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 approach for locally crowdsourcing interventions. Participants ($n = 1,822$) created nudges to encourage recipients to opt-in to receive notifications about the spread of COVID-19 in Kenya. Another set of participants ($n = 1,360$) predicted the effects of these nudges, and following preregistration I selected the top ten interventions for further testing. In two randomized field experiments, I evaluated the effects of all 1,822 nudges ($n = 40,911$) and the ten crowd-selected nudges ($n = 35,183$). The average nudge reduced uptake of the notification service by 19.6% relative to control ($p = 0.20$). However, the ten crowd-selected nudges increased adoption by 44.0%($p < 0.01$) or 39.4% ($p = 0.05$) compared to benchmark nudges. In a follow-up study ($n = 2,284$), only local forecasters similar to intervention recipients could predict which types of nudges would be effective. These results demonstrate the efficacy of crowdsourced interventions: local participants can produce effective nudges, and incentivized local forecasts provide a mechanism to harness the wisdom of 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, Katherine Milkman 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

Recent cross-country experiments show that local conditions significantly affect the efficacy of interventions (Banerjee et al., 2015; Dunning et al., 2019; Kizilcec et al., 2020; Blair et al., 2021; Slough et al., 2021; Legate et al., 2022). As a result, researchers and policymakers are increasingly calling for tailored interventions that suit the local context (Bates and Glennerster, 2017; Bryan et al., 2021; Szaszi et al., 2022). While foundational theories across the social sciences highlight the power of "local knowledge" to inform policy (Rousseau, 1782; Galton, 1907; Hayek, 1945; March and Simon, 1955; Arrow, 1971; Brown and Duguid, 1991; Fischer, 2000; Surowiecki, 2005), there is a shortage of practical guidance and causal evidence on how this can be done.

I introduce a new paradigm for producing policy interventions and demonstrate its effectiveness in a series of large-scale field experiments. In these studies, I elicit over 1,800 policies from people similar to intervention recipients, and then use incentivized predictions from crowds of local participants to prioritize interventions for further testing. The procedure is based on the intuition that people familiar with a particular setting possess rich information about the types of interventions that will be effective in their environment, but this information is often noisy (Kahneman et al., 2021) requiring the "wisdom of crowds" to extract promising policies (Galton, 1907; Surowiecki, 2005).

In two large-scale randomized field experiments with over 75,000 participants, I estimate the causal effects of (*i*) the full set of policies crowdsourced from participants, and (*ii*) the "crowd-selected" policies that were predicted to be most effective. This experimental procedure differs in important ways from the tremendous number of previous randomized policy experiments. First, traditional experiments evaluate the effects of a predefined set of interventions, whereas mine evaluates the effects of interventions elicited from local participants (following a pre-registered protocol). This bottom-up approach for designing interventions is appealing because it provides a mechanism for interventions to be tailored to the local setting.

Second, my experiments provide causal evidence on *how* to crowdsource policies, in contrast to previous research that has evaluated the effect of a single crowdsourced policy (Zhang et al., 2015; Pan et al., 2017; Tang et al., 2019). Specifically, I evaluate the impact of providing different randomized incentives to local policy designers, investigate the importance of crowd-selection to reduce noise in the crowdsourcing process, and test whether crowd forecasts need to be "local" to identify impactful policies. The empirical results from these experiments insights into how local knowledge can be leveraged to produce and identify high-impact policies, which complement other recent innovations in experimental social science (Banerjee et al., 2021; Milkman et al., 2021; Almaatouq et al., 2021; Duckworth and Milkman, 2022).

Concretely, participants designed light-touch nudges in the form of a short text message intervention aimed at increasing opt-in to a COVID-19 notification service in Kenya. Nudge interventions of this sort have gained widespread adoption among policymakers and researchers (Thaler, 2018; Duckworth and Milkman, 2022; DellaVigna and Linos, 2022). First, I recruited 1,822 Kenyan participants over online advertisements, and each participant created a single text message nudge. Then, I randomly assigned 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). Second, I developed a *crowd-selection* procedure that leveraged the wisdom of crowds to reduce noise in the crowdsourcing process. From a sample of 1,360 Kenyan participants, I collected 324,160 incentive-compatible forecasts of the causal effects of the locally designed nudges. Unlike previous studies that collected predictions of the causal effects of interventions (DellaVigna and Pope, 2018; Thomas et al., 2020; Milkman et al., 2021, 2022; Otis, 2022), I used these forecasts to select the ten nudge interventions with the largest predicted effects for additional testing.

Next, I conducted two large field experiments to estimate the causal effects of these crowdsourced nudges on recipients' willingness to opt into the COVID-19 notification service. In the first experiment ($n = 40,911$), I evaluated (i) the average impact of all 1,822 crowdsourced nudges relative to a control condition, and (ii) whether the efficacy of these nudges varies based on the incentives faced by the nudge producers. In the second experiment ($n =$35,183), I estimated 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 (two messages) and based on communications from the Kenyan Ministry of Health (one message).

I present four main results. First, the average treatment effect across all 1,822 nudges is, if anything, negative, reducing opt-in rates by 19.6% (0.21 percentage points (pp); p=0.20) compared to a control message. This result is not driven by a lack of incentives for participants to create persuasive content; those facing higher incentives for nudge efficacy spend significantly more time designing messages ($p < 0.001$) but produce messages that are no more effective. Second, I evaluate the effects of the 10 crowd-selected nudges that local participants predicted would have the largest causal effect. The crowd-prediction procedure allows me to aggregate the beliefs of many participants, reducing noise in the crowdsourcing process. These crowd-selected nudges increase opt-in to the COVID-19 notification service by 44.0% (0.41 pp; p<0.01) compared to the control group or by 39.4% (0.31 pp; p=0.05) relative to three benchmark conditions. Finally, using predictions from a new set of participants, I demonstrate that local Kenyan participants' predictions correlate strongly with observed experimental effects (cor=0.69, $\text{CI}_{95} = [0.33, 0.85]$), while nonlocal forecasters (from the U.S.) fail to identify what types of interventions will be effective (cor=-0.19, $\text{CI}_{95} = [-0.53, 0.20]$. Together, these results provide evidence on the effectiveness of crowdsourced and crowd-selected nudges: local participants can create effective nudges, and incentivized predictions from local crowds provide a mechanism to 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 Experimental design and results

Crowdsourcing nudges (Study 1). To generate a large set of crowdsourced nudge interventions I recruited 1,822 Kenyan participants over Facebook ads. Each participant designed a text message to increase opt-in to an SMS-based notification service that provides updates on the spread of COVID-19 in Kenya (For details on recruitment, see Appendix C). This service provides details on the positivity rate, deaths, and the distribution of cases across the country. Light-touch changes to communication materials are among the most used policy levers in behavioral science (DellaVigna and Linos, 2022), and delivering accurate real-time information on the spread of COVID-19 has been a key priority of the Kenyan Ministry of Health (Kenyan Ministry of Health, 2019, 2020), and is a cornerstone of pandemic control (Organization et al., 2017; Tumpey et al., 2018). Figure 1 provides an overview of the experimental design (Panel A), and a depiction of how nudges are displayed to participants (Panel B), and Panel A of Table A1 provides a randomly selected list of ten example messages.

