An illusion of predictability in scientific results

Sam Zhang,[1] Patrick R. Heck,[2] Michelle N. Meyer,[3] Christopher F. Chabris,[3] Daniel G. Goldstein,[4] and Jake M. Hofman[4, *]

[1] *Department of Applied Mathematics, University of Colorado, Boulder, CO*
[2] *Office of Research, Consumer Financial Protection Bureau, Washington, DC*
[3] *Department of Bioethics & Decision Sciences, Geisinger Health System, Danville, PA*
[4] *Microsoft Research, New York, NY*

Traditionally, scientists have placed more emphasis on communicating inferential uncertainty (i.e., the precision of statistical estimates) compared to outcome variability (i.e., the predictability of individual outcomes). Here we show that this can lead to sizable misperceptions about the implications of scientific results. Specifically, we present three pre-registered, randomized experiments where participants saw the same scientific findings visualized as showing only inferential uncertainty, only outcome variability, or both, and answered questions about the size and importance of findings they were shown. Our results, comprised of responses from medical professionals, professional data scientists, and tenure-track faculty, show that the prevalent form of visualizing only inferential uncertainty can lead to significant overestimates of treatment effects, even among highly trained experts. In contrast, we find that depicting both inferential uncertainty and outcome variability leads to more accurate perceptions of results while appearing to leave other subjective impressions of the results unchanged, on average.

Keywords: Statistics, uncertainty, science communication, visualization, experiments.

Much of science is concerned with making inferences about entire populations using only samples from them. For instance, a medical trial might compare the health of patients who were given an experimental treatment to those who received a placebo, or a social science study might contrast the economic mobility of individuals from different demographic groups. In each case the goal is to draw conclusions about the broader populations of interest, but this is often complicated by two factors: first, access to relatively small samples from these populations, and second, highly variable outcomes within each group. For example, a medical study might involve only a few dozen patients, and some patients who received the experimental treatment might have responded strongly to it while others did not.

Perhaps the most common solution to these problems is to focus on aggregate outcomes (e.g., averages within each group) instead of individual outcomes, and to report some measure of *inferential uncertainty* about them (e.g., how precisely we have estimated the average for each group). Reporting inferential uncertainty (typically through standard errors, confidence intervals, Bayesian credible intervals, or similar) has long been a cornerstone of statistics and constitutes a major part of introductory courses on the topic [1]. Quantifying inferential uncertainty is important for many reasons, from providing a plausible range of values for a quantity of interest to helping us avoid being misled by random variation in samples of data that may not accurately reflect trends in the underlying populations of interest.

At the same time, focusing on *only* aggregate outcomes and inferential uncertainty might lead us to overlook *outcome variability* (e.g., how much individual outcomes vary around averages for each group), often quantified by measures such as standard deviation or variance, and which is important for understanding effect sizes and the predictability of outcomes. Although there are systematic relationships between measures of inferential uncertainty and outcome variability, they capture two very different—but easily confused—concepts. Here we investigate the extent to which the pervasive focus on inferential uncertainty in scientific visualizations can produce illusory impressions about the size and importance of scientific findings, even among experts whose jobs involve creating and interpreting such results.

To highlight the difference between inferential uncertainty and outcome variability—and to see why focusing on the former might be misleading about the latter—consider the plot in the upper left panel of Figure 1, inspired by a highly-cited study on whether violent video games cause aggressive behavior [2]. In the study participants were randomly assigned to play either a violent or a non-violent video game, after which their behavior on an unrelated task was measured using a continuous aggressiveness score. The two black points in this plot show estimated average aggressiveness within each group and the error bars encode inferential uncertainty about those estimates (one standard error above and below the average). Compare this to the plot in the upper right panel of Figure 1, which depicts the same averages and error bars, but adds colored points to show individual outcomes as well.

