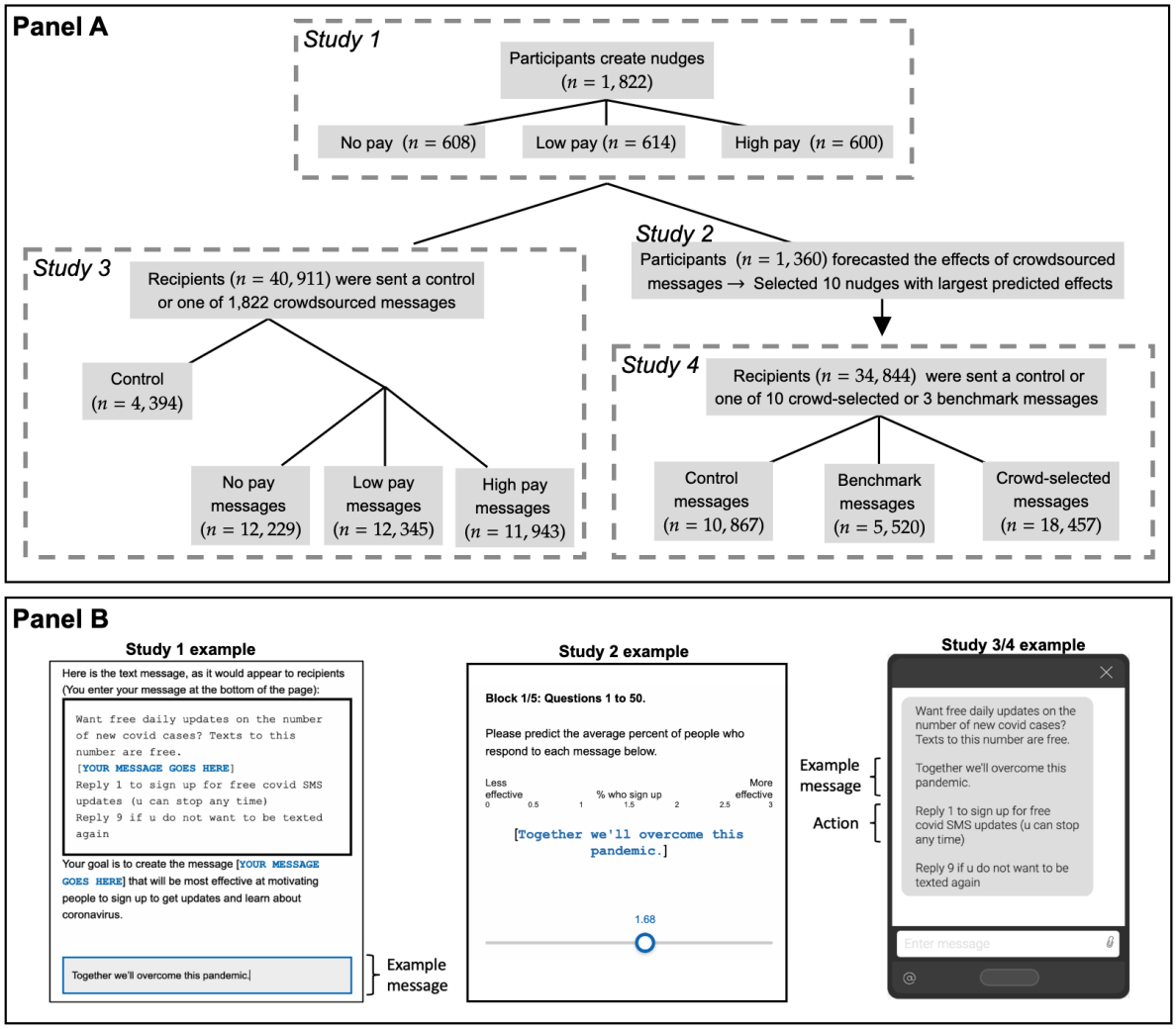
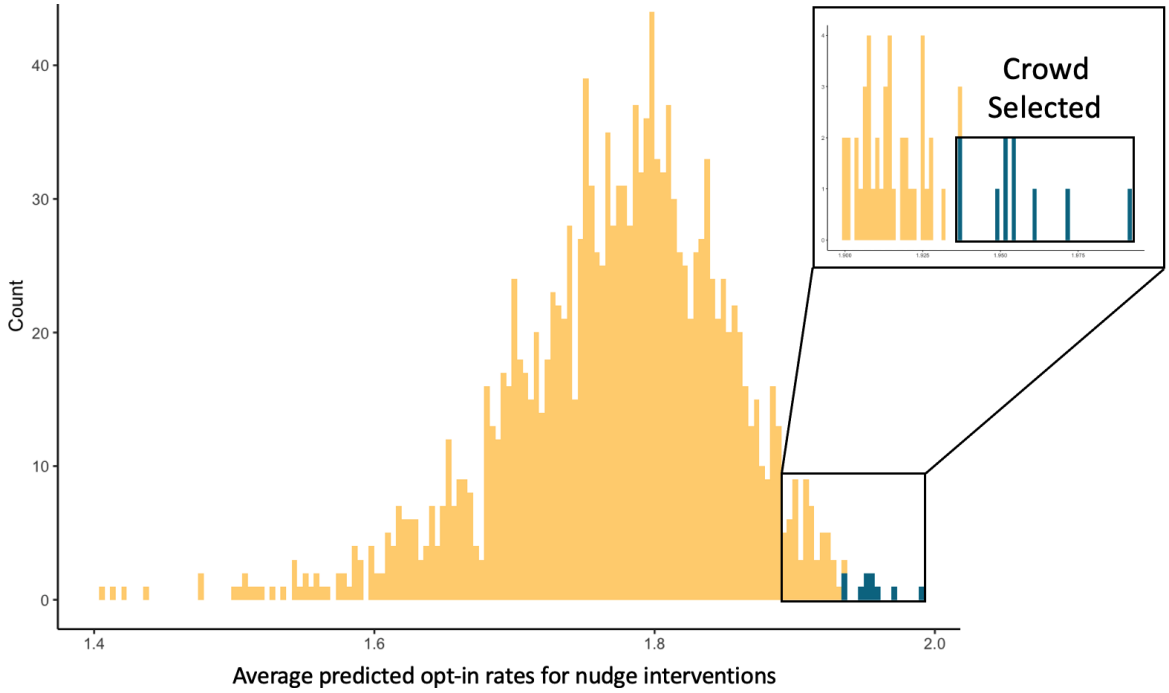
