Using in-survey randomized controlled trials to support future pandemic response*

Ben M. Tappin[¶] Luke B. Hewitt[§]

This paper is a pre-print and has not yet been published in a peer-reviewed journal

January 4, 2024
The latest version can be found here

Abstract

According to various sources the world is likely to witness another pandemic on the scale of COVID-19 in the future. How can the social and behavioral sciences contribute to a successful response? Here we conduct a cost-effectiveness analysis of an under-evaluated yet promising tool from modern social and behavioral science: the randomized controlled trial conducted in an online survey environment ("in-survey RCT"). Specifically, we analyze whether, in a pandemic context, a public health campaign that uses an in-survey RCT to pre-test two or more different message interventions — and then selects the top-performing one for their public outreach — has greater impact in expectation than a campaign which does not use this strategy. Our results are threefold. First, in-survey RCT pre-testing is plausibly cost-effective for public health campaigns with typical resources. Second, in-survey RCT pre-testing has potentially powerful returns to scale: for well-resourced campaigns, it looks highly cost-effective. Third, additional evidence for several key parameters could both confirm these patterns and further increase the cost-effectiveness of in-survey RCT pre-testing for public health campaigns. Together our results suggest in-survey RCT pre-testing can plausibly increase the impact of public health campaigns in a pandemic context and identify a research agenda to inform pandemic preparedness.

---

*We are grateful to Adam Bear and Stephen Gadsby for their helpful comments on an earlier draft of this paper (any remaining errors are ours); Sean Ellis, Greg Huber, Samantha Sinclair and Jon Green for supplying the raw data files used in some parts of our analysis; the Massachusetts Institute of Technology Supercloud team for the high-performance computing cluster used in some parts of our analysis; and the Leverhulme Trust for providing funding support to BMT. COI: We are founders of a public benefit organization that conducts behavior change research including in-survey RCTs. We made a concerted effort to be impartial while conducting the work in this paper, and we advocate that practitioners partner with academics rather than pay organizations for their services, which goes directly against the financial incentive of our organization. Nevertheless, readers may still want to be mindful of this COI when assessing our reasoning and conclusions.

[¶]Research fellow, Royal Holloway University of London. Corresponding author benmtappin@gmail.com.

[§]Research fellow, Stanford University.

The COVID-19 pandemic is estimated to have caused an excess 18–33 million deaths worldwide as of the end of 2023 [1]. According to various sources, there is a distinct possibility the world will witness another pandemic with similar or greater capacity for harm in the coming decades [2–6]. For example, the UK government's 2023 National Risk Register estimates a 5–25% likelihood of a reasonable worst-case pandemic in the next five years [6]. Such considerations have motivated a renewed focus on what societies can do to prepare for and respond to future public health emergencies [5,7–14], including advances in social and behavioral science. To that end, the World Health Organization recently designated public engagement and communication a key area for research and development to improve society's future pandemic response [5] (see also [15]).

In the early months of the COVID-19 pandemic, a collaboration of social and behavioral scientists published an article entitled, "Using social and behavioural science to support COVID-19 pandemic response" [16], which presented a narrative review of behavioral science theories and past empirical findings to help practitioners "align human behavior with [public health] recommendations." Much of the article focused on recommendations for public communication; advising on what communicators could say, and how they could say it, in order to best encourage uptake of the recommended behaviors. The article was highly influential, being cited thousands of times during the first two years of the pandemic and is said to have influenced the COVID-19 policy of governments [17].

In this paper, we analyze a contribution of modern social and behavioral science research that was omitted in the aforementioned influential review, but that nevertheless holds great promise to reliably improve the impact of public communication during a pandemic. The contribution we analyze is not a specific theory or past empirical findings, but rather a relatively widespread *method* of modern social and behavioral science research: the randomized controlled trial conducted in an online survey environment, or "in-survey RCT." In-survey RCTs involve recruiting people online to complete a survey, randomly assigning them to receive one of several different message interventions, and then measuring their beliefs, attitudes and/or behavioral intentions. Since online labor markets like Amazon's Mechanical Turk were popularized in the early 2010s [18,19], in-survey RCTs have become a widespread method for conducting social and behavioral science research [20–25]. They are typically much cheaper and faster to conduct than RCTs in the field or in a physical lab,

and allow for causal inference to be done with larger and often more diverse samples than would otherwise be available in lab settings. In many countries, online panels of survey respondents are now widely available for conducting in-survey RCTs.

The primary way in which in-survey RCTs are used in the social and behavioral sciences is to test hypotheses with the goal of advancing theory. For example, scientists often use in-survey RCTs to investigate questions like: "are messages of type X (e.g. narrative format) more effective than messages of type Y (e.g. non-narrative format) at changing people's beliefs, attitudes and/or behavior?" [26,27] Based on the findings of such studies, scientists derive and refine theories about what makes a message effective, and can use these theories to advise public health communicators on the types of messages they should develop. Indeed, this was the model of behavioral science followed by the influential review article described above [16]. Crucially, however, in-survey RCTs can be used for another, complementary, goal by public health communicators: to rigorously pre-test two (or more) specific messages they have developed — to identify which one is most effective — before rolling out the top-performing message in their public outreach [28].

This in-survey RCT pre-testing strategy promises to complement a theory-based approach to public health communication because research from across the social and behavioral sciences suggests that what makes a message "effective" is highly context-dependent and challenging to predict ex ante, even for domain experts [27,29–35]. For example, in a recent study of messages encouraging flu vaccination, behavioral scientists' predictions about the effects of the messages were uncorrelated with the actual effect of the messages [30]. Or consider a recent meta-analysis of 1,149 studies that examined the effectiveness of 30 different principles of message development [27]. The results showed that, while some principles (e.g. narrative format) were more effective on average *across contexts*, the effectiveness of each principle was highly variable *between contexts* — such that the opposite principle (non-narrative format) was in fact more effective in many contexts. Furthermore, it appears that many of our theories are currently unable to reliably predict this contextual variation [27,30,33]. Thus, in-survey RCT pre-testing promises to complement theory- and expert-based approaches to public health communication by enabling practitioners to rapidly and cheaply identify which one of several messages is most effective in a new, idiosyncratic, or dynamic context — such as a novel pandemic.

However, despite its promise, the value of in-survey RCT pre-testing for improving the impact of public health communication depends crucially on assumptions that have not been characterized and remain poorly understood. Moreover, behavioral scientists working in the COVID-19 context have expressed skepticism that survey-based testing methods can reliably inform public health communication [36]. In this paper, we thus systematically investigate whether and to what extent in-survey RCT pre-testing would reliably improve the impact of public health campaigns in a pandemic context; under what assumptions it would do so; and what existing evidence says about those assumptions.

Specifically, we conduct a cost-effectiveness analysis of in-survey RCT pre-testing for public health campaigns in the context of a pandemic. Importantly, we anchor our cost-effectiveness estimates to a recent meta-analysis of real public health campaigns that were conducted during the COVID-19 pandemic [37], the largest meta-analysis of its kind to date. That analysis examined 376 public health campaigns that ran on Facebook and Instagram during December 2020 to November 2021 and targeted people's attitudes and beliefs about the COVID-19 vaccines, finding that the average campaign spent $105,000 and had an estimated effect of 0.55 percentage points on people's attitudes and beliefs. Based on this, that analysis estimated a cost of $3.41 per incremental influence on attitudes/beliefs and $5.68 per incremental vaccination for the campaigns.

The key results of our cost-effectiveness analysis are as follows.

First, we find that in-survey RCT pre-testing is plausibly cost-effective for public health campaigns with a budget of $105,000 — equal to the typical campaign spend in the above meta-analysis — and could be robustly-so given more research. That is, a campaign that conducts in-survey RCT pre-testing has higher expected impact than a campaign that conducts no such pre-testing, despite both campaigns spending the same amount of money overall. On the basis of the above cost-per-impact estimates, this translates to hundreds or even thousands of additional attitudes/beliefs influenced and vaccinations received due to the campaign. Second, in-survey RCT pre-testing has potentially powerful returns to scale: for larger campaigns (e.g. $210,000–$420,000), in-survey RCT pre-testing looks highly cost-effective on our estimates, netting thousands of additional

attitudes/beliefs influenced and vaccinations received in expectation. Third, we demonstrate that possessing accurate knowledge of several key parameter values allows campaigns to optimize their in-survey RCT testing regime to further increase its returns. However, the evidence for several such parameter values is currently limited; if there was more evidence, in-survey RCT pre-testing could potentially deliver greater returns.

These results have two main implications.

The first implication of our results is that in-survey RCT pre-testing would likely have improved the impact of many public health campaigns that ran on social media during the COVID-19 pandemic. By extension, our results suggest this strategy would likely improve the impact of pandemic-related public health campaigns in the future. Furthermore, our results also point to the potential value of in-survey RCT pre-testing for public health campaigns in general, beyond social media campaigns conducted during pandemics.

The second implication of our results is to identify a clear agenda for future research that can enhance preparedness for, and response to, future public health emergencies. Specifically, to obtain better evidence on several key parameter values — and potentially further increase the returns to in-survey RCT pre-testing for public health campaigns — research can deploy a new study design: the "parallel-megastudy" design. This design combines two designs that are separately growing in popularity in the social and behavioral sciences, the in-survey megastudy [38,39] and the in-field megastudy [40,41], which involve testing many different interventions simultaneously in a survey or in the field, respectively. In the parallel-megastudy design, an in-survey megastudy and an in-field megastudy are conducted *in parallel* using the same set of intervention messages, allowing researchers to estimate the key parameters governing the returns to in-survey RCT pre-testing. We describe this design further in the Discussion section of this paper.

Cost-effectiveness analysis

Our cost-effectiveness analysis estimates whether a public health campaign that conducts an in-survey RCT pre-test — the "testing-campaign", for brevity — has greater impact in expectation than a campaign that does not use this pre-testing strategy — the

“naive-campaign” — where both types of campaign spend the same amount of money overall. At the core of our analysis is an expression that compares the expected impact of the testing-campaign to the expected impact of the naive-campaign. This expression is explained in full in Methods section 1; below we briefly give the intuition.

The expected impact of the testing-campaign depends on the trade-off between two quantities: (a) the expected improvement in message effectiveness from conducting the in-survey RCT pre-test versus (b) the financial cost of conducting the pre-test. This trade-off follows an intuitive logic: the more money that is spent on the pre-test, the less money that is left for actually running the campaign’s outreach — thus diminishing the number of people it is able to reach. However, if pre-testing allows the campaign to identify a sufficiently-more-effective message, then the diminished reach of their communication can be offset by a greater per-person impact among those exposed to it.

How these two quantities trade off in practice depends upon the values of various parameters. For example, suppose that there is very little variation between the effects of different intervention messages. In this case, using an in-survey RCT to identify which one of several different messages is most effective may require a very large sample of survey respondents to reliably identify which message is most effective. This will increase the cost of the in-survey RCT, all else equal, because it costs money to recruit survey respondents. If the cost is large, then identifying a more effective message may not compensate for it. The full list of parameters we consider in our analysis is in Methods section 2, but below we briefly describe three key parameters that determine the improvement in message effectiveness that is possible from in-survey RCT pre-testing.

First, there is the true variation in effect sizes across different messages. This parameter captures how much more effective some messages are than others. The larger the value of this parameter, the easier it is to identify more effective messages, all else equal, because there is greater variation between different messages’ effectiveness. For convenience, we operationalize this parameter as the true standard deviation in message effects divided by the effect size of the average message. For example, a variation parameter of 0.5 says that one standard deviation in message effects is equal to one-half of the effect size of the average message (for further detail, see Methods section 2).

Second, there is the true effect size of the average message in the survey environment. In our analysis, the larger the true effect size of the average message in-survey, the easier it is to identify more effective messages in the RCT, all else equal, as larger effect sizes are easier to detect. Third, there is the correlation between the true effects of messages in the survey environment and the true effects of those messages in the "field" — that is, when communicated in an actual campaign context. If this correlation is zero, then in-survey RCT pre-testing provides no signal about which message will be most effective in the field and thus provides no value. By contrast, if the correlation is positive, then in-survey RCT pre-testing can provide a reliable signal about which message is likely to be most effective in the field. The more positive the correlation, therefore, the more valuable in-survey RCT pre-testing is for improving campaign impact, all else equal, because the top-performing message in-survey is more likely to also be the top-performing message in the field.

For each unique combination of the parameter values, we simulate three-thousand in-survey RCT pre-tests to estimate the expected improvement in message effectiveness for the testing-campaign (details in Methods section 3). We consider two different testing "regimes" for how the testing-campaign decides on the number of intervention messages and survey respondents to include in its RCT. Under the "bare-minimum" testing regime, the campaign tests just two different messages, and uses a heuristic for recruitment: n=300 respondents per message. In contrast, under the "optimal" testing regime, we assume the campaign has perfect knowledge of the true parameter values and chooses the number of messages and survey respondents for the RCT that maximizes the expected increase in impact given this knowledge (see Methods section 3 for further details). Thus, the difference in performance between the two testing regimes illustrates the benefit of possessing perfectly accurate knowledge of the true parameter values.

Results

Returns to in-survey RCT pre-testing for a typical campaign

Figure 1 shows the expected impact of the testing-campaign over the expected impact of the naive-campaign across the full parameter space — which we call a "relative-impact curve" — where each campaign has a budget of $105,000. Recall that $105,000 is the

average campaign spend from the Athey et al. meta-analysis [37] of 376 real public health campaigns that ran on Facebook/Instagram through 2021. Figures 1a and 1b show the relative-impact curve under the bare-minimum and optimal testing regimes, respectively.

Across the majority of the parameter space, in-survey RCT testing is cost-effective for a budget of $105,000: the testing-campaign has higher expected impact than the naive-campaign, as indicated by a relative-impact curve greater than 1 (shown in red in Figure 1). However, in some parts of the parameter space, it is not cost-effective: the testing-campaign has lower expected impact than the naive-campaign, indicated by a relative-impact curve smaller than 1 (shown in gray in Figure 1). Indeed, the returns to in-survey RCT testing depend strongly on the true values of the parameters, ranging from the testing-campaign having *lower* impact than the naive-campaign to it having as much as +35% higher impact in expectation under the bare-minimum testing regime (Figure 1a) and +80% higher impact in expectation under the optimal testing regime (Figure 1b).

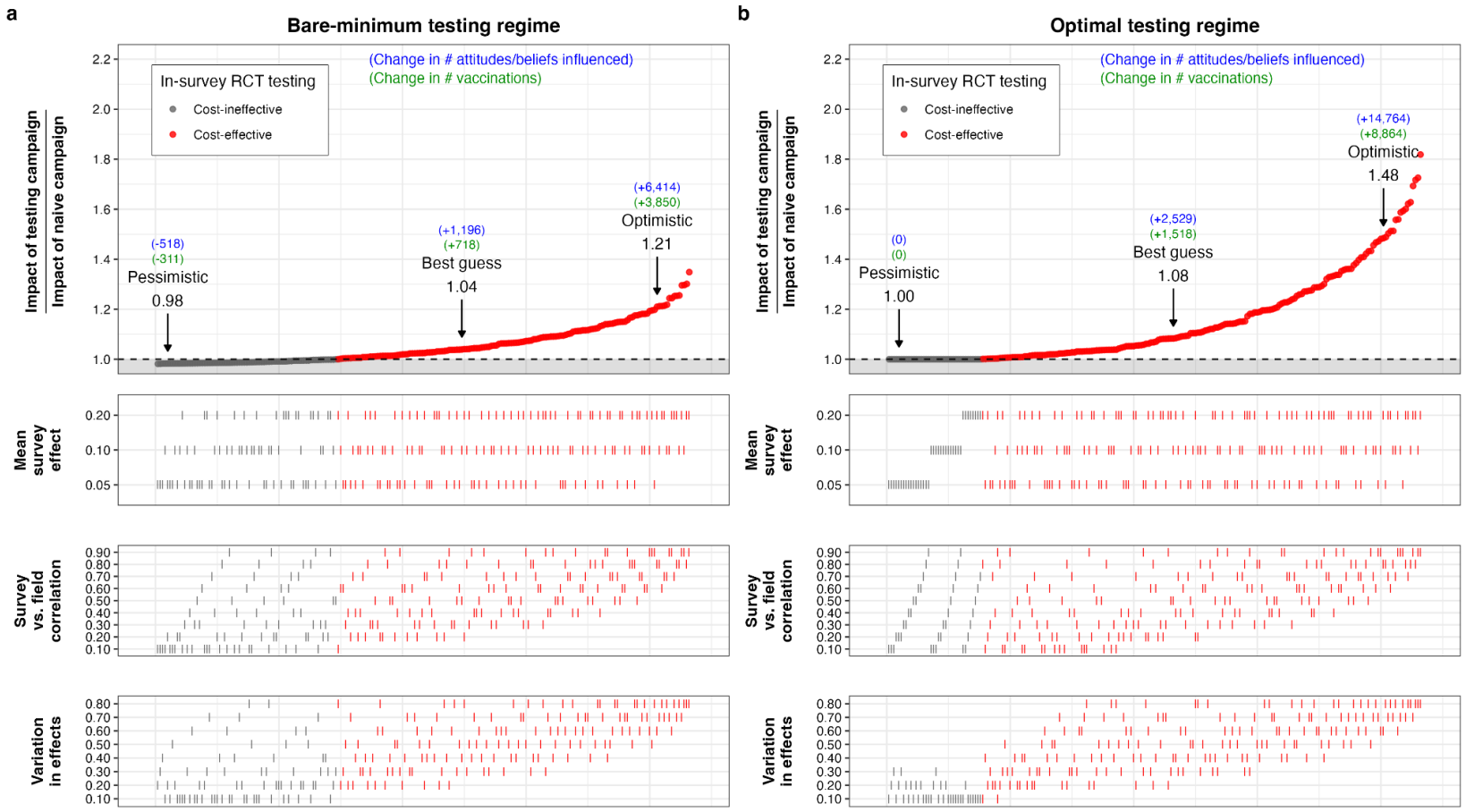


Figure 1. Relative-impact curves for a public health campaign with a budget of $105,000. Panel (a) shows the results for the bare-minimum testing regime and panel (b) shows the results for the optimal testing regime. The panels beneath the main top panel show the parameter values corresponding to the value of the relative-impact curve (similar to a "specification curve" [42]). There is a clear trend such that, when the survey-field correlation and variation-in-effects parameters are larger, the relative-impact curve is also larger — indicating that campaign impact receives a greater boost from in-survey RCT pre-testing.

