Aversion to pragmatic randomized controlled trials: Three survey experiments with clinicians and laypeople in the United States

Supplemental Materials

Table of Contents

Methods

In the main text, we grouped the vignettes thematically into three sets: “Lay Sentiments About pRCTs,” “Lay Sentiments About Covid-19 pRCTs,” and “Clinician Sentiments About Covid-19 pRCTs.” However, when we collected data, we grouped our vignettes differently such that we started with vignettes that we have used in previous published work and their respective Covid-19 derivatives, then we developed and tested novel Covid-19 specific vignettes separately, and then, again separately, we tested a Covid-19 vaccine vignette. We followed a similar pattern in our clinician sample: we first tested three Covid-19 specific vignettes (two which were derivatives of vignettes from our previous work, one which was new to this work) and then separately, we tested a Covid-19 vaccine vignette. These groupings are important for understanding how participants were randomly assigned to vignettes and why there are slight discrepancies (or large discrepancies in the case of the Best Vaccine vignette in the clinician sample[1]) in the number of participants in each vignette (see Table S1).

Table S1

Population, sample size, and dates of data collection for each vignette

Preregistration #	Vignette	Population	Sample size	Dates of data collection
1	Catheterization Safety Checklist	MTurk workers	343	August 13, 2020
	Intubation Safety Checklist	MTurk workers	347	August 13, 2020
	Best Anti-Hypertensive Drug	MTurk workers	357	August 13, 2020
	Best Corticosteroid Drug	MTurk workers	357	August 13, 2020
2	Masking Rules	MTurk workers	360	September 30-October 2, 2020
	School Reopening	MTurk workers	339	September 30-October 2, 2020
	Best Vaccine (ambiguous version)*	MTurk workers	350	September 30-October 2, 2020
	Ventilator Proning	MTurk workers	357	September 30-October 2, 2020
3	Intubation Safety Checklist	Clinicians	271	November 13-December 9, 2020
	Best Corticosteroid Drug	Clinicians	275	November 13-December 9, 2020
	Masking Rules	Clinicians	349	November 13-December 9, 2020
4	Best Vaccine	MTurk workers	450	January 8, 2021
5	Best Vaccine	Clinicians	1254	January 25-February 9, 2021

Note. Within each data collection batch, participants were randomly assigned to one of the vignettes. In the clinician sample (preregistration #3), clinicians saw all three vignettes in randomized order. The sample size reported here is the number of clinicians who saw that vignette first.

*Our first attempt at the Best Vaccine vignette included wording that unintentionally made the experiment condition less aversive. For this reason, this vignette is not included in the main analyses.

As shown in Table 1, in the first round of survey experiments (preregistration #1), the first set of lay participants were randomly assigned to read and respond to either Catheterization Safety Checklist, Best Anti-Hypertensive Drug, Intubation Safety Checklist, or Best Corticosteroid Drug. Then, in a second round of survey experiments (preregistration #2), a second, separate, set of lay participants were randomly assigned to read and respond to either Masking Rules, School Reopening, Ventilator Proning, or an unintentionally ambiguous version of Best Vaccine (results of which are reported in the SM). A third set of lay participants (preregistration #4) were recruited to read and respond to a correct version of Best Vaccine (no other vignette was included and, thus, no randomization was necessary). In the clinician sample, one set of clinicians (preregistration #3) was recruited to read and respond to Masking Rules, Intubation Safety Checklist, and Best Corticosteroid in a randomized order. All clinicians in this sample read and responded to all three vignettes. However, only their responses to the first vignette they read are considered for the purpose of the analyses presented in the main text. A second set of clinicians (preregistration #5) was recruited to read and respond to Best Vaccine (no other vignette was included and, thus, no randomization was necessary). However, because the clinician survey was fully anonymous, it is possible that there is some overlap between participants in the first and second clinician samples.

[1] The Best Vaccine vignette was combined with another study that required a sample size much larger than the sample sizes in our previous vignette studies to have adequate statistical power.

For clarity, in the main text of this article we used different names for the vignettes than those used in the preregistrations and in previous publications (see Table S2).

Table S2

Original vignette names from preregistrations and previous work and corresponding name in main text

Original vignette name	Main text vignette name Hospital
Safety Checklist (also called Checklist)	Catheterization Safety Checklist Best
Drug: Walk-In Clinic (also called Best Drug)	Best Anti-Hypertensive Drug
Checklist (Covid-19)	Intubation Safety Checklist
Best Drug (Covid-19)	Best Corticosteroid Drug
Ventilator Proning	Ventilator Proning
School Reopening	School Reopening
Mask Requirements	Masking Rules
Modified Covid-19 Vaccines	Best Vaccine
Vaccine Distribution	(not reported in main text)

Note. Vignette names in this article were changed from those in previous work and in our preregistrations in order to clarify the content for readers.

Preregistrations, sample sizes, and power analyses

Our research questions, power analyses and sample sizes, and analysis plans were all preregistered at Open Science Framework (OSF) before data collection. These sample size precommitments are copied from each preregistration document which can be found on OSF at https://osf.io/u945y/?view_only=a901fde13ddb423899074eb79964c6cd.

Preregistration 1 (Catheterization Safety Checklist, Best Anti-Hypertensive Drug, Intubation Safety Checklist, Best Corticosteroid Drug vignettes):

“We predict that, using a two-tailed, paired t-test with $\alpha = .05$ within each scenario, participants will rate the A/B test condition as significantly less appropriate than their own average rating of the two policy conditions, mean(A,B). This is the test for the “A/B Effect.” Recruiting 350 participants for each scenario provides 95% power to detect an effect as small as $d = 0.19$, which is substantially smaller than the effect sizes we have observed using the Hospital Safety Checklist and Best Drug: Walk-In Clinic vignettes in past research.”

Preregistration 2 (Ventilator Proning, School Reopening, Masking Rules, and Best Vaccine (initial ambiguous version) vignettes):

“We predict that, using a two-tailed, paired t-test with $\alpha = .05$ within each scenario, participants will rate the A/B test condition as significantly less appropriate than their own average rating of the two policy conditions, mean(A,B). This is the test for the “A/B Effect.” Recruiting 350 participants for each scenario provides 95% power to detect an effect as small as $d = 0.19$, which is substantially smaller than the effect sizes we have observed using the Hospital Safety Checklist and Best Drug: Walk-In Clinic vignettes in past research.”

Preregistration 3 (Clinicians; Intubation Safety Checklist, Best Corticosteroid Drug, and Masking Rules vignettes):

Note that because of time constraints around the possible starting dates of our clinician surveys, we launched this study before preregistering it, and we did not report an explicit power analysis before collecting the data. Because this study follows a similar structure to the studies above, however, it was reasonable to apply the previous sample size and power analysis considerations. We did, however, preregister our approach and research plan twice during this study: once during data collection, before any analyses had been conducted, and again after all data had been collected (but before analyzing any of them).

> Preregistration 3.1: "At the time of this preregistration, we have received 655 complete responses. No data have been explored or analyzed at this point. We will conduct an interim analysis on this dataset using the same analyses we have previously preregistered, and we may continue to collect more data from this population."
>
> Preregistration 3.2: "Data collection is now complete and we have closed the survey. On 11/24/2020, we conducted an interim analysis on 601 complete responses. Since then, we have received an additional 295 complete responses, to which we remain blind."

Preregistration 4 (Best Vaccine):

"We recruited 350 participants for the original Covid-19 vaccines study. Because we are running this study to determine whether even a small effect emerges, we will increase the sample size to 450 participants. This provides 80% power to detect an effect as small as d = 0.13 in a repeated- measures, two-tailed t-test, and 95% power to detect an effect as small as d = 0.17."

Preregistration 5 (Clinicians; Best Vaccine):

"Our previous survey of healthcare providers resulted in approximately 900 complete responses; we expect a similar response rate for this survey. This sample size provides 95% power to detect an effect as small as d = 0.12 using a two-tailed, repeated measures t-test. Even if we only receive 600 complete responses, we will have 95% power to detect an effect as small as d = 0.15."

Procedure and design

Several aspects of the procedure and experimental design were consistent across the studies reported here. Below, we describe these consistent features and note in specific studies where we deviated from them.

For the lay participant samples, we used the CloudResearch service to recruit crowd workers on Amazon Mechanical Turk (MTurk) to participate in a 3–5-minute survey experiment. These services provide samples that are broadly representative of the U.S. population and are well-accepted in social science research as providing as good or better-quality data than convenience samples such as student volunteers, with results that are similar to probability sampling methods [1,2]. Participants were excluded from recruitment in any of the studies reported here if they had participated in any of our previous studies on this topic. Across all laypeople vignettes, the completion rate of participants starting the survey was 91.5%. The Geisinger IRB determined that these anonymous surveys were exempt (IRB# 2017-0449).

For the clinician samples, we recruited healthcare providers (including physicians, physician assistants, nurse practitioners, and nurses) from a large health system in the Northeastern U.S via email. Each provider received either one or two emails about the study during the recruitment window. In the first clinician study (Intubation Safety Checklist, Best Corticosteroid Drug, and Masking Rules vignettes), we first tested the email recruitment system by sending out the survey invitation email to just 200 clinicians. Clinicians who completed the survey based on this survey invitation were included in the final sample. Then, all clinicians were sent the recruitment email on November 19, 2020, followed by a reminder email on December 3, 2020. In the second clinician study (Best Vaccine), the initial recruitment email was sent January 25, 2021, with the follow-up email sent February 2, 2021. In the first clinician study, 5,925 clinicians were emailed and 895 completed the survey. In the second clinician study, 6,993 clinicians were emailed and 1,254 completed the survey. In these samples, because survey responses were fully anonymous, we were not able to restrict participation based on our previous studies, so some participants who completed the Best Vaccine vignette may have earlier completed the Intubation Safety Checklist, Best Corticosteroid Drug, and Masking Rules vignettes.

In all cases, participants completed an online survey hosted by Qualtrics. After opening the survey, participants were evenly randomized to one of the possible vignettes being studied using the "evenly present elements" function in the survey flow of Qualtrics.[2,3] Qualtrics uses a least filled quota system which preferentially randomizes participants to the condition with the lowest count of responses at the time they enter the survey. The exact algorithm used by Qualtrics is proprietary. In the case of data collection batches 4 and 5, there was only one vignette being tested that all participants saw. At this point, we used the exact same procedure detailed in Heck et al. (2020) [4]. First, participants were instructed to read about several possible decisions made by different decision-makers[4], and to try to treat each decision as separate from the others. All scenarios contained a brief "background" text at the top of the page that summarized a problem, followed by three "situations," each of which detailed the decision-maker's choice to adopt intervention A, intervention B, or to run an A/B test by randomly assigning people to one of two test conditions. These conditions were presented in fully counterbalanced order; each participant received one of six possible orders (i.e., Situation 1 = A, Situation 2 = B, and Situation 3 = A/B; Situation 1 = A/B, Situation 2 = B, and Situation 3 = A; etc.…). At no point did we observe a meaningful effect of presentation order, so we collapsed across this variable for all analyses.

---

[2] For the clinician study of the Intubation Safety Checklist, Best Corticosteroid Drug, and Masking Rules vignettes, clinicians were randomly assigned to one of these three scenarios and then completed the remaining two scenarios in random order. For consistency with the rest of this project and with our previous survey experiment with clinicians regarding the A/B effect (3, Study 6), and in order to make the results from clinician samples comparable to those with lay samples (in which each participant only ever saw one scenario), we analyze data from this study as a between-subjects design where we only consider the first scenario that every participant completed. See the section "Order Effect in Clinician Study" elsewhere in this appendix for further analyses.

[3] The clinician version of the Best Vaccine vignette was combined with another study being conducted by a subset of researchers on this team. The materials for Best Vaccine were presented after the survey materials from the other study. Data from the other study are unrelated to the research questions tested here and will be reported separately.

For our primary outcome measures, participants were asked to rate the appropriateness of the decisions made in Situation 1, Situation 2, and Situation 3 (“How appropriate is the director's decision in Situation 1/2/3?”), using a 1-5 scale (1 = “Very inappropriate”, 2 = “Inappropriate”, 3 = “Neither inappropriate nor appropriate”, 4 =”Appropriate”, 5 = “Very appropriate”). Participants then specified a ranked order of the three decisions (“Among these three decisions, which decision do you think the director should make? Please drag and drop the options below into your preferred order from best to worst. You must click on at least one option before you can proceed.”), with 1 being the best decision and 3 being the worst. The last item on this page asked participants to explain why they chose these ratings and rankings in a couple of sentences (“In a couple of sentences, please tell us why you chose the ratings and rankings you chose.”).

Following these primary measures, participants completed standard demographic items on the next page. For MTurk participants, these were measures of sex, race/ethnicity, age, educational attainment, household income, religious belief or affiliation, whether they have a degree in a STEM field or not, and four items identifying political orientation and affiliation. As part of an ongoing study in our laboratory (whose results will be reported elsewhere), these participants were randomized to one of six conditions for this demographic questionnaire where we varied the option to select “prefer not to answer” and whether the items were mandatory, optional, or requested (but not required). For clinician participants, demographic items were mandatory response and were limited to the following: sex, sources of training in research methods and statistics, self-reported comfort with research methods and statistics, past experience with activities related to research methods and statistics (e.g., publishing a scientific paper or analyzing data), current involvement in research, position (e.g., doctor, physician assistant, nurse, medical student, etc.), length of time working in the medical field, and field of specialty.

After completing the survey, MTurk participants were given a completion code to receive payment ($0.40). Clinician participants were invited to enter into a lottery to win a $50 Amazon gift card by following a link to an independent survey where they could enter their email address. All participants were thanked for their participation and offered the opportunity to comment on the survey.

---

[4] In all vignettes, the protagonist (e.g., the hospital director or Dr. Jones) was male for ease of comparison to our previous work using these vignettes. Future work should examine the impact of the characteristics of the decision-maker on evaluations of their decisions regarding policy imposition and conducting RCTs.

Measures

We computed several variables to measure participants' sentiments about pRCTs.

Following Meyer et al. (2019) [3], we define an "A/B effect" as the difference between participants' mean policy rating and their rating of the A/B test—that is, the degree to which the policies are (on average) rated higher than the A/B test. We also report the percentage of participants whose mean policy rating is higher than their rating of the A/B test.