In principle there is no reason to prefer one of these plots to the other—in fact, given the sample sizes and a few distributional assumptions, one can calculate information about either inferential uncertainty or outcome variability from each. In practice, however, each of these representations has distinct visual features that empha-

[*] To whom correspondence should be addressed: jmh@microsoft.com

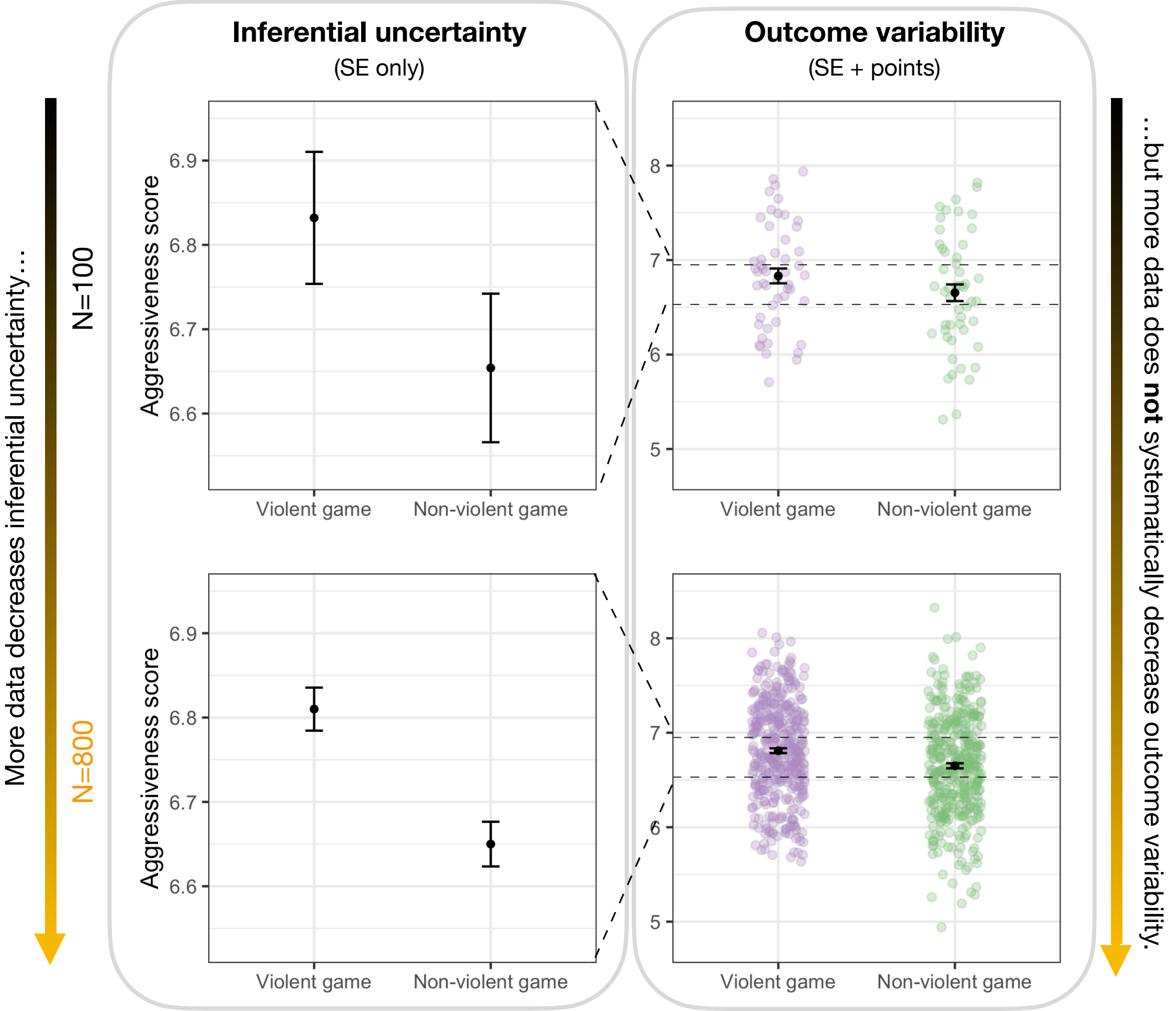


FIG. 1. **Inferential uncertainty vs. outcome variability** (Left) Estimated means and an error bar representing one standard error (SE) above and and one SE below the mean, for two conditions in an experiment. The SE is a measure of the uncertainty in our inference of the mean. (Right) Individual outcomes shown in addition to the same SEs on the left. With only 50 participants per condition (top), we have less confident estimates for the mean than when we have 400 participants per condition (bottom). However, more data does not systematically decrease the variability in the outcomes themselves.