Thus, an important question is: Which parameter values are most plausible? On the basis of existing evidence (see Methods section 4 for an evidence review), we determined a "best guess" set of parameter values as: an in-survey mean message effect size of 0.1 standard units; a correlation of 0.5 between survey and field effects; and a variation in the message effects of 0.4 — which, recall, should be interpreted in the following way: one standard deviation in message effects is equal to two-fifths of the mean effect (i.e. 0.4*0.1 = 0.04). Concretely, these parameter values imply that the true in-survey effects of most messages fall between 0.06 and 0.14 standard units (i.e. the mean ±1 SD). These parameter values also imply that, if one were to correctly identify the best of two different messages in a survey RCT, that message would also be the best message in the field ~66% of the time; far from perfect, but above chance (50%) (see Methods section 4 for details). In this scenario, for a $105,000 campaign, in-survey RCT testing under the bare-minimum testing regime increases campaign impact by 4% in expectation. At a cost of $3.41 per incremental influence on attitudes/beliefs and $5.68 per incremental vaccination, as estimated by Athey et al. [37], this implies an extra ~1,200 people's attitudes/beliefs influenced and ~700 extra people vaccinated, respectively (Figure 1a). Under an optimal testing regime, the expected increase in impact is 8%, implying an extra ~2,500 attitudes/beliefs influenced and ~1,500 extra people vaccinated (Figure 1b).

Nevertheless, there is of course uncertainty about whether the set of best-guess parameter values is correct — especially the values of the correlation and variation parameters, for which relevant prior research is limited (see Methods section 4). Therefore, we also consider more pessimistic and optimistic sets of parameter values.

Consider first a pessimistic scenario in which the mean message effect size is 0.05 standard units, the survey-field correlation is 0.2 and the variation in message effects is 0.2. This scenario implies that most in-survey message effects fall between just 0.04 and 0.06 standard units, and that correctly identifying the best of two different messages in-survey translates to identifying the best message in the field only ~56% of the time; that is, barely above chance (50%). In this unforgiving scenario, in-survey RCT testing is not cost effective for public health campaigns that have a budget of $105,000. Spending money to conduct the in-survey RCT test in an effort to identify a more effective message does not compensate for the less money subsequently available to actually run the public

outreach. Consequently, a campaign operating under the bare-minimum testing regime influences fewer attitudes/beliefs and results in fewer vaccinations than the naive-campaign in expectation (Figure 1a). A campaign operating under the optimal testing regime recognizes that testing is not cost-effective, and thus foregoes it — as a result, their expected impact is simply equal to that of the naive campaign (Figure 1b).

Now consider a more optimistic scenario in which the mean in-survey effect size is 0.2 standard units, the survey-field correlation is 0.8 and the variation in message effects is 0.6. This scenario implies that most in-survey message effects fall between 0.08 and 0.32 standard units, and that correctly identifying the best of two different messages in-survey translates to identifying the best message in the field ~79% of the time. In this scenario, in-survey RCT testing is highly cost-effective, resulting in thousands of extra attitudes/beliefs influenced and vaccinations received in expectation for a $105,000 campaign — under either of the testing regimes (Figure 1).

In summary, in-survey RCT pre-testing is cost-effective for a typical campaign given our best-guess assumptions about the true values of the parameters and is highly cost-effective on more optimistic assumptions. Furthermore, on pessimistic assumptions about the parameter values, the downside appears small enough for the bare-minimum regime (Figure 1a) that, even assuming each of the different parameter sets are equally likely to be true, campaigns with a budget of $105,000 would still benefit in expectation from conducting an in-survey RCT pre-test. In Appendix 1, we show this formally. In Appendix 1 we also show that, even if one puts somewhat higher probability on the pessimistic set of parameter values being true, in-survey RCT pre-testing with a bare-minimum testing regime remains cost-effective on our estimates. Nevertheless, it is important to also reiterate that, if the pessimistic set of parameter values is in fact true, then in-survey RCT pre-testing (as we operationalize it here) is not cost-effective for a public health campaign with a budget of $105,000. As a result, if campaigns are risk-averse — that is, they want to avoid the risk of performing worse than if they had gone without any RCT testing whatsoever — it is especially important for future research to produce more and better evidence to determine the true values of the parameters.

Returns to in-survey RCT pre-testing for smaller and larger campaigns

In our analysis thus far we have focused on a public health campaign with a budget of $105,000, the typical campaign spend in the Athey et al. meta-analysis [37] of COVID-19 social media campaigns. Now we consider campaigns with alternative budgets. The standard deviation in campaign spend in the aforementioned meta-analysis was approximately $327,000, implying that some public health campaigns were considerably better funded than others. This raises the important question of how the returns to in-survey RCT pre-testing scale with the resources available to the campaign.

To analyze this question, Figure 2 shows the estimated relative-impact curves for campaigns with three different budgets: $52,500 (Figure 2a), $210,000 (Figure 2b) and $420,000 (Figure 2c). To ease interpretation, the panels that show the corresponding parameter values are omitted from Figure 2, but can be viewed in full in Appendix 2.

Figure 2 implies that in-survey RCT pre-testing has powerful returns to scale. For example, for campaigns with a budget of $210,000 or $420,000, in-survey RCT testing is cost-effective under a wide range of parameter values, and is expected to net many thousands of extra attitudes/beliefs influenced and vaccinations received under the best-guess set of parameter values. For a $52,500 campaign, in contrast, whether in-survey RCT testing is cost-effective or not depends more strongly on the parameter values — yet, even for this budget, it is estimated to remain borderline cost-effective under the best-guess scenario. Furthermore, as with the results for a $105,000 campaign, the expected decrease in impact if the pessimistic assumptions are true is outweighed by the expected *increase* in impact if the best-guess or optimistic assumptions are true. Thus, in-survey RCT pre-testing with a bare-minimum testing regime is beneficial in expectation even if one assumes each of the different parameter sets are equally likely, and remains so even if one assumes the pessimistic set is somewhat more likely (Appendix 1).

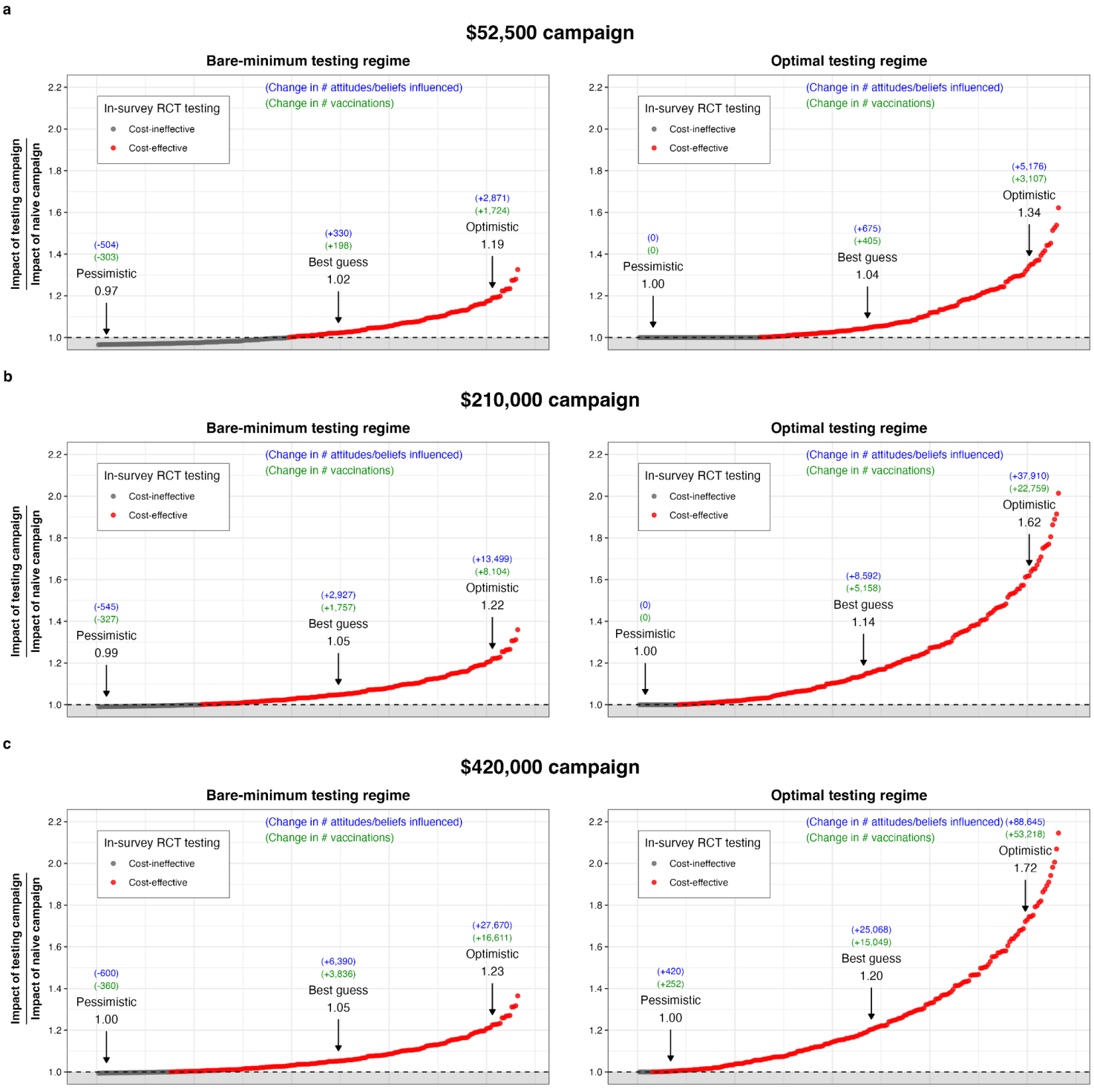


Figure 2. Relative-impact curves for campaigns with budgets other than $105,000. The full plots with the corresponding panels showing the parameter values are in Appendix 2.

Benefit of possessing accurate knowledge of the parameter values

Figures 1 and 2 show that the returns to the optimal testing regime are considerably larger than that of the bare-minimum testing regime. Therefore, public health campaigns seeking to maximize their impact would implement an optimal testing regime for their in-survey RCT pre-test. However, the advantage of the optimal regime relies on perfect knowledge of the true parameter values, which campaigns do not have. In reality, campaigns could optimize their testing regime for one set of parameter values when in fact the true values are quite different. In this section, we consider the consequences of such a mistake.

Specifically, we examine the returns to an in-survey RCT pre-test that is optimized for the best-guess values of the parameters when in fact the true values are different (e.g. more pessimistic or optimistic). Figure 3 shows these results for campaigns with each of the budgets we have considered thus far. (As with Figure 2, to ease interpretation, the panels that show the corresponding parameter values are omitted from Figure 3, but can be viewed in full in Appendix 3.) Figure 3 underscores the benefit of possessing accurate knowledge of the parameter values. While an in-survey RCT test optimized for the best-guess parameter values performs well in some areas of the parameter space, in other areas of the space it performs quite badly — in particular, when the true parameter values are on the pessimistic side. Indeed, when the true parameter values are pessimistic, the mistakenly-optimized testing regime performs worse even than the bare-minimum testing regime (compare Figure 3 vs. Figures 1 and 2). The reason for this is that the campaign significantly overspends on running the RCT pre-test. This illustrates that, while the gains from optimized-testing can be greater than for a non-optimized (e.g. bare-minimum) testing regime, the losses can also be greater. This in turn underscores the importance of possessing accurate knowledge of the parameter values and for future research to produce more and better evidence to determine their true values.

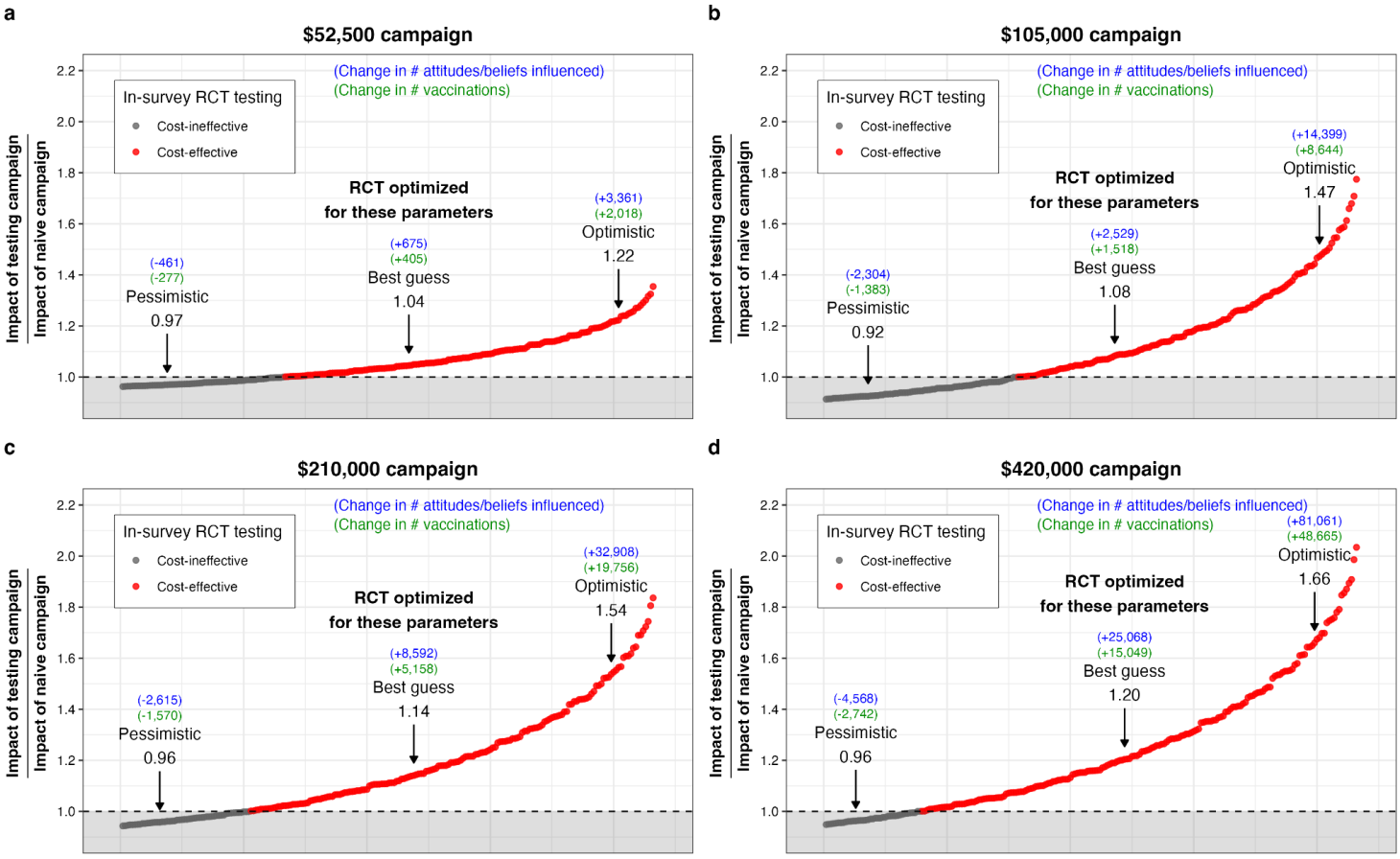


Figure 3. Relative-impact curves for public health campaigns when their in-survey RCT pre-testing regime is optimized for the best-guess set of parameter values. The full plots with the corresponding panels showing the parameter values are in Appendix 3.

Cost of RCT expertise

In a final analysis, we consider the fact that many public health campaigns may not have the expertise and/or infrastructure to conduct survey RCTs in-house, and so may need to partner with other actors in order to actually run an in-survey RCT pre-test. Such a partnership could impose further financial costs on the campaign, which will reduce the returns to in-survey RCT pre-testing. For example, on a consultant-model, campaigns could hire an organization with expertise in conducting in-survey RCTs. The cost of this service itself, over and above the cost of the survey respondents (which we already model in our analysis), is difficult to know in general — so we consider a range of additional

flat-cost possibilities: $1000, $3000, and $10,000. Notably, however, there are alternatives to the consultant-model which may be more appealing to public health campaigns.

For instance, the survey RCT designs considered in our analysis are relatively simple, involving just a handful of different intervention messages and simple random assignment; programming the survey RCT and analyzing its data is thus relatively straightforward. Indeed, many social and behavioral scientists working in academia regularly perform these tasks as part of their research, and do so with little difficulty. Therefore, an alternative model which is likely more cost-effective for public health campaigns is partnering with academia. Such partnerships could take a variety of forms, and reflect either bespoke one-offs or instead be institutionalized in academic-practitioner networks; an idea that has precedence [43] and was recently recommended by a review of the role of social/behavioral science in the COVID-19 response [17]. Such partnerships might involve an agreement whereby, in exchange for running the survey RCTs, the academic researchers have right-to-publish the results in a scientific journal. As well as proving more cost-effective for public health campaigns, institutionalized partnerships could also act as accelerators for scientific learning: centralizing large amounts of data from RCTs testing real public health communication interventions, which researchers could subsequently analyze using meta-analysis. This is a potentially fruitful model for increasing both the impact of public health communication, especially in the context of a novel pandemic, as well as scientific understanding — but we leave its fleshing out to future work.

Figure 4 shows the results of this analysis. Specifically, it shows the relative-impact scores for public health campaigns with different budgets, as a function of their RCT testing regime, the true parameter values, and the amount of money spent on hiring RCT expertise. Recall that relative-impact scores larger than 1 indicate in-survey RCT pre-testing is estimated to be cost-effective. Thus, the key takeaway of Figure 4 is that, for a typical campaign (budget $105,000), in-survey RCT pre-testing likely remains cost-effective under small ($1000) and moderate ($3000) expertise costs. However, under large costs ($10,000) it does not. In contrast, for campaigns with larger budgets, in-survey RCT pre-testing likely remains cost-effective even under large expertise costs. Finally, campaigns with smaller than average budgets ($52,500) can only tolerate small expertise costs before in-survey RCT pre-testing is likely no longer cost-effective on our estimates.