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 into the notification platform (as reference, study participants earned about 250 Kenyan Shillings ≈$2.50 per day). Both incentive conditions were effective at increasing effort; incentivized participants spent on average 0.78 more minutes (winsorized at the 5% level) creating their messages (0.32 sd, $p < 0.001$; see Figure A1 for robustness checks). There was no significant difference in completion time by incentive level ($p = 0.50$), suggesting that incentives were sufficiently large to induce effort to a point where marginal costs were high.

Crowd selection of nudges (Study 2). The previous study crowdsourced a large menu of interventions, but some of these interventions are going to be from producers of low ability or who exert minimal effort. In other words, crowdsourcing produces many nudges, but quality of crowdsourced content is likely to be variable. Study 2 extends the crowdsourcing process to crowd *selection* of nudges. I collected 324,160 forecasts of the causal effects of 1,496 messages in Study 1 from a new sample of 1,360 forecasters (see Appendix C for details). Following preregistration, the ten nudges with the largest predicted effects were selected for experimental evaluation in Study 4, as depicted in Figure 2.

Evaluating crowdsourced nudges. I ran two large experiments ($n_{\text{study3}} = 40,911$ and $n_{\text{study4}} = 35,183$) 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*, Panel C of Table A6 and Table A9 also provide evidence on the effectiveness of financial incentives for signing up for the information service.

The average effect of crowdsourced nudges (Study 3). 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 compared 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 =0.20).

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 effect of crowd-selected nudges (Study 4). 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 1,864 new recipients (18,665 participants total). I benchmark these crowd-selected nudges against (*i*) the control message from Study 3 ($n = 10,939$), and (*ii*) three additional messages (sent to a total of 5,579 participants) based on recent experimental literature (Legate et al., 2022) and a public health campaign from the Kenyan government (see Appendix E for details). Importantly, the purpose of this study is to estimate the aggregate impact of messages predicted to have the largest experimental effects relative to control and benchmark conditions. It is not meant to test the effects of individual messages, nor is it powered to make these comparisons. 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 44.0% over the pure control (0.41 pp; p<0.01), and a 39.4% (0.31 pp; p=0.05) improvement over the three benchmark messages. Comparing across studies the crowd-selected nudges outperform the full set of crowdsourced nudges by 0.61 pp ($p < 0.01$).

The value of local knowledge (Study 5). The results presented so far demonstrate that local forecasters can identify effective policies in their environment, but other forecasters may also be effective regardless of their contextual familiarity. I compared 278,810 forecasts from 1,138 non-local forecasters on Amazon Mechanical Turk (MTurk) in the US, to 280,770 forecasts from 1,146 local forecasters in Kenya. Both groups spent a similar amount of time providing forecasts (723 seconds for locals and 751 for non-locals), and previous research has shown that MTurk participants can provide accurate predictions of experimental results in the U.S. (DellaVigna and Pope, 2018).

For each group I first calculate the average predicted effects of messages which were grouped by research assistants into nine topics following a preregistered protocol (see Appendix G for detail). I then calculate the correlation between the predicted and experimentally estimated effects. If the results are driven by local knowledge, local forecasters in Kenya should show a higher correlation than nonlocal forecasters in the US. On the other hand, if the wisdom of the crowds drives the results regardless of forecasters' origins, we should observe similar levels of accuracy in both samples. In Panel A of figure 5 we see that forecasts from local participants show a strong correlation with the observed experimental effects (cor = 0.69, $\text{CI}_{95\%}$ = [0.33, 0.85]), whereas in Panel B predictions from nonlocal forecasters are negatively correlated with these effects (cor=-0.19, $\text{CI}_{95\%}$] = [0.53, 0.20]). This suggests that local knowledge plays a central role in the crowdsourcing process (methods described in Appendix G and Appendix table Table A10 provides robustness checks).

3 Discussion

Foundational theories in economics, sociology, and political science suggest that people possess valuable knowledge about local beliefs, preferences, and constraints. Motivated by this idea, this paper provides a large-scale experimental test of a new method for producing policy interventions, which involves (*i*) crowdsourcing a large set of interventions from individuals with local contextual knowledge, and (*ii*) using incentive compatible local forecasts of the causal effects of these nudges to prioritize interventions for testing.

My results highlight the value of *crowd-selection* in the crowdsourcing process: the average crowdsourced nudge is ineffective and if anything reduces opt-in to 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 44.0% (0.41 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. My findings clarify the important role of local knowledge in the crowd-selection process, as shown by the fact that only local predictions were positively correlated with experimental estimates of treatment effects.

Formalizing the crowdsourcing process into two stages—developing a choice set and selecting interventions from this choice set—suggests 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), testing different methods for eliciting predictions (DellaVigna et al., 2020) or allowing for communication between forecasters

(Becker et al., 2017) may result in 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.

4 Materials and methods

Study 1 design. 1,822 Kenyan participants recruited via Facebook passed pre-treatment screening (see Appendix C) and pre-registered exclusion criteria (see Table A4 for details on exclusion criteria in all five studies).

- **Performance incentives.** Participants were randomly assigned to one of three experimental conditions with performance incentives of 0, 4, or 10 Kenyan Shillings for each message recipient who signed up for the COVID-19 notification service. Almost all participants (99.23%) passed a comprehension check about the magnitude of their incentives. Treatment groups were balanced on covariates, as shown in Table A5.

Study 2 design. 1,360 Kenyan participants recruited over Facebook passed pre-registered exclusion criteria and provided predictions for 1,496 nudges from Study 1 (see Appendix D for details).

- **Forecast elicitation.** Participants predicted the effects of randomly selected messages on a slider scale bounded at 0 and 3, using a benchmark opt-in rate for the control group (1%) as a reference. Bonus payments were given based on prediction accuracy (see Appendix D). Panel B of Figure 1 shows the slider scale used to make predictions.
- **Aggregation and policy choice.** I pre-registered (see Table A4) 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 D).

Study 3 design. In Study 3, messages were sent to 40,911 participants randomly selected from the Busara Center for Behavioral Economics' sample pool via a five-digit SMS shortcode. The control text was identical to the crowdsourced text except it did not include an experimental message. See Appendix F for an example invitation. Each participant was assigned to receive only one experimental message, and then was given an opportunity to opt into the COVID-19 notification service. A reminder was sent if no response was received after 6 hours.

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 35,183 participants and were balanced across experimental conditions (see Table A8). Details on the benchmark messages and crowdsourced nudges can be found in Appendix F.

Study 5 design. 1,146 new Kenyan participants were recruited over Facebook and passed pre-registered exclusion criteria. A second sample of 1,138 forecasters from the U.S. were recruited over Amazon Mechanical Turk. The set of predicted nudges, exclusion criteria, forecast elicitation, and forecast aggregation were the same as in Study 2. See Appendix G for details.

Empirical strategy

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

$$Opt\ in_i \times 100 = \alpha + \beta \mathbf{T}_i + \varepsilon_i \tag{1}$$

where $Opt\ In_i \in \{0, 1\}$ takes a value of 1 if recipient i opts in to the COVID-19 notification service, α is the average opt-in rate in the control group, and ε_i is the error term. $\mathbf{T}_i$ is a vector of dichotomous treatment variables randomly assigned at the individual level and which varies by study, and β is the corresponding vector of regression coefficients.