size different notions of uncertainty and lends itself to different interpretations. In particular, the format on the left is designed to facilitate "inference by eye" [3–5], so that readers can deduce a range of plausible values for the average in each group and apply visual heuristics for hypothesis testing. By applying a rough heuristic, the lack of overlap of the error bars is taken as evidence against the idea that there is no difference in average aggressiveness scores between conditions. That said, displaying the error bar as a single visual object (arms capped off on either end), with the mean at its center, focuses perception on that object as itself a representation of the relevant data, and since the object is bounded at the ends of the error bar, such a display encourages the viewer to imagine that the underlying data must cluster more tightly around the mean than it actually does. (In fact, the majority of the individual data points would be above and below the range of the y-axis scale and thus not even visible.) As a result, one might look at the plot on the left and conclude that violent videogames cause aggressive behavior, and indeed popular outlets that covered this work featured strongly-worded headlines to this effect (see [6], for example).

The figure on the right contains all of the information present in the plot on the left, but simply adds points that show individual outcomes. This format was suggested by Gardner and Altman several decades ago [7] to place more emphasis on communicating sample size, outcome variability, and effect sizes. There have since been several efforts to popularize these types of plots [8–12], but they remain relatively uncommon and, to the best of our knowledge, have not been empirically tested. The dots draw some attention away from the object represented by the error bars, and the contrast (in color and intensity) between the dots and the bar make it easy to focus on either the individual points or the mean and error bar, to shift attention between them, and to see that the bar does not represent the entirety of the data, merely one particular facet of it.

Specifically, adding individual outcomes highlights that while there is relatively low inferential uncertainty in this study (i.e., the average in each group is precisely estimated), there is still a great deal of outcome variability within each group (i.e., individual outcomes vary quite a bit around their respective averages). So much so, in fact, that one has to rescale the y-axis just to accommodate the range of outcomes, providing some perspective for the difference in means between conditions. "Inference by eye" is still possible in this alternative representation, but it also makes clear that while violent video games may change aggressive behavior *on average*, the relationship is far from deterministic: knowing only if someone played a violent video game or not says relatively little about how aggressively they might behave.

Moreover, as depicted in the bottom row of Figure 1, this divergence between inferential uncertainty and outcome variability actually *grows* with sample size. For instance, if we were to conduct a larger study—as is more commonplace today compared to when the original study was done—and sample 800 participants instead of 100, we would get extremely precise estimates of averages in each condition (indicated by the small error bars in the bottom left panel), but, as the bottom right panel shows, collecting more data would not systematically decrease outcome variability.

As these examples demonstrate, differences in what visualizations emphasize might lead readers to different conclusions. So which one of these formats should we prefer when presenting statistical findings to readers, and how much does this choice matter? While there is a large body of literature in the fields of data visualization and human-computer interaction on different ways of depicting either inferential uncertainty or outcome variability [13–15], to the best of our knowledge there is little empirical work that compares the two. In many fields there is an emphasis on inference and hypothesis testing, and so plots displaying inferential uncertainty are the default and considered a "best practice" [16–18]. At the same time, there is an increasingly large body of research showing that people routinely make mistakes when making inferences based on such plots. For instance, when shown these plots, people often mis-estimate the range of plausible values for a parameter and draw incorrect conclusions related to hypothesis testing and the replicability of scientific findings [13, 14, 19–21]. As a result, it may be the case that plots designed to convey inferential uncertainty may in fact not be very effective for statistical inference.

Here we raise a different but potentially more important concern. Beyond being unreliable for traditional statistical inference tasks, the pervasive preference for communicating inferential uncertainty found in published work can lead to an "illusion of predictability" [22], whereby people underestimate the variability of outcomes and overestimate the size and importance of scientific findings. In particular, if a reader mistakes inferential uncertainty for outcome variability when viewing the plots like those on the left of Figure 1, they might be left with the impression that most outcomes fall within the depicted error bars and conclude that violent video games have an alarmingly strong effect on aggressive behavior, with predictable outcomes in each condition. The plots on the right hopefully avoid this confusion, showing that such a strong conclusion may not be warranted. In this example, seeing the comparison side-by-side should help clarify the distinction between inferential uncertainty and outcome variability. In practice, however, it is common for figures to depict only one type of uncertainty, a choice which is often not even explicitly stated [23, 24]. Moreover, there are many published examples where authors themselves mistake the two concepts, errantly labeling standard deviations as standard errors or vice versa [25]. This can leave the reader guessing as to what is being communicated—a task that is not helped by the fact that the terms involved sound similar (e.g., "standard error" versus "standard deviation"), or that they are often both depicted by the same visual marks in plots (e.g., error bars).

