
Supplementary information

Evidence from a statewide vaccination RCT shows the limits of nudges

In the format provided by the
authors and unedited

Supplementary Information

SI.1. Message creation

Despite intense research interest in facilitators of and factors associated with vaccine uptake over the last several decades, many systematic reviews find a lack of extensive evidence linking messaging interventions to actual vaccination.¹ If tried and true strategies for promoting seasonal flu, H1N1, HPV, or childhood vaccines had emerged from the literature, we might simply adapt them to the context of COVID-19 despite the pandemic's unique features (e.g., its unprecedented reach or unusual degree of politicization). Since they do not, we have conducted a two-step review process for generating candidate messages (see Figure S1). First, we searched for individual message test studies and judged the strength and relevance of the evidence using the following criteria, with the second in each pair indicating better evidence than the first: (i) measured outcome: self-reported/projected behavior (e.g., intention to get vaccinated) versus actual behavior (measured vaccine uptake); (ii) context: seasonal flu/HPV/other vaccination context versus COVID-19; (iii) random message assignment: no versus yes. Studies that did not meet at least one of the preferred criteria are not included. Most studies were conducted in the United States. Second, we considered the survey work on various demographic groups' concerns about COVID-19 vaccination (which are not uniform) to ensure that some messages in the set conceivably address these concerns. This process generated eight candidate messages, described below and listed in Table 1. The content of the messages was discussed with and approved by RIDOH's communications team and medical directors. Slight changes to wording were necessary to keep readability scores at a desirable level.

Although there is a lack of extensive evidence on the efficacy of particular message strategies, we note that a number of studies find increased intended² and actual³ vaccine uptake due to *any* message at all relative to control (no message) conditions. No studies to our

¹ "We were surprised to find that few randomized trials have successfully changed what people think and feel about vaccines, and those few that succeeded were minimally effective in increasing uptake" ([Brewer et al., 2018](#)). "Given the paucity of information on effective strategies to address vaccine hesitancy, when interventions are implemented, planning a rigorous evaluation of their impact on vaccine hesitancy/vaccine acceptance will be essential" ([Dubé et al., 2015](#)). "...given the complexity of vaccine hesitancy and the limited evidence available on how it can be addressed, identified strategies should be carefully tailored according to the target population, their reasons for hesitancy, and the specific context" ([Jarrett et al., 2015](#)). "More high quality research is needed to demonstrate the effects of messaging interventions on actual vaccine uptake" ([Lawes-Wickwar et al., 2021](#)). "Overall, there is a lack of good-quality primary studies [on risk messages], and existing interventions are suboptimal" ([Parsons et al., 2018](#)).

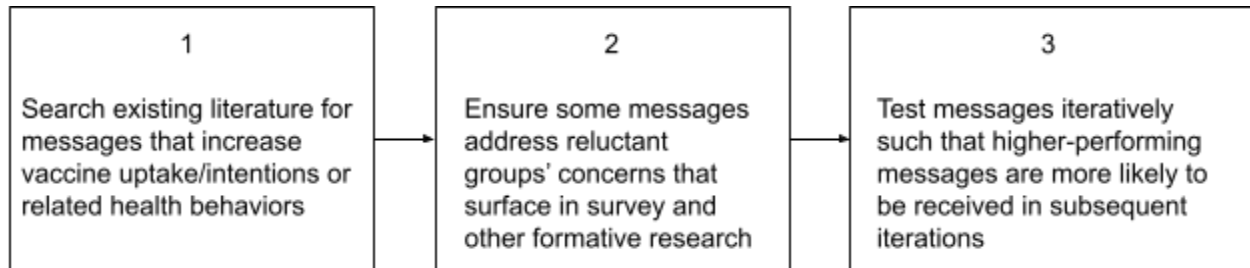
² [Argote et al., 2021](#). Messaging interventions that increase COVID-19 vaccine willingness in Latin America.

³ [Esteban-Vasallo et al., 2019](#). Effect of mobile phone text messaging for improving the uptake of influenza vaccination in patients with rare diseases. [Lee et al., 2020](#). Large-scale influenza vaccination promotion on a mobile app platform: A randomized controlled trial. [Regan et al., 2017](#). Randomized controlled trial of text message reminders for increasing influenza vaccination. [Yokum et al., 2018](#). Letters designed with behavioural science increase influenza vaccination in Medicare beneficiaries. But see [OES, 2021](#). Low-cost interventions to increase vaccination uptake.

knowledge show depressed rates of vaccination due to message reminders. Thus, while some messages may be more effective than others, there was no indication that any would have negative effects on vaccine uptake.

For the complete list of messages used, see Extended Data Table 1.

Figure S1. Message selection process.



SI.2. Message reweighting (adaptive design)

In assigning individual treatments, we employ the following strategy. First, for each week w , we draw N_w individuals from our remaining pool uniformly at random. For each of these N_w people, with probability ϵ_w , we assign them uniformly at random to one of our nine messaging arms. With probability $(1 - \epsilon_w)$, we utilize a Thompson Sampler to assign them one of the nine messaging arms. For this experiment, we took $\epsilon_1 = 1$ (since no information is available at time 1), $\epsilon_2 = 0.25$, and $\epsilon_3 = 0.33$.