Following Heck et al. (2020 [4]; see also Mislavsky et al., 2019 [5]), we define "experiment aversion" as the difference between participants' rating of their own lowest-rated policy and their rating of the A/B test. We also report the percentage of participants who express experiment aversion.

"Experiment rejection" (first reported in Heck et al., 2020 [4], but without this name) occurs when a participant rates the A/B test as inappropriate (1 or 2 on the 5-point scale) while also rating each policy as neutral or appropriate (3–5 on the scale).

A "reverse A/B effect" is the difference between participants' rating of the A/B test and their mean policy rating—that is, the degree to which the A/B test is rated higher than the policies (on average). We also report the percentage of participants whose rating of the A/B test is higher than their mean policy rating.

"Experiment appreciation" is the difference between participants' rating of the A/B test and their rating of their own highest-rated policy. We also report the percentage of participants who express experiment appreciation.

"Experiment endorsement" occurs when a participant rates the A/B as appropriate (4 or 5 on the 5-point scale) while also rating each intervention as neutral or inappropriate (1–3 on the scale).

In all cases where a *d*-value was calculated (i.e., A/B effect, experiment aversion, reverse A/B effect, experiment appreciation), we used Cohen's *d* recovered from the *t*-statistic, *n*, and correlation between the two measures being compared (Dunlap et al., 1996 [6], equation 3: $d = t_c[2(1-r)/n]^{½}$; see also http://jakewestfall.org/blog/index.php/category/effect-size/kewestfall.org [7]. To calculate this *d*-value, we use the following R code: effsize::cohen.d(x,y, paired = TRUE).

In Figures 1B, 2B, and 3B, we transformed participants A, B, and A/B ratings on the continuous 5-point Likert scale into a binary objected/did not object variable (where objecting was defined as assigning a rating of 1 or 2—"very inappropriate" or "somewhat inappropriate"— on the 1–5 scale). We do this only for visualization and do not perform any statistical analyses on this transformed objected/did not object variable. Instead, as is standard in social and moral psychology, we treated appropriateness ratings elicited on the 5-point Likert scale as continuous. Therefore, we use t-tests to test the differences between the ratings of the A/B test and the interventions (lowest, average, and highest). Other methodologies and statistical analyses like a discrete choice approach, in which participants would see and evaluation two of the three possible decisions (e.g., intervention A vs. A/B test) at a time, or the Stuart-Maxwell test, which requires a kxk matrix of categorical variables, would not be appropriate.

Vignettes

Our vignettes were inspired by discussions about the ethics of real-world RCTs (see Table S3).

Table S3

Literature calling for or reporting an RCT similar to what is proposed in each vignette

Vignette name	Relevant literature
Catheterization Safety Checklist	Pronovost et al. [8], Urbach et al. [9], Arriaga et al. [10]
Best Anti-Hypertensive Drug	ROMP Ethics Study [11], Sinnott et al. [12]
Intubation Safety Checklist	Turner et al. [13]
Best Corticosteroid Drug	Wagner et al. [14]
Ventilator Proning	Elharrar et al. [15], Sartini et al. [16], Caputo et al. [17]
School Reopening	Fretheim et al. [18, 19], Helsingen et al. [20], Angrist et al. [21], Kolata [22]
Masking Rules	Abaluck et al. [23], Jefferson et al. [24], Bundgaard et al. [25]
Best Vaccine	Bach [26]

The following section shows the exact vignette text that participants read in these studies (with the exception of the bolded titles, which are never shown to participants).

Catheterization Safety Checklist

(Originally from Heck et al. (2020) [4], adapted from Meyer et al. (2019) [2])

Background: Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections.

Situation 1

A hospital director wants to reduce these infections, so he decides to give each doctor who performs this procedure a new ID badge with a list of standard safety precautions for the procedure printed on the back. All patients having this procedure will then be treated by doctors with this list attached to their clothing.

Situation 2

A hospital director wants to reduce these infections, so he decides to hang a poster with a list of standard safety precautions for this procedure in all procedure rooms. All patients having this procedure will then be treated in rooms with this list posted on the wall.

Situation 3

A hospital director thinks of two different ways to reduce these infections, so he decides to run an experiment by randomly assigning patients to one of two test conditions. Half of patients will be treated by doctors who have received a new ID badge with a list of standard safety precautions for the procedure printed on the back. The other half will be treated in rooms with a poster listing the same precautions hanging on the wall. After a year, the director will have all patients treated in whichever way turns out to have the highest survival rate.

Best Anti-Hypertensive Drug

(Originally from Heck et al. (2020) [4], adapted from Meyer et al. (2019) [2])

Background: Several drugs have been approved by the US. Food and Drug Administration as safe and effective for treating high blood pressure. Doctor Jones works in a multi-doctor walk-in clinic where patients see whichever doctor is available. Some doctors in the clinic prescribe drug A for high blood pressure, while others prescribe drug B. Both drugs are affordable and patients can tolerate their side effects.
Situation 1

Doctor Jones wants to provide good treatment to his patients, so he decides that his patients who need high blood pressure medication will be prescribed drug A.

Situation 2

Doctor Jones wants to provide good treatment to his patients, so he decides that his patients who need high blood pressure medication will be prescribed drug B.

Situation 3

Doctor Jones thinks of two different ways to provide good treatment to his patients, so he decides to run an experiment by randomly assigning his patients who need high blood pressure medication to one of two test conditions. Half of patients will be prescribed drug A, and the other half will be prescribed drug B. After a year, he will only prescribe to new patients whichever drug has had the best outcomes for his patients.

Intubation Safety Checklist

Background: Some treatments for coronavirus (Covid-19) patients require a doctor to insert a plastic breathing tube into the throat. These treatments can save lives, but they can also lead to deadly fluid buildup in the lungs.

Situation 1

A hospital director wants to reduce these cases of fluid buildup, so he decides to give each doctor who performs this procedure a new ID badge with a list of standard safety precautions for the procedure printed on the back. All coronavirus patients having this procedure will then be treated by doctors with this list attached to their clothing.

Situation 2

A hospital director wants to reduce these cases of fluid buildup, so he decides to hang a poster with a list of standard safety precautions for this procedure in all procedure rooms. All coronavirus patients having this procedure will then be treated in rooms with this list posted on the wall.

Situation 3

A hospital director thinks of two different ways to reduce these cases of fluid buildup, so he decides to run an experiment by randomly assigning coronavirus patients who need a breathing tube to one of two test conditions. Half of patients will be treated by doctors who have received a new ID badge with a list of standard safety precautions for the procedure printed on the back. The other half will be treated in rooms with a poster listing the same precautions hanging on the wall. After two months, the director will have all patients treated in whichever way turns out to have the highest survival rate.

Best Corticosteroid Drug

Background: Several corticosteroids (a family of anti-inflammatory drugs) have been approved by the U.S. Food and Drug Administration as safe and effective for treating a variety of diseases. There is some evidence that corticosteroids can also help certain coronavirus (Covid-19) patients, and many doctors prescribe corticosteroids for these patients. Doctor Jones works in a multi-doctor emergency department where patients see whichever doctor is available. Some doctors in the emergency department prescribe corticosteroid A for coronavirus symptoms, while others prescribe corticosteroid B. Both corticosteroids are affordable and patients can tolerate their side effects.

Situation 1

Doctor Jones wants to provide good treatment to his patients, so he decides that his coronavirus patients who need medication will be prescribed corticosteroid A.

Situation 2

Doctor Jones wants to provide good treatment to his patients, so he decides that his coronavirus patients who need medication will be prescribed corticosteroid B.

Situation 3

Doctor Jones thinks of two different ways to provide good treatment to his coronavirus patients, so he decides to run an experiment by randomly assigning his patients who need medication to one of two test conditions. Half of coronavirus patients will be prescribed corticosteroid A, and the other half will be prescribed corticosteroid B. After two months, he will only prescribe to new coronavirus patients whichever corticosteroid has had the best outcomes for his patients.

Ventilator Proning

Background: Some coronavirus (Covid-19) patients have to be sedated and placed on a ventilator to help them breathe. Even with a ventilator, these patients can have dangerously low blood oxygenation levels, which can result in death. Current standards suggest that laying ventilated patients on their stomach for 12-16 hours per day can reduce pressure on the lungs and might increase blood oxygen levels and improve survival rates.

Situation 1

A hospital director wants to save as many ventilated Covid-19 patients as possible, so he decides that all of these patients will be placed on their stomach for 12-13 hours per day.

Situation 2

A hospital director wants to save as many ventilated Covid-19 patients as possible, so he decides that all of these patients will be placed on their stomach for 15-16 hours per day.

Situation 3

A hospital director thinks of two different ways to save as many ventilated Covid-19 patients as possible, so he decides to run an experiment by randomly assigning ventilated Covid-19 patients to one of two test conditions. Half of these patients will be placed on their stomach for 12-13 hours per day. The other half of these patients will be placed on their stomach for 15-16 hours per day. After one month, the director will have all ventilated Covid-19 patients treated in whichever way turns out to have the highest survival rate.

Best Vaccine (ambiguous version; results not reported in main analyses)

Background: Imagine that several vaccines have been approved by the U.S. Food and Drug Administration as safe and effective for preventing Covid-19. Vaccine A uses mRNA molecules to provide the cells with a blueprint for how to destroy the virus. Vaccine B uses deactivated or weakened coronavirus to help the body create an immune resistance to the disease. Both vaccines are affordable, similarly priced, and people can tolerate their side effects. However, people can only receive one of these two vaccines.

Situation 1

The director of public health for a state wants to reduce Covid-19 cases. So he decides that all clinics in the state will offer Vaccine A for free. People can get any other vaccine somewhere else, if they want.

Situation 2

The director of public health for a state wants to reduce Covid-19 cases. So he decides that all clinics in the state will offer Vaccine B for free. People can get any other vaccine somewhere else, if they want.

Situation 3

The director of public health for a state thinks of two different ways to reduce Covid-19 cases, so he decides to run an experiment by randomly assigning clinics in the state to one of two test conditions. Half of the clinics will offer Vaccine A for free, and the other half will offer Vaccine B for free. People can get any other vaccine somewhere else, if they want.[5] After six months, he will direct the state to offer whichever vaccine has resulted in the fewest cases of Covid-19.

Best Vaccine

Background: Imagine that several vaccines have been approved by the U.S. Food and Drug Administration as safe and effective for preventing Covid-19. Vaccine A uses mRNA molecules to provide the cells with a blueprint for how to destroy the virus. Vaccine B uses deactivated or weakened coronavirus to help the body create an immune resistance to the disease. Both vaccines are affordable, similarly priced, and people can tolerate their side effects.

Situation 1

The director of public health for a state wants to reduce Covid-19 cases. So he decides that all clinics in the state will offer Vaccine A for free.

Situation 2

The director of public health for a state wants to reduce Covid-19 cases. So he decides that all clinics in the state will offer Vaccine B for free.

Situation 3

The director of public health for a state thinks of two different ways to reduce Covid-19 cases, so he decides to run an experiment by randomly assigning clinics in the state to one of two test conditions. Half of the clinics will offer Vaccine A for free, and the other half will offer Vaccine B for free. After six months, he will direct the state to offer whichever vaccine has resulted in the fewest cases of Covid-19.

---

[5] This wording unintentionally implied that residents could choose their vaccine (by going elsewhere) if they did not wish to be subject to the official's decision (including policy implementation or A/B test); we suspect this had the effect of making the experiment condition less aversive, since people could effectively opt-out of it, and our goal in this research is to study pragmatic, real-world situations in which avoiding randomization is not a realistic option.

School Reopening

Background: This Fall, school districts must decide whether to reopen their doors to students, teachers, and staff despite the risks of spreading coronavirus (Covid-19). Many school and public health officials have decided to use a "hybrid model" of teaching that offers some of the benefits of face-to-face learning time while attempting to minimize the risks related to Covid-19.

Situation 1

A superintendent at a large school district wants to provide good education to his students while slowing the spread of Coronavirus. So, he decides that students will attend school according to an even-odd schedule. Students in even-numbered grades (e.g., 2nd grade, 4th grade, etc.) will attend school in the morning and learn remotely in the afternoons, while students in odd- numbered grades will attend school in the afternoon and learn remotely in the mornings.

Situation 2

A superintendent at a large school district wants to provide good education to his students while slowing the spread of Coronavirus. So, he decides that students will attend school according to an A-day/B-day schedule. Students in the A group will attend school in person on Monday, Tuesday, and Wednesday morning, and students in the B group will attend school in person on Wednesday afternoon, Thursday, and Friday. Students will learn remotely on the days they do not attend school.

Situation 3

A superintendent at a large school district thinks of two different ways to provide good education to his students while slowing the spread of Coronavirus. So, he decides to conduct an experiment by randomly assigning schools in the district to one of two test conditions. For half of schools, students will attend school according to an even-odd schedule. Students in even-numbered grades (e.g., 2nd grade, 4th grade, etc.) will attend school in the morning and learn remotely in the afternoons, while students in odd-numbered grades will attend school in the afternoon and learn remotely in the mornings. For the other half of schools, students will attend school according to an A-day/B-day schedule. Students in the A group will attend school in person on Monday, Tuesday, and Wednesday morning, and students in the B group will attend school in person on Wednesday afternoon, Thursday, and Friday. Students will learn remotely on the days they do not attend school. At the end of the semester, all schools will adopt, for future semesters when the pandemic threat level remains similar, whichever policy has resulted in the best combination of test scores on state aptitude tests and number of Covid-19 cases.

Masking Rules

Background: Public health officials have considered different rules about when and where people must wear masks or other face coverings to reduce the spread of coronavirus (Covid-19).
Increasing mask use can reduce the spread of the disease, but highly restrictive mask policies can substantially reduce compliance rates.

Situation 1

A state health department director wants to reduce coronavirus spread within his state, so he decides that all counties will require masks in all businesses and public buildings.

Situation 2

A state health department director wants to reduce coronavirus spread within his state, so he decides that all counties will require masks in all businesses, public buildings, and outdoor public spaces.