Recent work has shown evidence of this confusion among laypeople: in a series of large-scale, online experiments, participants overestimated the effectiveness of, and were willing to pay more for, the same hypothetical treatment when shown visualizations that depicted inferential uncertainty compared to outcome variability, even when controlling for other visual factors such as the scale of the y-axis [26, 27]. However, these studies' participants were laypeople (crowd workers), not professors or practitioners trained in statistics. In addition, the studies involved fictitious, low-stakes scenarios. There is good reason to imagine that these effects might disappear with appropriate training or in sufficiently consequential settings, in which case they would be of much less concern. Here we investigate the extent to which visual displays of inferential uncertainty versus outcome variability affect judgments by *experts* in more realistic, high-stakes scenarios. Specifically, we present a series of pre-registered, randomized experiments where experts saw the same scientific findings depicting different types of uncertainty and answered a series of questions about the size and importance of findings they were shown. Our results, comprised of responses from medical professionals, professional data scientists, and tenure-track academic faculty, show that the prevalent form of visualizing only inferential uncertainty can lead to significant overestimates of treatment effects, even among highly trained and knowledgeable experts. In contrast, we find that an alternative format that depicts both inferential uncertainty (by showing statistical estimates) and outcome variability (by also showing individual data points) leads to more accurate perceptions of results while appearing to leave other subjective impressions of the results unchanged, on average. We conclude with a discussion of how this relates to larger issues around practical vs. statistical significance, inference vs. prediction, and scientific communication.

RESULTS

We conducted three pre-registered experiments to investigate how the graphical communication of different types of uncertainty affects experts' perceptions of the size and importance of scientific findings. All three experiments used similar experimental setups but with different types of experts. In each experiment participants were shown the results of a study that compared a treatment group to a control group, where we randomly varied whether the figure in the study depicted inferential uncertainty (via standard errors), outcome variability (via standard deviations or individual data points), or both. After reviewing the study, participants were asked to estimate the effectiveness of the treatment shown in the study and make additional decisions based on the findings they saw. In all three studies we elicited perceived treatment effectiveness by asking for the probability that a randomly selected member of the treatment group had a higher (or lower) score than a randomly selected member of the control group, a number between 50% and 100% known as the common language effect size, probability of superiority, or AUC. We chose this measure because it was developed to aid in the communication of effect sizes and thus provides an easy and effective means of eliciting effect sizes from participants [28, 29]. Pre-registrations for the three experiments are available online.[30] The code to generate stimuli and run our experiment is available online along with the code and data to reproduce our analysis.[31] The "Materials and Methods" section and Section S2 of the Supporting Information contain full details of the experimental design and analyses.

Experiment 1: Medical providers

For our first experiment we recruited medical providers with prescribing privileges employed at a regional healthcare system. All participation was voluntary and no direct payment was made; instead we donated Thanksgiving meals to a local food bank for each completion of the study. We performed eleven 30-minute structured interviews with physicians about the task to ensure that it was realistic, familiar, and easy-to-understand (see SI S3). Those who participated were randomly assigned to see the results of a hypothetical trial for either blood pressure or COVID-19 medications. All participants saw the same information about the corresponding medication type, but some were randomly assigned to see accompanying figures depicting inferential uncertainty first (means and standard errors, in the "Saw SEs first" condition) while others were shown figures depicting outcome variability first (means and standard deviations, in the "Saw SDs first" condition) (Figure S1). Participants were then asked to estimate the probability of superiority for the medication they were shown along with how much it would be worth to patients. They were also asked to recall what the error bars in the figures represented and to provide a histogram of outcomes for patients in the treatment and control groups using a tool called Distribution Builder [32–34]. After completing these tasks, participants were shown "another scenario" for the same type of medication and repeated the entire process. Although not revealed to them, the second scenario was identical to the first except for the accompanying figure—those in the "Saw SEs first" condition were subsequently shown the same results but with SDs in the accompanying figure, whereas those in the "Saw SDs first" condition were shown SEs. The study concluded with a background survey to gauge participants' medical experience and statistical training.