- **Average effect of crowdsourced nudges (Study 3)**: T_i is a single variable equal to 1 if i receives a crowdsourced message (see Panel A of Table A6).
- **Effect of incentives (Study 3):** $\mathbf{T}_i$ is a vector of three dichotomous variables each equal to 1 if 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 (see Panel B of Table A6, or Table A7 for robustness checks).
- **Effect of crowd-selected and benchmark messages (Study 4)**: $\mathbf{T}_i$ is a vector of two dichotomous variables equal to 1 if i received one of the (a) ten crowdsourced messages 1 or (b) three benchmark messages. I also provide estimates at the message level. Alternative specifications control for recipient gender (see Table A9).

All estimation uses robust standard errors. When estimating the effects of individual crowd-selected nudges and benchmark nudges in Study 4 I also provide adjusted p-values which control for the false discovery rate (Benjamini and Hochberg, 1995), while noting that my substantive focus is on the pooled effects of these interventions.

Local versus nonlocal predictions (Study 5). In Study 5, I estimate the correlation between predicted and experimentally estimated effects of messages which were grouped into nine topics according to a pre-registered protocol (see Table A4). I use a bootstrap procedure to measure the stability of this correlation under alternative weighting and estimation strategies and for Spearman's rank-order correlation in Table A10. Bootstrapped confidence intervals are calculated by first generating bootstrapped crowds of forecasters by sampling n forecasters from the full sample of local forecasters ($n = 1,146$) and forecasters from the nonlocal forecasters ($n = 1,138$). For each crowd, I calculate the average predicted effect for each message, and aggregate predictions at the topic level. Finally, I calculate the average correlation between the predicted and observed effects of message topics. I repeat this procedure 20,000 times and use these bootstrapped distributions to define 95% confidence intervals around the estimated correlation.

References

Almaatouq, A., Becker, J., Bernstein, M., Botto, R., Bradlow, E., Damer, E., Duckworth, A., Griffiths, T., Hartshorne, J., Law, E., et al. (2021). Scaling up experimental social, behavioral, and economic science.

Arrow, K. J. (1971). The economic implications of learning by doing. In *Readings in the Theory of Growth*, pages 131–149. Springer.

Banerjee, A., Chandrasekhar, A. G., Dalpath, S., Duflo, E., Floretta, J., Jackson, M. O., Kannan, H., Loza, F. N., Sankar, A., Schrimpf, A., et al. (2021). Selecting the most effective nudge: Evidence from a large-scale experiment on immunization. Technical report, National Bureau of Economic Research.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., and Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, 348(6236):1260799.

Bates, M. A. and Glennerster, R. (2017). The generalizability puzzle. *Stanford Social Innovation Review*, 15(3):50–54.

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.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Blair, G., Weinstein, J. M., Christia, F., Arias, E., Badran, E., Blair, R. A., Cheema, A., Farooqui, A., Fetzer, T., Grossman, G., et al. (2021). Community policing does not build citizen trust in police or reduce crime in the global south. *Science*, 374(6571):eabd3446.

Brown, J. S. and Duguid, P. (1991). Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation. *Organization science*, 2(1):40–57.

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.

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

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

DellaVigna, S., Otis, N., and Vivalt, E. (2020). Forecasting the results of experiments: Piloting an elicitation strategy. In *AEA Papers and Proceedings*, volume 110, pages 75–79.

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

Duckworth, A. L. and Milkman, K. L. (2022). A guide to megastudies. *PNAS Nexus*, 1(5):pgac214.

Dunning, T., Grossman, G., Humphreys, M., Hyde, S. D., McIntosh, C., Nellis, G., Adida, C. L., Arias, E., Bicalho, C., Boas, T. C., et al. (2019). Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials. *Science advances*, 5(7):eaaw2612.

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

Fischer, F. (2000). *Citizens, experts, and the environment: The politics of local knowledge.* Duke University Press.

Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451.

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

Hayek, F. (1945). The use of knowledge in society. *The American Economic Review*, 35(4):519–530.

Kahneman, D., Sibony, O., and Sunstein, C. R. (2021). *Noise: A flaw in human judgment.* Little, Brown.

Kenyan Ministry of Health (2019). Utilizing the community health strategy to respond to covid 2019. Retrieved from `https://www.health.go.ke/wp-content/uploads/2020/04/Community-Response-to-COVID-2019_1.docx.pdf`.

Kenyan Ministry of Health (2020). National communication and community engagement strategy for coronavirus. Retrieved from `https://thecompassforsbc.org/project-examples/kenya-national-communication-and-community-engagement-strategy-coronavirus`.

Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., Turkay, S., Williams, J. J., and Tingley, D. (2020). Scaling up behavioral science interventions in online education. *Proceedings of the National Academy of Sciences*, 117(26):14900–14905.

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).

March, J. and Simon, H. (1955). Organizations.

Milkman, K. L., Gandhi, L., Patel, M. S., Graci, H. N., Gromet, D. M., Ho, H., Kay, J. S., Lee, T. W., Rothschild, J., Bogard, J. E., et al. (2022). A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proceedings of the National Academy of Sciences*, 119(6):e2115126119.

Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., Park, Y., Rai, A., Bazerman, M., Beshears, J., et al. (2021). Megastudies improve the impact of applied behavioural science. *Nature*, 600(7889):478–483.

Organization, W. H. et al. (2017). *Communicating risk in public health emergencies: a WHO guideline for emergency risk communication (ERC) policy and practice*. World Health Organization.

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

Pan, S. W., Stein, G., Bayus, B., Tang, W., Mathews, A., Wang, C., Wei, C., and Tucker, J. D. (2017). Systematic review of innovation design contests for health: spurring innovation and mass engagement. *BMJ innovations*, 3:227.

Rousseau, J.-J. (1782). *On the Social Contract.* Launette Aux Deux-Ponts: Chez Sanson Et Compagnie.

Slough, T., Rubenson, D., Levy, R., Alpizar Rodriguez, F., Bernedo del Carpio, M., Buntaine, M. T., Christensen, D., Cooperman, A., Eisenbarth, S., Ferraro, P. J., et al. (2021). Adoption of community monitoring improves common pool resource management across contexts. *Proceedings of the National Academy of Sciences*, 118(29):e2015367118.

Surowiecki, J. (2005). *The wisdom of crowds.* Anchor.

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.

Tang, W., Mao, J., Liu, C., Mollan, K., Zhang, Y., Tang, S., Hudgens, M., Ma, W., Kang, D., Wei, C., et al. (2019). Reimagining health communication: a non-inferiority randomized controlled trial of crowdsourced intervention in china. *Sexually transmitted diseases*, 46(3):172.