Situation 3

A state health department director thinks of two different ways to reduce coronavirus spread within his state, so he decides to run an experiment by randomly assigning counties within the state to one of two test conditions. Half of counties will require masks in all businesses and public buildings. The other half of counties will require masks in all businesses, public buildings, and outdoor public spaces. After one month, the director will require all counties to adopt whichever policy has led to the fewest cases of Covid-19 for as long as the pandemic threat level remains high.

Results

Sample demographics

Lay participants

Across all vignettes reported in the main text (i.e., excluding the initial ambiguous version of the Best Vaccine vignette), there were a total of 2,909 lay participants. They ranged in age from 18 to 88 years old (mean = 38.4, SD = 12.8) and the majority were White (74.6%) and female (55.9%). 35.7% had a 4-year college degree, 29.7% had some college, and 20.5% had a graduate degree. 21.3% of participants had a degree in a STEM field. The most frequently selected income level was between $20,000 and $40,000 (20.7%). A majority of participants reported being moderate, leaning liberal, or being liberal both generally and specifically with regards to social and economic issues. Similarly, a majority of participants reported being independent, leaning Democrat, or being Democrat in their political party affiliations. 37.7% of participants reported being non-religious. Of those who reported being religious, the most reported religion was Protestant (24.2%). See Table S4 for demographic breakdowns by vignette and in the combined lay participant sample.

Table S4

Demographics of lay participants by vignette

	Catheterization Safety Checklist	Best Anti-Hypertensive Drug	Intubation Safety Checklist	Best Corticosteroid Drug	Best Vaccine (first attempt)	Best Vaccine	School Reopening	Ventilator Proning	Masking Rules	All vignettes
Total N	343	357	346	357	350	450	339	357	360	2909
Age [Mean (SD)]	37.9 (12.9)	38.6 (12.9)	37.9 (12.4)	38.0 (12.7)	36.7 (12.0)	37.7 (12.6)	38.7 (13.0)	39.4 (12.7)	39.0 (12.8)	38.4 (12.8)
Sex (%)										
Male	51.3%	41.5%	48.1%	51.5%	36.6%	38.4%	39.2%	40.9%	39.7%	43.6%
Female	47.8%	58.0%	51.9%	48.2%	63.1%	60.9%	60.5%	58.8%	60.0%	55.9%
Other	0.6%	0.6%	0.0%	0.0%	0.3%	0.4%	0.3%	0.3%	0.3%	0.2%
Prefer not to answer	0.3%	0.0%	0.0%	0.3%	0.0%	0.2%	0.0%	0.0%	0.0%	0.2%
Race - select all that apply (%)										
Black/African-American	11.1%	5.0%	8.4%	10.1%	10.9%	11.3%	9.7%	6.7%	8.9%	9.0%
Hispanic or Latino	8.2%	8.4%	7.2%	8.4%	8.3%	5.6%	5.9%	9.5%	7.5%	7.5%
White	72.0%	78.7%	71.5%	72.0%	70.9%	72.7%	77.0%	77.6%	75.8%	74.6%
Asian	12.5%	8.7%	15.3%	12.6%	12.6%	13.3%	8.6%	7.0%	7.8%	10.8%
Other	1.2%	1.7%	1.2%	0.3%	3.4%	0.9%	1.8%	1.7%	2.2%	1.3%
Prefer not to answer	0.9%	0.6%	0.0%	0.6%	0.3%	0.9%	0.6%	0.3%	0.3%	0.5%
Education (%)										
Less than high school	0.6%	0.8%	0.3%	0.3%	0.6%	0.2%	0.3%	9.8%	0.8%	0.4%
High school degree	5.5%	7.8%	8.9%	9.2%	9.1%	10.2%	10.3%	29.4%	11.4%	9.2%
Some college	32.7%	32.2%	24.2%	28.0%	30.3%	32.0%	26.3%	33.6%	31.9%	29.7%
Four-year college degree	37.3%	35.6%	39.5%	35.9%	37.1%	35.8%	37.8%	3.1%	30.6%	35.7%
Some graduate school	4.4%	3.4%	4.6%	4.2%	4.6%	5.1%	4.4%	23.8%	4.7%	4.3%
Graduate degree	19.2%	19.9%	22.5%	22.1%	18.3%	16.2%	20.9%	0.3%	20.6%	20.5%
Prefer not to answer	0.3%	0.3%	0.0%	0.3%	0.0%	0.4%	0.0%	0.0%	0.0%	0.2%
Income (%)										
< $20,000	11.1%	8.4%	9.2%	7.6%	12.0%	9.3%	9.4%	11.2%	9.7%	9.5%
$20,000-$40,000	17.8%	22.1%	21.6%	25.8%	19.7%	20.2%	18.9%	19.0%	19.7%	20.7%
$40,000-$60,000	24.5%	18.8%	19.0%	20.2%	21.4%	20.4%	21.2%	19.9%	20.8%	20.6%
$60,000-$80,000	13.7%	17.4%	16.1%	17.9%	18.6%	17.8%	16.5%	19.3%	19.2%	17.3%
$80,000-$100,000	11.4%	13.7%	11.0%	9.5%	10.6%	12.2%	13.3%	8.4%	12.2%	11.5%
> $100,000	20.7%	18.5%	21.3%	17.4%	17.1%	18.7%	20.4%	19.6%	16.9%	19.1%
Prefer not to answer	0.9%	1.1%	0.9%	1.4%	0.3%	1.3%	0.3%	2.5%	1.4%	1.2%
No response	0.0%	0.0%	0.9%	0.3%	0.3%	0.0%	0.0%	0.0%	0.0%	0.1%
Political Ideology (%)										
Very liberal	12.2%	12.6%	13.0%	11.2%	10.6%	13.1%	12.7%	12.0%	12.8%	12.5%
Liberal	32.1%	30.3%	32.3%	35.9%	29.4%	31.1%	30.4%	30.8%	28.6%	31.4%
Moderate	29.2%	25.5%	28.2%	26.1%	31.1%	27.3%	27.7%	24.9%	28.3%	27.1%
Conservative	19.8%	20.2%	20.7%	17.1%	21.7%	18.7%	20.9%	21.3%	23.6%	20.2%
Very conservative	5.8%	10.6%	5.2%	9.5%	6.3%	8.9%	7.4%	9.8%	5.8%	7.9%
Prefer not to answer	0.9%	0.6%	0.3%	0.3%	0.9%	0.9%	0.6%	0.8%	0.8%	0.7%
No response	0.0%	0.3%	0.3%	0.0%	0.0%	0.0%	0.3%	0.3%	0.0%	0.1%

Table S4, continued

Demographics of lay participants by vignette

	Catheterization Safety Checklist	Best Anti-Hypertensive Drug	Intubation Safety Checklist	Best Corticosteroid Drug	Best Vaccine (first attempt)	Best Vaccine	School Reopening	Ventilator Proning	Masking Rules	All vignettes
Political ideology on social issues (%)										
Very liberal	18.7%	16.8%	19.6%	13.7%	17.7%	18.0%	17.7%	17.6%	17.5%	17.5%
Liberal	34.1%	33.3%	33.4%	40.3%	31.1%	30.4%	36.6%	34.2%	31.7%	34.1%
Moderate	21.6%	23.8%	23.9%	19.9%	26.0%	25.6%	19.8%	21.8%	23.3%	22.6%
Conservative	16.6%	15.4%	17.3%	17.1%	18.0%	16.0%	18.3%	16.0%	19.4%	17.0%
Very conservative	8.2%	10.4%	5.2%	8.4%	6.3%	9.1%	6.8%	9.8%	7.5%	8.2%
Prefer not to answer	0.9%	0.3%	0.6%	0.6%	0.9%	0.9%	0.6%	0.6%	0.6%	0.6%
No response	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%
Political ideology on economic issues (%)										
Very liberal	9.9%	12.0%	13.5%	11.2%	8.0%	13.8%	11.8%	10.4%	11.9%	11.9%
Liberal	28.3%	21.6%	27.1%	28.3%	24.9%	23.3%	27.7%	23.0%	19.7%	24.8%
Moderate	28.0%	27.5%	25.1%	25.2%	27.7%	28.4%	24.2%	27.5%	32.2%	27.3%
Conservative	23.0%	24.9%	24.8%	22.1%	30.9%	22.0%	24.2%	25.8%	26.4%	24.1%
Very conservative	9.3%	13.7%	8.6%	12.0%	7.4%	11.3%	11.2%	12.9%	9.2%	11.1%
Prefer not to answer	1.5%	0.3%	0.9%	1.1%	1.1%	0.9%	0.6%	0.6%	0.6%	0.8%
No response	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.3%	0.0%	0.0%	0.1%
Political party (%)										
Strong Democrat	14.9%	10.9%	12.4%	13.7%	12.0%	13.6%	13.0%	14.0%	12.8%	13.2%
Democrat	23.3%	22.7%	27.7%	28.9%	26.3%	24.4%	22.7%	21.0%	21.7%	24.1%
Independent (but lean Democrat)	15.7%	16.2%	14.7%	12.9%	13.4%	14.9%	17.4%	14.3%	15.8%	15.2%
Independent	15.7%	16.8%	17.6%	14.3%	16.9%	16.9%	13.6%	15.1%	18.1%	16.0%
Independent (but lean Republican)	7.0%	8.7%	7.8%	10.4%	9.4%	8.7%	10.6%	10.9%	10.6%	9.3%
Republican	16.3%	14.6%	14.1%	12.0%	13.1%	15.3%	15.6%	14.0%	13.9%	14.5%
Strong Republican	4.1%	8.4%	4.3%	7.3%	6.9%	4.9%	6.5%	9.0%	6.4%	6.3%
Prefer not to answer	2.9%	1.7%	1.4%	0.6%	2.0%	1.3%	0.3%	1.7%	0.8%	1.3%
No response	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%
Religion (%)										
Christian - Protestant	26.2%	24.6%	23.6%	21.0%	24.6%	24.2%	25.4%	24.4%	23.9%	24.2%
Christian - Catholic	17.5%	16.5%	15.9%	18.2%	17.7%	14.0%	17.1%	18.8%	15.3%	16.6%
Christian - Other	11.1%	11.2%	8.1%	11.2%	11.7%	11.1%	11.8%	10.9%	12.2%	11.0%
Jewish	2.6%	1.7%	1.7%	1.7%	1.7%	1.3%	1.8%	1.4%	2.5%	1.8%
Muslim	2.0%	0.8%	1.4%	0.6%	0.3%	0.9%	1.2%	1.1%	1.7%	1.2%
Buddhist	2.3%	1.4%	2.0%	1.7%	1.1%	2.0%	2.4%	0.6%	1.4%	1.7%
Hindu	1.2%	0.6%	2.6%	1.1%	1.7%	1.6%	0.3%	0.6%	0.6%	1.1%
Non-religious	32.7%	38.1%	40.9%	40.3%	36.6%	40.0%	35.4%	37.0%	36.4%	37.7%
Other	3.5%	3.6%	2.6%	3.4%	3.7%	3.8%	4.1%	3.4%	4.2%	3.6%
Prefer not to answer	0.9%	1.4%	1.2%	0.6%	0.9%	1.1%	0.6%	1.7%	1.9%	1.2%
No response	0.0%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%	0.3%	0.0%	0.1%
STEM degree (%)										
No	77.6%	77.0%	75.2%	76.8%	77.4%	80.7%	78.5%	78.4%	78.6%	77.9%
Yes	21.9%	22.1%	23.3%	22.4%	22.3%	18.7%	21.5%	20.2%	21.1%	21.3%
Prefer not to answer	0.6%	0.8%	1.4%	0.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.7%
No response	0.0%	0.0%	0.0%	0.0%	0.3%	0.7%	0.0%	0.3%	0.3%	0.1%

Clinicians

There were 2,149 clinician responses across all vignettes. In the clinician samples, survey responses were anonymous, so we could not restrict participation based on our previous studies so some participants who completed the Intubation Safety Checklist, Best Corticosteroid Drug, and Masking Rules vignettes may have also completed the Best Vaccine vignette. For this reason, demographics are reported separately by vignette in Table S5. Across vignettes, a majority of clinicians were female. Over 50% of participants in the sample were registered nurses, followed by physicians and physician assistants. Over 50% of participants in the sample reported that they had been in the medical field for over 10 years. The clinicians reported that they had received training in research methods and statistics via an average of 1.5 of the sources we listed, and that they engaged in an average of 2.5 research methods and statistics activities. Most clinicians reported being somewhat to moderately comfortable with research methods and statistics.