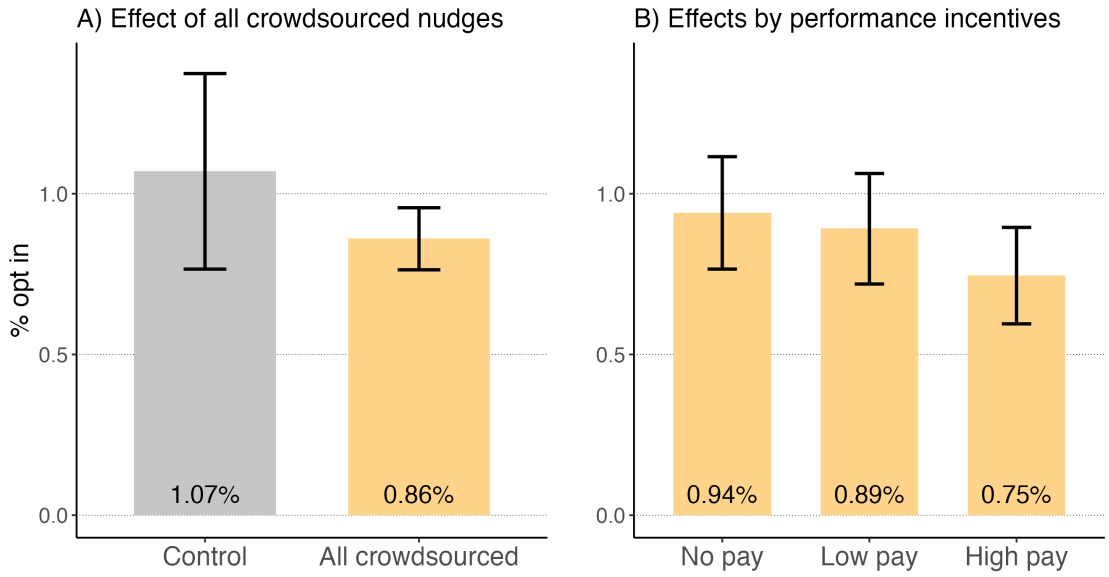