Specifically, before each assignment time t , we consider the m_{at} people who had been assigned to arm a by that point. Of those m_{at} individuals, let s_{at} be the number of individuals assigned to arm a who by time t had received a vaccine or scheduled an appointment for a vaccine according to RIDOH's records. Let $f_{at} = m_{at} - s_{at}$ be the number who had not. Let $p_{at} \sim \text{Beta}(\alpha + s_{at}, \beta + f_{at})$ be the posterior distribution at time t of the true probability p_a that an individual enrolled in arm a will receive a vaccine. Here we take a $\text{Beta}(\alpha, \beta)$ prior distribution for p_a with a hyperprior $\pi(\alpha, \beta) \propto (\alpha + \beta)^{-2.5}$.

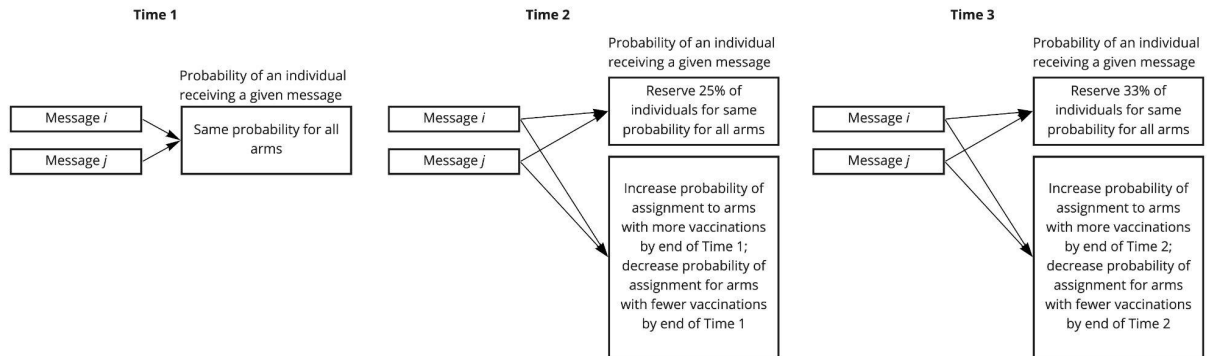
Then if an individual i is to be assigned by the Thompson sampler at time t , we draw realizations q_{ati} from each of the distributions p_{at} and assign individual i to $a_i = \text{argmax}_a q_{ati}$.

This strategy, which [Caria et al. \(2020\)](#) call Tempered Thompson Sampling, though we prefer the term ϵ -bounded Thompson Sampling, allows us to interpolate between gathering maximum information about each arm (uniform assignment to each arm) and maximizing expected reward at the cost of not gathering information about some arms (Thompson sampling).

Intuitively, the process resembles a Bayesian reasoner: if more people in the message 1 group than in the control group engage in the desired behavior during the first iteration, then during the second iteration the probability that anyone receives message 1 will increase slightly, thus making the design responsive to incoming evidence.

We note that while our ultimate outcome was *vaccination*, our Thompson Sampler utilized sign-ups as well. Some noise entered into this list, due to the deduplication process employed in matching individual testing records. This led to approximately 39% of those phone numbers which had a record of either a vaccination or a sign-up after the first week of the study to reappear in the eligible population after the second week of the study. However, a χ^2 test indicates that this noise was not significantly different between arms ($p = 0.83$). This error limits interpretation of the results to do with adaptation (see [SL4 EQ6](#)) but not those relevant to the primary research question of the study, which is answered through average treatment effects by message condition regardless of the likelihood that a given individual would receive a particular message.

Figure S2. Overview of ϵ -bounded Thompson sampler.



SI.3. Additional pre-registered analyses.

These analyses were registered at <https://osf.io/pkhae/>. All asymptotic tests were paired with permutation based tests. All results were the same so only asymptotic tests reported here. Below we use the R statistical analysis language and within it the `cmh_test()` function from the `coin` package (which enables both asymptotic and permutation based randomization inferences) and `lm_robust()` from the `estimatr` package (which produces randomization justified HC2 standard errors by default). Code for all analyses, including permutation tests, is available at <http://github.com/thepolicylab/thepolicylab/COVID-SMSExperiment>.