Table S5

Demographics of clinicians by vignette

	Intubation Safety Checklist	Best Corticosteroid Drug	Masking Rules	Best Vaccine
Total N	271	275	349	1254
Sex (%)				
Male	18.1%	22.5%	18.1%	18.7%
Female	81.9%	77.1%	81.4%	81.2%
Other	0.0%	0.4%	0.6%	0.2%
Source of research methods/statistics training - select all that apply (%)				
Undergraduate coursework	48.7%	49.5%	48.7%	47.4%
Professional school instruction	40.2%	31.3%	34.4%	34.4%
Postgraduate coursework	26.2%	20.7%	22.1%	21.1%
CME/CEU courses	27.7%	25.1%	24.1%	25.8%
Self-instruction via peer-reviewed literature	19.2%	15.6%	17.2%	21.3%
Other	7.0%	4.0%	3.2%	3.9%
Total number of research methods/statistics training [mean (SD)]	1.69 (1.22)	1.46 (1.02)	1.50 (1.13)	1.54 (1.16)
Comfort with research methods/statistics (%)				
Not at all	8.9%	12.7%	10.9%	11.1%
Somewhat	37.6%	44.4%	45.8%	46.6%
Moderately	39.5%	32.0%	32.7%	30.8%
Very	11.8%	9.1%	8.9%	9.9%
Extremely	2.2%	1.8%	1.7%	1.7%
Research methods/statistics activities - select all that apply (%)				
Read results of RCT in peer-reviewed journal article	81.2%	75.3%	71.9%	71.2%
Changed typical prescription/recommendation after personally reading results of RCT in peer-reviewed journal article	41.0%	33.1%	33.0%	39.8%
Published scientific paper in peer-reviewed journal	13.3%	12.4%	9.7%	12.0%
Conducted or worked on a team conducting an RCT	18.5%	20.0%	19.2%	17.1%
Took a course/class in statistics, biostatistics, research methods	73.1%	69.8%	69.1%	68.5%
Analyzed data for statistical significance outside of course require	23.6%	21.8%	19.2%	21.1%
Used statistical software	12.2%	11.6%	11.5%	9.3%
Total number of research methods/statistics activities [mean (SD)]	2.63 (1.69)	2.44 (1.71)	2.34 (1.66)	2.39 (1.72)
Currently involved in research (%)	10.7%	9.1%	9.7%	9.6%
Position (%)				
Doctor	14.8%	14.5%	12.6%	15.7%
Physician Assistant	12.5%	6.9%	9.5%	7.7%
Nurse Practitioner	6.3%	2.5%	4.3%	4.7%
Nurse (RN)	51.3%	57.1%	55.6%	52.8%
Nurse (LPN)	6.3%	9.5%	8.0%	15.6%
Nurse (Other)	1.8%	1.1%	1.4%	0.6%
Genetic Counselor	0.0%	0.0%	0.0%	0.0%
Non-prescribing clinician or staff without clinical credential	0.0%	0.0%	0.0%	0.0%
Medical student	5.2%	5.5%	4.6%	0.1%
Faculty or Professor	0.4%	0.7%	0.3%	0.3%
Other	1.5%	2.2%	3.7%	2.6%
Years in medical field (%)				
< 1 year	2.6%	2.9%	3.2%	2.8%
1-2 years	6.3%	5.5%	6.0%	5.8%
3-5 years	15.1%	11.3%	12.6%	13.6%
6-10 years	16.6%	14.2%	15.8%	15.8%
> 10 years	59.4%	66.2%	62.5%	62.0%

Note. Reported here are the demographics of the clinicians who saw the Intubation Safety Checklist, Best Corticosteroid Drug, or Masking Rules vignette first (responses to the Best Vaccine vignette were collected at a different time). All clinicians who participated in this study completed all vignettes but in randomized order. In the main text, we only analyze responses to the first vignette, so we report demographics similarly here.

Results presented in main text

In Figures S1-3, we show all individual appropriateness ratings (1 = very inappropriate, 5 = very appropriate) for intervention A, intervention B, and the A/B test across all vignettes.

Figure S1

Lay Sentiments About pRCTs

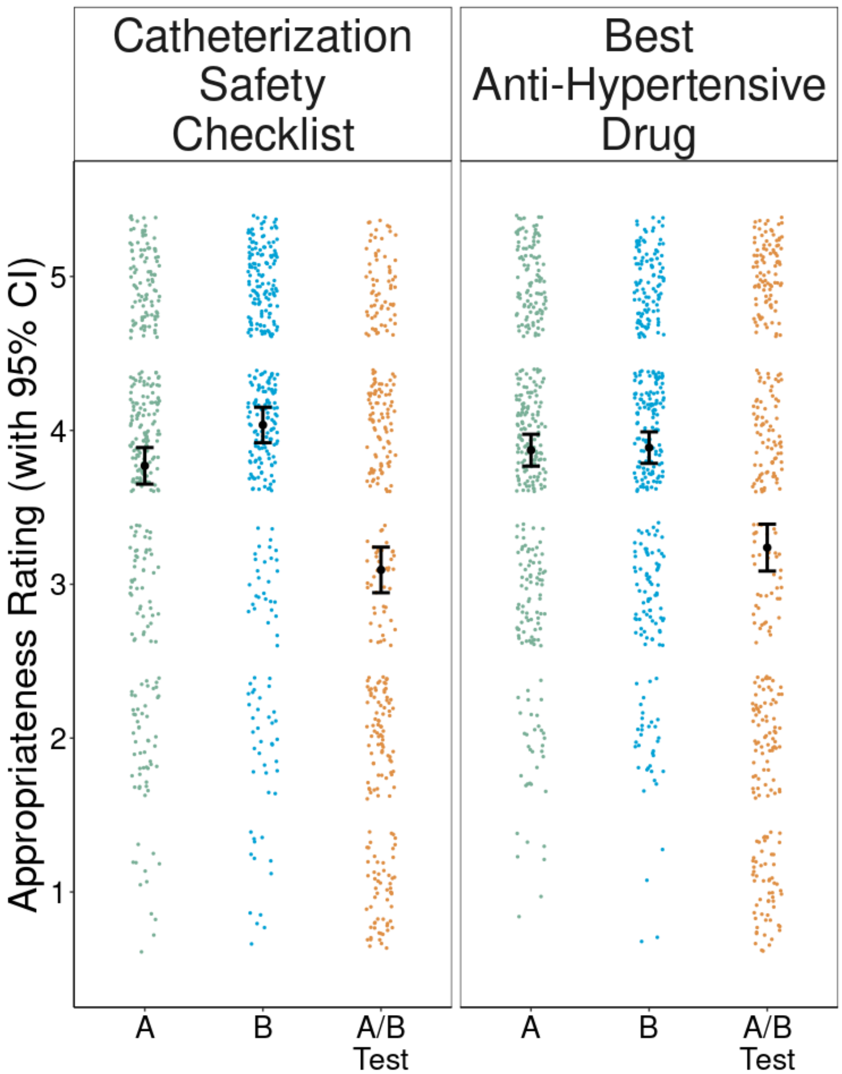


Figure S2

Lay Sentiments About Covid-19 pRCTs

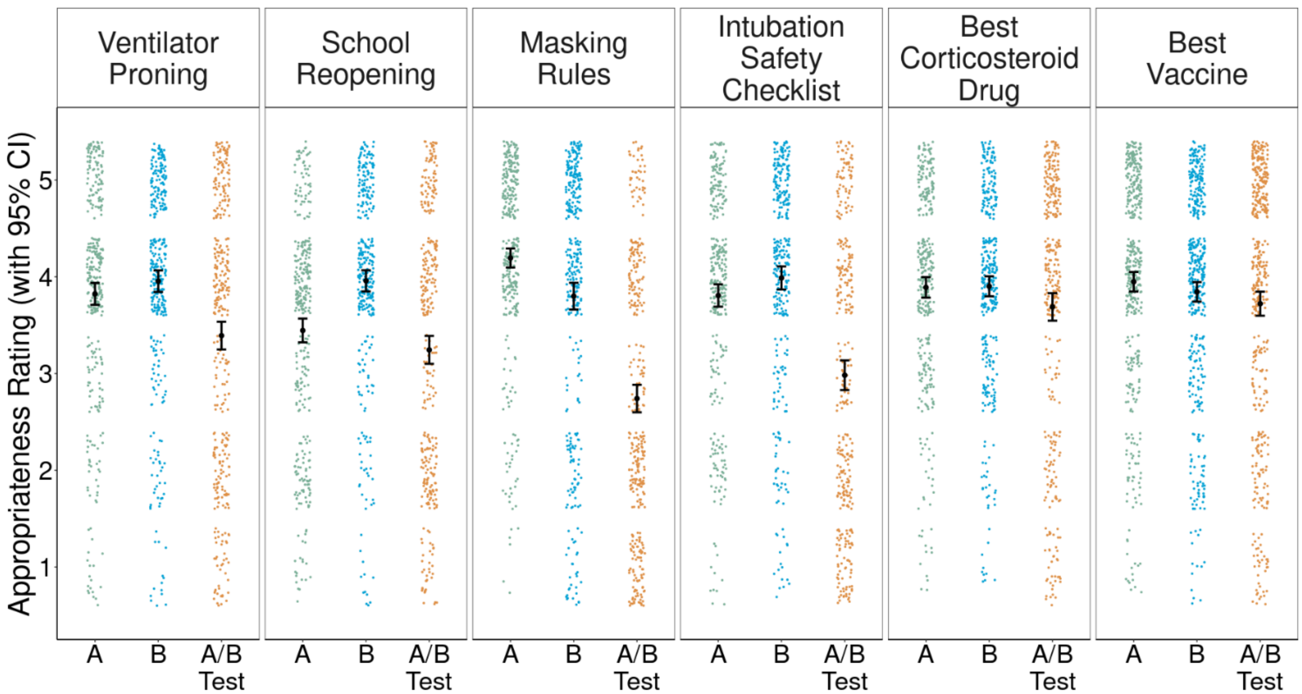
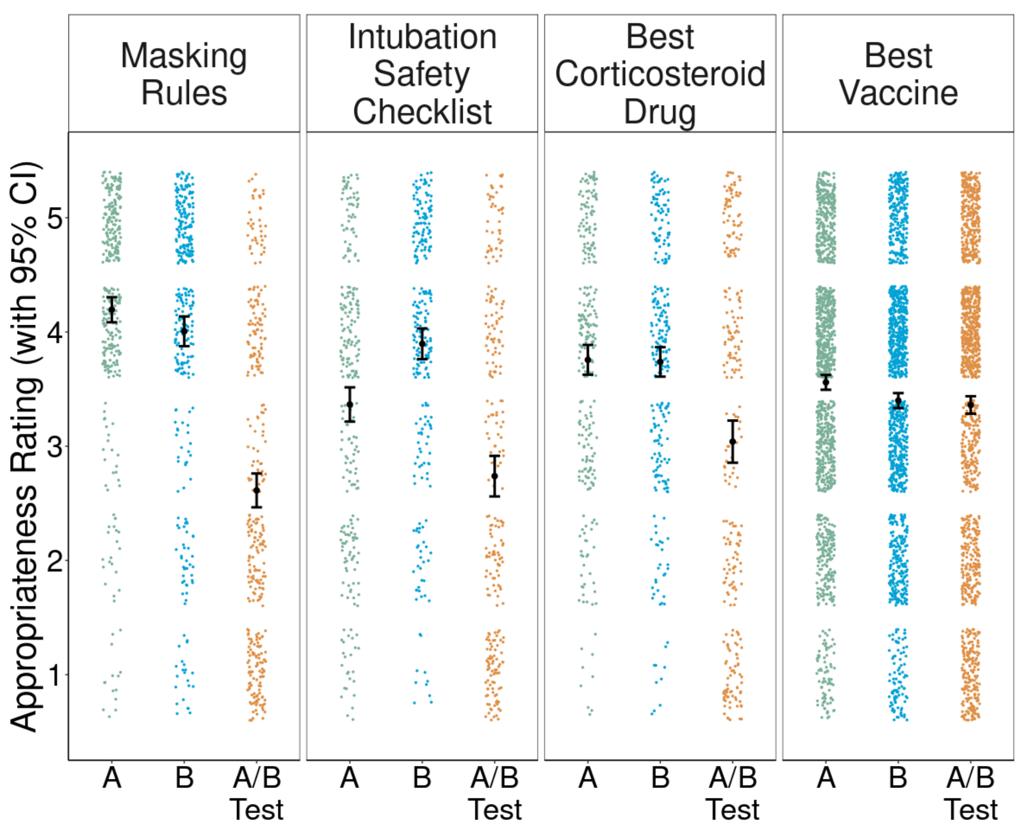


Figure S3

Clinician Sentiments About Covid-19 pRCTs



In Table S6A-C, we present the descriptive and inferential results for all vignettes discussed in the main text.

Table S6A

Descriptive and inferential results of ratings and rankings of interventions and experiment for all vignettes

		Descriptive Results			***Inferential Results***	
Vignette	**Variable**	**Mean (SD)**	**% Ranking Best**	**% Ranking Worst**	**Test Description**	**Test Outcome**
Lay Sentiments About pRCTs						
Catheterization Safety Checklist (n = 343 laypeople)					A/B Effect	t (342) = 9.74***, d = 0.69 ± .16
					Mean(A,B) > AB	58% ± 5%
	A	3.77 (1.12)	27%	32%	Reverse A/B effect	t (342) = -9.74***, d = -0.69 ± .16
	B	4.03 (1.09)	42%	21%	AB > Mean(A,B)	27% ± 4%
	AB	3.09 (1.40)	32%	48%	Experiment Aversion	t (342) = 3.70***, d = 0.25 ± .14
	Mean(A,B)	3.90 (0.84)	-	-	Min(A,B) > AB	41% ± 5%
	Min(A,B)	3.42 (1.16)	-	-	Experiment Appreciation	t (342) = -14.61***, d = -1.13 ± .20
	Max(A,B)	4.39 (0.81)	-	-	AB > Max(A,B)	15% ± 3%
					Experiment Rejection (A,B = 3,4,5; AB = 1,2)	28% ± 5%
					Experiment Endorsement (AB = 4,5; A,B = 1,2,3)	3% ± 1%
Best Anti-Hypertensive Drug (n = 357 laypeople)					A/B Effect	t (356) = 6.68***, d = 0.52 ± .16
					Mean(A,B) > AB	47% ± 5%
	A	3.87 (1.00)	25%	27%	Reverse A/B effect	t (356) = -6.68***, d = -0.52 ± .16
	B	3.89 (0.99)	25%	28%	AB > Mean(A,B)	31% ± 5%
	AB	3.24 (1.47)	50%	45%	Experiment Aversion	t (356) = 5.96***, d = 0.46 ± .16
	Mean(A,B)	3.88 (0.95)	-	-	Min(A,B) > AB	44% ± 5%
	Min(A,B)	3.82 (1.03)	-	-	Experiment Appreciation	t (356) = -7.26***, d = -0.57 ± .17
	Max(A,B)	3.94 (0.95)	-	-	AB > Max(A,B)	29% ± 4%
					Experiment Rejection (A,B = 3,4,5; AB = 1,2)	34% ± 5%
					Experiment Endorsement (AB = 4,5; A,B = 1,2,3)	18% ± 4%

Note. The A/B Effect refers to the difference between the average rating of the two interventions and the rating of the A/B test. Mean(A,B) > AB is the percentage of people whose average intervention rating was higher than their rating of the A/B test. The Reverse A/B Effect refers to difference between the rating of the A/B test and the average rating of the two interventions. AB > Mean(A,B) is the percentage of people who rating of the A/B test was higher than their average intervention rating. Experiment Aversion refers to the difference between the rating of the A/B test and the lowest-rated intervention. Min(A,B) > AB is the percentage of people whose lowest-rated intervention is rated higher than their rating of the A/B test. Experiment Appreciation refers to the difference between the rating of the highest-rated intervention and the rating of the A/B test. AB > Max(A,B) is the percentage of people whose rating of the A/B test is higher than the rating of their highest-rated intervention. Experiment Rejection is the percentage of people who rated interventions A and B as "neither inappropriate nor appropriate" or more appropriate while rating the A/B test as "very" or "somewhat" inappropriate. Experiment Endorsement is the percentage of people who rated the A/B test as "very" or "somewhat" appropriate while rating interventions A and B as "neither inappropriate nor appropriate" or less appropriate.