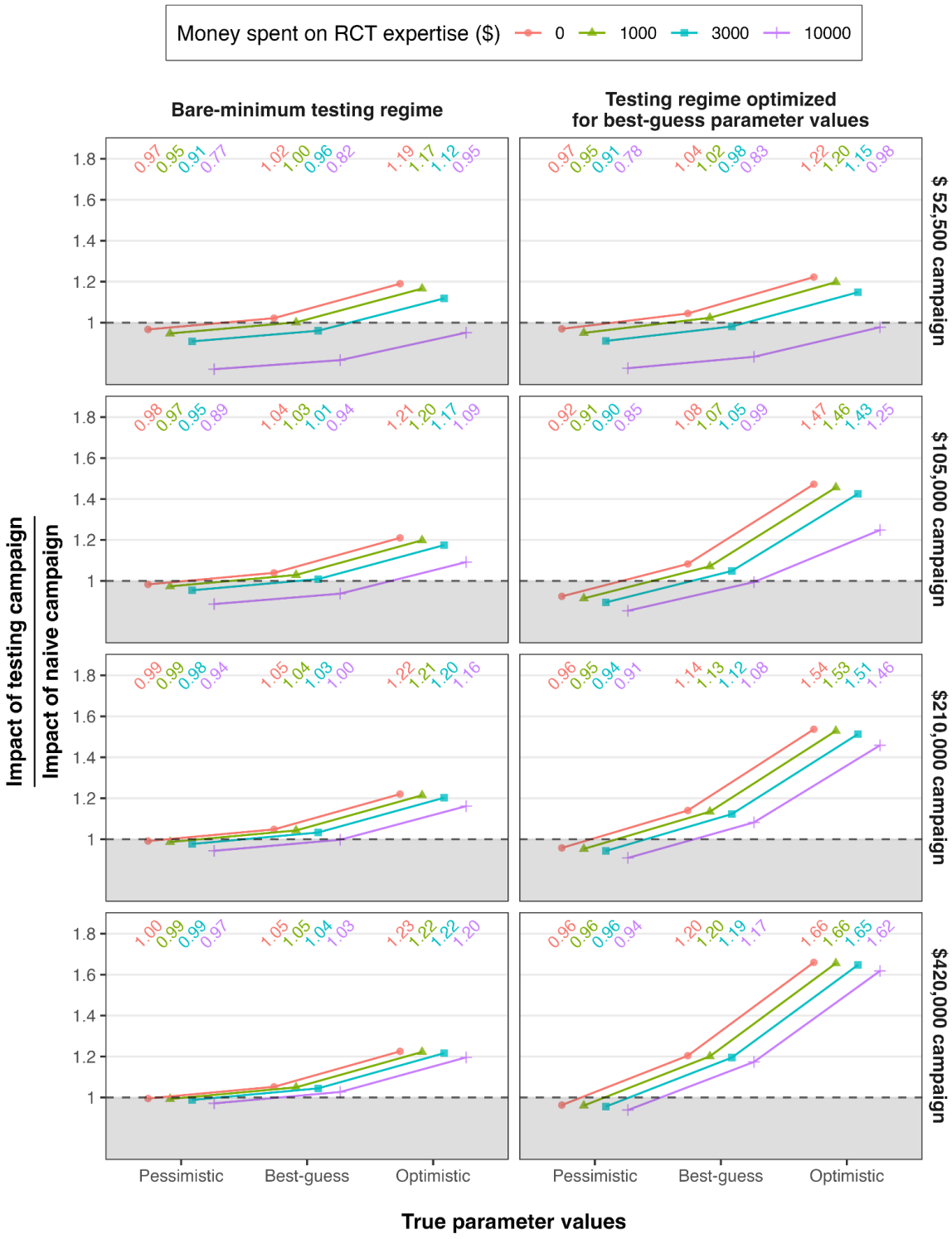


Figure 4. Relative-impact score as a function of different campaign budgets (facet rows), RCT testing regimes (facet columns), true values of the parameters (x axis), and amounts of money spent on hiring RCT expertise (colors). Relative-impact scores <1 indicate that in-survey RCT pre-testing is not cost-effective; scores >1 indicate that it is cost-effective.

Discussion

Following the worst of the COVID-19 pandemic, the World Health Organization designated public engagement and communication one of several priority areas in need of further research and development to improve society's response to future pandemics [5] (see also [15]). In this paper, we responded to this call by conducting a systematic cost-effectiveness analysis of in-survey RCT pre-testing as a strategy for improving the impact of public health campaigns in a pandemic context. Our results suggest this strategy would likely have improved the impact of many public health campaigns that ran on social media during the COVID-19 pandemic. By extension, our results suggest this strategy would likely improve the impact of similar public health campaigns in the future.

Importantly, our results rigorously quantify the importance of several parameters for the returns to in-survey RCT pre-testing; most notably, the variation in message effects, and the correlation between in-survey effects and those effects in the field (i.e. in an actual public health campaign). These parameters strongly determine the improvement in impact that is possible from in-survey RCT pre-testing. Moreover, possessing good information about the values of these parameters allows the RCT pre-testing regime to be optimized — thereby enabling further improvements in impact. While there is some existing evidence that can speak to the plausible values of these parameters, it could be much improved (see Methods section 4). To improve this evidence, future research in public health communication should deploy a new study design: the "parallel-megastudy" design. This design combines two existing designs, the in-survey megastudy [38,39] and the in-field megastudy [40,41], both of which involve testing many different interventions simultaneously. In the parallel-megastudy design, an in-survey and in-field megastudy are conducted *in parallel using the same set of intervention messages*. As a result, the parallel-megastudy design allows researchers to estimate the two most key parameters that we identify in our analysis as governing the returns to in-survey RCT pre-testing: the variation in message effects and the correlation between in-survey effects and in-field effects. The parallel-megastudy design can thus rapidly advance scientific understanding of the returns to in-survey RCT pre-testing for public health communication.

Our results are anchored to the details of hundreds of public health campaigns that ran on Facebook and Instagram during the COVID-19 pandemic [37], and therefore apply most readily to similar contexts — namely, social media campaigns conducted during a novel pandemic. However, here we highlight some broader implications for public health campaigns in other contexts. On the basis of our results, it is plausible that public health campaigns whose goal is to encourage vaccination for other diseases (e.g. flu, HPV, etc.) or other health behaviors (e.g. smoking cessation, contraceptive use, clinic visits, etc.) could improve their impact via in-survey RCT pre-testing. This is especially likely to be the case for campaigns with budgets larger than $200,000, owing to the comparatively cheap cost of running an in-survey RCT. Importantly, the parameters that may be most likely to vary across different public health contexts, and thus most affect the generalizability of our results, are: (1) the cost of developing messages to test in an RCT, as well as (2) the in-survey average effect size of the messages, (3) the variation in their effect sizes, and (4) the correlation between survey and field effect sizes. In contexts where the values of these parameters may be expected to depart considerably from our analysis — especially in a more costly/pessimistic direction — our results should be generalized with caution.

Now we highlight some more general limitations of our analysis and results.

First, readers should not mistake potentially large *relative* increases in campaign impact due to in-survey RCT testing (e.g. 20%) for large *absolute* increases in campaign impact. On the contrary, our analysis shows that, even for the best-resourced campaigns operating under optimistic assumptions, the expected increase in impact from in-survey RCT testing is limited to tens of thousands of additional attitudes/beliefs influenced and vaccinations received. In other words, even in the best case, it is clear that in-survey RCT testing can only form a small part of a successful pandemic response. Indeed, more broadly, we concur with other scholars that the potential impact of interventions encouraging individual-level behavior change — such as public health campaigns — typically comes a distant second to the impact of policy- and other system-level interventions [44]. While interventions encouraging individual-level behavior change can cost-effectively contribute to a successful outcome, this asymmetry in impact should be kept in mind [45].

Second, our analysis does not detract from the importance of field RCTs for evaluating the impact of public health campaigns. While our results suggest that in-survey RCT pre-testing can improve the impact of such campaigns, field RCTs are necessary to understand the magnitude of the campaigns' impact in the real world, and whether it is worth the cost of running them to begin with. If zero people are willing or able to watch a public health campaign's video on social media, for example, then no amount of in-survey RCT pre-testing of the video's content will be able to improve the impact of the campaign.

Third, to compute the additional numbers of vaccinations expected due to in-survey RCT testing, our analysis relied on the cost per incremental vaccination estimated by Athey and colleagues [37] in their the meta-analysis of COVID-19 social media campaigns. Their estimate assumes that changing people's self-reported attitudes, beliefs and/or intentions to get vaccinated converts to actual vaccinations at a rate of 0.6. In other words, if ten people were to report being in favor of the COVID-19 vaccinations when previously they were opposed, we should expect six of them to actually get vaccinated. This discount rate represents the well-known "intention-behavior gap" [46,47]. If the discount rate is smaller than 0.6, then our estimates of the number of additional vaccinations should also be correspondingly shrunk. Notably, estimates of the discount rate from other research studies (conducted in various different contexts) are between 0.33 and 0.55 [46–49].

Fourth, another assumption made in our analysis for which there is currently no relevant evidence either way is that the variation in message effects in the field is *proportional* to the variation in message effects in the survey — even if the absolute effect sizes are much smaller in the field (as is to be expected). If the variation in message effects in the field is proportionally larger than in the survey, our analysis will underestimate the returns to in-survey RCT pre-testing; if the reverse is true, our analysis will overestimate its returns. Notably, the parallel-megastudy design we describe above can also bring evidence to bear on this question, further highlighting the value of this design for advancing scientific understanding of the returns to in-survey RCT pre-testing for public health communication.

To conclude, we reiterate that various sources suggest the next pandemic is a question of "when" not "if" [2–6]. Understanding how the social and behavioral sciences can contribute to a successful response is a worthwhile goal — one that we aimed to advance here.

Methods

1 Expression for cost-effectiveness analysis

In this section we describe the expression used in the cost-effectiveness analysis. To ease interpretation, before writing out the full expression we will explain its constituent parts. Consider first the impact of a naive-campaign, given by the following expression:

$$\frac{budget - cost\ of\ creating\ one\ message}{CPI} \tag{1}$$

The "CPI" term represents the cost-per-person-influenced of the campaign; for example, $3.41, as estimated by Athey et al. [37]. The budget is the overall budget of the campaign; for example, $105,000, the average campaign spend in the aforementioned meta-analysis. Thus, expression 1 estimates how many people would be influenced by the naive-campaign, where there is no in-survey RCT pre-testing; the campaign simply creates one message and spends their remaining budget disseminating it in their public outreach. For instance, if creating a message (e.g. a brief social media video) costs $1000, then, on the above numbers, expression 1 would imply that (105000 – 1000) / 3.41 = 30,499 people would have their attitudes/beliefs influenced by the naive-campaign in expectation.

Now consider the impact of a testing-campaign, given by the following expression:

$$\frac{budget - cost\ of\ testing}{CPI} \times message\ improvement \tag{2}$$

Here, the campaign's budget gets modified by the cost of testing — which is equal to the cost of creating *two or more* messages and running an in-survey RCT pre-test — before being divided by the CPI. The resulting number then gets multiplied by the expected improvement in message effectiveness achieved from the campaign conducting the in-survey RCT test and selecting the message with the largest point-estimate for disseminating in their public outreach. For example, let us assume that the cost of testing is $5,000 and allows the campaign to identify messages that are 10% more effective than average in expectation. Using the same campaign budget and CPI as in the example for

expression 1, expression 2 implies that ((105000 – 5000) / 3.41) * 1.1 = 32,258 people would have their attitudes/beliefs influenced by the testing-campaign in expectation.

Taking the ratio of the two expressions above (and simplifying the result) gives the key expression for the relative impact score of the testing-campaign vs. the naive-campaign:

$$\frac{(budget - cost\ of\ testing)}{(budget - cost\ of\ one\ message)} \times message\ improvement \quad (3)$$

If the output of expression 3 is greater than 1, then in-survey RCT pre-testing improves campaign impact in expectation and is therefore cost-effective. As expression 3 shows, and as discussed in the main text, whether or not in-survey RCT pre-testing improves campaign impact in expectation depends on the trade-off between two quantities: the expected improvement in message effectiveness from doing the pre-test versus the cost of doing the pre-test. This trade-off follows an intuitive logic, as described in the main text.

2 Description of analysis parameters

Methods Table 1 shows the key parameters and their corresponding values examined in our analysis. The subsections below explain each parameter in the table.

Methods Table 1. Parameters and values examined in our cost-effectiveness analysis.

Parameter	Values analyzed
Campaign budget	$52.5k, $105k, $210k, $420k
Cost per message (e.g., production of a brief video)	$1000
Cost per survey respondent	$0.75
Cost of RCT expertise	$0, $1000, $3000, $10,000
True mean effect size in-survey (in standardized units)	0.05, 0.1, 0.2
True variation in effects across messages (operationalized as the standard deviation in message effects divided by the average message effect)	0.1–0.8
True correlation between survey and field effects	0.1–0.9

Campaign budget

As described in the main text, in order to choose campaign budgets for our analysis, we referred to a large meta-analysis of 376 real public health campaigns that ran on social media through 2020 and 2021 that targeted people's beliefs and attitudes about the COVID-19 vaccines [37]. The average campaign spend was approximately $105,000, with a standard deviation of approximately $327,000, implying that some campaigns had substantially larger budgets than average. Thus we consider a range of budgets.

Cost per message

There are various different messaging formats that campaigns could use for their public health communications, with plausibly different production costs for the messages. In line with our focus on the aforementioned meta-analysis of COVID-19 social media campaigns, we assume the messages are brief social-media-style videos. The cost of producing one such video is likely to depend upon where it is commissioned. According to the popular marketplace Upwork (https://www.upwork.com/), many "social media videographers" charge between $50 and $100 per hour. Therefore, we assume a per-video cost of $1000, which translates to 10-20 hours of work by a social media videographer.

Cost per survey respondent

For the cost per survey respondent, we refer to popular survey providers used by behavioral scientists, such as Prolific (https://www.prolific.co/) and Cloud Research (https://www.cloudresearch.com/). For a 3-minute survey that pays $11 per hour, the cost per survey respondent on Prolific is approximately $0.75, inclusive of their service fee (on Cloud Research the figure is similar). Thus we assume a cost of $0.75 per respondent. Notably, this cost is for a convenience sample not a national probability sample.

Cost of RCT expertise

This parameter is explained in detail in the main text; thus, we refer readers there. Note that, for the results presented in Figures 1–3, the RCT expertise cost is assumed to be $0.

True mean effect size in-survey

This parameter refers to the true (i.e. latent) mean effect size in the survey environment, not the *estimated* mean effect size in a particular survey RCT. This distinction is important insofar as the estimated mean effect size in a particular survey RCT need not equal the true mean effect size due to sampling variability. We consider values of 0.05, 0.1 and 0.2 standard units, which are all considered small/tiny effect sizes by conventional academic standards [50]. In Methods section 4 we describe the evidence that informs these choices.

True variation in effects

This parameter refers to the true variation in effect sizes across messages, normalized by the effect size of the average message (i.e. the true mean effect size) — sometimes called a “coefficient of variation”. This normalization is particularly convenient because it allows us to more easily compare and aggregate the variation in effect sizes estimated in different studies (described later in Methods section 4). We consider values that range from 0.1 to 0.8 in increments of 0.1. A value of 0.1 implies that one standard deviation in message effects is equal to one-tenth of the average message effect; in other words, it implies that the messages barely vary from the average effect. By contrast, a value of 0.8 implies that one standard deviation in message effects is equal to four-fifths of the average message effect; and, thus, that we should expect a substantial minority of messages to have effects that are in the opposite direction to that of the average

message — so-called "backlash" effects. Given that backlash appears uncommon [51–53], a value of 0.8 is a reasonable upper bound on the likely variation in message effects. In Methods section 4 we describe further evidence that informs these choices.

True correlation between survey and field effects

This parameter refers to the correlation between the true effects of messages in the survey environment and the true effects of those messages in the field (i.e. an actual campaign). There are several reasons why this correlation may be less than 1.

For example, there may be between-person heterogeneity in the effects of different messages (e.g. among highly-educated respondents, message X is more effective than message Y, whereas among less educated respondents the reverse is true). If the survey sample of respondents does not represent the target population for the campaign (e.g. highly-educated respondents are overrepresented in-survey), then the most effective message in the survey may not be the most effective message in the field, thereby diminishing the survey-field correlation in message effects and the value of in-survey RCT pre-testing. (Notably, high-quality evidence suggests that between-person heterogeneity in message effects tends to be small [54,55], though even small absolute differences could still have important implications for campaign impact [56].) Another reason why the survey-field correlation may be less than 1 is that surveys can typically only measure self-reported outcomes (e.g. vaccination intentions), while campaigns are typically interested in changing actual behavior (e.g. vaccination uptake). It could be that the messages' effectiveness is not perfectly aligned across these outcome variables (e.g. message X performs better on self-reported outcomes than message Y, but vice versa for behavioral outcomes). This could also diminish the survey-field correlation in message effects. Notably, there are other reasons why the survey-field correlation may be less than 1 which, for brevity, we do not discuss in detail here (e.g. uneven decay in treatment effects across messages, attention driving treatment effects in the field more than in-survey, etc.).

We consider a range of positive values for the survey-field correlation: from 0.1 to 0.9, in increments of 0.1. We do not consider negative values because a correlation of 0.1 is small enough that in-survey RCT testing would not be cost-effective for most campaigns. In Methods section 4 we describe evidence that can speak to the value of this parameter.

3 Simulating in-survey RCT pre-tests

For each unique set of parameter values, we simulate three-thousand in-survey RCT tests. The purpose of this simulation is to estimate the expected improvement in message effectiveness from running the RCT and selecting the message with the largest estimated effect. The procedure for each of the simulated RCTs proceeds in the following steps.

First, we draw X messages from a bivariate gaussian distribution, where X is an integer between 2 and 20. The two dimensions of the bivariate distribution correspond to the true message effects in-survey vs. the true effects of the messages in the field. The distribution has a mean, equal to the true mean effect size in-survey (e.g. 0.1), as well as a covariance matrix that captures both (a) the true variation in the message effects and (b) the true correlation between the two (in-survey vs. in-field) dimensions of the effects.