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

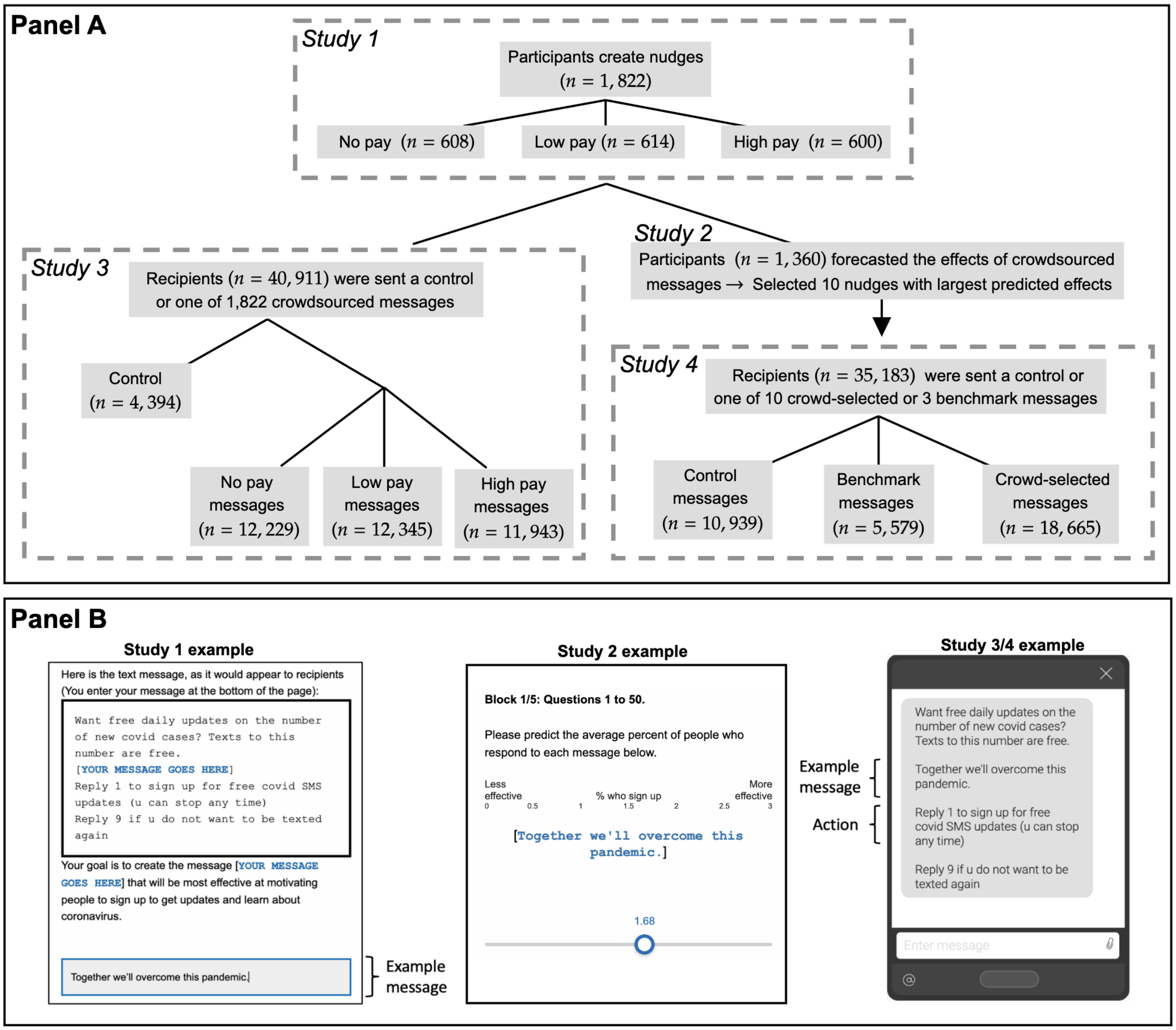
Thaler, R. H. (2018). Nudge, not sludge.

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.

Tumpey, A. J., Daigle, D., and Nowak, G. (2018). Communicating during an outbreak or public health investigation. *The CDC field epidemiology manual*, pages 243–259.

Zhang, Y., Kim, J. A., Liu, F., Tso, L. S., Tang, W., Wei, C., Bayus, B. L., and Tucker, J. D. (2015). Creative contributory contests (ccc) to spur innovation in sexual health: Two cases and a guide for implementation. *Sexually transmitted diseases*, 42(11):625.

Figure 1: Overview of main studies



Panel A

Study 1

Participants create nudges (n = 1,822)

No pay (n = 608)

Low pay (n = 614)

High pay (n = 600)

Study 2

Participants (n = 1,360) forecasted the effects of crowdsourced messages → Selected 10 nudges with largest predicted effects

Study 3

Recipients (n = 40,911) were sent a control or one of 1,822 crowdsourced messages

Control (n = 4,394)

No pay messages (n = 12,229)

Low pay messages (n = 12,345)

High pay messages (n = 11,943)

Study 4

Recipients (n = 35,183) were sent a control or one of 10 crowd-selected or 3 benchmark messages

Control messages (n = 10,939)

Benchmark messages (n = 5,579)

Crowd-selected messages (n = 18,665)

Panel B

Study 1 example

Here is the text message, as it would appear to recipients (You enter your message at the bottom of the page):

```
Want free daily updates on the number
of new covid cases? Texts to this
number are free.
[YOUR 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
```

Your goal is to create the message [YOUR MESSAGE GOES HERE] that will be most effective at motivating people to sign up to get updates and learn about coronavirus.

Together we'll overcome this pandemic.

Study 2 example

Block 1/5: Questions 1 to 50.

Please predict the average percent of people who respond to each message below.

[Together we'll overcome this pandemic.]

Study 3/4 example

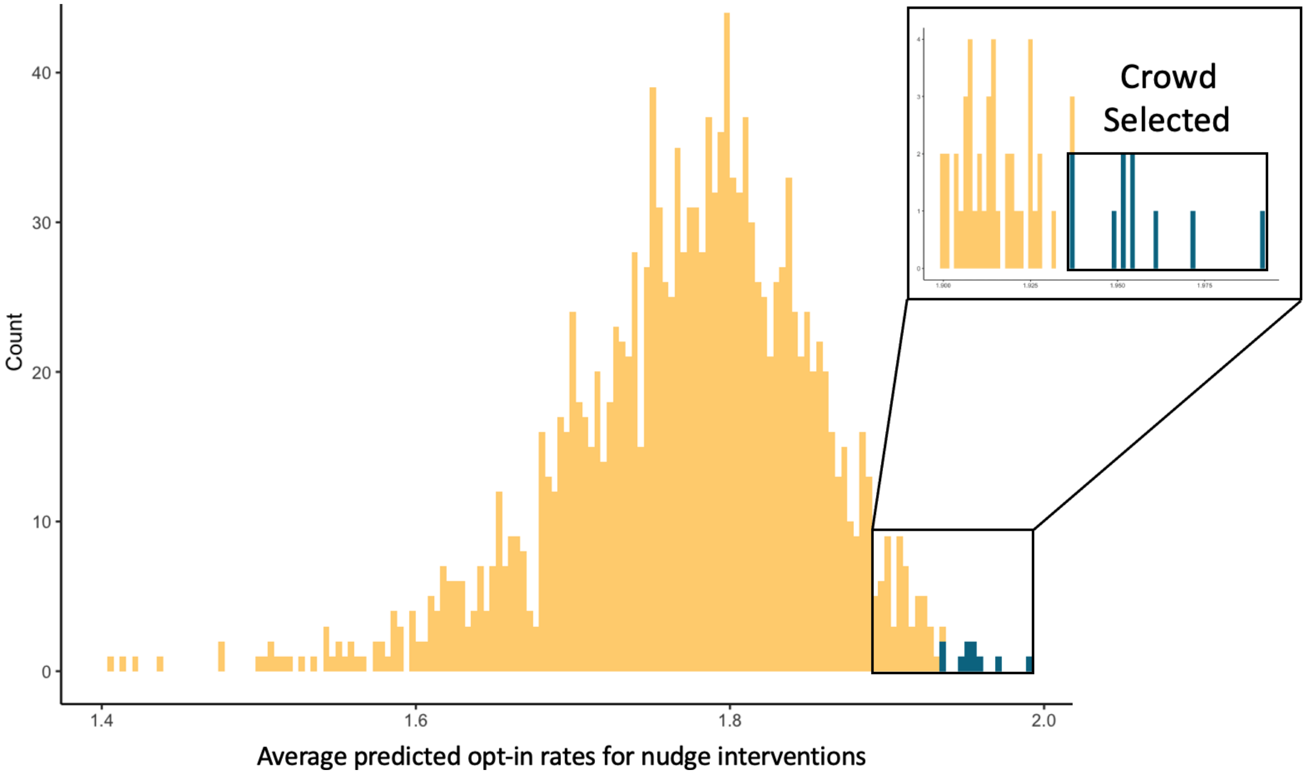
Want free daily updates on the number of new covid cases? Texts to this number are free.