*p < .05

**p < .01

***p < .001

Table S6B

Descriptive and inferential results of ratings and rankings of interventions and experiment for all vignettes

	Descriptive Results				***Inferential Results***	
Vignette	**Variable**	**Mean (SD)**	**% Ranking Best**	**% Ranking Worst**	**Test Description**	**Test Outcome**
Lay Sentiments About Covid-19 pRCTs						
Intubation Safety Checklist ($n = 346$ laypeople)					A/B Effect	$t(345) = 10.69$***, $d = 0.75 \pm .16$
					Mean(A,B) > AB	58% ± 5%
	A	3.81 (1.10)	29%	29%	Reverse A/B effect	$t(345) = -10.69$***, $d = -0.75 \pm .16$
	B	3.99 (1.13)	43%	19%	AB > Mean(A,B)	25% ± 4%
	AB	2.98 (1.46)	29%	52%	Experiment Aversion	$t(345) = 5.28$***, $d = 0.35 \pm .14$
	Mean(A,B)	3.90 (0.88)	-	-	Min(A,B) > AB	45% ± 5%
	Min(A,B)	3.46 (1.19)	-	-	Experiment Appreciation	$t(345) = -14.94$***, $d = -1.14 \pm .19$
	Max(A,B)	4.34 (0.84)	-	-	AB > Max(A,B)	14% ± 3%
					Experiment Rejection (A,B = 3,4,5; AB = 1,2)	31% ± 5%
					Experiment Endorsement (AB = 4,5; A,B = 1,2,3)	4% ± 2%
Best Corticosteroid Drug ($n = 357$ laypeople)					A/B Effect	$t(356) = 2.28$*, $d = 0.17 \pm .15$
					Mean(A,B) > AB	34% ± 5%
	A	3.89 (1.03)	17%	32%	Reverse A/B effect	$t(356) = -2.28$*, $d = -0.17 \pm .15$
	B	3.90 (1.00)	18%	37%	AB > Mean(A,B)	38% ± 5%
	AB	3.69 (1.37)	65%	31%	Experiment Aversion	$t(356) = 1.55$, $p = .123$, $d = 0.12 \pm .15$
	Mean(A,B)	3.90 (0.99)	-	-	Min(A,B) > AB	31% ± 5%
	Min(A,B)	3.83 (1.04)	-	-	Experiment Appreciation	$t(356) = -2.99$**, $d = -0.23 \pm .15$
	Max(A,B)	3.96 (0.98)	-	-	AB > Max(A,B)	35% ± 5%
					Experiment Rejection (A,B = 3,4,5; AB = 1,2)	22% ± 4%
					Experiment Endorsement (AB = 4,5; A,B = 1,2,3)	17% ± 4%
Best Vaccine ($n = 450$ laypeople)					A/B Effect	$t(449) = 2.41$*, $d = 0.15 \pm .12$
					Mean(A,B) > AB	34% ± 4%
	A	3.95 (1.09)	26%	27%	Reverse A/B effect	$t(449) = -2.41$*, $d = -0.15 \pm .12$
	B	3.84 (1.09)	19%	39%	AB > Mean(A,B)	36% ± 4%
	AB	3.72 (1.34)	55%	34%	Experiment Aversion	$t(449) = 0.61$, $p = .546$, $d = 0.04 \pm .12$
	Mean(A,B)	3.90 (1.03)	-	-	Min(A,B) > AB	29% ± 4%
	Min(A.B)	3.77 (1.13)	-	-	Experiment Appreciation	$t(449) = -4.06$***, $d = -0.25 \pm .12$
	Max(A,B)	4.03 (1.04)	-	-	AB > Max(A,B)	32% ± 4%
					Experiment Rejection (A,B = 3,4,5; AB = 1,2)	17% ± 3%
					Experiment Endorsement (AB = 4,5; A,B = 1,2,3)	13% ± 3%

Note. The A/B Effect refers to the difference between the average rating of the two interventions and the rating of the A/B test. Mean(A,B) > AB is the percentage of people whose average intervention rating was higher than their rating of the A/B test. The Reverse A/B Effect refers to difference between the rating of the A/B test and the average rating of the two interventions. AB > Mean(A,B) is the percentage of people who rating of the A/B test was higher than their average intervention rating. Experiment Aversion refers to the difference between the rating of the A/B test and the lowest-rated intervention. Min(A,B) > AB is the percentage of people whose lowest-rated intervention is rated higher than their rating of the A/B test. Experiment Appreciation refers to the difference between the rating of the highest-rated intervention and the rating of the A/B test. AB > Max(A,B) is the percentage of people whose rating of the A/B test is higher than the rating of their highest-rated intervention. Experiment Rejection is the percentage of people who rated interventions A and B as "neither inappropriate nor appropriate" or more appropriate while rating the A/B test as "very" or "somewhat" inappropriate. Experiment Endorsement is the percentage of people who rated the A/B test as "very" or "somewhat" appropriate while rating interventions A and B as "neither inappropriate nor appropriate" or less appropriate.

*$p < .05$

**$p < .01$

Table S6B, continued

Descriptive and inferential results of ratings and rankings of interventions and experiment for all vignettes

Descriptive Results					***Inferential Results***	
Vignette	**Variable**	**Mean (SD)**	**% Ranking Best**	**% Ranking Worst**	**Test Description**	**Test Outcome**
Lay Sentiments About Covid-19 pRCTs						
					A/B Effect	t(338) = 6.42***, d = 0.39 ± .12
					Mean(A,B) > AB	46% ± 5%
	A	3.45 (1.15)	17%	46%	Reverse A/B effect	t(338) = -6.42***, d = -0.39 ± .12
	B	3.96 (1.03)	53%	14%	AB > Mean(A,B)	28% ± 5%
School	AB	3.24 (1.36)	30%	40%	Experiment Aversion	t(338) = 0.47, p = .638, d = 0.03 ± .12
Reopening	Mean(A,B)	3.70 (0.90)	-	-	Min(A,B) > AB	28% ± 5%
(n = 339	Min(A,B)	3.28 (1.15)	-	-	Experiment Appreciation	t(338) = -11.25***, d = -0.75 ± .15
laypeople)	Max(A,B)	4.12 (0.91)	-	-	AB > Max(A,B)	15% ± 3%
					Experiment Rejection (A,B = 3,4,5; AB = 1,2)	19% ± 4%
					Experiment Endorsement (AB = 4,5; A,B = 1,2,3)	4% ± 2%
					A/B Effect	t(356) = 6.07***, d = 0.42 ± .14
					Mean(A,B) > AB	45% ± 5%
	A	3.82 (1.09)	21%	33%	Reverse A/B effect	t(356) = -6.07***, d = -0.42 ± .14
	B	3.96 (1.07)	36%	25%	AB > Mean(A,B)	31% ± 5%
Ventilator	AB	3.39 (1.38)	43%	42%	Experiment Aversion	t(356) = 2.63**, d = 0.17 ± .13
Proning	Mean(A,B)	3.89 (0.96)	-	-	Min(A,B) > AB	36% ± 5%
(n = 357	Min(A,B)	3.61 (1.11)	-	-	Experiment Appreciation	t(356) = -8.927***, d = -0.64 ± .16
laypeople)	Max(A,B)	4.17 (0.99)	-	-	AB > Max(A,B)	22% ± 4%
					Experiment Rejection (A,B = 3,4,5; AB = 1,2)	23% ± 4%
					Experiment Endorsement (AB = 4,5; A,B = 1,2,3)	6% ± 2%
					A/B Effect	t(359) = 14.55***, d = 1.07 ± .18
					Mean(A,B) > AB	68% ± 5%
	A	4.19 (0.95)	44%	14%	Reverse A/B effect	t(359) = -14.55***, d = -1.07 ± .18
	B	3.80 (1.34)	38%	27%	AB > Mean(A,B)	21% ± 4%
Masking	AB	2.74 (1.38)	18%	59%	Experiment Aversion	t(359) = 7.63***, d = 0.56 ± .15
Rules	Mean(A,B)	4.00 (0.91)	-	-	Min(A,B) > AB	50% ± 5%
(n = 360	Min(A,B)	3.47 (1.22)	-	-	Experiment Appreciation	t(359) = -20.85***, d = -1.57 ± .22
laypeople)	Max(A,B)	4.53 (0.84)	-	-	AB > Max(A,B)	8% ± 2%
					Experiment Rejection (A,B = 3,4,5; AB = 1,2)	38% ± 5%
					Experiment Endorsement (AB = 4,5; A,B = 1,2,3)	3% ± 1%

Note. The A/B Effect refers to the difference between the average rating of the two interventions and the rating of the A/B test. Mean(A,B) > AB is the percentage of people whose average intervention rating was higher than their rating of the A/B test. The Reverse A/B Effect refers to difference between the rating of the A/B test and the average rating of the two interventions. AB > Mean(A,B) is the percentage of people who rating of the A/B test was higher than their average intervention rating. Experiment Aversion refers to the difference between the rating of the A/B test and the lowest-rated intervention. Min(A,B) > AB is the percentage of people whose lowest-rated intervention is rated higher than their rating of the A/B test. Experiment Appreciation refers to the difference between the rating of the highest-rated intervention and the rating of the A/B test. AB > Max(A,B) is the percentage of people whose rating of the A/B test is higher than the rating of their highest-rated intervention. Experiment Rejection is the percentage of people who rated interventions A and B as "neither inappropriate nor appropriate" or more appropriate while rating the A/B test as "very" or "somewhat" inappropriate. Experiment Endorsement is the percentage of people who rated the A/B test as "very" or "somewhat" appropriate while rating interventions A and B as "neither inappropriate nor appropriate" or less appropriate.

*p < .05

**p < .01

***p < .001

Table S6C

Descriptive and inferential results of ratings and rankings of interventions and experiment for all vignettes

		Descriptive Results			*Inferential Results*	
Vignette	**Variable**	**Mean (SD)**	**% Ranking Best**	**% Ranking Worst**	**Test Description**	**Test Outcome**
Clinician Sentiments About Covid-19 pRCTs						
					A/B Effect	t (270) = 9.00***, d = 0.71 ± .17
					Mean(A,B) > AB	57% ± 6%
Intubation Safety Checklist (n = 271 clinicians)	A	3.37 (1.26)	19%	32%	Reverse A/B effect	t (270) = -9.00***, d = -0.71 ± .17
	B	3.90 (1.12)	53%	14%	AB > Mean(A,B)	23% ± 5%
	AB	2.74 (1.49)	28%	54%	Experiment Aversion	t (270) = 3.98***, d = 0.30 ± .15
	Mean(A,B)	3.63 (0.96)	-	-	Min(A,B) > AB	43% ± 6%
	Min(A.B)	3.14 (1.23)	-	-	Experiment Appreciation	t (270) =-12.70***, d = -1.08 ± .21
	Max(A,B)	4.12 (1.01)	-	-	AB > Max(A,B)	16% ± 4%
					Experiment Rejection (A,B = 3,4,5; AB = 1,2)	28% ± 5%
					Experiment Endorsement (AB = 4,5; A,B = 1,2,3)	6% ± 2%
					A/B Effect	t (274) = 6.59***, d = 0.52 ± .17
					Mean(A,B) > AB	48% ± 6%
Best Corticosteroid Drug (n = 275 clinicians)	A	3.76 (1.10)	28%	28%	Reverse A/B effect	t (274) = -6.59***, d = -0.52 ± .17
	B	3.74 (1.09)	23%	26%	AB > Mean(A,B)	27% ± 5%
	AB	3.04 (1.56)	49%	46%	Experiment Aversion	t (274) = 6.18***, d = 0.49 ± .17
	Mean(A,B)	3.75 (1.08)	-	-	Min(A,B) > AB	46% ± 6%
	Min(A,B)	3.71 (1.11)	-	-	Experiment Appreciation	t (274) = -6.93***, d = -0.55 ± .17
	Max(A,B)	3.79 (1.08)	-	-	AB > Max(A,B)	26% ± 5%
					Experiment Rejection (A,B = 3,4,5; AB = 1,2)	34% ± 5%
					Experiment Endorsement (AB = 4,5; A,B = 1,2,3)	15% ± 4%

Note. The A/B Effect refers to the difference between the average rating of the two interventions and the rating of the A/B test. Mean(A,B) > AB is the percentage of people whose average intervention rating was higher than their rating of the A/B test. The Reverse A/B Effect refers to difference between the rating of the A/B test and the average rating of the two interventions. AB > Mean(A,B) is the percentage of people who rating of the A/B test was higher than their average intervention rating. Experiment Aversion refers to the difference between the rating of the A/B test and the lowest-rated intervention. Min(A,B) > AB is the percentage of people whose lowest-rated intervention is rated higher than their rating of the A/B test. Experiment Appreciation refers to the difference between the rating of the highest-rated intervention and the rating of the A/B test. AB > Max(A,B) is the percentage of people whose rating of the A/B test is higher than the rating of their highest-rated intervention. Experiment Rejection is the percentage of people who rated interventions A and B as "neither inappropriate nor appropriate" or more appropriate while rating the A/B test as "very" or "somewhat" inappropriate. Experiment Endorsement is the percentage of people who rated the A/B test as "very" or "somewhat" appropriate while rating interventions A and B as "neither inappropriate nor appropriate" or less appropriate.