Of the 221 participants who fully completed the study, we removed 58 participants who indicated that they had completed the study more than once, which occurred due to a web browser incompatibility in our experiment code, leaving 163 participants.[35] Most participants were medical doctors with at least some experience with randomized controlled trials (Fig. S9).

Comparing participants' probability of superiority estimates for the first scenario they saw, we find that average estimates were substantially higher for those who saw SEs (depicting inferential uncertainty) first compared to those who saw SDs (depicting outcome variability) first (Fig. 2A). With both medications, participants who saw SEs overestimated the size of the effects they were shown, whereas those who saw SDs underestimated those same effects. For the blood pressure medication, with a true value of 72%, average estimates were 88.5% for SEs vs. 65.9% for SDs ($t(65.09) = 6.83, p < .001$). For the COVID-19 medication, with a true value of 76%, average estimates were 85.9% for SEs vs. 67% for SDs ($t(77.90) = 6.35, p < .001$).

As shown in Fig. 2A, this roughly 20 percentage point difference in average estimates between conditions is largely driven by a sizeable and statistically significant difference in extreme responses, defined as estimates that exceeded 90%, from those who saw SEs compared to those who saw SDs ($t(65.94) = 6.27, p < .001$). For both scenarios, the majority of participants who saw SEs made extreme responses, whereas a small minority of those who saw SDs did so. A within-subjects analysis comparing how much each participant's estimate of the probability of superiority changed between the two scenarios they saw reveals a similar pattern (Fig. S3). Responses to the recall question indicate that these differences are likely due to participants mistaking SE error bars for showing outcome variability instead of inferential uncertainty (Fig. S6), with only 36% of participants who saw SEs first correctly recalling the type and meaning of the error bars they were shown, which is worse than chance. This is consistent with the responses we collected via Distribution Builder, depicted in Fig. 2C, which show that those who saw SEs first generated narrower outcome distributions

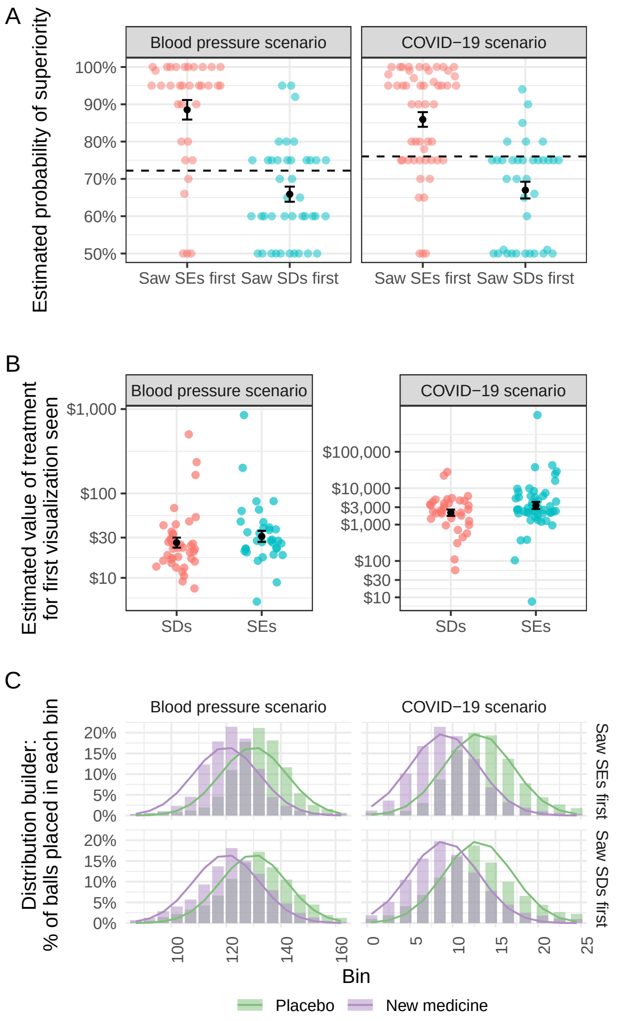


FIG. 2. **Results for medical providers** (A) The estimated probability of superiority of each treatment is depicted for each provider, with black dot and error bars signifying the mean and one SE above and below the mean, as well as dashed lines representing the true underlying effect size. (B) The perceived value of treatment is shown with a logarithmic y axis, and the black dot and error bars depict the mean and one SE above and below the mean. (C) Perceived distributions (bars) versus actual distributions (line) of the effectiveness of treatment.