Figure 2: Illustration of crowd selection (Study 2)



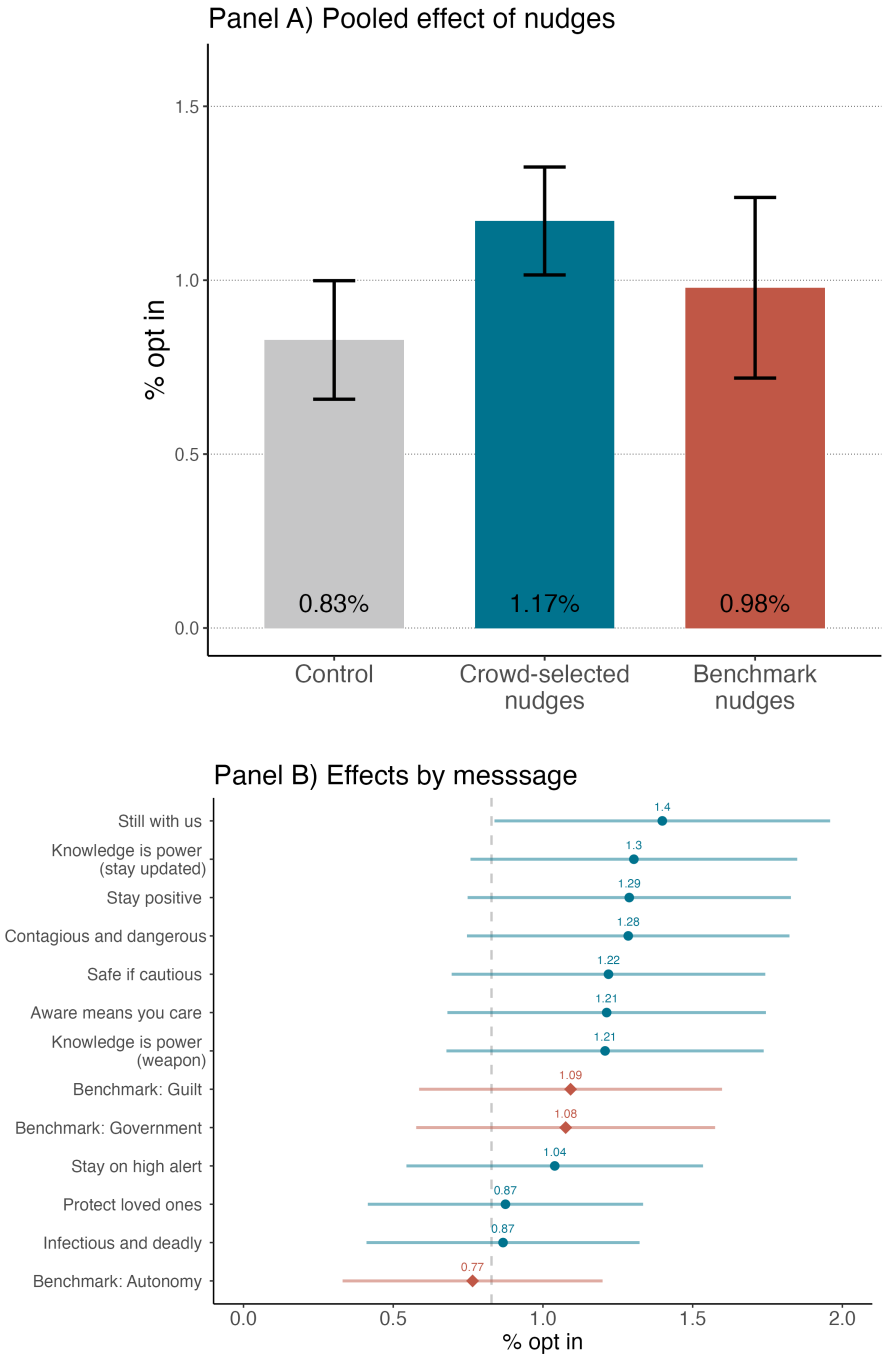
This figure displays the distribution of average predicted effects for 1,496 nudges from Study 2. 1,360 forecasters provided a total of 324,160 predictions. The ten "crowd-selected" nudges with the largest predicted effects are displayed in blue.

Figure 3: Effects of crowdsourced nudges (Study 3)



Panel A displays average opt-in rates for recipients who receive a control message ($n = 4,394$) or one of 1,822 crowdsourced nudges ($n = 36,517$). Panel B depicts the average opt-in rates by randomly assigned performance incentives for producing effective nudges, ($n_{\text{no pay}} = 12,229$; $n_{\text{low pay}} = 12,345$; $n_{\text{high pay}} = 11,943$). Error bars display 95% confidence intervals clustered at the message level.