*p < .05
**p < .01
***p < .001

Table S6C, continued

Descriptive and inferential results of ratings and rankings of interventions and experiment for all vignettes

Descriptive Results					*Inferential Results*	
Vignette	**Variable**	**Mean (SD)**	**% Ranking Best**	**% Ranking Worst**	**Test Description**	**Test Outcome**
Clinician Sentiments About Covid-19 pRCTs						
					A/B Effect	t(348) = 16.50***, d = 1.27 ± .20
					Mean(A,B) > AB	72% ± 5%
	A	4.19 (1.05)	39%	15%	Reverse A/B effect	t(348) = -16.50***, d = -1.27 ± .20
	B	4.01 (1.24)	44%	22%	AB > Mean(A,B)	16% ± 3%
Masking Rules (n = 349 clinicians)	AB	2.61 (1.41)	17%	62%	Experiment Aversion	t(348) = 9.72***, d = 0.74 ± .17
	Mean(A,B)	4.10 (0.88)	-	-	Min(A,B) > AB	57% ± 5%
	Min(A,B)	3.58 (1.20)	-	-	Experiment Appreciation	t(348) = -22.58***, d = -1.74 ± .24
	Max(A,B)	4.62 (0.82)	-	-	AB > Max(A,B)	6% ± 2%
					Experiment Rejection (A,B = 3,4,5; AB = 1,2)	43% ± 5%
					Experiment Endorsement (AB = 4,5; A,B = 1,2,3)	2% ± 1%
					A/B Effect	t(1253) = 2.50*, d = 0.10 ± .07
					Mean(A,B) > AB	35% ± 3%
	A	3.56 (1.17)	27%	28%	Reverse A/B effect	t(1253) = -2.50*, d = -0.10 ± .07
	B	3.40 (1.18)	17%	39%	AB > Mean(A,B)	34% ± 3%
Best Vaccine (n = 1254 clinicians)	AB	3.36 (1.38)	56%	33%	Experiment Aversion	t(1253) = -0.89, p = .375, d = -0.03 ± .07
	Mean(A,B)	3.48 (1.09)	-	-	Min(A,B) > AB	29% ± 2%
	Min(A,B)	3.32 (1.18)	-	-	Experiment Appreciation	t(1253) = -5.49***, d = -0.22 ± .08
	Max(A,B)	3.64 (1.16)	-	-	AB > Max(A,B)	30% ± 2%
					Experiment Rejection (A,B = 3,4,5; AB = 1,2)	20% ± 2%
					Experiment Endorsement (AB = 4,5; A,B = 1,2,3)	20% ± 2%

Note. The A/B Effect refers to the difference between the average rating of the two interventions and the rating of the A/B test. Mean(A,B) > AB is the percentage of people whose average intervention rating was higher than their rating of the A/B test. The Reverse A/B Effect refers to difference between the rating of the A/B test and the average rating of the two interventions. AB > Mean(A,B) is the percentage of people who rating of the A/B test was higher than their average intervention rating. Experiment Aversion refers to the difference between the rating of the A/B test and the lowest-rated intervention. Min(A,B) > AB is the percentage of people whose lowest-rated intervention is rated higher than their rating of the A/B test. Experiment Appreciation refers to the difference between the rating of the highest-rated intervention and the rating of the A/B test. AB > Max(A,B) is the percentage of people whose rating of the A/B test is higher than the rating of their highest-rated intervention. Experiment Rejection is the percentage of people who rated interventions A and B as "neither inappropriate nor appropriate" or more appropriate while rating the A/B test as "very" or "somewhat" inappropriate. Experiment Endorsement is the percentage of people who rated the A/B test as "very" or "somewhat" appropriate while rating interventions A and B as "neither inappropriate nor appropriate" or less appropriate.

*p < .05

**p < .01

***p < .001

Comparisons to previously published work

To compare these results to our previous findings reporting sentiments about experiments, as we do in the main text, please refer to Heck et al. (2020) [4]. For example, in the Results section "Lay Sentiments About pRCTs," we say, "these levels of experiment aversion near the height of the pandemic were slightly (but not significantly) higher than those we observed among similar laypeople in 2019 (41% ± 5% in 2020 vs. 37% ± 6% in 2019 for Catheterization Safety Checklist, p = .31 ; 44% ± 5% in 2020 vs. 40% ± 6% in 2019 for Best Anti-Hypertensive Drug, p = .32)." We extracted the percentage of participants who were experiment averse in 2019 from Heck et al. (2020) [4]. We then performed a two-sample z-test for proportions to compare the 2019 and 2020 proportions. As noted in the main text, we did not find a significant difference between the percentage of people who were experiment averse in 2019 and the percentage of people who were experiment averse in the current studies which took place in 2020 and 2021 (Catheterization Safety Checklist: $\chi^2(1) = 1.034$, p = .309, Anti- Hypertensive Drug: $\chi^2(1) = 0.998$, p = .318).

Results not presented in the main text

Results of Best Vaccine vignette (initial ambiguous version)

The only vignette which showed no A/B Effect was the initial ambiguous version of Best Vaccine (see Table S6D). The two versions of Best Vaccine both presented a public health official's decision to either distribute an mRNA-based vaccine to every county in their state, distribute an inactivated-virus vaccine to every county, or run an experiment in which counties are randomized to receive one of the two vaccine types. However, in version 1, the wording unintentionally implied that residents could choose their vaccine (by going elsewhere) if they did not wish to be subject to the official's decision (including intervention implementation or A/B test), while in version 2 we eliminated this possible interpretation; we suspect this had the effect of making the experiment condition in version 1 less aversive, since people could effectively opt- out of it, and our goal in this research is to study pragmatic, real-world situations in which avoiding randomization is typically not a realistic option.

Table S6D

Descriptive and inferential results of ratings and rankings of interventions and experiment for all vignettes

		Descriptive Results			***Inferential Results***	
Vignette	**Variable**	**Mean (SD)**	**% Ranking Best**	**% Ranking Worst**	**Test Description**	**Test Outcome**
Best Vaccine (initial ambiguous version; n = 350 laypeople)					A/B Effect	t (349) = -0.72, p = .473, d = -0.05 ± .15
					Mean(A,B) > AB	33% ± 5%
	A	3.58 (1.08)	21%	29%	Reverse A/B effect	t (349) = 0.72, p = .473, d = 0.05 ± .15
	B	3.47 (1.10)	21%	40%	AB > Mean(A,B)	45% ± 5%
	AB	3.59 (1.37)	58%	31%	Experiment Aversion	t (349) = -2.28*, d = -0.17 ± .15
	Mean(A,B)	3.53 (1.02)	-	-	Min(A,B) > AB	29% ± 5%
	Min(A,B)	3.38 (1.11)	-	-	Experiment Appreciation	t (349) = -0.84, p = .399, d = -0.07 ± .15
	Max(A,B)	3.67 (1.05)	-	-	AB > Max(A,B)	40% ± 5%
					Experiment Rejection (A,B = 3,4,5; AB = 1,2)	21% ± 4%
					Experiment Endorsement (AB = 4,5; A,B = 1,2,3)	24% ± 4%

Order effect in clinician study

For the clinician study of the Catheterization Safety Checklist, Best Anti-Hypertensive Drug, and Masking Rules vignettes, participants were randomly assigned to one of these three vignettes and then completed the remaining two vignettes in random order. For consistency with the rest of this project and with our previous approach (Meyer et al., 2019) [3], we analyze data from this study as a between-subjects design where we only consider the first vignette that every participant completed.

While conducting an interim analysis on the data for this study, we observed an intriguing and unexpected order effect of presentation.

For the first 601 complete responses we received, we observed an effect of presentation order on participants' appropriateness ratings of the A/B test condition within the Best Anti-Hypertensive Drug vignette. Participants who received the Best Anti-Hypertensive Drug vignette first rated the A/B test an average of 2.95 (SD = 1.57), participants who received this vignette second rated the A/B test an average of 3.48 (SD = 1.39), and participants who received this vignette last rated the A/B test an average of 3.78 (SD = 1.41). This suggests that participants who read about other policies and A/B tests before considering the Best Anti-Hypertensive Drug vignette found the A/B test in the Best Anti-Hypertensive Drug vignette to be less objectionable than participants who received this vignette earlier in the survey. The relationship between presentation order (1, 2, or 3) and appropriateness rating of the A/B test was r = .23. This order effect did not emerge for the other two vignettes or for ratings of either intervention (A or B).

After observing this order effect but before examining any additional data, we preregistered this order effect with the goal of replicating it in an independent sample. 294 new participants completed the study after this interim analysis, and we analyzed the data from this sample independently from the sample that generated the order effect. Table S7 displays ratings of the A/B condition within each scenario grouped by the order in which participants received them.

The order effect observed with the Best Anti-Hypertensive Drug A/B test condition replicated (*r* = .15), as did the absence of any similar order effect for the other conditions.

Table S7

Ratings of A/B test in Clinician Sample

Exploratory Sample (N = 601)	**Best Corticosteroid Drug A/B Rating (SD)**	**Intubation Safety Checklist A/B Rating (SD)**	**Masking Rules A/B Rating (SD)**
Target Scenario First	2.95 (1.57)	2.79 (1.49)	2.63 (1.43)
Target Scenario Second	3.48 (1.39)	2.53 (1.35)	2.66 (1.44)
Target Scenario Last	3.78 (1.41)	2.78 (1.38)	2.57 (1.29)

Confirmatory Sample (N=294)	**Best Corticosteroid Drug A/B Rating (SD)**	**Intubation Safety Checklist A/B Rating (SD)**	**Masking Rules A/B Rating (SD)**
Target Scenario First	3.22 (1.54)	2.63 (1.50)	2.58 (1.38)
Target Scenario Second	3.49 (1.51)	2.76 (1.39)	2.38 (1.42)
Target Scenario Last	3.77 (1.33)	2.69 (1.15)	2.51 (1.38)

Heterogeneity in experiment aversion

In both the lay participant sample and the clinician sample, associations between demographic variables, including educational attainment, having a degree in a STEM field, years of experience in the medical field, and role in the healthcare system, and sentiment about pRCTs (e.g., A/B effect, experiment aversion, experiment appreciation) are consistently small ($r < |.13|$, therefore explaining less than 2% of the variance; Tables S8–11).

In the lay sample, women show larger AB and experiment aversion effects (e.g., larger difference between mean intervention rating/lowest-rated intervention rating and AB test rating; r = .067–.068, $p < .001$) and a smaller experiment appreciation effect (e.g., smaller difference between AB test and highest-rated intervention rating; r = –.064, $p < .001$). Lay participants who are more conservative (in general and with respect to social and economic issues) or more likely to be strong Republicans show lower levels of an AB effect and experiment aversion (i.e., smaller difference between mean intervention rating/lowest-rated intervention rating and AB test rating; all *r*s < –.094, *p*s < .0001). These participants also show significantly more experiment appreciation, though the strength of the association is weaker (*r*s = .037–.046, $p < .0001$).
Finally, we find that people who are non-religious show a larger degree of experiment aversion ($r = .061$, $p < .001$; they also show a larger AB effect, $r = .051$, but $p = .007$ which is greater than $p < .005$, the standard proposed in Benjamin et al. (2018)[17] for exploratory analyses without a priori hypotheses). For all other variables, we find no significant associations between the individual difference measures and experiment sentiments (all *r*s < |.051|, all *p*s > .005).

In the clinician sample, the strongest association was between self-reported comfort with research methods and statistics and experiment aversion—clinicians who report being more comfortable with research methods and statistics are more likely to appreciate the A/B test (r = .070, $p = .001$).

Table S8

Correlations between lay participant characteristics and sentiments about experiments

	Size of A/B effect		A/B effect		Size of experiment aversion		Experiment aversion		Experiment rejection		Size of experiment appreciation		Experiment appreciation		Experiment endorsement	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Age	-0.008	0.662	-0.020	0.286	-0.020	0.270	-0.038	0.043	-0.046	0.012	-0.004	0.809	-0.016	0.389	-0.033	0.073
Sex (1 = male, 2 = female)	0.068	<.001	0.048	0.010	0.067	<.001	0.039	0.035	0.059	0.002	-0.064	<.001	-0.071	<.001	-0.036	0.053
Race (0 = all other, 1 = Nonhispanic White)	-0.004	0.814	-0.017	0.360	-0.001	0.945	-0.016	0.388	0.003	0.867	0.007	0.706	0.001	0.937	-0.012	0.533
Education	0.047	0.011	0.033	0.075	0.049	0.008	0.051	0.006	0.029	0.114	-0.042	0.024	-0.023	0.216	-0.019	0.298
Income	0.020	0.293	0.005	0.787	0.020	0.273	0.011	0.571	0.005	0.777	-0.017	0.353	-0.025	0.184	-0.026	0.158
Political Ideology (1 = Very Liberal, 5 = Very Conservative)	-0.114	< .0001	-0.087	< .0001	-0.118	< .0001	-0.101	< .0001	-0.091	< .0001	0.101	<.0001	0.043	0.022	0.045	0.015
Political Ideology (Social) (1 = Very Liberal, 5 = Very Conservative)	-0.123	< .0001	-0.099	< .0001	-0.128	< .0001	-0.118	< .0001	-0.106	< .0001	0.109	<.0001	0.039	0.036	0.052	0.005
Political Ideology (Economic) (1 = Very Liberal, 5 = Very Conservative)	-0.094	< .0001	-0.065	<.001	-0.095	< .0001	-0.082	< .0001	-0.073	< .0001	0.085	<.0001	0.046	0.013	0.040	0.031
Political Party (1 = Strong Democrat, 7 = Strong Republican)	-0.096	< .0001	-0.073	< .0001	-0.098	< .0001	-0.075	< .0001	-0.075	< .0001	0.087	<.0001	0.037	0.050	0.035	0.063
Conservatism (mean of z-scored Political Ideology, Political Ideology (Social), Political Ideology (Economic), and Political Party)	-0.117	<.0001	-0.089	< .0001	-0.121	< .0001	-0.103	< .0001	-0.095	< .0001	0.105	<.0001	0.045	0.015	0.047	0.012
Non-religious (0 = Religious (any religion), 1 = non-religious)	0.051	0.007	0.027	0.150	0.061	<.001	0.049	0.009	0.046	0.015	-0.036	0.053	-0.013	0.496	-0.021	0.266
STEM degree (0 = no, 1 = yes)	0.023	0.208	0.016	0.399	0.027	0.154	0.026	0.157	0.027	0.142	-0.019	0.318	0.016	0.403	0.024	0.205

Note. Size of the A/B effect refers to the magnitude of the difference between the mean intervention rating and the A/B test rating. A/B effect refers to the presence or absence of an A/B effect -- people who have a positive difference between their mean intervention rating and their A/B test rating show the A/B effect, people who have no difference or a negative difference between their mean intervention rating and their A/B test rating do not show an A/B effect. Size of experiment aversion refers to the magnitude of the difference between the worst intervention rating and the A/B test rating. Experiment aversion refers to the presence or absence of experiment aversion -- people who have a positive difference between their rating of their least-preferred intervention and their A/B test rating are experiment averse, people who have no difference or a negative difference are not experiment averse. Experiment rejection refers to the presence or absence of experiment rejection -- people who rate interventions A and B as "neither inappropriate nor appropriate" or more appropriate while rating the A/B test as "very" or "somewhat" inappropriate reject the experiment. Size of experiment appreciation refers to the magnitude of the difference between the A/B test rating and the best intervention. Experiment appreciation refers to the presence or absence of experiment appreciation -- people who have a positive difference between their rating of the A/B test and their rating of their most-preferred intervention are experiment appreciative. Experiment endorsement refers to the presence or absence of experiment endorsement -- people who rate the A/B test as "very" or "somewhat" appropriate while rating interventions A and B as "neither inappropriate nor appropriate" or less appropriate endorse the experiment.