We then randomly assign each of Y simulated respondents to one of the messages, where Y is an integer between 500 and 20,000. We also assign some of the simulated respondents to a control group, to account for the fact that it is desirable for campaigns to include a control group (that receives no message) to check that their messages have an effect in the intended direction. Specifically, each message group is assigned Y/(M+1) respondents, where M is the number of messages in the simulation. After respondents have been assigned, we add noise to the true in-survey effect of each message to simulate sampling variability. The noise assumes that 10% of the variance in the outcome variable can be explained by adjusting for pre-treatment covariates (e.g. age, gender, etc.).

In a final step, we first identify the message with the largest *estimated* effect in the survey. Then, we consult the vector that contains the *true* effects of the messages, and select the true in-field effect size of that message. In other words, we assume that the message that gets selected on the basis of the survey RCT exerts an effect in the actual campaign that is equal to its true in-field effect size (similar to previous work [57]).

After three-thousand simulated RCTs, we take the mean of the three-thousand selected true in-field effects, and divide it by the true mean in-field effect. This ratio thus tells us the

expected improvement in message effectiveness from performing the in-survey RCT (compared to the naive-campaign that did not run any in-survey RCT test). For example, if the ratio is 1.2, this tells us that the in-survey RCT test procedure identifies messages that are 20% more effective than the average message in expectation. Finally, with this ratio in hand, we use expression 3 (see Methods section 1) to compute the expected impact of the testing-campaign and the naive-campaign given all the other parameters in question.

For the bare-minimum testing regime, the number of messages and survey respondents in the RCT is always the same, irrespective of the parameter values in question: n messages = 2 and n respondents per message = 300. By contrast, for the optimal testing regime, for each unique combination of parameter values we perform a grid search over the joint distribution of n messages (2–20) and n respondents (500–20,000) to find the combination of messages and respondents that maximizes the expected impact of the testing-campaign. Notably, if the expected impact of the testing-campaign under the optimal testing regime is smaller than that of the naive-campaign, we assume that the testing-campaign decides not to do any testing. In such cases, the expected impact of the testing-campaign is simply equal to that of the naive-campaign. In practice this means that the optimal testing regime, because it assumes perfect knowledge of the parameter values, can never generate less impact than the naive-campaign in our analysis.

4 Evidence for parameter value sets

In this section we describe the evidence underlying our determination of the best-guess, pessimistic, and optimistic sets of parameter values.

True mean effect size in-survey

To determine plausible values for this parameter, we relied primarily on a 2022 systematic review of RCTs that evaluated interventions to increase COVID-19 vaccine uptake [58]. We examined all of the studies in that review that used an in-survey RCT design to evaluate one or more message interventions on people's self-reported attitudes, beliefs, and/or behavioral intentions related to the COVID-19 vaccines. This amounted to 25 studies. To these 25 studies we added a further 11 studies, all of which focused on COVID-19 outcomes, and one meta-analysis of public health communication on various non-COVID

outcomes. These extra studies were identified using snowball sampling and our knowledge of the literature. We therefore examined 37 effect sizes in total.

For each study, we extracted the average effect size across the interventions and, where necessary/possible, standardized the effect size by dividing by the standard deviation of the outcome variable. In many cases, this required a back-of-the-envelope calculation. For example, some studies reported their effect sizes in percentage points, but did not report the standard deviation of the outcome variable, meaning we could not calculate the standardized effect size directly. In cases like this, we took a maximally conservative approach and used the standard deviation of a uniform distribution over a 0-1 binary scale (equal to 0.5), ensuring that we would err on the side of underestimating the standardized effect size (Appendix Table 1 provides further detail about the studies and our calculations). In addition, some of the studies lacked a "pure" control group that received no relevant information; instead, people in the control group received baseline relevant information. This also renders our standardized effect size estimate conservative. Some studies did not report the actual point estimates of the treatment effects — opting to display them in figures only — so we approximated the estimates based on the figures. Finally, where studies included multiple relevant outcome variables, we took the mean across the estimated treatment effects for each outcome variable.

Across the 37 extracted effect sizes, the overall mean standardized effect size is 0.12 and the median is 0.08. Notably, we do not compute a precision-weighted average because one of the studies [49] has a sample size (~484,000) that is several orders of magnitude larger than any of the others. It is desirable to avoid letting that study dominate the average effect size here because that would unduly privilege the very specific outcome variable and intervention type used in that study over the many possible outcome variables and intervention types that may be relevant in a pandemic context. In summary, therefore, on the basis of this evidence, we assume a best-guess mean effect size of 0.1 standard units. For the pessimistic and optimistic values, we halve and double the best-guess value, respectively, giving values of 0.05 (pessimistic) and 0.20 (optimistic). These effect sizes are all considered small or tiny by conventional academic standards [50].

True variation in effects

In contrast to the average message effect size, there is less evidence from prior work that can speak to the variation in effects across messages. This is because the evidence for this parameter must meet more demanding criteria. Specifically, to be informative for the context in which campaigns would perform an in-survey RCT test, the message effects must be estimated on the same outcome variable, the same types of people, and under broadly similar background conditions. This rules out meta-analyses of public health communication (and most other types of communication), since meta-analyses nearly always combine studies which differ from one another along one of the aforementioned dimensions. Generally these differences will inflate the variation in message effects because, for example, some outcome variables or types of people are more receptive to interventions. Thus, excluding meta-analyses renders our estimate of the variation in message effects conservative. To obtain estimates of the variation that are relevant for the context in which campaigns would perform an in-survey RCT test, the effects of multiple different messages must ideally be estimated within the same study.

Furthermore, because our quantity of interest is the variation across message effects, studies that include just several different messages can only offer extremely uncertain estimates of this variation; akin to estimating the standard deviation of a variable for which there are only several data points — each of which is itself measured with uncertainty. For this reason, studies should estimate the effects of more than a handful of different messages, ideally many more. Such studies will typically demand large sample sizes of survey respondents. Lastly, even when such studies have been conducted, they must have actually reported an estimate of the variation in message effects.

We were unable to locate any studies of public health communication that met all of these criteria. In particular, we could not find any such studies that reported an estimate of the variation in message effects, even though some had investigated more than a handful of different messages. Thus, for the current paper, we sought to identify relevant studies and re-analyze their data ourselves in order to estimate the variation in message effects. To that end, we drew on the aforementioned systematic review of COVID-19 interventions [58], as well as snowball sampling and our knowledge of the literature, to identify 14 studies of public health communication that investigated at least five different messages. Of these

studies, however, one did not contain a control group and so could not be included in our re-analysis because a control group is necessary to estimate each of the message effects. Another study included some messages that tried to *dis*courage vaccination — this study was also excluded from our re-analysis as it had fewer than 5 messages aimed at encouraging vaccination. A further two studies we identified used highly overlapping data, so we only included the data from one study in our re-analysis. Finally, while some of the studies had publicly available data, many did not; and we were unable to access the data of one study despite contacting the corresponding author listed on the article. This left 10 studies whose data we could re-analyze here (detailed in Appendix Table 2).

We used the following strategy to estimate the variation in message effects for each study. First, we computed the average treatment effect of each message relative to the study's control group. We then conducted a random-effects meta-analysis [59] of the treatment effects for a given study, which provided an estimate of: (1) the standard deviation (SD) in message effects, taking into account the error with which each message effect is estimated, as well as (2) the mean message effect (i.e. the effect size of the average message). Finally, for each study, we normalized the estimated SD by dividing it by the estimated mean message effect — thus, allowing us to sensibly aggregate the estimated SDs across studies. (Without this normalization, the estimated SDs in message effects are not comparable across studies because different studies use different outcome variables, and, for example, some outcomes could have larger estimated SDs simply because of their scale). If a study included multiple outcomes, we conducted the above analysis for each outcome and then computed the mean normalized SD across outcomes. Further details about these analyses are reported in Appendix Table 2.

Across the 10 re-analyzed studies, the mean normalized SD is 0.59 and the median is 0.27. In other words, across these studies, one standard deviation in message effects is estimated to be equal to between one-quarter and three-fifths of the size of the average message effect. However, as expected, there is substantial heterogeneity in the SDs across studies (Appendix Table 2). This is likely due in part to the small numbers of messages in each study; the median number of messages investigated was just 7. Thus, the estimated SDs are likely to be highly heterogeneous between studies due to sampling variability alone, and, as a result, the average SD across studies is correspondingly

uncertain. In addition, most of the studies examined messages that were simply lines of text. However, in many public health communication contexts, such as the COVID-19 campaigns that ran on social media in 2020/21 [37], the messages are likely to be professionally-produced videos (this is why, in our cost-effectiveness analysis, we assume a cost-per-message of $1000; to capture video production costs). This is important because the true variation in message effects may be larger when the messages in question differ not only in their text content but also in their video/audio content.

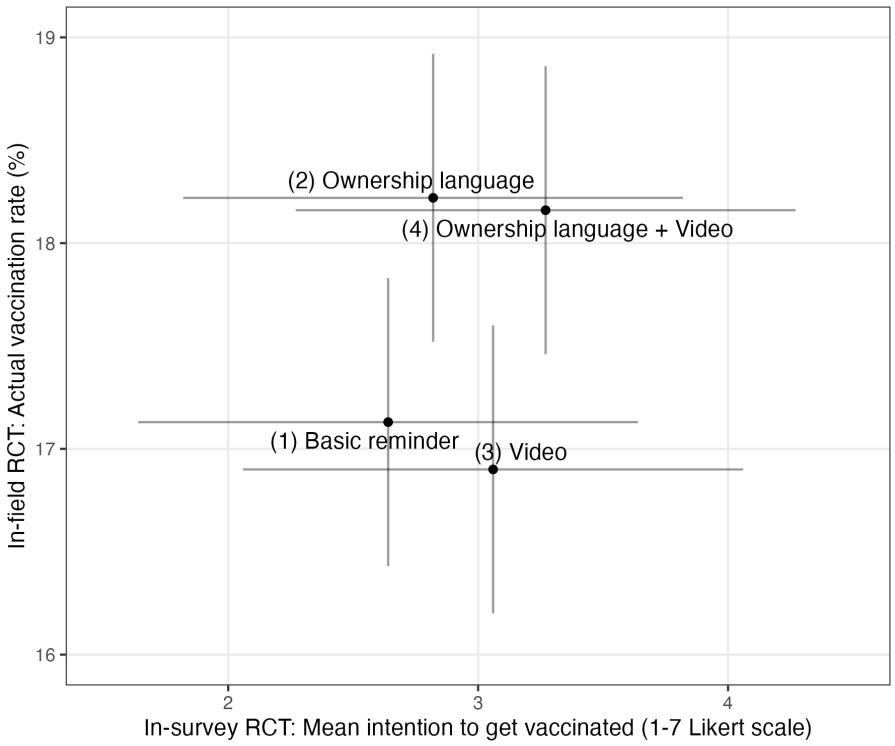
To account for these issues — the small samples of messages and lack of video/audio content in each study — we supplemented our re-analysis with 3 additional studies from the domain of political communication, sourced via our knowledge of the literature [54,57,60]. The details of these studies are also reported in Appendix Table 2. Crucially, these additional studies had unusually large samples of messages — the median number of messages investigated was 59 — and the messages in each study were all short videos with audio content. Furthermore, each of these studies already reports an estimate of the SD in message effects, obviating the need for us to re-analyze their data. The mean normalized SD across these three studies is 1.23 and the median is 0.95. Thus, while these studies should receive less weight given that their focus is political communication rather than public health communication, their estimates suggest that variation in message effects may indeed be larger when messages differ in video/audio content.

We consider this evidence together with our estimate of the average variation across the studies of public health communication from earlier (i.e. mean = 0.59, median = 0.27). Thus, overall we settle on a best-guess value of 0.4 for the variation parameter. A value of 0.4 implies that the true standard deviation in message effects is equal to two-fifths of the effect size of the average message. For example, if the effect size of the average message is 0.1 standard units, a variation parameter of 0.4 implies an SD of 0.4*0.1 = 0.04; meaning that most message effects will fall between 0.06 and 0.14 (i.e. $0.1 \pm$ SD = 0.1 ± 0.04). By extension, these numbers imply that most messages will have true effects that are small by conventional academic standards and not too dissimilar to the mean message effect. For the pessimistic and optimistic values of this parameter, we halve and 1.5x the best-guess value, respectively, giving values of 0.2 (pessimistic) and 0.6 (optimistic).

True correlation between survey and field effects

Evidence for this parameter is the most demanding of all the parameters we consider in our analysis. In order to estimate the correlation between the effects of messages in-survey and those same messages in the field, relevant studies are those that have conducted a survey RCT and field RCT *in parallel using the same set of messages*, and ideally using a large sample of messages to minimize sampling variability. Because field RCTs that study a large number of messages are resource-intensive, the pool of potentially relevant studies is small at the outset. Prominent field RCTs that include a large number of different public health messages [30,61,62] typically do not conduct a parallel in-survey RCT and, as a result, typically cannot estimate the survey-field correlation.

We located just one study of public health communication that conducted a field RCT and survey RCT in parallel, using a set of four messages [36]. In the study, respondents in the field RCT were randomized to receive (via cell phone text) one of four different treatment messages encouraging them to schedule an appointment for their COVID-19 vaccination. The messages consisted of either (1) a basic reminder, (2) a reminder that used "ownership" language, (3) a basic reminder with video content or (4) the ownership-reminder with the video content. The y axis of Methods Figure 1 shows the estimated vaccination rate in each of these message groups from the field RCT (copied from Figure 2b in ref. [36]); the x axis shows people's self-reported vaccination intentions in each of the message groups from the corresponding in-survey RCT in which vaccination intentions were measured (from Figure 5b in the supplement of ref. [36]). We approximate confidence intervals from visual inspection of the standard errors as we could not locate the standard errors/confidence intervals on the means reported in numeric form.



Methods Figure 1. Message effects from survey RCT and field RCT reported by ref. [36]. Note that the displayed confidence intervals are approximate (see in-text).

Methods Figure 1 could be interpreted as providing evidence against a positive correlation between survey and field message effects because, in the survey RCT, the messages with the video content performed better on average, whereas those with the ownership language did not. By contrast, in the field RCT, the reverse pattern was observed. Nevertheless, closer inspection of the results suggests this interpretation is not quite right. In particular, in the survey RCT the top performing message on respondents' vaccination intentions was message #4. In the field RCT, this message was also a top performer on increasing vaccination rates; its point estimate was close to that of the "winning" message (#2). The implication is that, had a public health campaign selected the top performing message from the survey RCT (#4) for running in their outreach to increase vaccination uptake, it would have been a reasonably good choice. By contrast, had they tried to infer a general principle of message development — such as "video content performs best" — they would have been misled. But, recall that the goal of in-survey RCT pre-testing (as considered in the current article) is not to infer such

principles; rather, it is to select the top-performing individual message from among several different messages tested. Added to this interpretational ambiguity, furthermore, is the fact that the evidence in Methods Figure 1 for the value of the survey-field correlation is extremely limited by the small set of only four messages — all of whose effects are estimated with relatively large amounts of noise. As a result, we are cautious to conclude much from this evidence about the likely value of the survey-field correlation parameter.

In an effort to gather more evidence, we looked to larger studies conducted outside the domain of public health communication [63–65]. These studies point toward at least a moderate positive survey-field correlation. We briefly describe them below.

Hainmueller et al. [63] studied data from Switzerland in which some municipalities used referendums to vote on the naturalization applications of immigrants. In the referendums, voters received a leaflet with a short description of the applicant, including information about their attributes, such as age, sex, education, and so on, and then cast a secret ballot to accept or reject individual applicants one at a time. Voters decided over thousands of immigrants with varying characteristics, allowing the authors to causally identify how much each particular attribute affected the probability of being accepted or rejected by voters in a real-world setting. Ten years later, the authors conducted survey experiments in which survey respondents completed a hypothetical referendum task, choosing whether to accept or reject hypothetical immigrant profiles based on similar attributes as in the real referendums. The authors then used estimates of attribute importance from their survey data to generate predicted probabilities of acceptance for each of the immigrant applications from the real referendums. These survey-based predictions were correlated at 0.5 (on average) with the probabilities of acceptance for each application generated by the model that was fitted to the actual referendum data. In other words, the survey-based estimates were correlated with the field-based estimates at an average of 0.5, even despite a ten year gap between the two sets of estimates being collected.

Another piece of evidence comes from Coppock and Green [64], who examined paired survey and field effect sizes from 12 different studies of political behavior phenomena. They estimated an overall rank-order correlation of 0.73 between the pairs of estimates.

A final piece of evidence comes from O'Keefe [65]. One mechanism through which in-survey message effects may be poorly correlated with field effects is that the outcome variable in a survey is typically self-reported — e.g. behavioral *intentions* — whereas the outcome in the field is the actual behavior. O'Keefe conducted a meta-analysis of 317 studies in which two messages were compared (e.g. a loss-framed message vs. a gain-framed message) on different outcome variables: self-reported outcomes (intentions, attitudes) and behavioral outcomes. He examined how often the direction of the difference between messages was the same on the different types of outcome, such as the loss-framed message having a larger effect than the gain-framed message on behavioral *intentions* as well as on *behavior*. He found that, in 82% of possible comparisons (49/60), the direction of the difference between messages was the same on the attitude outcome as it was on the behavioral outcome; while, in 94% of possible comparisons (102/109), the direction was the same on the behavioral intention outcome as it was on the behavioral outcome.

These percentages (82%, 94%) imply correlations of at least 0.85 between the message effects estimated on the self-reported outcomes and those estimated on the behavioral outcomes. To determine this, we conducted a simple simulation in which we sampled two "messages" from a bivariate normal distribution, and selected the message with the highest value on the first dimension. We then recorded whether this was also the message with the highest value on the second dimension; that is, whether the rank order of the message values was the same on both dimensions. When the true correlation between messages is set to 0.85, the sampled messages have the same rank order approximately 82% of the time. The O'Keefe study thus suggests that the survey-field correlation is not much attenuated by the fact that in-survey RCTs rely on self-reported outcomes.