RQ0: Is there any effect of condition assignment?

```
rq0_asym <- cmh_test(vaccinatedF ~ messageF | date_sentF, data = wrkdat,
distribution = asymptotic())
rq0_asym
Asymptotic Generalized Cochran-Mantel-Haenszel Test
data: vaccinatedF by
      messageF (message_0, message_1, message_2, message_3, message_4,
message_5, message_6, message_7, message_8)
      stratified by date_sentF
chi-squared = 13, df = 8, p-value = 0.1
pvalue(rq0_asym)
[1] 0.1153
```

RQ1: Is there an effect of receiving a message as opposed to not receiving a message?

```
rq1_asym <- cmh_test(vaccinatedF ~ not_controlF | date_sentF, data = wrkdat,
distribution = asymptotic())
rq1_asym
Asymptotic Generalized Cochran-Mantel-Haenszel Test
data: vaccinatedF by not_controlF (0, 1)
      stratified by date_sentF
chi-squared = 1.2, df = 1, p-value = 0.3
pvalue(rq1_asym)
[1] 0.2663
rq1_est <- difference_in_means(vaccinated ~ not_controlF, blocks =
date_sentF, data = wrkdat)
rq1_est
Design: Blocked
      Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
not_controlF -0.00147  0.001363 -1.079  0.2806 -0.004141 0.001201 142402
```

RQ2: Does any given message differ from control (focal tests)?

```
rq2_est <- lm_robust(vaccinated ~ messageF, weights = IPW_weight_multarm,  
data = wrkdat)
```

```
rq2_est
```

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.02060861	0.001312	15.71145	1.406e-55	0.018038	0.0231795	142419
messageFmessage_1	-0.00208199	0.001842	-1.13002	2.585e-01	-0.005693	0.0015291	142419
messageFmessage_2	-0.00002266	0.001898	-0.01194	9.905e-01	-0.003742	0.0036968	142419
messageFmessage_3	0.00232052	0.002012	1.15339	2.488e-01	-0.001623	0.0062638	142419
messageFmessage_4	-0.00222375	0.001863	-1.19395	2.325e-01	-0.005874	0.0014267	142419
messageFmessage_5	0.00015034	0.001886	0.07972	9.365e-01	-0.003546	0.0038467	142419
messageFmessage_6	-0.00260175	0.001719	-1.51333	1.302e-01	-0.005971	0.0007679	142419
messageFmessage_7	-0.00316289	0.001824	-1.73444	8.284e-02	-0.006737	0.0004113	142419
messageFmessage_8	-0.00012584	0.001844	-0.06826	9.456e-01	-0.003739	0.0034876	142419

Adding the fixed effects estimates (biased, but more precise/statistically powerful) No substantive difference. This was not what we pre-registered: we pre-registered using the unbiased block-size weighting approach above. We report using the above analysis but present the fixed effects FYI.

```
rq2_fe_est <- lm_robust(vaccinated ~ messageF, fixed_effects = ~date_sentF,  
data = wrkdat)
```

```
rq2_fe_est
```

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
messageFmessage_1	-0.0022322	0.002007	-1.1125	0.26595	-0.006165	0.0017006	142407
messageFmessage_2	-0.0002051	0.001966	-0.1044	0.91688	-0.004058	0.0036474	142407
messageFmessage_3	0.0005308	0.002000	0.2654	0.79071	-0.003390	0.0044514	142407
messageFmessage_4	-0.0029133	0.002013	-1.4469	0.14792	-0.006860	0.0010330	142407
messageFmessage_5	-0.0006247	0.001874	-0.3334	0.73881	-0.004297	0.0030474	142407
messageFmessage_6	-0.0027722	0.001567	-1.7690	0.07689	-0.005844	0.0002992	142407
messageFmessage_7	-0.0034071	0.001902	-1.7915	0.07321	-0.007135	0.0003204	142407
messageFmessage_8	0.0004312	0.002011	0.2144	0.83020	-0.003510	0.0043723	142407

Verifying the p-values above with a permutation based cmh test for each message versus control

```
test_msgs <- function(msg1, msg2) {  
  ## msg1 and msg2 are strings indicating message assignment in messageF  
  effect_test <- cmh_test(vaccinatedF ~ messageF | date_sentF,  
    data = wrkdat,  
    subset = wrkdat$messageF %in% c(msg1, msg2),  
    distribution = approximate(nresample = 10000, parallel = "multicore",  
ncpu = 6)  
  )  
  return(pvalue(effect_test)[1])  
}
```

```

}
message_test_ps <- sapply(levels(wrkdat$messageF)[-1], function(msg) {
  test_msgs("message_0", msg)
})

```

```

message_test_ps
message_1 message_2 message_3 message_4 message_5 message_6 message_7 message_8
0.2778    1.0000    0.7040    0.1866    0.9335    0.1025    0.0502    0.8985

```

We specified that we would report adjusted p-values, although it is not necessary since we are not reporting any discoveries. We did the adjustment and show it in the Github repository but do not report them here.

RQ3: Does epistemic humility help?

Message 4 vs. 3 (CMH test, difference of proportions estimator). Only very small differences between those two arms.

```

rq3_est <- difference_in_means(vaccinated ~ messageF, blocks = date_sent,
data = wrkdat, subset = wrkdat$messageF %in% c("message_3", "message_4"))
rq3_est
Design: Blocked

```

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
messageFmessage_4	-0.003194	0.002074	-1.54	0.1235	-0.00726	0.0008707	22046

```

rq3_test <- test_msgs("message_3", "message_4")
rq3_test
[1] 0.1083

```

RQ5: How do social proof and appeals to the family interact?