Table S9

Means and percentages of sentiments about experiments by demographic variable in lay participants

	Size of A/B effect		A/B effect	Size of experiment aversion		Experiment aversion	Experiment rejection	Size of experiment appreciation		Experiment appreciation	Experiment endorsement
	mean	SD	%	mean	SD	%	%	mean	SD	%	%
Sex											
Male	0.479	1.620	45.6	0.183	1.650	35.7	23.2	-0.775	1.730	25.0	9.8
Female	0.703	1.630	50.4	0.408	1.680	39.5	28.4	-0.998	1.710	19.1	7.8
Other	0.571	1.880	28.6	0.429	1.810	28.6	28.6	-0.714	1.980	28.6	0.0
Prefer not to answer	0.900	1.880	60.0	0.800	1.920	40.0	20.0	-1.000	1.870	20.0	0.0
Race											
Black/African-American	0.504	1.597	49.8	0.149	1.647	37.2	21.8	-0.858	1.681	21.5	9.6
Hispanic or Latino	0.692	1.646	50.2	0.429	1.675	38.8	28.8	-0.954	1.726	20.1	7.8
White	0.601	1.631	47.7	0.309	1.671	37.2	26.2	-0.893	1.724	21.7	8.4
Asian	0.594	1.634	47.1	0.296	1.645	39.2	26.1	-0.892	1.757	23.2	10.5
Other	0.679	1.730	48.7	0.256	1.831	38.5	23.1	-1.103	1.818	25.6	5.1
Prefer not to answer	1.200	1.623	60.0	0.933	1.624	40.0	33.3	-1.467	1.767	13.3	6.7
Education											
Less than high school	1.580	1.440	75.0	1.330	1.610	58.3	41.7	-1.830	1.400	0.0	0.0
High school degree	0.403	1.550	42.2	0.093	1.650	30.6	22.0	-0.713	1.610	20.9	9.0
Some college	0.524	1.690	47.5	0.216	1.720	36.3	25.2	-0.831	1.790	24.2	10.2
Four-year college degree	0.643	1.620	48.7	0.361	1.650	38.4	26.7	-0.925	1.710	21.4	8.0
Some graduate school	0.673	1.600	50.0	0.379	1.640	37.9	28.2	-0.968	1.700	20.2	6.5
Graduate degree	0.713	1.590	50.6	0.419	1.620	41.7	27.8	-1.010	1.690	19.8	8.2
Prefer not to answer	0.750	1.720	50.0	0.667	1.750	33.3	16.7	-0.833	1.720	16.7	0.0
Income											
< $20,000	0.672	1.570	47.8	0.380	1.650	37.7	26.8	-0.964	1.640	17.4	6.9
$20,000-$40,000	0.480	1.700	46.6	0.215	1.730	37.1	25.0	-0.745	1.790	27.8	10.8
$40,000-$60,000	0.592	1.630	49.4	0.220	1.670	36.9	25.4	-0.930	1.750	20.5	8.9
$60,000-$80,000	0.629	1.620	49.5	0.376	1.640	38.0	27.4	-0.883	1.710	20.9	10.5
$80,000-$100,000	0.741	1.520	50.0	0.488	1.530	41.3	27.2	-0.994	1.640	18.9	6.0
> $100,000	0.608	1.620	47.2	0.302	1.680	37.5	25.7	-0.914	1.700	21.0	7.4
Prefer not to answer	0.861	1.940	47.2	0.556	2.080	38.9	36.1	-1.170	1.930	19.4	2.8
No response	-0.250	0.866	25.0	-0.500	1.000	0.0	0.0	0.000	0.816	25.0	0.0

Table S9, continued

Means and percentages of sentiments about experiments by demographic variable in lay participants

	Size of A/B effect		A/B effect	Size of experiment aversion		Experiment aversion	Experiment rejection	Size of experiment appreciation		Experiment appreciation	Experiment endorsement
	mean	SD	%	mean	SD	%	%	mean	SD	%	%
Political Ideology											
Very liberal	0.888	1.740	54.3	0.590	1.780	44.1	31.1	-1.190	1.830	19.8	6.1
Liberal	0.753	1.650	51.6	0.491	1.680	42.3	29.8	-1.010	1.740	20.2	8.2
Moderate	0.557	1.570	47.5	0.247	1.600	36.2	25.4	-0.867	1.670	21.1	8.1
Conservative	0.380	1.600	43.8	0.058	1.650	33.1	21.4	-0.703	1.700	25.0	11.2
Very conservative	0.307	1.520	39.0	0.026	1.570	27.7	18.6	-0.589	1.500	24.2	9.5
Prefer not to answer	0.684	1.680	57.9	0.263	1.560	31.6	21.1	-1.110	1.940	21.1	15.8
No response	0.625	0.750	50.0	0.250	0.957	50.0	50.0	-1.000	0.816	0.0	0.0
Political Ideology (Social)											
Very liberal	0.927	1.720	55.7	0.628	1.760	46.3	33.3	-1.230	1.810	19.1	5.5
Liberal	0.714	1.610	51.2	0.445	1.640	41.1	28.5	-0.983	1.710	20.9	8.2
Moderate	0.498	1.600	45.2	0.205	1.660	35.2	25.0	-0.791	1.680	22.1	9.4
Conservative	0.321	1.590	42.5	-0.016	1.630	30.6	19.8	-0.658	1.710	25.1	12.1
Very conservative	0.362	1.500	40.6	0.059	1.550	28.9	18.8	-0.665	1.590	22.6	8.0
Prefer not to answer	0.528	1.540	55.6	0.222	1.560	33.3	11.1	-0.833	1.650	16.7	11.1
No response	-1.000	NA	0.0	-2.000	NA	0.0	0.0	0.000	NA	0.0	0.0
Political Ideology (Economic)											
Very liberal	0.795	1.760	49.4	0.514	1.770	40.5	28.6	-1.080	1.870	19.9	6.7
Liberal	0.800	1.630	53.8	0.512	1.670	43.7	31.5	-1.090	1.730	18.9	7.8
Moderate	0.594	1.600	48.2	0.307	1.650	38.0	25.5	-0.882	1.670	21.4	8.4
Conservative	0.401	1.580	44.2	0.076	1.620	33.5	22.4	-0.726	1.710	25.5	10.4
Very conservative	0.435	1.600	42.9	0.165	1.650	30.7	21.7	-0.705	1.660	22.7	9.6
Prefer not to answer	0.783	1.540	65.2	0.435	1.530	39.1	21.7	-1.130	1.660	13.0	8.7
No response	-1.000	0.000	0.0	-1.500	0.707	0.0	0.0	0.500	0.707	50.0	0.0
Political Party											
Strong Democrat	0.869	1.710	54.6	0.582	1.720	43.9	28.7	-1.160	1.820	19.6	7.6
Democrat	0.701	1.630	50.7	0.411	1.690	39.7	29.9	-0.990	1.700	19.9	6.7
Independent (but lean Democrat)	0.755	1.620	51.9	0.470	1.640	42.0	29.6	-1.040	1.730	21.0	8.6
Independent	0.468	1.590	43.7	0.173	1.630	34.0	23.3	-0.762	1.670	22.1	9.2
Independent (but lean Republican)	0.437	1.720	42.4	0.144	1.730	33.9	24.7	-0.731	1.830	28.8	14.8
Republican	0.387	1.550	44.8	0.076	1.610	33.4	20.9	-0.699	1.640	22.5	8.8
Strong Republican	0.432	1.500	44.0	0.130	1.570	32.6	20.7	-0.734	1.580	21.7	7.6
Prefer not to answer	0.615	1.580	56.4	0.282	1.490	41.0	23.1	-0.949	1.790	20.5	10.3
No response	-1.000	NA	0.0	-2.000	NA	0.0	0.0	0.000	NA	0.0	0.0

Table S9, continued

Means and percentages of sentiments about experiments by demographic variable in lay participants

	Size of A/B effect		A/B effect	Size of experiment aversion		Experiment aversion	Experiment rejection	Size of experiment appreciation		Experiment appreciation	Experiment endorsement
	mean	SD	%	mean	SD	%	%	mean	SD	%	%
Religion											
Christian - Protestant	0.515	1.620	45.9	0.212	1.680	34.9	24.3	-0.818	1.700	22.5	10.0
Christian - Catholic	0.483	1.510	46.7	0.176	1.550	34.4	21.6	-0.790	1.610	20.7	6.4
Christian - Other	0.589	1.650	48.3	0.298	1.690	37.3	25.4	-0.881	1.740	22.9	9.7
Jewish	0.868	1.720	54.7	0.453	1.840	43.4	32.1	-1.280	1.770	13.2	7.6
Muslim	0.357	1.700	45.7	-0.057	1.800	28.6	20.0	-0.771	1.780	31.4	17.1
Buddhist	0.840	1.690	54.0	0.520	1.570	48.0	32.0	-1.160	1.940	24.0	14.0
Hindu	-0.129	1.550	38.7	-0.452	1.570	29.0	16.1	-0.194	1.620	35.5	19.4
Non-religious	0.704	1.650	49.9	0.435	1.680	40.7	28.5	-0.973	1.750	21.1	8.0
Other	0.673	1.780	49.0	0.337	1.810	40.4	31.7	-1.010	1.880	22.1	8.7
Prefer not to answer	1.090	1.570	58.8	0.794	1.650	41.2	38.2	-1.380	1.600	11.8	0.0
No response	1.250	1.770	50.0	1.000	1.410	50.0	50.0	-1.500	2.120	0.0	0.0
STEM degree											
No	0.587	1.620	47.9	0.289	1.650	37.2	25.6	-0.885	1.720	21.3	8.4
Yes	0.680	1.680	49.8	0.397	1.740	40.3	28.5	-0.963	1.750	22.9	10.0
Prefer not to answer	0.400	1.510	40.0	0.200	1.510	30.0	15.0	-0.600	1.570	25.0	0.0
No response	0.250	1.060	50.0	-0.500	0.707	0.0	0.0	-1.000	1.410	0.0	0.0

Note. If there is an NA in the SD column, that indicates that there was only 1 respondent in that group so there is no variability in responses to report.

Size of the A/B effect refers to the magnitude of the difference between the mean intervention rating and the A/B test rating. A/B effect refers to the presence or absence of an A/B effect -- people who have a positive difference between their mean intervention rating and their A/B test rating show the A/B effect, people who have no difference or a negative difference between their mean intervention rating and their A/B test rating do not show an A/B effect. Size of experiment aversion refers to the magnitude of the difference between the worst intervention rating and the A/B test rating. Experiment aversion refers to the presence or absence of experiment aversion -- people who have a positive difference between their rating of their least-preferred intervention and their A/B test rating are experiment averse, people who have no difference or a negative difference are not experiment averse. Experiment rejection refers to the presence or absence of experiment rejection -- people who rate interventions A and B as "neither inappropriate nor appropriate" or more appropriate while rating the A/B test as "very" or "somewhat" inappropriate reject the experiment. Size of experiment appreciation refers to the magnitude of the difference between the A/B test rating and the best intervention. Experiment appreciation refers to the presence or absence of experiment appreciation -- people who have a positive difference between their rating of the A/B test and their rating of their most-preferred intervention are experiment appreciative. Experiment endorsement refers to the presence or absence of experiment endorsement -- people who rate the A/B test as "very" or "somewhat" appropriate while rating interventions A and B as "neither inappropriate nor appropriate" or less appropriate endorse the experiment.