Together we'll overcome this pandemic.

Reply 1 to sign up for free covid SMS updates (u can stop any time)

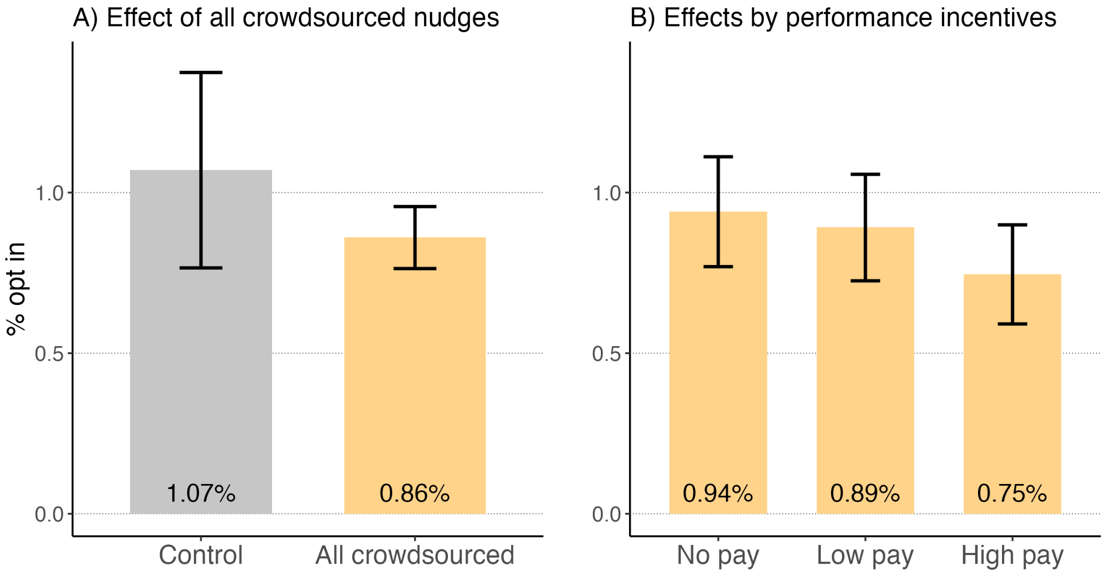
Reply 9 if u do not want to be texted again

Figure 2: Crowd selection (Study 2)



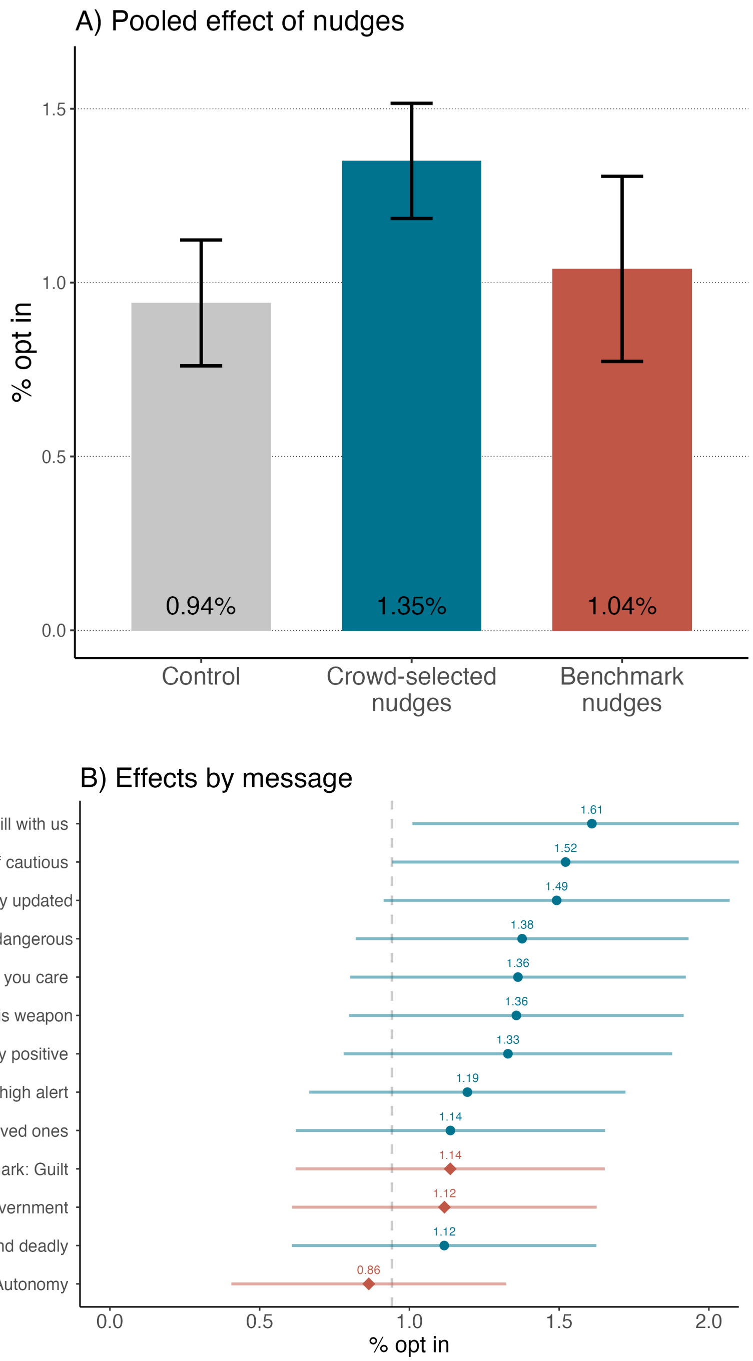
This figure displays the distribution of average predicted effects for 1,496 nudges from Study 2. 1,360 forecasters provided a total of 324,160 predictions. The ten "crowd-selected" nudges with the largest predicted effects are displayed in blue.