We will test the overall hypothesis of no difference between 6 (family concern), 7 (social proof), and 8 (family concern + social proof). If we reject this, we test 6 versus 8 and 7 versus 8.

```

rq5_overall <- cmh_test(vaccinatedF ~ messageF | date_sentF, data = wrkdat,
subset = wrkdat$messageF %in% c("message_6", "message_7", "message_8"))
rq5_overall

```

Asymptotic Generalized Cochran-Mantel-Haenszel Test

```

data: vaccinatedF by
      messageF (message_6, message_7, message_8)
      stratified by date_sentF
chi-squared = 5.1, df = 2, p-value = 0.08

```

RQ6: Did adaptive randomization increase vaccinations over fixed randomization?

We also will report the effect of using adaptive randomization versus fixed randomization on total vaccinations — since we withheld 25% of each of the three weeks experimental pools for

fixed randomization and adapted the other 100 – 25%. Our aim in this study was to (1) learn about which messages worked best but also (2) increase vaccination. The fixed randomization maximized statistical power to detect effects whereas the adaptive randomization increased the numbers of people exposed to more effective messages.

```
rq6_est <- difference_in_means(vaccinated ~ is_chosen_from_uniform, blocks =  
date_sentF, data = wrkdat, subset = wrkdat$date_sent >= "2021-06-02")  
rq6_est  
Design: Blocked
```

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper
DF						
is_chosen_from_uniform	0.0009958	0.0008211	1.213	0.2252	-0.0006135	0.002605
102410						

```
rq6_cmh_perm <- cmh_test(vaccinatedF ~ factor(is_chosen_from_uniform) |  
date_sentF, data = wrkdat, subset = wrkdat$date_sent >= "2021-06-02",  
distribution = approximate(nresample = 10000, parallel = "multicore", ncpus =  
6))  
rq6_cmh_perm  
Approximative Generalized Cochran-Mantel-Haenszel Test  
data: vaccinatedF by  
      factor(is_chosen_from_uniform) (FALSE, TRUE)  
      stratified by date_sentF  
chi-squared = 1.5, p-value = 0.2
```

SI.4. Exploratory analyses, pre-registered

EQ1: Do explicit appeals to the safety of vaccines increase responses in areas with higher proportions of Black or Latinx people? Message 2 vs. control

We cannot detect any simple linear differential effect of pct black or latinx on the message 2 versus control comparison.

```
wrkdat3_eq1 <- wrkdat3 %>%
  filter(messageF %in% c("message_0", "message_2") & zcta != "00000") %>%
  droplevels()

make_weights <- function(dat) {
  block_m_each <- with(dat, table(date_sentF, messageF, exclude = c()))
  block_prob_each <- block_m_each / rowSums(block_m_each)
  declared_randomization <- declare_ra(blocks = dat$date_sentF, block_m_each
= block_m_each, conditions = sort(unique(dat$messageF)))
  IPW_weight <- 1 / obtain_condition_probabilities(declaration =
declared_randomization, assignment = dat$messageF)
  stopifnot(all.equal(sort(unique(1 / IPW_weight)),
sort(unique(block_prob_each))))
  return(IPW_weight)
}

wrkdat3_eq1$IPW_eq1 <- make_weights(wrkdat3_eq1)
eq1_blk_estA <- lm_robust(vaccinated ~ messageF * pct_any_blk, data =
wrkdat3_eq1, weights = IPW_eq1)

eq1_blk_estA
              Estimate Std. Error   t value Pr(>|t|) CI Lower CI Upper   DF
(Intercept)    0.0188928    0.002016   9.370867 7.829e-21  0.014941 0.022845 23245
messageFmessage_2 -0.0002165    0.002783  -0.077775 9.380e-01 -0.005672 0.005239 23245
pct_any_blk     0.0538385    0.018998   2.833943 4.602e-03  0.016602 0.091075 23245
messageFmessage_2:pct_any_blk -0.0002156    0.026120  -0.008253 9.934e-01 -0.051413 0.050982 23245
eq1_lat_estA <- lm_robust(vaccinated ~ messageF * pct_hisp, data =
wrkdat3_eq1, weights = IPW_eq1)
```

```
eq1_lat_estA
              Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
(Intercept)    0.018012    0.001849   9.7392 2.264e-22  0.014387 0.021637 23245
messageFmessage_2  0.001080    0.002534   0.4262 6.700e-01 -0.003887 0.006048 23245
pct_hisp        0.034689    0.009438   3.6756 2.378e-04  0.016191 0.053187 23245
messageFmessage_2:pct_hisp -0.007924    0.012534  -0.6322 5.272e-01 -0.032492 0.016643 23245
```

EQ2: Does the implication of choice through emphasis on a conspicuous advantage increase responses in areas with higher proportions of Republican people? Message 3 vs. control