Table S10

Correlations between clinician characteristics and sentiments about experiments

	Size of A/B effect		A/B effect		Size of experiment aversion		Experiment aversion		Experiment rejection		Size of experiment appreciation		Experiment appreciation		Experiment endorsement	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Sex (1 = male, 2 = female)	0.016	0.453	0.016	0.457	0.000	0.991	-0.011	0.619	-0.021	0.326	-0.030	0.165	-0.026	0.226	-0.032	0.134
Number of research methods/statistics training units	-0.005	0.812	0.000	0.992	0.000	0.999	0.016	0.471	0.017	0.428	0.010	0.659	0.019	0.382	0.010	0.643
Comfort with research methods/statistics	-0.036	0.100	-0.018	0.410	-0.039	0.071	-0.021	0.335	-0.016	0.446	0.030	0.165	0.070	0.001	0.045	0.035
Number of research methods/statistics activities	-0.019	0.375	-0.022	0.301	-0.006	0.796	0.006	0.778	0.020	0.360	0.031	0.157	0.041	0.056	0.023	0.279
Currently involved in research	-0.002	0.912	-0.012	0.570	-0.009	0.691	-0.016	0.470	-0.022	0.309	-0.004	0.870	-0.024	0.267	0.009	0.693
Position (0 = non-prescriber, 1 = prescriber)	0.033	0.121	0.029	0.176	0.040	0.061	0.042	0.050	0.052	0.016	-0.025	0.250	-0.020	0.347	-0.021	0.338
Years in medicine	0.016	0.452	-0.004	0.865	0.011	0.599	-0.007	0.734	0.006	0.792	-0.020	0.362	0.029	0.185	-0.003	0.879

Note. Size of the A/B effect refers to the magnitude of the difference between the mean intervention rating and the A/B test rating. A/B effect refers to the presence or absence of an A/B effect -- people who have a positive difference between their mean intervention rating and their A/B test rating show the A/B effect, people who have no difference or a negative difference between their mean intervention rating and their A/B test rating do not show an A/B effect. Size of experiment aversion refers to the magnitude of the difference between the worst intervention rating and the A/B test rating. Experiment aversion refers to the presence or absence of experiment aversion -- people who have a positive difference between their rating of their least-preferred intervention and their A/B test rating are experiment averse, people who have no difference or a negative difference are not experiment averse. Experiment rejection refers to the presence or absence of experiment rejection -- people who rate interventions A and B as "neither inappropriate nor appropriate" or more appropriate while rating the A/B test as "very" or "somewhat" inappropriate reject the experiment. Size of experiment appreciation refers to the magnitude of the difference between the A/B test rating and the best intervention. Experiment appreciation refers to the presence or absence of experiment appreciation -- people who have a positive difference between their rating of the A/B test and their rating of their most-preferred intervention are experiment appreciative. Experiment endorsement refers to the presence or absence of experiment endorsement -- people who rate the A/B test as "very" or "somewhat" appropriate while rating interventions A and B as "neither inappropriate nor appropriate" or less appropriate endorse the experiment.

Table S11

Means and percentages of sentiments about experiments by demographic variable in clinician sample

	Size of A/B effect		A/B effect	Size of experiment aversion		Experiment aversion	Experiment rejection	Size of experiment appreciation		Experiment appreciation	Experiment endorsement
	mean	SD	%	mean	SD	%	%	mean	SD	%	%
Sex											
Male	0.456	1.800	43.9	0.270	1.800	38.5	28.2	-0.642	1.890	26.5	17.2
Female	0.529	1.750	45.9	0.271	1.750	37.2	25.8	-0.786	1.890	23.6	14.2
Other	0.000	1.870	40.0	0.000	1.870	40.0	20.0	0.000	1.870	20.0	20.0
Source of research methods/statistics training											
Undergraduate coursework	0.483	1.755	44.2	0.258	1.753	37.7	26.5	-0.707	1.870	25.0	14.1
Professional school instruction	0.571	1.767	46.0	0.314	1.756	38.2	27.1	-0.828	1.916	22.8	14.7
Postgraduate coursework	0.624	1.818	49.4	0.402	1.809	41.5	29.4	-0.847	1.936	24.5	14.5
CME/CEU courses	0.463	1.788	47.1	0.217	1.767	38.6	26.6	-0.708	1.925	25.7	16.7
Self-instruction via peer-reviewed literature	0.333	1.820	41.2	0.097	1.798	32.9	23.2	-0.569	1.949	27.3	16.6
Other	0.722	1.902	46.7	0.478	1.915	41.1	32.2	-0.967	1.986	22.2	14.4
Comfort with research methods/statistics											
Not at all	0.682	1.760	45.8	0.432	1.780	37.7	26.3	-0.932	1.870	18.2	12.7
Somewhat	0.516	1.710	45.7	0.282	1.690	37.8	26.8	-0.750	1.840	22.5	14.0
Moderately	0.482	1.770	46.5	0.237	1.770	38.3	26.6	-0.727	1.880	26.8	15.1
Very	0.491	1.910	43.9	0.203	1.900	34.0	23.1	-0.778	2.070	29.2	17.9
Extremely	0.105	2.020	31.6	-0.079	2.050	28.9	23.7	-0.289	2.100	26.3	23.7
Research methods/statistics activities											
Read results of RCT in peer-reviewed journal article	0.521	1.772	45.5	0.284	1.762	38.0	27.2	-0.758	1.898	24.7	15.0
Changed typical prescription/recommendation after personally reading results of RCT in peer-reviewed journal article	0.430	1.813	43.3	0.217	1.814	36.8	26.3	-0.643	1.921	26.6	16.7
Published scientific paper in peer-reviewed journal	0.530	1.692	43.3	0.339	1.681	38.2	29.9	-0.720	1.802	22.8	13.4
Conducted or worked on a team conducting an RCT	0.371	1.745	42.9	0.114	1.725	35.1	20.9	-0.628	1.902	25.8	16.3
Took a course/class in statistics, biostatistics, research methods	0.505	1.775	45.0	0.277	1.770	37.8	27.3	-0.732	1.892	25.4	15.2
Analyzed data for statistical significance outside of course requirement	0.470	1.781	43.7	0.251	1.766	36.7	26.2	-0.690	1.912	26.2	15.4
Used statistical software	0.588	1.803	49.3	0.389	1.795	42.5	31.7	-0.787	1.915	26.7	14.9

Table S11, continued

Means and percentages of sentiments about experiments by demographic variable in clinician sample

	Size of A/B effect		A/B effect	Size of experiment aversion		Experiment aversion	Experiment rejection	Size of experiment appreciation		Experiment appreciation	Experiment endorsement
	mean	SD	%	mean	SD	%	%	mean	SD	%	%
Currently involved in research											
Yes	0.526	1.740	47.4	0.316	1.720	39.7	29.2	-0.737	1.860	27.3	13.9
No	0.512	1.760	45.3	0.265	1.760	37.2	25.9	-0.759	1.890	23.8	14.9
Position											
Doctor	0.556	1.730	45.5	0.374	1.720	39.9	28.7	-0.738	1.840	23.1	13.7
Physician Assistant	0.757	1.780	53.0	0.508	1.780	44.3	34.4	-1.010	1.890	21.9	13.1
Nurse Practitioner	0.500	1.910	45.9	0.184	1.970	36.7	25.5	-0.816	2.030	23.5	14.3
Nurse (RN)	0.436	1.720	43.8	0.181	1.720	35.2	23.9	-0.690	1.850	25.3	15.1
Nurse (LPN)	0.410	1.790	42.1	0.150	1.760	33.5	22.6	-0.669	1.960	24.8	17.3
Nurse (Other)	1.180	1.910	65.0	0.800	1.910	55.0	35.0	-1.550	2.060	10.0	10.0
Genetic Counselor	---	---	---	---	---	---	---	---	---	---	---
Non-prescribing clinician or staff without clinical credential	---	---	---	---	---	---	---	---	---	---	---
Medical student	1.170	1.770	65.2	0.935	1.790	56.5	45.7	-1.410	1.830	15.2	8.7
Faculty or Professor	1.120	2.050	62.5	0.875	2.030	50.0	37.5	-1.380	2.200	25.0	12.5
Other	0.727	2.000	45.5	0.618	1.980	41.8	32.7	-0.836	2.060	25.5	16.4
Years in medical field											
< 1 year	0.582	1.540	47.5	0.377	1.540	39.3	32.8	-0.787	1.660	24.6	8.2
1-2 years	0.560	1.720	48.4	0.333	1.710	41.3	29.4	-0.786	1.840	23.8	14.3
3-5 years	0.392	1.570	44.8	0.140	1.570	36.0	21.3	-0.643	1.690	23.4	13.6
6-10 years	0.423	1.730	43.3	0.205	1.760	36.5	24.6	-0.641	1.830	26.4	15.1
> 10 years	0.555	1.820	45.9	0.303	1.810	37.5	27.1	-0.807	1.950	23.7	15.3

Note. Size of the A/B effect refers to the magnitude of the difference between the mean intervention rating and the A/B test rating. A/B effect refers to the presence or absence of an A/B effect -- people who have a positive difference between their mean intervention rating and their A/B test rating show the A/B effect, people who have no difference or a negative difference between their mean intervention rating and their A/B test rating do not show an A/B effect. Size of experiment aversion refers to the magnitude of the difference between the worst intervention rating and the A/B test rating. Experiment aversion refers to the presence or absence of experiment aversion -- people who have a positive difference between their rating of their least-preferred intervention and their A/B test rating are experiment averse, people who have no difference or a negative difference are not experiment averse. Experiment rejection refers to the presence or absence of experiment rejection -- people who rate interventions A and B as "neither inappropriate nor appropriate" or more appropriate while rating the A/B test as "very" or "somewhat" inappropriate reject the experiment. Size of experiment appreciation refers to the magnitude of the difference between the A/B test rating and the best intervention. Experiment appreciation refers to the presence or absence of experiment appreciation -- people who have a positive difference between their rating of the A/B test and their rating of their most-preferred intervention are experiment appreciative. Experiment endorsement refers to the presence or absence of experiment endorsement -- people who rate the A/B test as "very" or "somewhat" appropriate while rating interventions A and B as "neither inappropriate nor appropriate" or less appropriate endorse the experiment.

References

1. Germine L, Nakayama K, Duchaine BC, Chabris CF, Chatterjee G, Wilmer JB. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. Psychon Bull Rev. 2012;19(5):847–57.

2. Simons DJ, Chabris CF. Common (mis)beliefs about memory: A replication and comparison of telephone and mechanical turk survey methods. PLoS One. 2012;7(12):e51876.

3. Meyer MN, Heck PR, Holtzman GS, et al. Objecting to experiments that compare two unobjectionable policies or treatments. Proceedings of the National Academy of Sciences 2019;116(22):10723–8.

4. Heck PR, Chabris CF, Watts DJ, Meyer MN. Objecting to experiments even while approving of the policies or treatments they compare. Proceedings of the National Academy of Sciences 2020;117(32):18948–50.

5. Mislavsky R, Dietvorst BJ, Simonsohn U. The minimum mean paradox: A mechanical explanation for apparent experiment aversion. Proceedings of the National Academy of Sciences 2019;116(48):23883–4.

6. Dunlap WP, Cortina JM, Vaslow JB, Burke MJ. Meta-analysis of experiments with matched groups or repeated measures designs. Psychological Methods 1996;1:170–7.

7. Westfall J. effect size | Cookie Scientist [Internet]. 2016;Available from: http://jakewestfall.org/blog/index.php/category/effect-size/

8. Pronovost P, Needham D, Berenholtz S, et al. An Intervention to Decrease Catheter-Related Bloodstream Infections in the ICU. New England Journal of Medicine 2006;355(26):2725– 32.

9. Urbach DR, Govindarajan A, Saskin R, Wilton AS, Baxter NN. Introduction of Surgical Safety Checklists in Ontario, Canada. New England Journal of Medicine 2014;370(11):1029–38.

10. Arriaga AF, Bader AM, Wong JM, et al. Simulation-Based Trial of Surgical-Crisis Checklists. New England Journal of Medicine 2013;368(3):246–53.

11. The ROMP Ethics Study [Internet]. ROMP Ethics Study. Available from: https://www.iths.org/rompethics/

12. Sinnott S-J, Tomlinson LA, Root AA, et al. Comparative effectiveness of fourth-line anti- hypertensive agents in resistant hypertension: A systematic review and meta-analysis. Eur J Prev Cardiol 2017;24(3):228–38.

13. Turner JS, Bucca AW, Propst SL, et al. Association of Checklist Use in Endotracheal Intubation With Clinically Important Outcomes: A Systematic Review and Meta-analysis. JAMA Network Open 2020;3(7):e209278.

14. Wagner C, Griesel M, Mikolajewska A, et al. Systemic corticosteroids for the treatment of COVID-19: Equity-related analyses and update on evidence. Cochrane Database of Systematic Reviews 2022;(11). Available from: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD014963.pub2/full

15. Elharrar X, Trigui Y, Dols A-M, et al. Use of Prone Positioning in Nonintubated Patients With COVID-19 and Hypoxemic Acute Respiratory Failure. JAMA 2020;323(22):2336–8.

16. Sartini C, Tresoldi M, Scarpellini P, et al. Respiratory Parameters in Patients With COVID- 19 After Using Noninvasive Ventilation in the Prone Position Outside the Intensive Care Unit. JAMA 2020;323(22):2338–40.

17. Caputo ND, Strayer RJ, Levitan R. Early Self-Proning in Awake, Non-intubated Patients in the Emergency Department: A Single ED's Experience During the COVID-19 Pandemic. Academic Emergency Medicine 2020;27(5):375–8.

18. Fretheim A, Flatø M, Steens A, et al. COVID-19: we need randomised trials of school closures. J Epidemiol Community Health 2020;74(12):1078–9.

19. Fretheim A. School opening in Norway during the COVID-19 pandemic.

20. The TRAiN study group, Helsingen LM, Løberg M, et al. Randomized Re-Opening of Training Facilities during the COVID-19 pandemic [Internet]. Public and Global Health; 2020. Available from: http://medrxiv.org/lookup/doi/10.1101/2020.06.24.20138768

21. Angrist N, Bergman P, Brewster C, Matsheng M. Stemming Learning Loss During the Pandemic: A Rapid Randomized Trial of a Low-Tech Intervention in Botswana [Internet]. 2020;Available from: https://papers.ssrn.com/abstract=3663098

22. Kolata G. Did Closing Schools Actually Help? [Internet]. The New York Times. 2020;Available from: https://www.nytimes.com/2020/05/02/sunday-review/coronavirus- school-closings.html

23. Abaluck J, Kwong LH, Styczynski A, et al. Impact of community masking on COVID-19: A cluster-randomized trial in Bangladesh. Science 2021;375(6577):eabi9069.

24. Jefferson T, Dooley L, Ferroni E, et al. Physical interventions to interrupt or reduce the spread of respiratory viruses. Cochrane Database of Systematic Reviews [Internet] 2023;(1). Available from: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD006207.pub6/full?s=08

25. Bundgaard H, Bundgaard JS, Raaschou-Pedersen DET, et al. Effectiveness of Adding a Mask Recommendation to Other Public Health Measures to Prevent SARS-CoV-2 Infection in Danish Mask Wearers. Ann Intern Med 2021;174(3):335–43.

26. Bach PB. We can't tackle the pandemic without figuring out which Covid-19 vaccines work the best [Internet]. STAT. 2020;Available from: https://www.statnews.com/2020/09/24/big- trial-needed-determine-which-covid-19-vaccines-work-best/