In summary, there is very limited evidence regarding the value of the correlation between in-survey message effects and the effects of those messages in the field (i.e. in an actual public health campaign). However, the little evidence that does exist is either inconclusive or points towards a moderate-to-strong correlation. Thus, considering this evidence altogether, we settle on a best-guess correlation of 0.5 between in-survey and in-field message effects for our context. To give an intuitive sense of what this means, a correlation of 0.5 implies that, if one were to correctly identify the best of two different messages in a survey RCT, that message would also be the best message in the field

approximately 66% of the time (this percentage is determined using the simulation approach described in the previous paragraph). Reflecting the limited evidence, we use wide pessimistic and optimistic values for the correlation: 0.2 and 0.8, respectively.

References

1. Estimated cumulative excess deaths during COVID. *Our World in Data* https://ourworldindata.org/grapher/excess-deaths-cumulative-economist-single-entity .
2. Juan, C. What Comes After COVID—Asterisk. https://asteriskmag.com/issues/2/what-comes-after-covid (2023).
3. Marani, M., Katul, G. G., Pan, W. K. & Parolari, A. J. Intensity and frequency of extreme novel epidemics. *Proc. Natl. Acad. Sci.* **118**, e2105482118 (2021).
4. Inglesby, T. Opinion | How to Prepare for the Next Pandemic. *The New York Times* (2023).
5. World Health Organization. How global research can end this pandemic and tackle future ones. https://www.who.int/publications/m/item/how-global-research-can-end-this-pandemic-and-tackle-future-ones (2022).
6. UK Government. *National Risk Register 2023*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1175834/2023_NATIONAL_RISK_REGISTER_NRR.pdf (2023).
7. No government can address the threat of pandemics alone – we must come together. *GOV.UK* https://www.gov.uk/government/speeches/no-government-can-address-the-threat-of-pandemics-alone-we-must-come-together.
8. Vora, N. M. *et al.* Want to prevent pandemics? Stop spillovers. *Nature* **605**, 419–422 (2022).
9. Dzau, V. & Yadav, P. The influenza imperative: we must prepare now for seasonal and pandemic influenza. *Lancet Microbe* **4**, e203–e205 (2023).

10. Sen. Risch, J. E. [R-I. S.2297 - 117th Congress (2021-2022): International Pandemic Preparedness and COVID-19 Response Act of 2021. http://www.congress.gov/ (2021).
11. Preventing catastrophic pandemics. *80,000 Hours* https://80000hours.org/problem-profiles/preventing-catastrophic-pandemics/.
12. CEPI | New Vaccines For A Safer World. *CEPI* https://cepi.net/.
13. Betsch, C. *et al.* A call for immediate action to increase COVID-19 vaccination uptake to prepare for the third pandemic winter. *Nat. Commun.* **13**, 7511 (2022).
14. Pandemic preparedness. *Nature* https://www.nature.com/collections/jaacfgeief (2022).
15. Nuzzo, J. B. & Ledesma, J. R. Why Did the Best Prepared Country in the World Fare So Poorly during COVID? *J. Econ. Perspect.* **37**, 3–22 (2023).
16. Bavel, J. J. V. *et al.* Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* **4**, 460–471 (2020).
17. Ruggeri, K. *et al.* A synthesis of evidence for policy from behavioural science during COVID-19. *Nature* 1–14 (2023) doi:10.1038/s41586-023-06840-9.
18. Rand, D. G. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *J. Theor. Biol.* **299**, 172–179 (2012).
19. Berinsky, A. J., Huber, G. A. & Lenz, G. S. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Polit. Anal.* **20**, 351–368 (2012).
20. Chandler, J. & Shapiro, D. Conducting Clinical Research Using Crowdsourced Convenience Samples. *Annu. Rev. Clin. Psychol.* **12**, 53–81 (2016).
21. Peyton, K., Huber, G. A. & Coppock, A. The Generalizability of Online Experiments Conducted During the COVID-19 Pandemic. *J. Exp. Polit. Sci.* **9**, 379–394 (2022).
22. Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J. & Litman, L. Online panels in

social science research: Expanding sampling methods beyond Mechanical Turk. *Behav. Res. Methods* **51**, 2022–2038 (2019).

23. Fowler, C., Jiao, J. & Pitts, M. Frustration and ennui among Amazon MTurk workers. *Behav. Res. Methods* (2022) doi:10.3758/s13428-022-01955-9.
24. Robinson, J., Rosenzweig, C., Moss, A. J. & Litman, L. Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. *PLOS ONE* **14**, e0226394 (2019).
25. Coppock, A. & McClellan, O. Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Res. Polit.* **6**, 2053168018822174 (2019).
26. Shen, F., Sheer, V. C. & Li, R. Impact of Narratives on Persuasion in Health Communication: A Meta-Analysis. *J. Advert.* **44**, 105–113 (2015).
27. O'Keefe, D. J. & Hoeken, H. Message Design Choices Don't Make Much Difference to Persuasiveness and Can't Be Counted On—Not Even When Moderating Conditions Are Specified. *Front. Psychol.* **12**, (2021).
28. O'Keefe, D. Evidence-based advertising using persuasion principles: Predictive validity and proof of concept. *Eur. J. Mark.* **50**, 294–300 (2016).
29. Dimant, E., Clemente, E. G., Pieper, D., Dreber, A. & Gelfand, M. Politicizing mask-wearing: predicting the success of behavioral interventions among republicans and democrats in the U.S. *Sci. Rep.* **12**, 7575 (2022).
30. Milkman, K. L. & et al. A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proc. Natl. Acad. Sci.* **119**, e2115126119 (2022).
31. Druckman, J. N. A Framework for the Study of Persuasion. *Annu. Rev. Polit. Sci.* (2021) doi:10.2139/ssrn.3849077.
32. Blumenau, J. & Lauderdale, B. E. The Variable Persuasiveness of Political Rhetoric.

Am. J. Polit. Sci. (2021).

33. Rode, J. B. *et al.* Influencing climate change attitudes in the United States: A systematic review and meta-analysis. *J. Environ. Psychol.* **76**, 101623 (2021).
34. Bowen, D. Simple models predict behavior at least as well as behavioral scientists. Preprint at https://doi.org/10.48550/arXiv.2208.01167 (2022).
35. Broockman, D., Kalla, J., Caballero, C. & Easton, M. Political practitioners poorly predict which messages persuade the public. Preprint at https://doi.org/10.31219/osf.io/8un6a (2023).
36. Dai, H. *et al.* Behavioural nudges increase COVID-19 vaccinations. *Nature* **597**, 404–409 (2021).
37. Athey, S., Grabarz, K., Luca, M. & Wernerfelt, N. Digital public health interventions at scale: The impact of social media advertising on beliefs and outcomes related to COVID vaccines. *Proc. Natl. Acad. Sci.* **120**, e2208110120 (2023).
38. Jan G. Voelkel & et al. Megastudy identifying effective interventions to strengthen Americans' democratic attitudes. *Strengthening Democracy Challenge* https://www.strengtheningdemocracychallenge.org/paper (2022).
39. Vlasceanu, M., Doell, K., Bak-Coleman, J. & Bavel, J. J. V. Addressing Climate Change with Behavioral Science: A Global Intervention Tournament in 63 Countries. Preprint at https://doi.org/10.31234/osf.io/cr5at (2023).
40. Milkman, K. L. & at al. Megastudies improve the impact of applied behavioural science. *Nature* **600**, 478–483 (2021).
41. Duckworth, A. L. & Milkman, K. L. A guide to megastudies. *PNAS Nexus* **1**, pgac214 (2022).
42. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nat. Hum. Behav.* 1–7 (2020) doi:10.1038/s41562-020-0912-z.

43. McManus, J., Constable, M., Bunten, A. & Chadborn, T. *Improving people's health: Applying behavioural and social sciences to improve population health and wellbeing in England*. (2018).
44. Chater, N. & Loewenstein, G. The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behav. Brain Sci.* 1–60 (2022) doi:10.1017/S0140525X22002023.
45. Hagmann, D., Ho, E. H. & Loewenstein, G. Nudging out support for a carbon tax. *Nat. Clim. Change* **9**, 484–489 (2019).
46. Sheeran, P. Intention—Behavior Relations: A Conceptual and Empirical Review. *Eur. Rev. Soc. Psychol.* **12**, 1–36 (2002).
47. Webb, T. L. & Sheeran, P. Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychol. Bull.* **132**, 249–268 (2006).
48. Rhodes, R. E. & Dickau, L. Experimental evidence for the intention–behavior relationship in the physical activity domain: A meta-analysis. *Health Psychol.* **31**, 724–727 (2012).
49. Moehring, A. *et al.* Providing normative information increases intentions to accept a COVID-19 vaccine. *Nat. Commun.* **14**, 126 (2023).
50. Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front. Psychol.* **4**, (2013).
51. Guess, A. & Coppock, A. Does Counter-Attitudinal Information Cause Backlash? Results from Three Large Survey Experiments. *Br. J. Polit. Sci.* 1–19 (2020) doi:10.1017/S0007123418000327.
52. Tappin, B. M. & Gadsby, S. Biased belief in the Bayesian brain: A deeper look at the evidence. *Conscious. Cogn.* **68**, 107–114 (2019).

53. Wood, T. & Porter, E. The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Polit. Behav.* **41**, 135–163 (2019).
54. Coppock, A., Hill, S. J. & Vavreck, L. The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. *Sci. Adv.* **6**, eabc4046 (2020).
55. Coppock, A. *Persuasion in Parallel*. (University of Chicago Press, 2022).
56. Tappin, B. M., Wittenberg, C., Hewitt, L. B., Berinsky, A. J. & Rand, D. G. Quantifying the potential persuasive returns to political microtargeting. *Proc. Natl. Acad. Sci.* **120**, e2216261120 (2023).
57. Hewitt, L. *et al.* How experiments help campaigns persuade voters: evidence from a large archive of campaigns' own experiments. *Am. Polit. Sci. Rev.* (2023).
58. Batteux, E., Mills, F., Jones, L. F., Symons, C. & Weston, D. The Effectiveness of Interventions for Increasing COVID-19 Vaccine Uptake: A Systematic Review. *Vaccines* **10**, 386 (2022).
59. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.* **36**, 1–48 (2010).
60. Hewitt, L. & Tappin, B. M. Rank-heterogeneous effects of political messages: Evidence from randomized survey experiments testing 59 video treatments. Preprint at https://doi.org/10.31234/osf.io/xk6t3 (2022).
61. Milkman, K. L. & et al. A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proc. Natl. Acad. Sci.* **118**, e2101165118 (2021).
62. Patel, M. S. *et al.* A Randomized Trial of Behavioral Nudges Delivered Through Text Messages to Increase Influenza Vaccination Among Patients With an Upcoming Primary Care Visit. *Am. J. Health Promot.* **37**, 324–332 (2023).

63. Hainmueller, J., Hangartner, D. & Yamamoto, T. Validating vignette and conjoint survey experiments against real-world behavior. *Proc. Natl. Acad. Sci.* **112**, 2395–2400 (2015).
64. Coppock, A. & Green, D. P. Assessing the Correspondence between Experimental Results Obtained in the Lab and Field: A Review of Recent Social Science Research. *Polit. Sci. Res. Methods* **3**, 113–131 (2015).
65. O'Keefe, D. J. Persuasive Message Pretesting Using Non-Behavioral Outcomes: Differences in Attitudinal and Intention Effects as Diagnostic of Differences in Behavioral Effects. *J. Commun.* **71**, 623–645 (2021).

Appendix: Using in-survey randomized controlled trials to support future pandemic response

Ben M. Tappin Luke B. Hewitt

Contents

1 Computing weighted-average relative impact scores **2**

2 Relative-impact curves for campaigns with different budgets **5**

3 Benefit of possessing accurate knowledge of the parameter values **7**

4 Reviewing evidence for parameter values **12**
4.1 True mean effect size in-survey 12
4.2 True variation in effect sizes 26

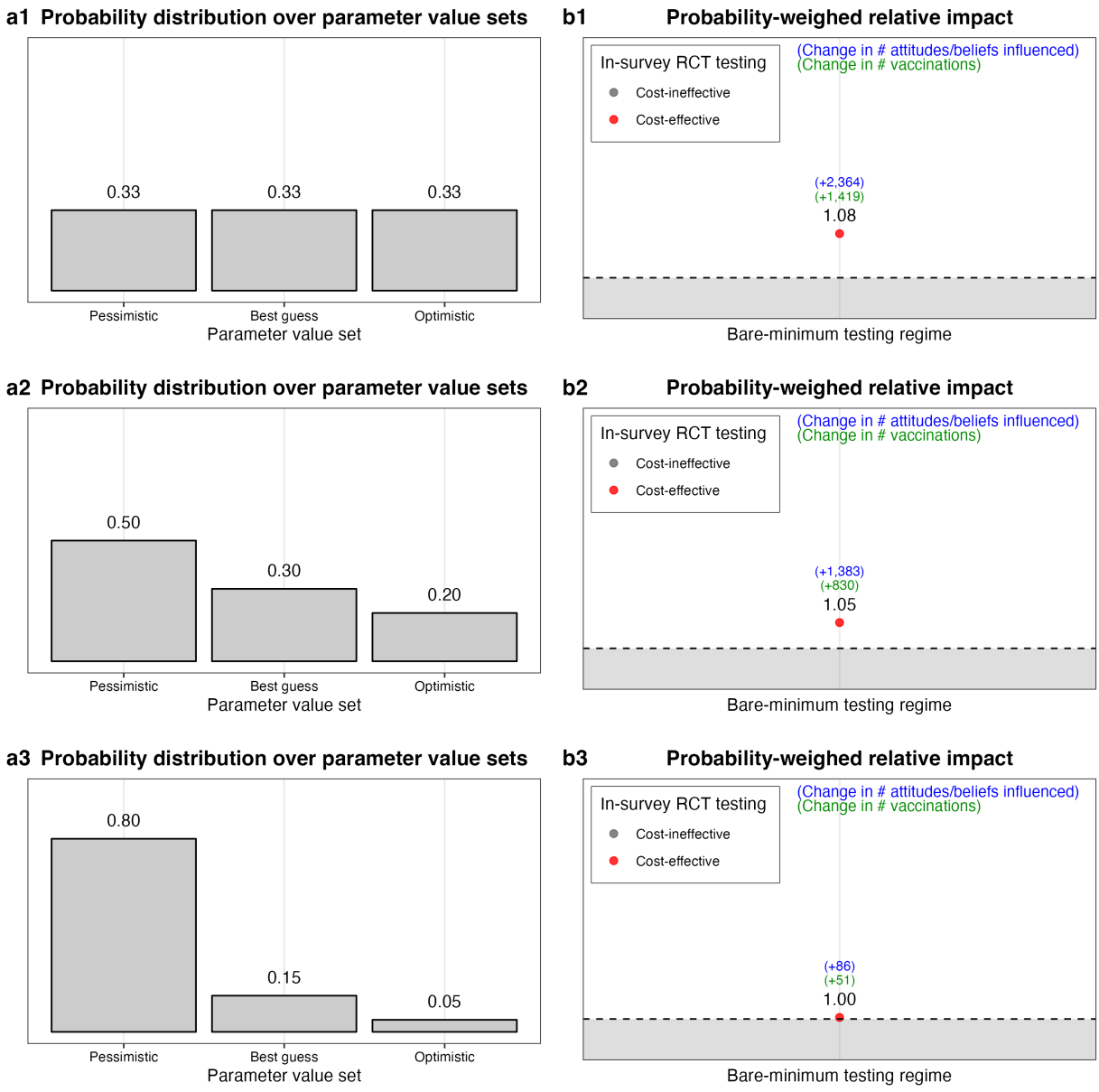
5 References **32**

1 Computing weighted-average relative impact scores

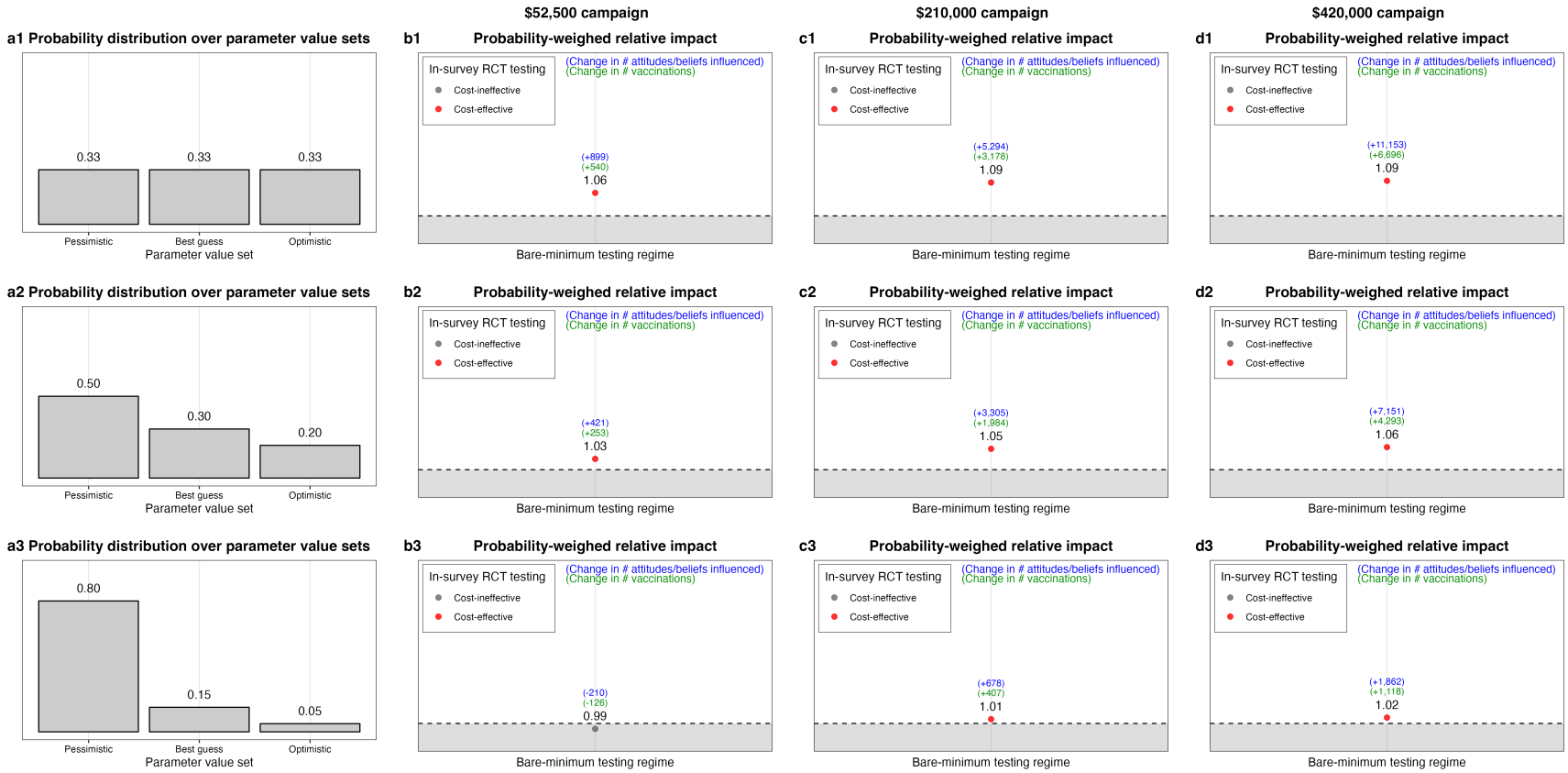
Given that we have uncertainty about which set of parameter values is correct, Appendix Figure 1 shows the implication of assuming different distributions of uncertainty for a $105,000 campaign. In Appendix Figure 1a1, for example, we represent each set of parameter values as being equally likely by assigning each a probability of 1/3. We then compute a weighted-average relative impact of the testing-campaign over the naive-campaign across the three sets of parameter values, using the assigned probabilities as the weights. The result shows that the bare-minimum testing regime is expected to increase campaign impact on average, netting at least one-thousand extra attitudes/beliefs influenced and vaccinations received (Appendix Figure 1b1).