No detectable difference in effects.

```

wrkdat3_eq2 <- wrkdat3 %>%
  filter(messageF %in% c("message_0", "message_3") & zcta != "00000") %>%
  droplevels()
wrkdat3_eq2$IPW_eq2 <- make_weights(wrkdat3_eq2)

```

```

eq2_gop_estA <- lm_robust(vaccinated ~ messageF * pct_gop, data =
wrkdat3_eq2, weights = IPW_eq2)
eq2_gop_estA

```

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.033399	0.00480	6.9582	3.541e-12	0.02399	0.042808	22768
messageFmessage_3	-0.002249	0.00661	-0.3402	7.337e-01	-0.01520	0.010707	22768
pct_gop	-0.027623	0.01238	-2.2321	2.562e-02	-0.05188	-0.003366	22768
messageFmessage_3:pct_gop	0.009509	0.01705	0.5576	5.771e-01	-0.02392	0.042935	22768

EQ3: Do explicit appeals to ease of access increase responses in areas with higher proportions of Black or Latinx people? Message 5 vs. control

No detectable differences. Magnitude of moderation is not small given this phenomenon (on order of 1 or 2 pts, but negative).

```

wrkdat3_eq3 <- wrkdat3 %>%
  filter(messageF %in% c("message_0", "message_5") & zcta != "00000") %>%
  droplevels()
wrkdat3_eq3$IPW_eq3 <- make_weights(wrkdat3_eq3)

```

```

eq3_blk_estA <- lm_robust(vaccinated ~ messageF * pct_any_blk, data =
wrkdat3_eq3, weights = IPW_eq3)
eq3_blk_estA

```

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.018567	0.002043	9.0861	1.098e-19	0.01456	0.022572	25930
messageFmessage_5	0.001946	0.002743	0.7095	4.780e-01	-0.00343	0.007322	25930
pct_any_blk	0.051307	0.019179	2.6752	7.473e-03	0.01372	0.088898	25930
messageFmessage_5:pct_any_blk	-0.026276	0.025041	-1.0493	2.940e-01	-0.07536	0.022805	25930

```

eq3_lat_estA <- lm_robust(vaccinated ~ messageF * pct_hisp, data =
wrkdat3_eq3, weights = IPW_eq3)

```

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.017433	0.001874	9.3041	1.455e-20	0.013760	0.02111	25930
messageFmessage_5	0.001536	0.002512	0.6114	5.410e-01	-0.003388	0.00646	25930
pct_hisp	0.034860	0.009734	3.5814	3.424e-04	0.015782	0.05394	25930
messageFmessage_5:pct_hisp	-0.011863	0.012585	-0.9426	3.459e-01	-0.036531	0.01280	25930

EQ4: Does epistemic humility increase responses in areas with higher proportions of either Black or Latinx people or Republican people? Message 4 versus 3

No detectable differences in effect.

```

wrkdat3_eq4 <- wrkdat3 %>%
  filter(messageF %in% c("message_3", "message_4") & zcta != "00000") %>%
  droplevels()

wrkdat3_eq4$IPW_eq4 <- make_weights(wrkdat3_eq4)

eq4_gop_estA <- lm_robust(vaccinated ~ messageF * pct_gop, data =
wrkdat3_eq4, weights = IPW_eq4)
eq4_gop_estA

```

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.031424	0.004622	6.7992	1.079e-11	0.02237	0.040483	21573
messageFmessage_4	-0.002244	0.006677	-0.3360	7.369e-01	-0.01533	0.010844	21573
pct_gop	-0.018053	0.011944	-1.5115	1.307e-01	-0.04146	0.005358	21573
messageFmessage_4:pct_gop	-0.003277	0.017226	-0.1902	8.491e-01	-0.03704	0.030487	21573

```

eq4_blk_estA <- lm_robust(vaccinated ~ messageF * pct_any_blk, data =
wrkdat3_eq4, weights = IPW_eq4)
eq4_blk_estA

```

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.019125	0.002084	9.17906	4.730e-20	0.015041	0.02321	21573
messageFmessage_4	-0.003302	0.002909	-1.13507	2.564e-01	-0.009003	0.00240	21573
pct_any_blk	0.067888	0.020681	3.28267	1.030e-03	0.027352	0.10842	21573
messageFmessage_4:pct_any_blk	-0.001599	0.028836	-0.05547	9.558e-01	-0.058120	0.05492	21573

```

eq4_lat_estA <- lm_robust(vaccinated ~ messageF * pct_hisp, data =
wrkdat3_eq4, weights = IPW_eq4)
eq4_lat_estA

```

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.018922	0.001878	10.0774	7.851e-24	0.015242	0.022603	21573
messageFmessage_4	-0.002755	0.002664	-1.0345	3.009e-01	-0.007976	0.002465	21573
pct_hisp	0.038470	0.009790	3.9294	8.542e-05	0.019280	0.057660	21573
messageFmessage_4:pct_hisp	-0.004235	0.013966	-0.3032	7.617e-01	-0.031610	0.023140	21573

EQ5: Is there a day-of-week effect? Proportions of vaccinations collapsed across all messages by day.