Figure 4: Effects of crowd-selected nudges (Study 4)



Panel A displays average opt-in rates for recipients who receive a control message ($n = 10,867$), the ten crowdsourced nudges that participants predicted would have the highest causal effect ($n = 18,457$), or and the three benchmark nudges ($n = 5,520$). Error bars display 95% confidence intervals clustered at the message level. Panel B depicts the average opt-in rates of the thirteen messages, which are listed in Panels C and D of Table A1. Blue circles represent the ten crowdsourced nudges, and red diamonds depict the three benchmark nudges. The vertical dashed line displays the control mean. Error bars represent 95% confidence intervals.

SUPPORTING INFORMATION

The Efficacy of Crowdsourced Nudges: Experimental Evidence

Nicholas G. Otis

A Appendix tables

Table A1: Nudges from Studies 1-4

Panel A: Random sample of crowdsourced messages
Let us get our lives back.
COVID is real and the only way to prove this and stay informed is by getting the updates.
COVID 19 is a deadly disease. You all deserve to get updates, all you need to do is to subscribe.
Please sign up and get updates and learn more about corona virus.
Information clears all doubts, get informed!
COVID is real, stay alert.
In need of COVID. 19 case updates in the country?
Together we'll overcome this pandemic.
Sign up as directed below to remain updated each day.
Health is wealth.
Panel B: Messages containing health tips or repeating invitations
Repeating invitation:
Would you want to sign up for free COVID SMS updates?
Get registered for daily COVID sms updates by replying with 1.
Health tips:
Wash your hands and sanitize at all times.
Observe social distance, wear your mask and stay at home.
Panel C: Benchmark messages
Benchmark: Government. Be your neighbors' keeper and get informed! A healthy community leads to a healthier country.
Benchmark: Guilt. If you don't get information you're making a mistake and putting yourself and others at risk.
Benchmark: Autonomy. You can help to stop COVID. 19 by choosing to get information about the number of new cases.
Panel D: Crowd. selected messages
Still with us. Corona is still with us, let's keep adhering to the Ministry of Health protocols by getting daily updates.
Contagious and dangerous. Corona virus is a contagious and dangerous disease therefore one is advised to take vaccination.
Infectious and deadly. COVID. 19 is highly infectious and deadly. Get new infection updates and protect ur loved ones.
Aware means you care. Get to know how to stay safe during the period, being aware means you care.
Stay positive. Kindly sign up to get daily knowledge and information about COVID 19 updates daily. Stay positive.
Knowledge is power (weapon). Knowledge is power, and knowledge about coronavirus is the first weapon in fighting the disease.
Knowledge is power (stay updated). Knowledge is power, get COVID 19 daily cases and stay updated, sign up for updates via free sms.
Stay on high alert. Stay on high alert concerning the changing trends of the COVID pandemic.
Protect loved ones. Stay safe, protect yourself and your loved ones. Wear a mask while in public.
Safe if cautious. We can all be safe if we take COVID precautions seriously. Let's get our guard rolling soon.
Panel E: Financial incentives
Payment of 5 Kenyan Shillings. We will send you a bonus of KES 5 airtime tomorrow if you sign up to receive our updates.

Panel A presents a random sample of ten crowdsourced messages. Panel B contains examples of messages included in Study 1 and 3 but excluded from the forecasting exercise in Study 2. Panel C depicts the three benchmark nudges in Study 4. Panel D lists the ten crowd-selected nudges from Studies 2 and 4. Panel E lists the financial incentive condition from Studies 3 and 4. Abbreviations for experimental conditions used in Studies 3 and 4 are presented in italics.

Table A2: Sociodemographics and balance (Study 1)

		By incentive condition			
	Full sample	No pay	Low pay	High pay	F-stat.
	(1)	(2)	(3)	(4)	(5)
Panel A) Stratification variables					
% with borderline test message	38.36	37.83	39.58	37.67	0.29
	(1.14)	(1.97)	(1.98)	(1.98)	
% above median comprehension	0.66	65.13	65.15	66.33	0.13
	(0.01)	(1.93)	(1.92)	(1.93)	
Panel B) Other sociodemographic variables					
% female	45.01	42.60	44.46	48.00	1.83
	(1.17)	(2.01)	(2.01)	(2.04)	
% completed college	56.2	55.26	58.79	54.50	1.31
	(1.16)	(2.02)	(1.99)	(2.03)	
log(monthly income+1)	800.53	807.39	799.09	795.03	0.27
	(7.01)	(12.07)	(12.12)	(12.26)	
Panel C) Post-treatment variables					
% messages with tips or repeat invitation	17.89	18.26	18.57	16.83	0.36
	(0.9)	(1.57)	(1.57)	(1.53)	
$n_{\text{producers}}$	1822	608	614	600	