We can also consider other uncertainty distributions to examine how robust the returns to in-survey RCT testing are against pessimism. For example, if we assign greater probability to the pessimistic scenario (0.5) than either the best-guess (0.3) or optimistic (0.2) scenarios, the weighted-average relative impact is still greater than 1 (Appendix Figure 1b2). Even when we assume the pessimistic scenario is substantially more likely (0.8) than either of the other two scenarios (0.15 and 0.05, respectively), Appendix Figure 1a3, in-survey RCT pre-testing remains cost-effective under the bare-minimum testing regime (Appendix Figure 1b3). In summary, these results suggest that, even under conservative assumptions, in-survey RCT pre-testing is likely cost-effective for public health campaigns with a budget of $105,000.

Finally, we consider the impact of incorporating uncertainty over the parameter values for campaigns with different budgets. Appendix Figure 2 shows the weighted-average relative-impact estimates for each campaign budget and reinforces the above results: for the larger campaigns, in-survey RCT testing is clearly and robustly cost-effective — even if one places a large amount of probability (0.8) on the pessimistic set of parameter values. For the smaller campaign, in-survey RCT testing is robust against some pessimism, but is not cost-effective under stronger pessimism.



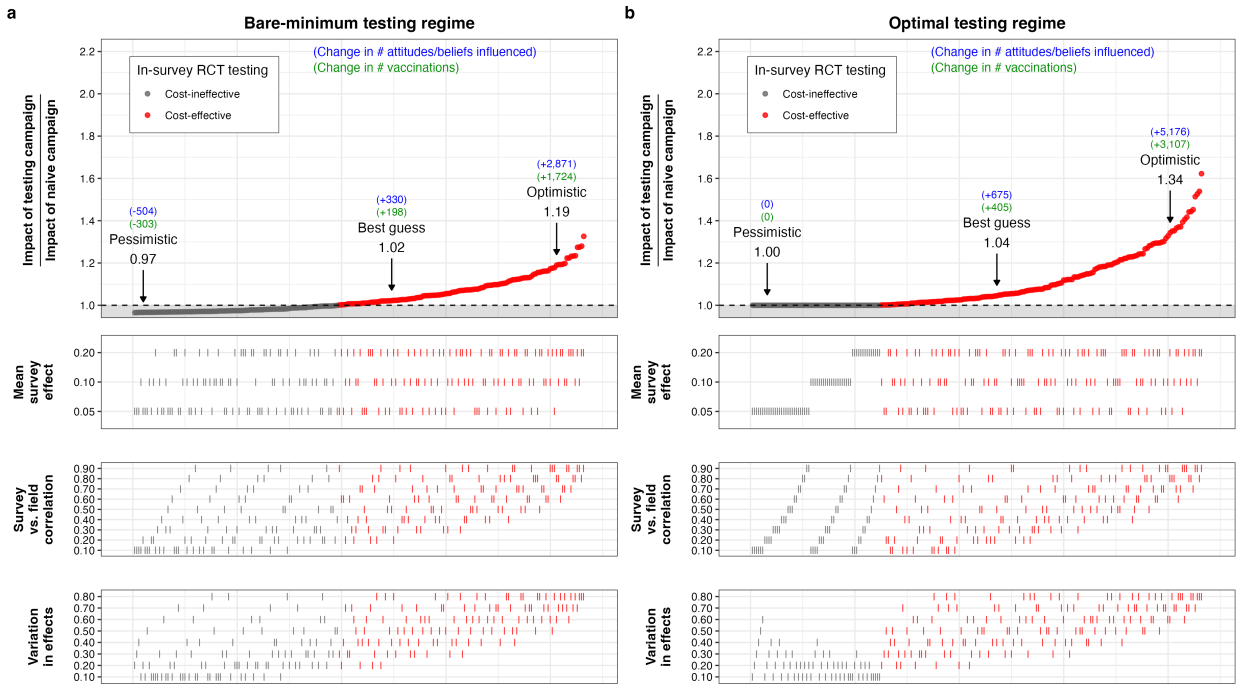
Appendix Figure 1: Weighted-average relative-impact score for a campaign using the bare-minimum testing regime with a budget of $105,000.



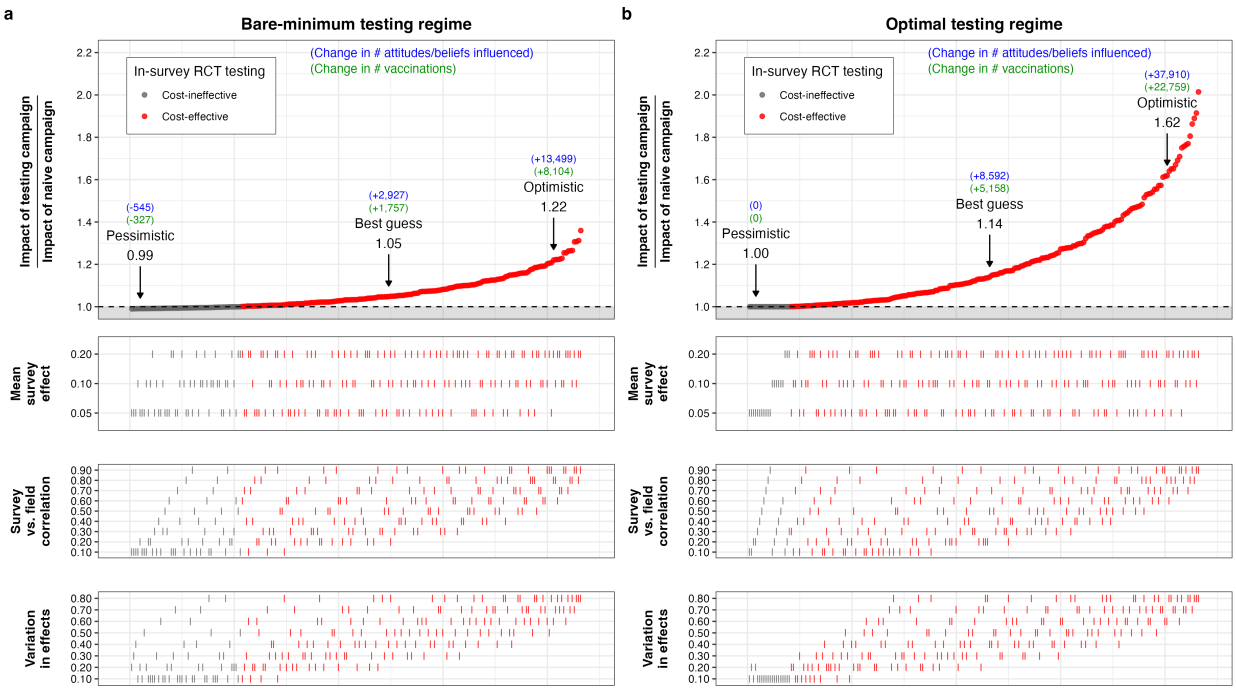
Appendix Figure 2: Weighted-average relative-impact score for campaigns using the bare-minimum testing regime with different budgets.

2 Relative-impact curves for campaigns with different budgets

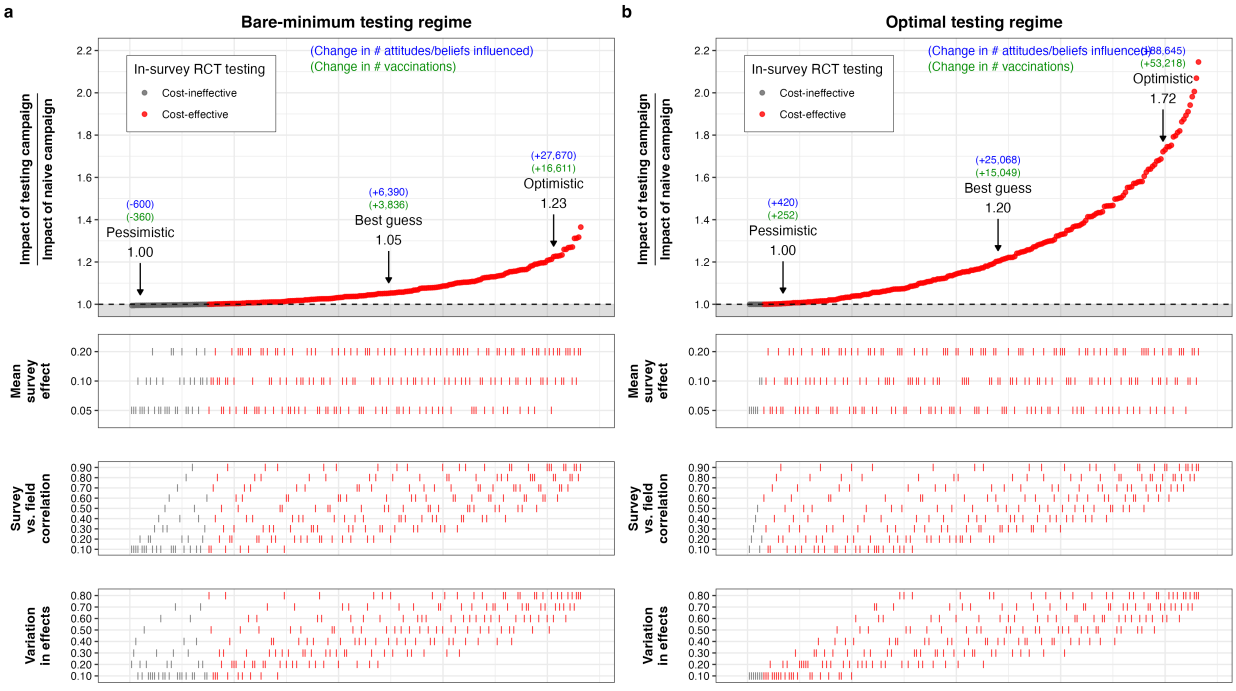
Appendix Figures 3, 4 and 5 show, respectively, the estimated relative-impact curves for campaigns with budgets of $52,500, $210,000 and $420,000, with lower panels showing the corresponding parameter values.



Appendix Figure 3: Relative-impact curves for a campaign with a budget of $52,500.



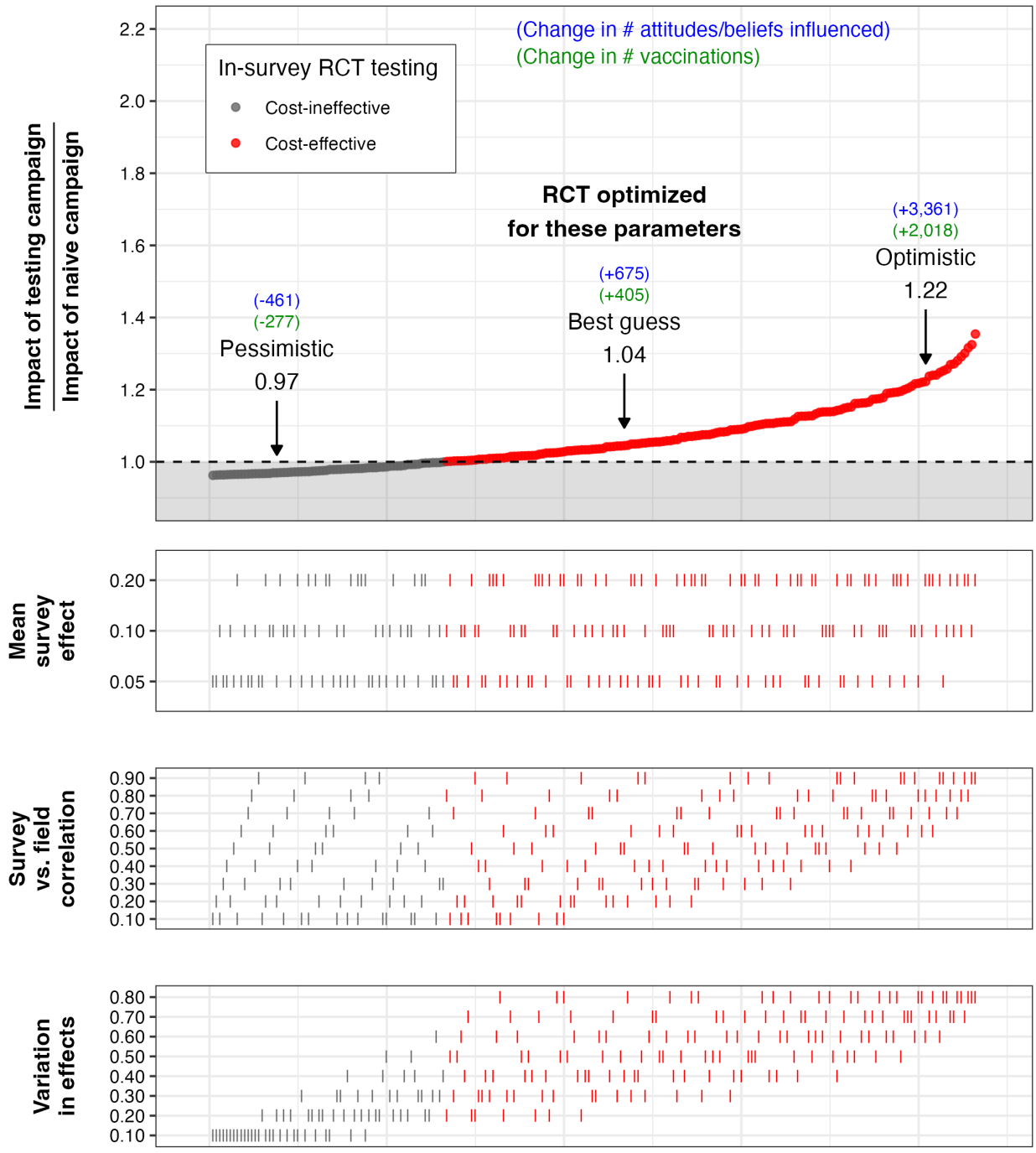
Appendix Figure 4: Relative-impact curves for a campaign with a budget of $210,000.



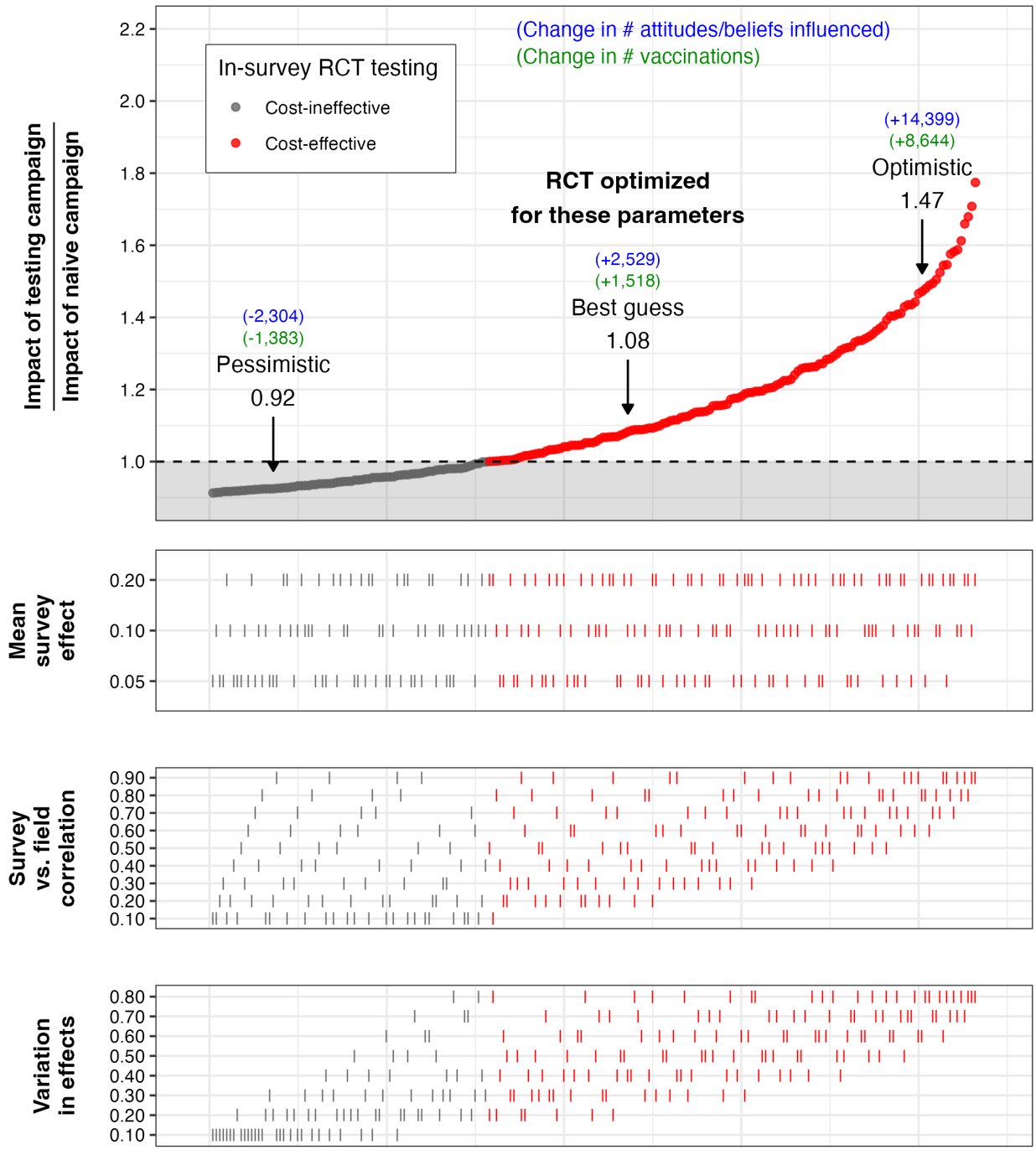
Appendix Figure 5: Relative-impact curves for a campaign with a budget of $420,000.