Since the randomization to message occurred **within a given date** and we have relatively few weeks, it is difficult to disentangle day of week effects from date effects. So, we only present descriptive information here.

```

table(weekdays(wrkdat3$date_sent))

```

Friday	Monday	Thursday	Tuesday	Wednesday
33626	23621	33619	17940	33622

```

wrkdat3$weekday_sent <- weekdays(wrkdat3$date_sent)

wrkdat3_weekday <- wrkdat3 %>%
  group_by(weekday_sent) %>%
  summarize(

```

```

    prop_vac = mean(vaccinated),
    prop_vac_in_week = mean(vac_in_week), nweek = n()
  )
}

wrkdat3_weekday
# A tibble: 5 x 4
  weekday_sent prop_vac prop_vac_in_week nweek
  <chr>         <dbl>         <dbl> <int>
1 Friday       0.0181         0.00803 33626
2 Monday       0.00923        0.00703 23621
3 Thursday     0.0195         0.00943 33619
4 Tuesday      0.0294         0.0109 17940
5 Wednesday    0.0218         0.0102 33622

```

EQ6: Is there an iteration effect? Some people were randomly assigned to have 3 weeks to schedule a vaccination and others only 1 week before the study ended. We explore whether there is a difference here.

The following table shows that we have no strong arguments against the claim that our messages were the same as control in regards either vaccination at all or vaccination within a week, regardless of whether the messages were sent in the first, second, or third weeks of the study. We present raw p-values here because this is exploratory work and do not adjust because we have so few small p-values: for example, we are not interpreting the effect of message_7 in week 1 or the effect of message_3 in week 3 below as discoveries.

```

test_msgs2 <- function(msg1, msg2, the_iteration, thefmla = vaccinatedF ~
messageF | date_sentF) {
  ## msg1 and msg2 are strings indicating message assignment in messageF
  effect_test <- cmh_test(thefmla,
    data = wrkdat3,
    subset = wrkdat3$messageF %in% c(msg1, msg2) & wrkdat3$iteration ==
the_iteration,
    distribution = asymptotic())
  # no difference when we used permutation tests, for speed switching to
asymptotic tests
  # approximate(nresample = 10000, parallel = "multicore", ncpu = 6)
  )
  return(pvalue(effect_test)[1])
}

msg_by_iteration <- as_tibble(expand.grid(iteration = 1:3, messageF =
levels(wrkdat3$messageF)[-1], stringsAsFactors = FALSE))

msg_by_iteration <- msg_by_iteration %>%
  rowwise() %>%
  mutate(p_vs_ctrl = test_msgs2("message_0", messageF, iteration)) %>%
  arrange(iteration, messageF)

```

```
msg_by_iteration <- msg_by_iteration %>%
  rowwise() %>%
  mutate(p_vac_week_vs_ctrl = test_msgs2("message_0", messageF, iteration,
thefmla = vac_in_weekF ~ messageF | date_sentF))
```

```
msg_by_iteration <- msg_by_iteration %>% mutate(p_vac_week_vs_ctrl =
ifelse(p_vac_week_vs_ctrl == p_vs_ctrl, NA, p_vac_week_vs_ctrl))
print(msg_by_iteration, n = 100)
```

A tibble: 24 x 4

Rowwise:

	iteration	messageF	p_vs_ctrl	p_vac_week_vs_ctrl
	<int>	<chr>	<dbl>	<dbl>
1	1	message_1	0.414	0.624
2	1	message_2	0.777	0.776
3	1	message_3	0.485	0.554
4	1	message_4	0.0950	0.151
5	1	message_5	0.432	0.999
6	1	message_6	0.174	0.368
7	1	message_7	0.00849	0.185
8	1	message_8	0.911	0.458
9	2	message_1	0.601	0.188
10	2	message_2	0.685	0.655
11	2	message_3	0.870	0.127
12	2	message_4	0.510	0.215
13	2	message_5	0.369	0.357
14	2	message_6	0.741	0.250
15	2	message_7	0.732	0.298
16	2	message_8	0.289	0.0548
17	3	message_1	0.541	0.731
18	3	message_2	0.988	0.893
19	3	message_3	0.0394	0.127
20	3	message_4	0.321	0.203
21	3	message_5	0.944	0.680
22	3	message_6	0.318	0.517
23	3	message_7	0.567	0.523
24	3	message_8	0.405	0.469

Nor is there strong evidence that “any message” was better than control, even when we assess the relationships for each iteration separately:

```
rq8_iteration1_test <- cmh_test(vaccinatedF ~ not_controlF | date_sentF, data
= wrkdat3, subset = wrkdat3$iteration == 1)
rq8_iteration1_test
```