This table tests for balance among participants creating nudge interventions by randomly assigned incentives for nudge efficacy. Panel A lists displays balance on two pre-registered stratification variables: *Borderline message,* equal to 1 if participants create a test message in a screening survey that only marginally passed pre-specified message rules (see Appendix B for a list of rules and Panel B of Table A1 for examples of messages violating these rules), and *Above median comprehension* which is equal to 1 if a participant is above median comprehension on a set of 11 comprehension questions in the screening survey. Panel B depicts balance on additional sociodemographic variables. The final variable (Panel C) is the proportion of experimental participants who created a message that was excluded from Study 2 (measured post treatment), either because it focuses on providing health tips or repeats the control invitation.

Table A3: Effects of crowdsourced messages (Study 3)

	Effect on opt in (pp) (1)	p-value (2)	n_{messages} (3)	$n_{\text{recipients}}$ (4)
Reference: Control mean	1.07			4394
Panel A) Average crowdsourced nudges				
All crowdsourced nudges	-0.21	0.32	1822	36517
	[-0.95,0.53]			
Panel B) Effects by incentive condition				
No financial incentives	-0.13	0.35	608	12229
	[-0.69,0.44]			
Low financial incentives	-0.18	0.29	614	12345
	[-0.73,0.36]			
High financial incentives	-0.32	0.13	600	11943
	[-0.78,0.15]			
Panel C) Financial incentives				
Payment of 5 Kenyan Shillings	0.96	0.03		1180
	[0.10,1.83]			

Panel A reports the average effect of 1,822 crowdsourced nudges. Panels B pools messages randomly by randomly assigned incentive conditions. Panel C looks at the effect financial incentives for opting into the notification service. 95% confidence intervals are presented in brackets. In Panel A and B these intervals are generated using the wild bootstrap clustered at the message level. Pane C uses confidence intervals generated from robust standard errors.

Table A4: Robustness check on effect of incentives (Study 3)

	Effect on opt in (pp) (1)	p-value (2)	Effect on opt in (pp) (3)	p-value (4)	n_{messages} (5)	$n_{\text{recipients}}$ (6)
Panel A) All crowdsourced nudges						
Reference: No financial incentives	0.94				608	12229
Low financial incentives	-0.05	0.69	-0.05	0.70	614	12345
	[-0.30,0.20]		[-0.30,0.20]			
High financial incentives	-0.20	0.09	-0.20	0.09	600	11943
	[-0.42,0.03]		[-0.42,0.03]			
Panel B) Excluding nudges with tips or repeating invitation						
Reference: No financial incentives	0.94				497	9994
Low financial incentives	-0.02	0.88	-0.02	0.90	500	10108
	[-0.29,0.25]		[-0.29,0.25]			
High financial incentives	-0.12	0.36	-0.12	0.36	499	9960
	[-0.37,0.13]		[-0.37,0.13]			
Controls	None		Strata			

Panels A looks at the effects of nudges pooled by randomly assigned incentive conditions for message producers either without controlling for pre-registered producer stratification variables (whether the producer was above median on comprehension checks or produced a message was considered 'borderline" based on pre-registered exclusion criteria. Panel B excludes 327 nudges that were excluded from the crowdsourcing/crowd-selection exercise. 95% confidence intervals generated using the wild bootstrap clustered at the message level are presented in brackets.

Table A5: Recipient balance (Study 4)

	% female (1)	$n_{\text{recipients}}$ (2)
Control mean	56.41	10867
	(0.48)	
Benchmark messages		
Benchmark: Government	55.78	1859
	(1.15)	
Benchmark: Guilt	55.05	1831
	(1.16)	
Benchmark: Autonomy	57.87	1830
	(1.15)	
Crowd-selected messages		
Still with us	54.81	1859
	(1.15)	
Contagious and dangerous	55.46	1868
	(1.15)	
Infectious and deadly	55.80	1846
	(1.16)	
Aware means you care	55.18	1814
	(1.17)	
Stay positive	56.04	1863
	(1.15)	
Knowledge is power (weapon)	56.70	1822
	(1.16)	
Knowledge is power (stay updated)	55.35	1841
	(1.16)	
Stay on high alert	56.95	1828
	(1.16)	
Protect loved ones	56.48	1829
	(1.16)	
Safe if cautious	55.70	1887
	(1.14)	
F-statistic		0.56
n_{total}		34844

This table presents message-level balance by recipient gender for Study 2. Standard errors are displayed in parentheses. Dull messages are listed in Panels C and D of Table A1.