Figure 3: Effects of crowdsourced nudges (Study 3)



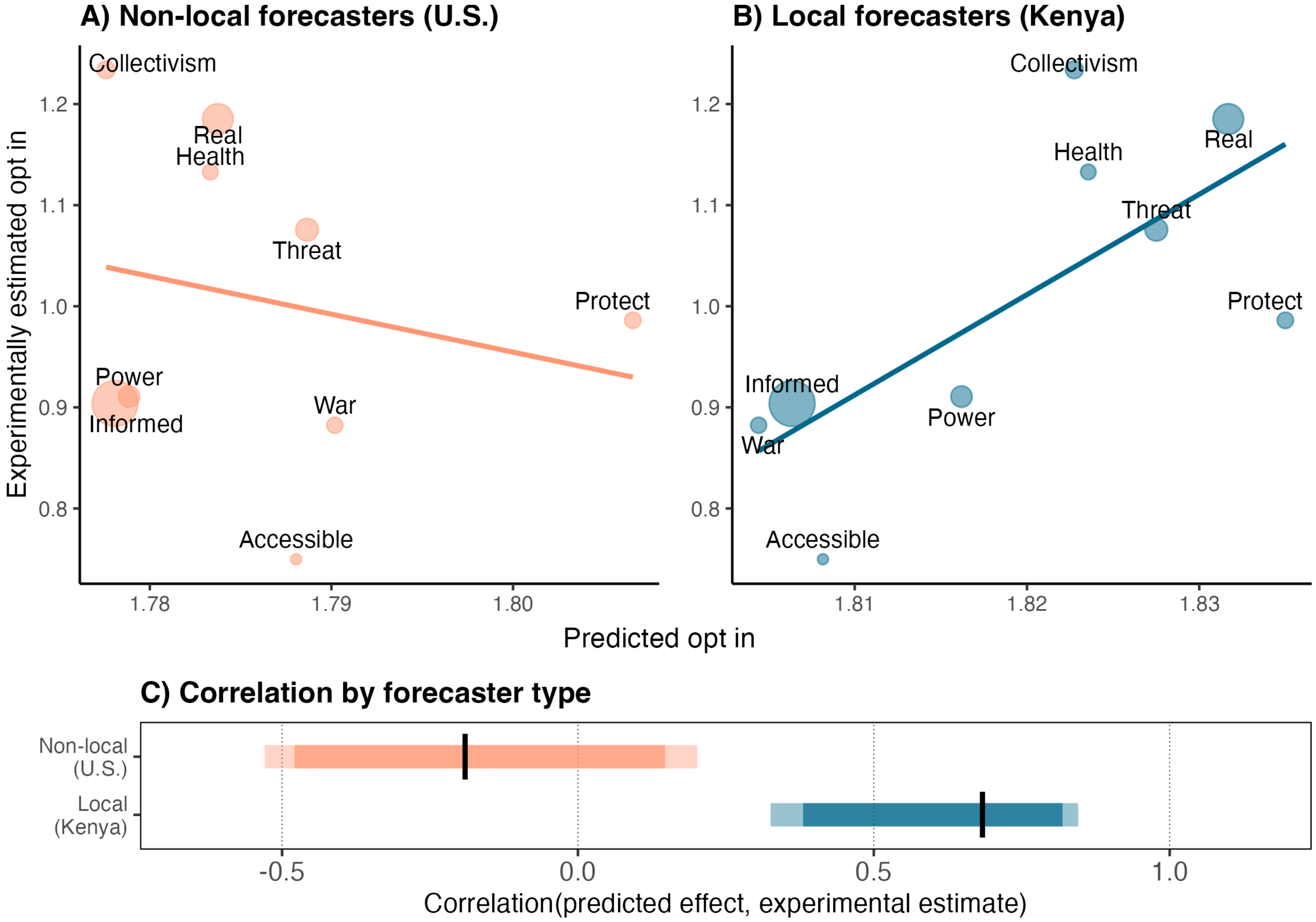
Panel A displays average opt-in rates for recipients who receive a control message ($n = 4,394$) or one of 1,822 crowdsourced nudges ($n = 36,517$). Panel B depicts the average opt-in rates by randomly assigned performance incentives for producing effective nudges, ($n_{\text{no pay}} = 12,229$; $n_{\text{low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

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



Panel A displays average opt-in rates for recipients who receive a control message ($n = 10,939$), the ten crowd-selected nudges that participants predicted would have the highest causal effect ($n = 18,665$), or and the three benchmark nudges ($n = 5,579$). 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 around the conditional mean for each group (Panel A) and the estimated causal effect of each message (Panel B).

Figure 5: Topic forecasts (Study 5)



Panels A and B present linear models of the relationship between the average predicted effect of messages grouped by topic (x-axis) and the average experimentally estimated effect by message topic. For details on topics, see Appendix G. In Panel A 1,146 participants from Kenya provide 155,507 forecasts, and in Panel B 1,138 participants from the nonlocal forecasters (from the U.S.) provide 154,210 forecasts. Following pre-registration, this analysis does not include messages that do not belong to any of the nine pre-registered topics. Point size denotes the number of recipients per topic. Panel C depicts the correlation between predicted and observed message effects by topic. Dark and light bars are 90% and 95% confidence intervals, which were generated by taking 20,000 bootstrapped samples of local and nonlocal forecasters and calculating the correlation with the average predicted experimental effect by topic for each bootstrapped sample. Table A10 provides several robustness checks.

Online Appendix

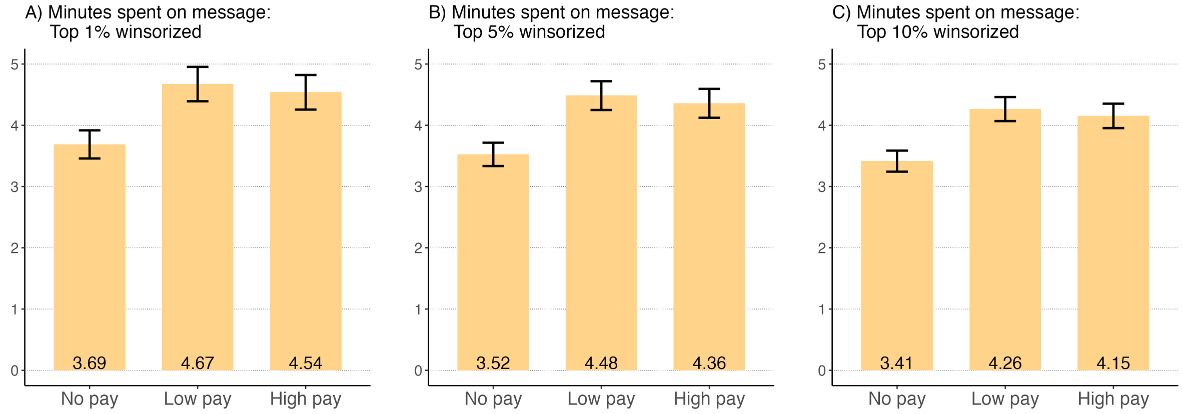
The Efficacy of Crowdsourced Nudges:

Experimental Evidence

Nicholas G. Otis

A Appendix figures

Figure A1: Effects of incentives on message time (Study 1)



This figure depicts the effects of randomly assigned incentives faced by the 1,822 message producers on time spent creating a message in minutes. Number of minutes is winsorized at the top 1%, 5% and 10% level in panels A, B, and C respectively. The low-pay ($n = 614$) and high-pay ($n = 600$) conditions received a bonus of four and ten Kenyan Shillings for each randomly assigned recipient that received their message and opted into the notification service. The no-pay condition ($n = 608$) received no performance incentives. Error bars display 95% confidence intervals.

B Appendix tables

Table A1: Overview of messages

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 invitation
Repeating invitation: (a) Want to receive free covid updates? (b) Get registered for daily covid sms updates by replying with 1.
Health tips: (a) Wash your hands and sanitize at all times. (b)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 weapon. Knowledge is power, and knowledge about coronavirus is the first weapon in fighting the disease.
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: 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: Summary of samples sizes (Studies 1-5)

	Participants (1)	Messages produced/received/forecasted (2)	Forecasts (3)
Study 1			
Message producers	1822	1822	
Study 2			
Local forecasters (Kenyan)	1360	1496	324160
Study 3			
Message recipients (Pooled)	40911		
Message recipients (Control)	4394	1	
Message recipients (Crowdsourced)	36517	1822	
Study 4			
Message recipients (Pooled)	35183		
Message recipients (Control)	10939	1	
Message recipients (Benchmark)	5579	3	
Message recipients (Crowd selected)	18665	10	
Study 5			
Local forecasters (Kenyan)	1146	1496	280770
Non-local forecasters (U.S.)	1138	1496	278810

This table provides details on the sample sizes for studies 1-5. Column 1 presents the number of participants (message producers in Study 1, forecasters in studies 2 and 5, and recipients in studies 3 and 4. Column 2 displays the number of messages produced (in Study 1), received (in Study 2) and the number of forecasts (Column 3).