Asymptotic Generalized Cochran-Mantel-Haenszel Test

data: vaccinatedF by not_controlF (0, 1)

stratified by date_sentF

chi-squared = 1, df = 1, p-value = 0.3

```
rq8_iteration2_test <- cmh_test(vaccinatedF ~ not_controlF | date_sentF, data
= wrkdat3, subset = wrkdat3$iteration == 2)
```

```
rq8_iteration2_test
```

```
Asymptotic Generalized Cochran-Mantel-Haenszel Test
data: vaccinatedF by not_controlF (0, 1)
      stratified by date_sentF
chi-squared = 0.0055, df = 1, p-value = 0.9
rq8_iteration3_test <- cmh_test(vaccinatedF ~ not_controlF | date_sentF, data
= wrkdat3, subset = wrkdat3$iteration == 3)
rq8_iteration3_test
```

```
Asymptotic Generalized Cochran-Mantel-Haenszel Test
data: vaccinatedF by not_controlF (0, 1)
      stratified by date_sentF
chi-squared = 0.57, df = 1, p-value = 0.5
## Also looking at vaccinations within a week for the first iteration
rq9_iteration1_test <- cmh_test(vac_in_weekF ~ not_controlF | date_sentF,
data = wrkdat3, subset = wrkdat3$iteration == 1)
rq9_iteration1_test
```

```
Asymptotic Generalized Cochran-Mantel-Haenszel Test
data: vac_in_weekF by not_controlF (0, 1)
      stratified by date_sentF
chi-squared = 0.36, df = 1, p-value = 0.6
rq9_iteration2_test <- cmh_test(vac_in_weekF ~ not_controlF | date_sentF,
data = wrkdat3, subset = wrkdat3$iteration == 2)
rq9_iteration2_test
```

```
Asymptotic Generalized Cochran-Mantel-Haenszel Test
data: vac_in_weekF by not_controlF (0, 1)
      stratified by date_sentF
chi-squared = 1.9, df = 1, p-value = 0.2
rq9_iteration3_test <- cmh_test(vac_in_weekF ~ not_controlF | date_sentF,
data = wrkdat3, subset = wrkdat3$iteration == 3)
rq9_iteration3_test
```

```
Asymptotic Generalized Cochran-Mantel-Haenszel Test=
data: vac_in_weekF by not_controlF (0, 1)
      stratified by date_sentF
chi-squared = 0.34, df = 1, p-value = 0.6
```

SI.5. Exploratory Analyses, not pre-registered

Effects on vaccination within a week

The experiment ran during a time of national campaigns in favor of vaccination. The control group in our experiment would have been exposed to this, and thus, might have gotten vaccinated for reasons other than a nudge from a text message.

No strong evidence that people were likely to be vaccinated within a week in “any message” versus control or versus any given message.

```
rq7_test <- cmh_test(vac_in_weekF ~ not_controlF | date_sentF, data = wrkdat)
rq7_test
```

Asymptotic Generalized Cochran-Mantel-Haenszel Test

```
data: vac_in_weekF by not_controlF (0, 1)
      stratified by date_sentF
chi-squared = 0.00041, df = 1, p-value = 1
rq7a_test <- cmh_test(vac_in_weekF ~ messageF | date_sentF, data = wrkdat)
rq7a_test
```

Asymptotic Generalized Cochran-Mantel-Haenszel Test

```
data: vac_in_weekF by
      messageF (message_0, message_1, message_2, message_3, message_4,
message_5, message_6, message_7, message_8)
      stratified by date_sentF
chi-squared = 8.7, df = 8, p-value = 0.4
rq7a_est <- lm_robust(vac_in_week ~ messageF, weights = IPW_weight_multarm,
data = wrkdat)
rq7a_est
```

	Estimate	Std. Error	t value	Pr(> t)	CI Lower	CI Upper	DF
(Intercept)	0.009127748	0.0009266	9.851260	6.884e-23	0.0073117	0.010944	142419
messageFmessage_1	0.000124368	0.0013438	0.092548	9.263e-01	-0.0025095	0.002758	142419
messageFmessage_2	0.000262444	0.0013718	0.191314	8.483e-01	-0.0024263	0.002951	142419
messageFmessage_3	0.002280338	0.0014842	1.536381	1.244e-01	-0.0006287	0.005189	142419
messageFmessage_4	-0.001289688	0.0012946	-0.996239	3.191e-01	-0.0038270	0.001248	142419
messageFmessage_5	-0.000005737	0.0013428	-0.004273	9.966e-01	-0.0026376	0.002626	142419
messageFmessage_6	-0.000254321	0.0011945	-0.212902	8.314e-01	-0.0025956	0.002087	142419
messageFmessage_7	-0.000963766	0.0013199	-0.730162	4.653e-01	-0.0035508	0.001623	142419
messageFmessage_8	0.000933308	0.0013295	0.702014	4.827e-01	-0.0016724	0.003539	142419

SI.6. Sources of noise in the dataset

As with any study drawing on government data, noise in our initial dataset is inevitable. Dai and colleagues independently estimated base rates of vaccination in Rhode Island during the study period adjusted for the randomization scheme used here and found a rate of 5.67%, almost three times the rate in our data. This raises two questions: (1) Could the discrepancy reflect some systematic difference between our population and the true population of vaccine-hesitant people that undermines our interpretation of the null result? And (2) could the discrepancy reflect a problem with the initial dataset such that the null is explained by insufficient power to detect a true effect?