Table A6: Effects of crowd-selected nudges

	Effect on opt in (pp)	p-value	Effect on opt in (pp)	p-value	$n_{\text{recipients}}$
Reference: Control mean	0.83				10867
Panel A) Pooled effects					
Crowd-selected nudges	0.34		0.34		18457
	[0.13,0.54]	0.01	[0.12,0.53]	0.01	
Benchmark nudges	0.15		0.15		5520
	[-0.17,0.32]	0.25	[-0.16,0.30]	0.30	
Panel B) Message effects					
Benchmark: Government	0.25	0.35	0.24	0.36	1859
	[-0.27,0.77]		[-0.28,0.76]		
Benchmark: Guilt	0.26	0.32	0.25	0.34	1831
	[-0.26,0.78]		[-0.27,0.77]		
Benchmark: Autonomy	-0.06	0.81	-0.06	0.83	1830
	[-0.58,0.46]		[-0.58,0.46]		
Still with us	0.57	0.03	0.56	0.03	1859
	[0.05,1.09]		[0.04,1.08]		
Contagious and dangerous	0.46	0.08	0.45	0.09	1868
	[-0.05,0.97]		[-0.06,0.96]		
Infectious and deadly	0.04	0.88	0.03	0.90	1846
	[-0.48,0.56]		[-0.49,0.55]		
Aware means you care	0.38	0.15	0.38	0.16	1814
	[-0.14,0.90]		[-0.14,0.90]		
Stay positive	0.46	0.08	0.46	0.08	1863
	[-0.06,0.98]		[-0.05,0.97]		
Knowledge is power (weapon)	0.38	0.15	0.38	0.15	1822
	[-0.14,0.90]		[-0.14,0.90]		
Knowledge is power (stay updated)	0.48	0.07	0.47	0.08	1841
	[-0.04,1.00]		[-0.05,0.99]		
Stay on high alert	0.21	0.43	0.21	0.42	1828
	[-0.31,0.73]		[-0.31,0.73]		
Protect loved ones	0.05	0.86	0.05	0.86	1829
	[-0.47,0.57]		[-0.47,0.57]		
Safe if cautious	0.39	0.14	0.38	0.14	1887
	[-0.12,0.90]		[-0.13,0.89]		
Panel C) Financial incentives					
Payment of 5 Kenyan Shillings	1.77	0.00	1.77	0.00	1850
	[1.25,2.29]		[1.25,2.29]		
Controls	None		Gender		

Panel A reports the average effect of the 10 crowd-selected nudges and the three benchmark nudges. Panels B and C look at the average effects of individuals messages. 95% confidence intervals are presented in bracket. In Panel A these intervals are generated using the wild bootstrap clustered at the message level. Panels B and C use confidence intervals generated from robust standard errors.

B Study 1 details

Recruitment. Participants were recruited using a Facebook advertisement which led them to a short screening survey designed to identify inattentive respondents.

Exclusion criteria: I pre-registered on the AEA RCT registry (AEARCTR-0006414) that I would screen out participants (a) who failed any of four attention checks, (b) who tried to take the survey multiple times, or who (c) had below a secondary education. Participants who answer the four screening questions correctly could choose to end the survey and earn a payment of 20 Kenyan Shillings or could continue and create a test text message that would not be distributed. Participants were told that they could be invited to participate in a second survey if their message passed the following rules.

1. Messages shouldn't contain any false information.
2. Messages should motivate people to sign up for updates but shouldn't include health information/tips.
3. Messages shouldn't offer financial incentives.
4. Messages shouldn't be repeated. Don't change just a few words. Write a different and new message.

Following pre-registration, messages were independently evaluated by research assistants. Rules 1 and 3 are designed to avoid misinformation and deceit. Rule 2 is designed to avoid a common misconception identified during piloting (several people simply listed health tips like "wash your hands" that are not related to the notification service). Rule 4 is designed to discourage participants to avoid simply re-typing the control message text.

Messages were reviewed by research assistants and were classified as "include","exclude" or "borderline" (which *partially* violate these rules). For example, the first part of the message "Avoid large indoor gatherings. Together, we can save lives." violates Rule 2 ("Avoid large indoor gatherings." is a health tip), but the second part ("Together, we can save lives.") does not. Similarly, the message "Are the government efforts in battling COVID working? Sign up to find out" potentially violates Rule 1, since we only provide information on COVID cases, which has an ambiguous mapping to the efficacy of government programs. Borderline messages are independently reviewed by an additional research assistant. If the message is classified as *borderline* or *pass* during this review stage, the message is included. Otherwise, the message is excluded.