Table A3: Topic coding

Name (Abbreviation)	Keywords/phrases	Example message	n_{messages}
(1)	(2)	(3)	(4)
Threat	kill; death, deadly; die; danger; dangerous	Covid kills, follow instructions.	122
Covid is real (Real)	covid is real; still {here,there,with us}; exists	COVID-19 is still with us.	169
Accessible info (Accessible)	easy; comfort; fingers; fingertips; instant	Information on COVID at your fingertips.	53
Collectivism	together; unite; join hands	Only united can we beat this pandemic.	93
Protect others (Protect)	protect; save lives; save a life; loved ones safe; families safe	Come learn more and save lives.	85
War	war; fight; battle; enemy; weapon; conquer	Let's fight this pandemic once and for all.	83
Be informed (Informed)	{be,stay,keep} {informed,in the know,updated}	Always be in the know.	257
Knowledge is power (Power)	{knowledge,information} is power	Knowledge is power.	114
Health is important (Health)	health matters; health is wealth	Health is wealth.	80

This table lists the nine message topics that were identified by research assistants following a preregistered protocol (see Table A4 for links to study preregistration materials)). Column 1 presents abbreviations used in figures in and tables. Column 2 lists common phrases or keywords in each topic. Column 3 provides an example message. Column 4 provides a count of the number of messages in each topic. For details on the topic coding procedure see Appendix G.

Table A4: Preregistration overview

Study	Outcomes		Exclusion criteria		Other		Preregistration
Study 1	Opt in	✓	Producer	✓	None		https://www.socialscienceregistry.org/trials/6414
	Time	X	Message	✓			https://www.socialscienceregistry.org/versions/167037/docs/version/file
							https://www.socialscienceregistry.org/versions/167049/docs/version/file
Study 2	Predicted opt in	✓	Forecaster	✓	# chosen messages	✓	https://aspredicted.org/PGF_X7V
Study 3	Opt in	✓	None		None		https://www.socialscienceregistry.org/trials/6414
Study 4	Opt in	✓	None		None		https://aspredicted.org/PGF_X7V
Study 5	Predicted opt in	✓	Forecaster	✓	Topic coding	✓	https://aspredicted.org/CNP_8LQ https://www.socialscienceregistry.org/versions/167046/docs/version/file

This table summarizes preregistration details from Studies 1 to 5. Check marks (✓) denote that a outcome, exclusion criteria, or other study feature was preregistered. The final column contains a link to the preregistration page and documents.

Table A5: Sociodemographics and balance (Study 1)

		By incentive condition			
	Full sample (1)	No pay (2)	Low pay (3)	High pay (4)	F-stat. (5)
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)	
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 in Kenyan Shillings +1)	8.01	8.07	7.99	7.95	0.27
	(0.07)	(0.12)	(0.12)	(0.12)	
C) Post-treatment variables					
% messages with tips or repeating 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 across randomly assigned incentive conditions. Panel A lists displays balance on two pre-registered stratification variables: *Borderline message* is equal to 1 if participants create a test message in a screening survey that only marginally passed pre-registered message rules (see Appendix C 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 A6: 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
A) Average crowdsourced nudges				
All crowdsourced nudges	-0.21 (0.16)	0.20	1822	36517
B) Effects by incentive condition				
No financial incentives	-0.13 (0.18)	0.47	608	12229
Low financial incentives	-0.18 (0.18)	0.31	600	12345
High financial incentives	-0.32 (0.17)	0.06	614	11943
C) Financial incentives				
Payment of 5 Kenyan Shillings	0.96 (0.44)	0.03		1180

Panel A reports the average effect of 1,822 crowdsourced nudges. Panels B pools messages by randomly assigned incentive conditions. Panel C looks at the effect financial incentives for opting into the notification service. Robust standard errors are presented in parentheses.

Table A7: Robustness check on effect of incentives (Study 3)

	Effect on opt in (pp)	p-value	Effect on opt in (pp)	p-value	n_{messages}	$n_{\text{recipients}}$
	(1)	(2)	(3)	(4)	(5)	(6)
A) All crowdsourced nudges						
Reference: No financial incentives	0.94				608	12229
Low financial incentives	-0.05	0.68	-0.05	0.69	600	12345
	(0.12)		(0.12)			
High financial incentives	-0.20	0.10	-0.19	0.10	614	11943
	(0.12)		(0.12)			
B) Excluding nudges with tips or repeating invitation						
Reference: No financial incentives	0.91				497	9994
Low financial incentives	-0.02	0.88	-0.02	0.89	499	10108
	(0.13)		(0.13)			
High financial incentives	-0.12	0.37	-0.12	0.37	500	9960
	(0.13)		(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. Robust standard errors are presented in parentheses.

Table A8: Recipient balance (Study 4)

	% female (1)	$n_{\text{recipients}}$ (2)
Control mean	56.11	10939
	(0.47)	
Benchmark messages		
Benchmark: Government	55.90	1-880
	(1.15)	
Benchmark: Guilt	55.14	1848
	(1.16)	
Benchmark: Autonomy	57.48	1851
	(1.15)	
Crowd-selected messages		
Still with us	54.77	1864
	(1.15)	
Contagious and dangerous	55.27	1889
	(1.14)	
Infectious and deadly	55.50	1881
	(1.15)	
Aware means you care	54.33	1835
	(1.16)	
Stay positive	55.72	1881
	(1.15)	
Knowledge is weapon	56.41	1842
	(1.16)	
Stay updated	55.03	1877
	(1.15)	
Stay on high alert	56.70	1843
	(1.15)	
Protect loved ones	55.82	1847
	(1.16)	
Safe if cautious	55.14	1906
	(1.14)	
F-statistic		0.6
n_{total}		37035