SI.6.1. Does the base rate discrepancy suggest underlying characteristics of our population that undermine our interpretation?

A possible explanation for the discrepancy is that noise inherent in our dataset depressed the observed base rate. Almost certainly, some phone numbers were entered erroneously by participants or testing site staff, while others reflected individuals tested in Rhode Island but vaccinated elsewhere. Providence, the largest city, has a transient student population, and many Rhode Islanders work in neighboring states, both of which could affect where an individual got vaccinated. These factors would effectively increase the initial sample size of our study by making it appear that we had more phone numbers of unvaccinated residents than we actually did, thus decreasing our observed base rate by increasing our denominator.

We also note that some people share phone numbers, especially those in group quarters such as nursing homes. Only the first person vaccinated with a particular phone number would count in our numerator (if no one with that phone number had been vaccinated by the start of our study) or denominator. Of course, people would not receive text messages at all if they provided a land line—an issue we return to in section SI.6.2.

Another possibility is that design features contributed to the discrepancy. As far as we could tell, Dai et al.'s estimate did not remove people under 18, a population that was excluded from our study but could and did get vaccinated during the study period. To the extent that minors were helping to drive the state's topline numbers but were not in our study, our base rate would be lower than the state's; our own calculations suggest this was the case to a small degree. Moreover, as noted in the original manuscript, some vaccination sites like Veterans hospitals report aggregate numbers but not individual vaccination information to the state, which would contribute to topline vaccination rates without being matched to an individual. Despite these sources of uncertainty in our initial dataset—a problem for most research using administrative data—the entries were randomly assigned. So invalid phone numbers or people who got vaccinated out of state were no more or less likely to appear in treatment or control conditions. Moreover, we do not see how our population of ostensibly vaccine-hesitant individuals is

confounded by omitting adolescents⁴, veterans,⁵ or VA hospital staff⁶, since these groups have been less—or at best, equally—likely to get vaccinated relative to the general population. If anything, their exclusion from our population should push the result toward seeing a treatment effect rather than not.

In sum, while there are several, non-exclusive, possibilities as to why our base rate is lower than Dai and colleagues' estimate, none seemingly account for the observed lack of message effects and therefore threaten our interpretation that the critical time window for nudges had passed.

SI.6.2. If the base rate discrepancy were due to invalid entries, would it change the result?

Random assignment aside, we might ask whether invalid entries and/or artificial inflation of sample size meant the true population size did not provide sufficient power to detect an effect. Let us consider the worst-case scenario where two-thirds of our initial dataset was invalid.

In this scenario, the number of outcomes coded as non-vaccinations would have decreased by two-thirds in each arm. Excluding these theoretical bad entries would increase the base rate while decreasing sample size. Although Milkman et al.⁷ were sufficiently powered to find a ~2% increase in flu vaccinations due to the message we used as a base text with a sample size ($N = 47,306$) roughly equivalent to one-third of ours ($142,428/3 = 47,476$), we nonetheless thought it appropriate to examine the issue further with an example. Consider that we recorded 300 vaccinations in the “Access” message condition and about 265 vaccinations in the control condition during the study period. Now we may imagine an experiment with this outcome and roughly 15,000 people in the treatment condition and 15,000 in the control condition. Without blocking or other complex design features, this would yield an effect estimate of about $.002 = (300/15000) - (265/15000)$ and $p = .14$ using Neyman-style randomization inference from the R command `lm_robust()` (see <https://github.com/thepolicylab/COVID-SMSExperiment> for code and analysis). What would we see if instead of 15,000 total observations, we had only one-third that number? We would increase our effect size $(300/5000) - (265/5000) = .007$, but the loss of sample size would not appreciably improve our statistical power ($p = .13$). In order for the reduction of our denominator to make an increased effect size that is statistically significant, we would need to reduce the sample to 600 people in each arm $(300 / 600) - (265 / 600) = .06$, $p = .04$. Although we have noted several likely sources of noise above, it is extremely unlikely that 96% ($1 - 600/15000 = .96$) of our initial entries were invalid due to these sources. Even if

⁴ Scherer et al. (2021). [Acceptability of adolescent COVID-19 vaccination among adolescents and parents of adolescents — United States, April 15–23, 2021](#). *MMWR*.

⁵ Thorpe et al. (2021). [Communicating about COVID-19 vaccine development and safety](#). *medRxiv*.

⁶ Meyer et al. (2021). [Trends in health care worker intentions to receive a COVID-19 vaccine and reasons for hesitancy](#). *JAMA*.

⁷ Milkman et al. (2021). [A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment](#). *PNAS*.

two-thirds of our entries were invalid, which we also consider unlikely, a true effect would not be masked by insufficient power.