Stratification. I pre-register that random assignment will to the three different incentive

conditions will be stratified on two variables:

- **Attention.** A median split of a larger set of 11 attention checks included in the screening survey.
- **Borderline messages.** Whether the respondents' message was identified as *borderline*.

After random assignment, I recontacted participants over email and phone with the main message production survey.

Message exclusion criteria. I pre-register that messages will be evaluated by research assistants who are blind to the experimental condition the participant was assigned to. They will be asked to assess whether each message violates either of the following criteria (the examples below were provided to research assistants):
1) Messages cannot contain any false information. Here are four examples of messages violating this rule:

- *By careful, coronavirus will make you sterile.*
- *We will send you information on which of your friends have been vaccinated.*
- *Call us at 12345 to receive free advice on coronavirus.*
- *Hot water and lemon will boost your immunity and keep you safe from corona.*

2) Messages cannot offer financial incentives. Here are two examples of messages violating this rule:

- *We will pay you KSh 100 if you sign up.*
- *Don't wait! Sign up today and receive a cash prize.*

Message edits. Participants are informed in the survey we would make the following changes to their messages:

- *We will correct spelling and punctuation (you can still use abbreviations like u for you).*
- *We will replace messages in ALL CAPS with correct capitalization. We won't change capitalization if only a FEW words are in CAPS.*
- *We will remove emojis (do not include emojis).*
- *Messages will be sent in English (Kiswahili messages will be translated).*

Additionally, the survey clarified that participants whose messages reference the *reply* options in the text incorrectly will be corrected. For example, in the message *For peace of mind, text back 0 to begin receiving updates.* the respondent incorrectly listed as the

number the recipient needed to text to receive notifications instead of the correct reply option of 1.

C Study 2 details

Recruitment. I recruited a new sample of participants from Facebook. After applying pre-registered a attention check (AsPredicted #106631)[1], my sample consists of 1,360 forecasters who started the survey and predicted the effects of at least one message. In total these forecasters provided 324,160 predictions.

Incentives for accuracy. Accurate forecasts were incentivized such that participants would receive a bonus payment for one randomly selected forecast based on formula

$$9 - (\text{predicted opt in} - \text{observed opt in}|\text{message})^2.$$

In addition to displaying this equation, the survey emphasized that more accurate predictions correspond to larger bonuses.

Sample of messages in forecasting survey. Messages evaluated in Study 3 were based on a strict set of pre-registered exclusion criteria (AEARCTR-0006414). However, I also included a list of Rules (see Appendix B) that Study 1 participants were supposed to abide by when creating messages. For Study 2, I chose to exclude messages that were coded by research assistants as mainly (a) providing health advice, or (b) repeating the invitation text. This reduces the set of predicted nudges from 1,822 to 1,496. Panel B of Table A1 depicts examples of these excluded messages (Panel B of Table A1 depicts examples of these excluded messages). Panel B of Table A4 shows that there is no meaningful difference in comparison of the incentive conditions when excluding these messages. Additionally, there is no difference in opt-in rates between messages included in Study 2 (average opt in=0.86) and those excluded in study 2 (average opt in=0.84; p-value on difference=0.83).

Reliability. Of the 250 predicted messages, the first 245 are random draws without replacement from the sample of 1,496 described above. These predictions are used in calculating crowd-selection. To measure reliability, messages 246-250 are identical to

[1]This page contains 3 typos. In Question 4, the financial payments condition provides a payment of 5 Kenyan Shillings (this is true in Studies 3 and 4), not 4 Kenyan Shillings. I also repeat the attention-based exclusion check (Question 6), and repeat the word "recipients" in Question 7. The registration can be viewed here: `https://aspredicted.org/blind.php?x=PGF_X7V`.

messages 1-5. How reliable are forecasters? For each individual, I calculate the average absolute difference (AAD) between the predicted opt-in rates for messages 1-5 and messages 246-250. If forecasters' predictions are purely noise, we would expect AAD to be the same for messages (1-5 and 246-250) and (1-5 and 240-245). Consistent with predictions containing some information, I find that average AAD is significantly lower for the repeated messages than for the five preceding arbitrary messages (difference= -0.028 ($p<0.001$)). However, the average AAD across participants is 0.56 percentage points, indicating considerable noise in individuals' predictions. This provides some suggestive insights into why the crowd-forecasts perform well, and why Study 1 messages perform poorly: messages individuals produce and the predictions they provide are quite noisy, and a procedure for aggregating information is needed to identify effective interventions.