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

Table A9: Effects of crowd-selected nudges

	Effect on opt in (pp)	p-value	Effect on opt in (pp)	p-value	$n_{\text{recipients}}$
	(1)	(2)	(3)	(4)	(5)
Reference: Control mean	0.94				10939
A) Pooled effects					
Benchmark nudges	0.10	0.55	0.10	0.56	5579
	(0.16)		(0.16)		
Crowd-selected nudges	0.41	0.00	0.40	0.00	18665
	(0.13)		(0.12)		
B) Message effects					
Benchmark: Government	0.18	0.50	0.17	0.51	1880
	(0.26)	[0.54]	(0.26)	[0.56]	
Benchmark: Guilt	0.19	0.46	0.19	0.48	1848
	(0.26)	[0.54]	(0.26)	[0.56]	
Benchmark: Autonomy	-0.08	0.74	-0.07	0.76	1851
	(0.23)	[0.74]	(0.23)	[0.76]	
Still with us	0.67	0.03	0.65	0.03	1864
	(0.31)	[0.2]	(0.31)	[0.23]	
Contagious and dangerous	0.43	0.13	0.43	0.13	1889
	(0.28)	[0.29]	(0.28)	[0.3]	
Infectious and deadly	0.17	0.50	0.17	0.52	1881
	(0.26)	[0.54]	(0.26)	[0.56]	
Aware means you care	0.42	0.14	0.41	0.15	1835
	(0.29)	[0.29]	(0.29)	[0.3]	
Stay positive	0.39	0.17	0.38	0.17	1881
	(0.28)	[0.29]	(0.28)	[0.3]	
Knowledge is weapon	0.42	0.14	0.42	0.14	1842
	(0.29)	[0.29]	(0.28)	[0.3]	
Stay updated	0.55	0.06	0.54	0.07	1877
	(0.29)	[0.22]	(0.29)	[0.23]	
Stay on high alert	0.25	0.35	0.25	0.34	1843
	(0.27)	[0.54]	(0.27)	[0.54]	
Protect loved ones	0.20	0.46	0.20	0.46	1847
	(0.26)	[0.54]	(0.26)	[0.56]	
Safe if cautious	0.58	0.05	0.57	0.05	1906
	(0.30)	[0.22]	(0.30)	[0.23]	
C) Financial incentives					
Payment of 5 Kenyan Shillings	1.65	0.00	1.66	0.00	1852
	(0.38)	[0]	(0.38)	[0]	
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 individual messages. Robust standard errors are presented in parentheses. Because Panels B and C involve a total of 14 experimental comparisons, I present adjusted in square brackets which denote the smallest false discovery rate that the null hypothesis of no treatment effect is rejected under Benjamini and Hochberg (1995).

Table A10: Forecast accuracy among local and nonlocal forecasters

	Correlation(predicted effect, experimental estimate)			
	Pearson		Spearman	
	Kenya	U.S.	Kenya	U.S.
	(1)	(2)	(3)	(4)
A) Mean forecast				
Unweighted	0.69	-0.19	0.65	-0.27
	(0.33,0.85)	(-0.53,0.20)	(0.80,0.82)	(-0.60,0.27)
Weighted	0.75	-0.04	0.74	-0.03
	(0.84,0.86)	(-0.39,0.30)	(0.84,0.86)	(-0.61,0.34)
B) OLS forecast				
Unweighted	0.76	-0.37	0.73	-0.47
	(0.86,0.88)	(-0.62,0.01)	(0.85,0.88)	(-0.62,0.08)
Weighted	0.81	-0.19	0.79	-0.27
	(0.88,0.90)	(-0.50,0.16)	(0.87,0.90)	(-0.62,0.23)
$n_{forecasters}$	1138	1146	1138	1146
$n_{forecasts}$	154210	155507	154210	155507

This table presents the correlation between predicted and experimentally estimated effects of messages grouped at the topic level. For details on topics, see Appendix G. Following pre-registration, this analysis does not include messages that do not belong to any of the nine pre-registered topics. In Panel A, forecasts take the form of the average predicted experimental effect of messages in a given topic, and experimental estimates are the average opt-in rates in a given topic. In Panel B, I regress forecasts and opt-in rates on a vector of nine dummy variables (one for each topic) and estimate the correlation between predicted and observed coefficients (messages may belong to more than one topic). Weighted correlations weight topics by the number of recipients who receive messages in that topic, and unweighted correlations give each topic equal weight. Results for local forecasters (from Kenya) are presented in columns 1 and 3, and results for forecasters from the U.S. are presented in columns 2 and 4. Columns 1 and 2 present Pearson correlation coefficients, and columns 3 and 4 present Spearman correlation coefficients. Parentheses present bootstrapped 95% confidence intervals, which were generated by taking 20,000 bootstrapped samples of forecasters in each group and calculating the average predicted experimental effect by topic for each bootstrapped sample.

C 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; see Table A4 for links to study preregistration materials) 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, and they were informed that this message would not be distributed. Participants were also 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, adherence to these messages was independently evaluated by research assistants. Rules 1 and 3 are designed to avoid misinformation and deceit. Rule 2 is designed to avoid a misconception about the purpose of the messages that was 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 from simply re-typing the control message text.

Researcher assistants classified messages 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 violate a rule. Borderline messages were independently reviewed by an additional research assistant. If the message was classified as *borderline* or *pass* during this independent review stage, the message is included. Otherwise, the message is excluded.

Stratification. I pre-registered that random assignment to the three different incentive conditions would 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 our research team recontacted participants over email and text message with an invitation to the main survey where they are asked to create a message.

Message exclusion criteria. The following pre-registered exclusion criteria were applied to messages produced by participants in the follow-up survey who designed a message. Adherence to these criteria were assessed by research assistants who were blind to the experimental condition the participant was assigned to (the examples below were also 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 were 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 incorrectly will be corrected. For example, in the message *For peace of mind, text* *back 0 to begin receiving updates.* the respondent incorrectly listed 0 as the number the recipient needed to text to receive notifications instead of the correct reply option of 1.

D Study 2 details

Recruitment. I recruited a new sample of Kenyan participants over Facebook. After applying pre-registered a attention check (AsPredicted #106631; see Table A4 for links to study preregistration materials)[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 (in Kenyan Shillings) for one randomly selected forecast based on the equation

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

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 set of pre-registered exclusion criteria (AEARCTR-0006414). However, I also included a list of Rules (see Appendix C) 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 provides examples of excluded messages, and Panel B of Table A7 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).

[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. See Table A4 for links to study preregistration materials.

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, see Table A4 for links to study preregistration materials).

E 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 network failures 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.
[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 [`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, Kakamega1, Kilifi 1, and Mombasa 1.`").

F Study 4 details

Undelivered messages. Messages were sent to (35,183+18,878)=54,061 participants, of which 18,878 messages were not received by study participants (a campaign from the

Kenyan government to reduce the number of active mobile numbers in Kenya likely accounts for the difference in failure rates between Study 3 and 4, as mentioned on the study pre-registration page (see Table A4 for links to study preregistration materials). As with Study 3 this represents pre-treatment attrition, and these failed texts 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 "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."

G Study 5 details

Study five has two stages. The first stage involves identifying topics among the same set of 1,496 messages used in Study 2. In the second stage, I collected two new samples of forecasts of the causal effects of these messages from participants in (*i*) the U.S. and (*ii*) Kenya. The protocol used to identify topics, and the analysis of forecasts of messages which were grouped into topics were both pre-registered (see Table A4 for links to study preregistration materials).

Topic coding. Message topics were identified by a team of Kenyan research assistants following a pre-registered topic coding protocol. Research assistants independently read through messages, identifying common topics, themes, or phrases. They then discussed these patterns and converged on a set of nine groups of messages (they were to come up with about ten groups) following preregistration. Next, they classified whether each message belonged to each topic. Details on message topics and keywords can be found in Table A3.

Recruitment (Kenya). My sample consists of a new set of $1,146$ local forecasters (from Kenya) (after applying pre-registered exclusion criteria), who were recruited and made predictions following the same procedure as Study 2. These forecasters made a total of $280,770$ forecasts, of which 155,507 forecasts of messages assigned to one of the nine topic groups.

Recruitment (U.S.). My sample of US participants were recruited over Amazon Mechanical Turk. These participants faced the same pre-registered exclusion criteria as the Kenyan sample and faced accuracy incentives based on a convex loss function similar to the one used in Study 2, but which provided a bonus of up to \$0.27. In total 1,138 forecasters provided 278,810 forecasts, of which 154,210 forecasts were of messages assigned to one of the nine topic groups.