3 Benefit of possessing accurate knowledge of the parameter values

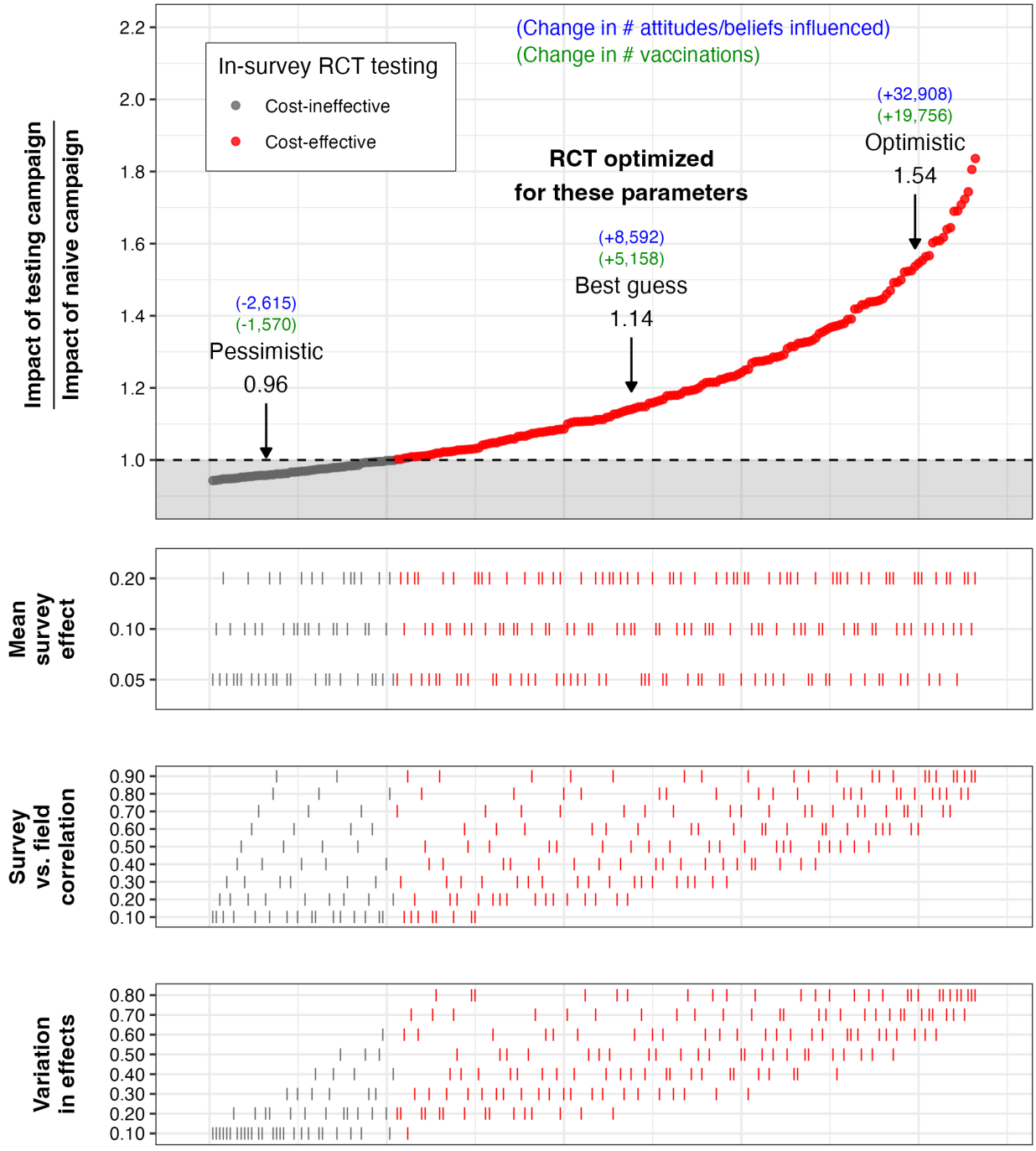
Appendix Figures 6, 7, 8 and 9 show, respectively, the estimated relative-impact curves for campaigns with budgets of $52,500, $105,000, $210,000 and $420,000 when their RCT testing regime is optimized for the best-guess set of parameter values, with lower panels showing the corresponding parameter values.



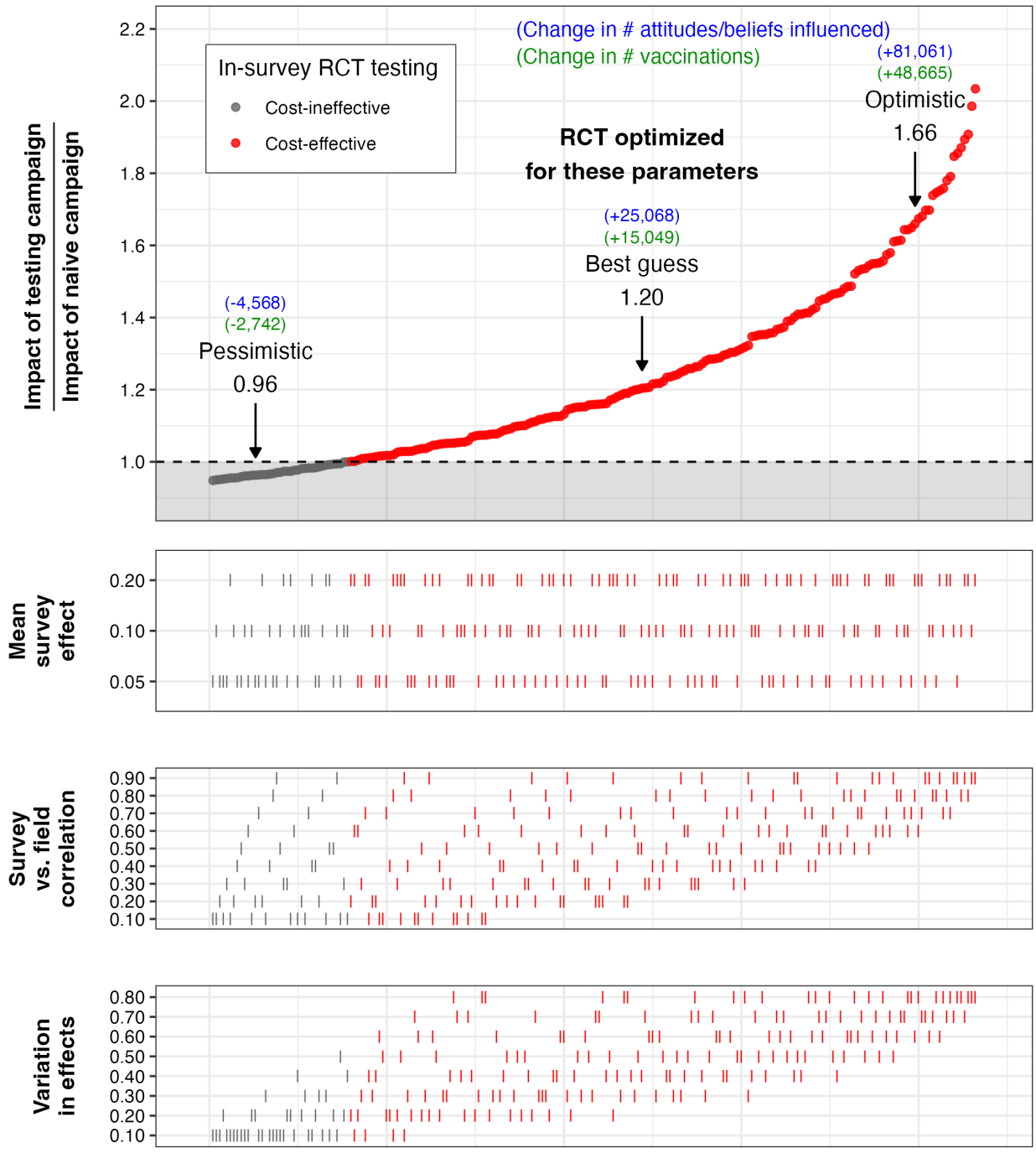
Appendix Figure 6: Relative-impact curve for a $52,500 campaign when their in-survey RCT pre-testing regime is optimized for the best-guess set of parameter values.



Appendix Figure 7: Relative-impact curve for a $105,000 campaign when their in-survey RCT pre-testing regime is optimized for the best-guess set of parameter values.



Appendix Figure 8: Relative-impact curve for a $210,000 campaign when their in-survey RCT pre-testing regime is optimized for the best-guess set of parameter values.



Appendix Figure 9: Relative-impact curve for a $420,000 campaign when their in-survey RCT pre-testing regime is optimized for the best-guess set of parameter values.

4 Reviewing evidence for parameter values

Appendix Table 1 reports the studies we reviewed to estimate the true mean effect size in-survey. Appendix Table 2 reports the studies we reviewed and re-analyzed to estimate the true variation in treatment effects. For those studies that were sourced from the systematic review conducted by Batteux et al. (2022), we refer to that review for the full references. The full references of the remaining papers are supplied at the end of this Appendix.

4.1 True mean effect size in-survey

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey.

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
Batteux et al. 2021	Online RCT	328	2	No	0.21	An information treatment caused a greater decrease in vaccination intentions (standardized difference = 0.2) and perceived effectiveness (0.22) when it conflicted with a prior announcement that was certain vs. uncertain. We take the mean of these two values.	Batteux et al. 2022 systematic review
Behavioural Insights Team	Discrete choice experiment	4085	NA	No	NA	We were unable to determine the standardized ATE because there is no control group in the design of the experiment (discrete choice experiment).	Batteux et al. 2022 systematic review
Chen et al. 2021	Online RCT	413	8	No	NA	We were unable to determine the standardized ATE; there is no control group; supplement and data are not accessible.	Batteux et al. 2022 systematic review
Craig 2021	Discrete choice experiment	1153	NA	No	NA	We were unable to determine the standardized ATE because there is no control group in the design of the experiment (discrete choice experiment).	Batteux et al. 2022 systematic review

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey. *(continued)*

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
Dai et al. 2022	Online RCT	3181	1	Yes	0.16	There was one informational video intervention conducted in the survey RCT, and the outcome variable was intention to schedule an appointment and reported desire for the vaccine. They report Cohen's d for the video treatment as 0.16 (see Extended Data Table 6), averaging across all online RCTs and outcomes.	Batteux et al. 2022 systematic review
Davis et al. 2021	Online RCT	481	3	Yes	0.54	The covid-information-only treatment caused an increase in covid vaccination intentions of 0.39 SDs; the treatments that also contrasted this with the flu vaccine increased covid intentions by 0.68 SDs (estimates taken from the paper's abstract). We take the mean of these two values.	Batteux et al. 2022 systematic review
Duch et al. 2021	Online RCT	1628	3	Yes but not pure	NA	The outcome variable in this study is survey click-through-rate after viewing a video, which is a sufficiently different estimand to most of the other studies that we do not include it.	Batteux et al. 2022 systematic review

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey. *(continued)*

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
Freeman et al. 2021	Online RCT	15000	9	Yes but not pure	0.10	We compute the standardized ATE by first taking the reported condition marginal means, standard errors and sample sizes from the paper, and using these numbers to reverse-engineer the SD of the outcome variable. We then divide the reported ATE by this SD (thus standardizing it). The outcome variable is a vaccine hesitancy scale. Among the full sample the resulting standardized ATE is -0.02; among the strongly vaccine hesitant it is -0.18. We take the mean of these two values and convert to positive magnitude.	Batteux et al. 2022 systematic review
Han et al. 2021	Online RCT	1497	4	Yes	0.09	The treatments in this study weren't really directional in nature - that is, they weren't aimed at encouraging a particular behavior - rather they simply emphasized scientific uncertainty about COVID-19 in different ways. We thus should not expect them to have strong effects on the outcome variables. There were a variety of outcome variables examined. For the most relevant outcomes, "Intentions for COVID-19 Risk-Reducing Behaviors" and "Vaccination", reported standardized ATE magnitudes ranged from approximately zero to 0.18 (see Figure 2C and 2D and accompanying text). We thus take the midpoint of this range.	Batteux et al. 2022 systematic review

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey. *(continued)*

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
Kerr et al. 2021	Online RCT	4300	4	Yes	0.05	This paper reports two studies and investigates many different outcome variables. Most estimated effects are and statistically non-significant, but some treatment effects vs. control are statistically significant and in the 0.2 SD to 0.3 SD region (see Figures 1 and 4 and accompanying text). We err on the side of there being some nonzero but very small effect: 0.05.	Batteux et al. 2022 systematic review
McPhedran et al. 2021	Discrete choice experiment	2012	NA	No	NA	We were unable to determine the standardized ATE because there is no control group in the design of the experiment (discrete choice experiment).	Batteux et al. 2022 systematic review
Moehring et al. 2022	Online RCT	484000	1	Yes	0.03	This paper reports variations on a social norm treatment, but it is basically always a similar idea. They report an average effect of approximately 0.035 on a five point scale of vaccination intentions. The SD of a uniform distribution over 1-5 is approximately 1.4, so we compute the standardized ATE as 0.035/1.4 = 0.025.	Batteux et al. 2022 systematic review

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey. *(continued)*

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
Motta et al. 2021	Online RCT	7000	3	Yes	0.05	This paper estimates the effects of several different treatments, with effect sizes ranging from 0 to approximately 5pp. We thus take the midpoint of this range (2.5pp) as the average effect. The SD of a uniform 0-1 distribution is approximately 0.5 so we compute the standardized ATE as 0.025/0.5 = 0.05.	Batteux et al. 2022 systematic review
Palm et al. 2021	Online RCT	1123	6	Yes	0.20	In this paper there are two positive-focused treatment conditions, dubbed Safe and Effective and Willing, which had estimated effects of 0.36 and 0.43 respectively on a 1 to 7 scale. The SD of a uniform 1-7 distribution is approximately 2, so we compute the standardized ATEs as 0.36/2 = 0.18 and 0.43/2 = 0.22 respectively and then take the mean of these two values.	Batteux et al. 2022 systematic review
Pink et al. 2021	Online RCT	1480	2	Yes	0.03	This paper exposed US Republicans to in- or out-party cues or a control condition. Estimated effects of in-party vs. control condition ranged from approximately zero to 2.5pp across outcome variables. We take the midpoint of this range (1.25pp) as the effect size. The SD of a uniform 0-1 distribution is approximately 0.5 so we calculate the standardized ATE as 0.0125/0.5 = 0.025.	Batteux et al. 2022 systematic review

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey. *(continued)*

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
Serra-Garcia & Szech 2023	Online RCT	2100	Gradations	Yes	0.16	This paper explored financial incentives and opt-in vs. opt-out schemes on COVID-19 vaccination intention and demand for tests. For the financial incentive treatment, the randomization is graded i.e. increasing through dollar amounts. The effect for financial incentive is nonlinear and thus difficult to interpret: small incentives caused a decrease in intention/demand, but larger incentives caused an increase. The effects for the opt-out (vs. opt-in) condition ranged from 4pp to 12pp across different specifications for the intention/demand outcomes (see Table 1). We thus take the midpoint of these values (8pp) as the average effect. The SD of a uniform 0-1 distribution is approximately 0.5, so we compute a standardized ATE of 0.08/0.5 = 0.16.	Batteux et al. 2022 systematic review

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey. *(continued)*

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
Sinclair et al. 2023	Online RCT	661	5	Yes	0.06	This paper examined vaccination intentions and scores on a vaccine hesitancy scale. On the intentions outcome, the ATEs ranged from -0.04 to +0.34 points on a five point Likert scale - we take the midpoint (0.15) as the average effect - with an SD of approximately 1.25 (see Table 1). Thus, we compute a standardized ATE of 0.15/1.25 = 0.12. The effects on the vaccine hesitancy scale are approximately zero in the aggregate, so we halve the overall standardized ATE to 0.06.	Batteux et al. 2022 systematic review
Sprengholz & Betsch 2020	Online RCT	576	1	Yes	0.41	This paper reports that "participants in the herd immunity communication condition reported a mean likeliness to get vaccinated of 16.14 or 79.9% (SD = 4.67 or 24.6 percentage points), compared to 13.92 or 68.0% (SD = 6.25 or 32.9 percentage points) for those who received no information about herd immunity." This gives an ATE of 2.2 scale points, with an average SD of 5.46; equating to a standardized ATE of 2.2/5.46 = 0.41. Note that the disease was fictitious.	Batteux et al. 2022 systematic review