Crowd choice. For each message, I calculate the average predicted causal effect from participants. Following pre-registration I select the ten messages with the highest predicted effects for evaluation in Study 4 (AsPredicted #106631).

D Study 3 details

Undelivered messages. Messages were sent to a total of (40,911+11,978)=52,889 participants, however the SMS platform failed to deliver 11,978 messages. This is likely because participants' numbers had been deactivated or their phones were off. This constitutes pre-randomization attrition, and these failed numbers are excluded from my analytic sample.

Invitation message and treatment. All participants were sent the following invitation:

```
Want free daily updates on the number of new covid cases?
Texts to this number are free.
[Nudge message goes here]
Reply 1 to sign up for free covid SMS updates (u can stop any time)
Reply 9 if u do not want to be texted again
```

For control participants the section [Nudge message goes here] is left blank. Only participants who received an invitation were able to opt into the notification service.

Notification example. Individuals who opted in received the following information on the number of new COVID-19 cases and deaths across the country (e.g., "Here's the latest information from the Ministry of Health: 53 people tested positive to COVID-19 from a sample size of 4,071 tested in the last 24 hours (1.3% positivity), and 0 died. Positive cases by county: Nairobi 22, Trans Nzoia 13, Nakuru 6, Kericho 4, Taita Taveta 3, Busia 1, Homa Bay 1, Kakamega 1, Kilifi 1, and Mombasa 1.").

E Study 4 details

Undelivered messages. Messages were sent to (34,844+26,125)=60,969 participants, of which 26,125 messages were not received by study participants (an aggressive campaign from the Kenyan government reduce the number of active mobile numbers in Kenya likely accounts for the difference in failure rates between Study 3 and 4. I mentioned this on the study pre-registration page (AsPredicted #106631). As with Study 3 this represents pre-treatment attrition, and these phone numbers are excluded.

Crowd-selected messages: Participants in the *crowd-selected* messages condition were randomly assigned to one of following ten messages:

1. *Still with us.* Corona is still with us, let's keep adhering to the Ministry of Health protocols by getting daily updates.
2. *Contagious and dangerous.* Corona virus is a contagious and dangerous disease therefore one is advised to take vaccination.
3. *Infectious and deadly.* COVID-19 is highly infectious and deadly. Get new infection updates and protect ur loved ones.
4. *Aware means you care.* Get to know how to stay safe during the period, being aware means you care.
5. *Stay positive.* Kindly sign up to get daily knowledge and information about COVID 19 updates daily. Stay positive.
6. *Knowledge is power (weapon).* Knowledge is power, and knowledge about coronavirus is the first weapon in fighting the disease.
7. *Knowledge is power (stay updated).* Knowledge is power, get COVID 19 daily cases and stay updated, sign up for updates via free sms.
8. *Stay on high alert.* Stay on high alert concerning the changing trends of the COVID pandemic.
9. *Protect loved ones.* Stay safe, protect yourself and your loved ones. Wear a mask

while in public.

10. *Safe if cautious.* We can all be safe if we take COVID precautions seriously. Let's get our guard rolling soon.

Benchmark messages. In addition to the control group, I test the effects of the crowd-selected messages against three benchmark messages:

1. *Benchmark: Guilt.* If you don't get information you're making a mistake and putting yourself and others at risk.
2. *Benchmark: Autonomy.* You can help to stop COVID-19 by choosing to get information about the number of new cases.
3. *Benchmark: Government.* Be your neighbors? keeper and get informed! A healthy community leads a healthier country.

Two of these nudges were from a large-scale behavioral science experiment conducted in 89 countries to increase willingness to social distance (Legate et al., 2022), which tested the effects of messages emphasizing either autonomy and personal choice or guilt and shame. My benchmark messages are based on the study materials from these experiments. For example, in the original study the Guilt condition (called "Controlling" in their study states "If you haven't engaged in social distancing, you are making a mistake and putting yourself and others at risk", and the Autonomy condition states "You can support global efforts to curb transmission of COVID-19 by choosing to stay at home.". The third benchmark is based on communications from a COVID-19 vaccination campaign run by the Kenyan Ministry of Health, which tweeted: "Be your neighbors' keeper and encourage them to get fully vaccinated today! A healthy community leads to a healthier country."