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey. *(continued)*

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
Sprengholz et al. 2021	Online RCT	2400	3	Unclear	NA	The outcome in this study is reactance (negative psychological reaction) towards vaccination and the treatments are not aiming to reduce it. It is also difficult to discern standardized effect sizes because the authors report primarily on interactions and the supplement doesn't make it clear whether the reported simple effects are standardized or unstandardized. We omit this study.	Batteux et al. 2022 systematic review
Sprengholz et al. 2022	Online RCT	782	1	Yes	0.07	This paper examined a legal incentive vs. no incentive condition to get vaccinated, for zero financial compensation. The point estimate on willingness to get vaccinated was 3.7pp higher on average in the legal incentive condition. Given an SD of 0.5 for a uniform 0-1 scale, this implies a standardized ATE of 0.037/0.5 = 0.074.	Batteux et al. 2022 systematic review
Strickland et al. 2021	Online RCT	1366	Various	Unclear	NA	Four experiments. It is difficult to discern the effect sizes of message exposure because the analysis is primarily an AUC analysis. We omit this study.	Batteux et al. 2022 systematic review

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey. *(continued)*

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
Taber et al. 2021	Online RCT	850	Various	No	NA	This paper reports two experiments: one on lottery structure and the other on framing loss/gain. The treatments are continuous through e.g. lottery structure. Not easy to discern relevant standardized ATEs. We omit this study.	Batteux et al. 2022 systematic review
Thorpe et al. 2022	Online RCT	361	2	Yes	0.00	This paper studied 4 outcome variables and 2 different treatments, giving 8 treatment effects. The corresponding estimated ATEs are all null effects whose point estimates bounce around all over the place (see Table 2). We code this as a treatment effect of zero overall.	Batteux et al. 2022 systematic review
Trueblood et al. 2022	Online RCT	1000	3	Yes	0.02	This paper examines three treatments where the outcome is how long people would wait for the vaccine. We take the mean of the three treatment effects (0.4155, -0.2133, -0.0427), which are measured on an 11-point outcome scale, and divide this mean by the approximate SD of a uniform distribution over 1-11 (i.e. 3.14). Thus giving an overall standardized ATE of 0.02.	Batteux et al. 2022 systematic review

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey. *(continued)*

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
Witus & Larson 2022	Online RCT	1632	3	Yes	0.15	This paper examines vaccination intentions and three treatments. We take the mean of the predicted probabilities of each treatment effect on "definitely" getting the vaccine (0.11, 0.075, 0.04; see Figure 1) and divide this mean by the SD of a 0-1 uniform distribution (i.e. 0.5). Thus, we compute the standardized ATE as 0.15.	Batteux et al. 2022 systematic review
Bartos et al. 2022	Online RCT	2000	1	Yes	0.09	This paper reports a longitudinal experiment. The treatment informs people of the consensus among doctors regarding the COVID-19 vaccines. Estimated treatment effects on beliefs and self-reported vaccination status range from 3pp to 6pp. We take the midpoint of this range (4.5pp) and divide it by the SD of a uniform distribution over a binary 0-1 variable (0.5); resulting in a standardized ATE of 0.045/0.5 = 0.09.	Snowball sampling / knowledge of literature
Wittenberg et al. 2021	Online RCT	3343	24	Yes	0.25	This paper studies dozens of treatment videos targeting COVID-19 beliefs, attitudes and behavioral intentions. The outcome variable is unique to each video. The overall average standardized ATE is reported as 0.25.	Snowball sampling / knowledge of literature

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey. *(continued)*

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
Shen et al. 2015	Lab and field RCTs	9330	25	Yes but not pure	0.13	This paper reports a meta-analysis of 25 studies, all of which are related to public health communication but not COVID-19 specifically. They include studies that compare the effect of narrative information against a control group that receives non-narrative statistical or factual information. The authors report effect size r, which we convert to Cohen's d here: https://www.escal.site/	Snowball sampling / knowledge of literature
Jordan et al. 2021	Online RCT	988	3	Yes	0.23	This paper reports several studies, however only study 1 contains a control group. Study 1 reports standardized ATEs of 0.17, 0.20 and 0.33. We take the mean of these values. Outcomes are COVID-19 related.	Snowball sampling / knowledge of literature
Kaufman et al. 2022	Online RCT	463	4	Yes	0.05	This paper examines treatments to encourage parents to covid-19 vaccinate their children. The primary outcome variable is probability the respondent answers "Definitely or probably will get a COVID-19 vaccine for child", coded 1 if so and 0 otherwise. Treatment effects are (in pp): -3.9, -0.8, 6.9, 7.8 (see Table 3). We take the mean of these values and divide by the SD of a uniform distribution over 0-1 (i.e. 0.5); giving a standardized ATE of 0.025/0.5 = 0.05.	Snowball sampling / knowledge of literature

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey. *(continued)*

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
Green et al. 2022	Online RCT	24682	5	Yes	0.07	This paper reports a study investigating the effects of five treatments to reduce COVID-19 vaccine resistance. They report the following effect magnitudes of each treatment (in pp): 5, 5, 3, 3, 2. We take the mean of these values and then divide by a uniform distribution over 0-1 (i.e. 0.5); thus giving a standardized ATE of 0.072.	Snowball sampling / knowledge of literature
Bokemper et al. 2021 study 1	Online RCT	855	6	Yes	0.08	This study examines several different COVID-19 outcomes and six treatments. We take the mean of the six treatment effects across all outcome variables (see Supplementary Table on the OSF) and divide it by the SD of a uniform distribution over 0-1 (0.5) to get the overall standardized ATE.	Snowball sampling / knowledge of literature
Bokemper et al. 2021 study 2	Online RCT	2419	5	Yes	0.02	This study examines several different COVID-19 outcomes and five treatments. We take the mean of the five treatment effects across all outcome variables (see Supplementary Table on the OSF) and divide it by the SD of a uniform distribution over 0-1 (0.5) to get the overall standardized ATE.	Snowball sampling / knowledge of literature

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey. *(continued)*

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
Bokemper et al. 2022 study 1	Online RCT	2568	10	Yes but not pure	0.03	This study examined various COVID-19 outcomes and ten treatments. Note that the control group was not a pure control but baseline persuasive information. We take the mean across treatment effects for each outcome variable, then divide by the SD of that outcome variable (reported in appendix table S2) to get the standardized overall ATE for each outcome. We then compute the overall standardized ATE by taking the mean across the standardized ATE for each outcome variable.	Snowball sampling / knowledge of literature
Bokemper et al. 2022 study 2	Online RCT	6000	3	Yes	0.04	This study examined five COVID-19 outcome variables and three treatments. We take the mean of the three treatment effects across all outcome variables (see appendix table S5) and divide it by the SD of a uniform distribution over 0-1 (0.5) to get the overall standardized ATE.	Snowball sampling / knowledge of literature

Appendix Table 1: Studies reviewed to estimate the true mean effect size in-survey. *(continued)*

Author and study (where relevant)	Design	Sample size	# Treatments	Control group	ATE (in SDs)	Notes	Source
James et al. 2021 study 1	Online RCT	4361	11	Yes	0.17	This study examined three COVID-19 outcomes and eleven treatments. We take the mean across treatment effects for each outcome variable, then divide by the SD of that outcome variable (reported in appendix table S1) to get the standardized overall ATE for each outcome. We then compute the overall standardized ATE by taking the mean across the standardized ATE for each outcome variable. Note that the vaccination intention outcome is the combined version.	Snowball sampling / knowledge of literature
James et al. 2021 study 2	Online RCT	5014	6	Yes	0.08	This study examined three COVID-19 outcome variables and six treatments. We take the mean of the six treatment effects across all outcome variables (see appendix table S2) and divide it by the SD of a uniform distribution over 0-1 (0.5) to get the overall standardized ATE.	Snowball sampling / knowledge of literature

4.2 True variation in effect sizes

Appendix Table 2: Studies reviewed and re-analyzed to estimate the true variation in treatment effects.

Author	Design	Sample size	# Treatments	Control group	Scaled SD	Notes	Domain	Source
James et al. 2021 study 1	Online RCT	4361	11	Yes	0.30	Greg Huber (author) provided the data over email. We first estimated the treatment effects from the response-level data. We then used random effects meta-analysis to estimate mean and SD in treatment effects, and then scaled the SD by dividing by the mean. We computed the scaled SD for each of three outcomes in the paper and then averaged across the scaled SDs to compute the overall scaled SD.	health, covid	Snowball sampling / knowledge of literature
James et al. 2021 study 2	Online RCT	5014	6	Yes	0.20	Greg Huber (author) provided the data over email. We first estimated the treatment effects from the response-level data. We then used random effects meta-analysis to estimate mean and SD in treatment effects, and then scaled the SD by dividing by the mean. We computed the scaled SD for each of three outcomes in the paper and then averaged across the scaled SDs to compute the overall scaled SD.	health, covid	Snowball sampling / knowledge of literature
Freeman et al. 2021	Online RCT	15000	9	Yes but not pure	NA	Paper says contact author for data. No reply to email attempts.	health, covid	Batteux et al. 2022 systematic review

Appendix Table 2: Studies reviewed and re-analyzed to estimate the true variation in treatment effects. *(continued)*

Author	Design	Sample size	# Treatments	Control group	Scaled SD	Notes	Domain	Source
Bokemper et al. 2022 study 1	Online RCT	2568	10	Yes but not pure	1.11	Data were publicly available from the Harvard dataverse. We first estimated the treatment effects from the response-level data. We then used random effects meta-analysis to estimate mean and SD in treatment effects, and then scaled the SD by dividing by the mean. We computed the scaled SD for each of four outcomes, as per the paper and appendix, and then averaged across the scaled SDs to compute the overall scaled SD.	health, covid	Snowball sampling / knowledge of literature
Green et al. 2022	Online RCT	24682	5	Yes	0.41	Jon Green (author) provided the data over email. We first estimated the treatment effects from the response-level data. We then used random effects meta-analysis to estimate mean and SD in treatment effects, and then scaled the SD by dividing by the mean. We computed the scaled SD for the primary seven-point outcome variable.	health, covid	Snowball sampling / knowledge of literature

Appendix Table 2: Studies reviewed and re-analyzed to estimate the true variation in treatment effects. *(continued)*

Author	Design	Sample size	# Treatments	Control group	Scaled SD	Notes	Domain	Source
Bokemper et al. 2021 study 1	Online RCT	855	6	Yes	1.73	Data were publicly available from the Harvard dataverse. We first estimated the treatment effects from the response-level data. We then used random effects meta-analysis to estimate mean and SD in treatment effects, and then scaled the SD by dividing by the mean. We computed the scaled SD for each of three outcomes, as per the paper, and then averaged across the scaled SDs to compute the overall scaled SD.	health, covid	Snowball sampling / knowledge of literature
Bokemper et al. 2021 study 2	Online RCT	2419	5	Yes	0.00	Data were publicly available from the Harvard dataverse. We first estimated the treatment effects from the response-level data. We then used random effects meta-analysis to estimate mean and SD in treatment effects, and then scaled the SD by dividing by the mean. We computed the scaled SD for each of three outcomes, as per the paper, and then averaged across the scaled SDs to compute the overall scaled SD.	health, covid	Snowball sampling / knowledge of literature

Appendix Table 2: Studies reviewed and re-analyzed to estimate the true variation in treatment effects. *(continued)*

Author	Design	Sample size	# Treatments	Control group	Scaled SD	Notes	Domain	Source
Sinclair et al. 2023	Online RCT	661	5	Yes	1.60	Samantha Sinclair (author) provided the data over email. We first estimated the treatment effects from the response-level data. We used random effects meta-analysis to estimate mean and SD in treatment effects, and then scaled the SD by dividing by the mean. We computed the scaled SD for each of three outcomes, as per the paper, and then averaged across the scaled SDs to compute the overall scaled SD.	health, covid	Batteux et al. 2022 systematic review
Palm et al. 2021	Online RCT	1123	6	Yes	NA	The treatments in this study point in different directions i.e. some of the messages are casting doubt on the vaccines while others are encouraging vaccination. Thus, there are fewer than 5 treatments in the same direction. Ineligible.	health, covid	Batteux et al. 2022 systematic review
Chen et al. 2021	Online RCT	413	8	No	NA	No control group, ineligible.	health, covid	Batteux et al. 2022 systematic review
Milkman et al. 2022	Field RCT	689693	22	Yes	0.24	Data were publicly available from the Open Science Framework. Data was aggregated to the condition-level, so we first used logistic regression with multiple trials to estimate ATEs in log-odds space and then used random effects meta-analysis to estimate the mean and SD in treatment effects. Finally we scaled the SD by dividing by the mean.	health, non-covid	Snowball sampling / knowledge of literature

Appendix Table 2: Studies reviewed and re-analyzed to estimate the true variation in treatment effects. *(continued)*

Author	Design	Sample size	# Treatments	Control group	Scaled SD	Notes	Domain	Source
Milkman et al. 2023	Field RCT	3662548	8	Yes	0.16	Data were publicly available from the Open Science Framework. We first estimated the treatment effects from the response-level data. We then used random effects meta-analysis to estimate the mean and SD in treatment effects and then we scaled the SD by dividing by the mean. Note this paper was not publicly available at the time of analysis, but was shared with us by the BCFG team.	health, covid	Snowball sampling / knowledge of literature
Milkman et al. 2021	Field RCT	47306	19	Yes	0.17	Data were publicly available from the Open Science Framework. Data was aggregated to the condition-level, so we first used logistic regression with multiple trials to estimate ATEs in log-odds space and then used random effects meta-analysis to estimate the mean and SD in treatment effects. Finally we scaled the SD by dividing by the mean.	health, non-covid	Snowball sampling / knowledge of literature
Coppock et al. 2020	Online RCT	34000	49	Yes	2.23	This paper reports two outcome variables: candidate favorability and vote choice. Estimated ATEs are reported as 0.0492 and 0.0072 respectively. Estimated SD(ATE) is reported as 0.0682 and 0.0222 respectively. Thus, the scaled SDs are 1.39 and 3.08 respectively. We take the mean of these values to compute the overall scaled SD.	politics	Snowball sampling / knowledge of literature

Appendix Table 2: Studies reviewed and re-analyzed to estimate the true variation in treatment effects. *(continued)*

Author	Design	Sample size	# Treatments	Control group	Scaled SD	Notes	Domain	Source
Hewitt et al. 2023	Online RCT	500000	617	Yes	0.52	This paper reports two outcome variables, candidate favorability and vote choice, across three electoral contexts: 2018 downballot, 2020 downballot and 2020 presidential. As per the paper, 0.52 is reported as the average scaled SD across outcomes and contexts.	politics	Snowball sampling / knowledge of literature
Hewitt & Tappin 2022	Online RCT	40000	59	Yes	0.95	This papers reports on two policy issues and two broad sets of argument types (for vs. against each issue). The estimated SD in treatment effects and the mean treatment effect are reported as follows (with implied scaled SD in parentheses): For-UBI: 0.06 / 0.11 = 0.55; Against-UBI: 0.08 / 0.12 = 0.67; For-USCA: 0.07 / 0.03 = 2.33; Against-USCA: 0.03 / 0.12 = 0.25. We take the mean of these four values to compute the overall scaled SD.	politics	Snowball sampling / knowledge of literature

5 References

Bartoš, V., Bauer, M., Cahlíková, J., & Chytilová, J. (2022). Communicating doctors' consensus persistently increases COVID-19 vaccinations. *Nature*, 606(7914), 542-549.

Batteux, E., Mills, F., Jones, L. F., Symons, C., & Weston, D. (2022). The effectiveness of interventions for increasing COVID-19 vaccine uptake: a systematic review. *Vaccines*, 10(3), 386.

Bokemper, S. E., Gerber, A. S., Omer, S. B., & Huber, G. A. (2021). Persuading US White evangelicals to vaccinate for COVID-19: Testing message effectiveness in fall 2020 and spring 2021. *Proceedings of the National Academy of Sciences*, 118(49), e2114762118.

Bokemper, S. E., Huber, G. A., James, E. K., Gerber, A. S., & Omer, S. B. (2022). Testing persuasive messaging to encourage COVID-19 risk reduction. *PloS one*, 17(3), e0264782.

Coppock, A., Hill, S. J., & Vavreck, L. (2020). The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. *Science Advances*, 6(36), eabc4046.

Green, J., Druckman, J. N., Baum, M. A., Lazer, D., Ognyanova, K., Simonson, M. D., ... & Perlis, R. H. (2023). Using general messages to persuade on a politicized scientific issue. *British Journal of Political Science*, 53(2), 698-706.

Hewitt, L. et al. (2023). How experiments help campaigns persuade voters: evidence from a large archive of campaigns' own experiments. *American Political Science Review*.

Hewitt, L., & Tappin, B. M. (2022). Rank-heterogeneous effects of political messages: evidence from randomized survey experiments testing 59 video treatments. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/xk6t3.

James, E. K., Bokemper, S. E., Gerber, A. S., Omer, S. B., & Huber, G. A. (2021). Persuasive messaging to increase COVID-19 vaccine uptake intentions. *Vaccine*, 39(49), 7158-7165.

Jordan, J. J., Yoeli, E., & Rand, D. G. (2021). Don't get it or don't spread it: Comparing self-interested versus prosocial motivations for COVID-19 prevention behaviors. *Scientific reports*, 11(1), 20222.

Kaufman, J., Steffens, M. S., Hoq, M., King, C., Marques, M. D., Mao, K., ... & Danchin, M. (2023). Effect of persuasive messaging about COVID-19 vaccines for 5-to 11-year-old children on parent intention to vaccinate. *Journal of Paediatrics and Child Health*, 59(4), 686-693.

Milkman, K. L., Gandhi, L., Patel, M. S., Graci, H. N., Gromet, D. M., Ho, H., ... & Duckworth, A. L. (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., Patel, M. S., Gandhi, L., Graci, H. N., Gromet, D. M., Ho, H., ... & Duckworth, A. L. (2021). A megastudy of text-based nudges encouraging patients

to get vaccinated at an upcoming doctor's appointment. *Proceedings of the National Academy of Sciences*, 118(20), e2101165118.

Shen, F., Sheer, V. C., & Li, R. (2015). Impact of narratives on persuasion in health communication: A meta-analysis. *Journal of advertising*, 44(2), 105-113.

Wittenberg, C., Tappin, B. M., Berinsky, A. J., & Rand, D. G. (2021). The (minimal) persuasive advantage of political video over text. *Proceedings of the National Academy of Sciences*, 118(47), e2114388118.