(and higher implied probabilities of superiority) overall compared to those who saw SDs first ($t(160) = -4.92$, $p < .001$).

Interestingly, however, these large differences in perceived effectiveness were not reflected in estimates of how much participants thought patients would value these treatments, perhaps because there are strong conventions for how much different medications should cost regardless of effectiveness. Participants who first saw SEs were not willing to pay a median price that was significantly different from participants who first saw SDs for either the blood pressure medication ($Z = 1.38, p = 0.084$) or the COVID-19 medication ($Z = 0.848, p = 0.198$). Further analyses, including participant feedback, is available in the Supporting Information, Section S1 A.

Overall the results of this experiment demonstrate that even medical professionals can be misled by common visualizations that depict inferential uncertainty. That said, this experiment has several limitations. First, many of the participants in our studies indicated only moderate training in and comfort with statistics, and so perhaps we would expect different results from experts with more rigorous statistical backgrounds. Second, although the figures that displayed outcome variability directly through SDs curbed extremely high estimates of effect sizes, estimates were on average *below* the true effect size when outcome variability was displayed using SDs. Third, we tested only one true underlying effect size in this study. We designed our second experiment to address these concerns by targeting experts with more statistical training, exploring alternative formats that depict both inferential uncertainty and outcome variability, and testing a wide range of true underlying effect sizes.

Experiment 2: Data scientists

For our second experiment we recruited professional data scientists at a large software company. All participation was voluntary and no direct payment was made; instead we donated one set of personal protective equipment to the United Nations COVID-19 relief effort on behalf of each participant who completed the study. Those who participated saw a one-page extended abstract based on the violent video game study described above. All participants saw the same abstract, but some were randomly assigned to see an accompanying figure depicting only inferential uncertainty (means and standard errors in the "SE only" condition, as in the lower left of Figure 1) whereas others saw both inferential uncertainty and outcome variability (means, standard errors, and individual outcomes in the "SE + points" condition, as in the lower right of Figure 1). We designed the latter to test whether this format, originally proposed by Gardner and Altman [7], would lead to more accurate perceptions of effect sizes than SEs or SDs alone. Then we asked participants for their editorial judgments on the abstract, including the overall appeal of the work, the sufficiency of the sample size used, and whether they would accept the extended abstract if they were a journal editor, all on 5-point Likert scales. Following this, we asked participants to estimate the size of the effect presented in the abstract, measured by the probability that someone who played a violent video game displayed more aggressive behavior than someone who played a non-violent video game (the probability of superiority), and to recall what the error bars in the figure represented. Finally, we had participants repeat probability of superiority estimates for five randomly generated figures of the same type that they saw in the abstract to explore how estimates change with the true underlying effect size.

A total of 175 participants finished Part 1 of the experiment, 161 participants finished Part 2, and 138 participants completed the post-experiment background survey. The majority of participants had upwards of three years of experience working in data science and reasonable prior experience with statistics (Fig. S10, middle). As per our pre-registration, we removed 2 participants who indicated that they had none of the prior experience in statistics or scientific literature that we screened for.

In line with our first experiment, we find that on average participants who saw only inferential uncertainty (in the SE only condition) made substantially higher probability of superiority estimates compared to those who saw both inferential uncertainty and outcome variability (in the SE + points condition) ($t(159.36) = 6.34, p < .001$). Moreover, responses in the SE + points condition were well calibrated to the true effect size of 59% ($\text{mean} = 60.6\%, \text{SD} = 12.6\%$), whereas we once again find overestimation with the conventional SE only format ($\text{mean} = 76.4\%, \text{SD} = 19.9\%$). This more than 15 percentage point difference in average estimates is apparent in extreme responses as well: only 6% of participants in the SE + points condition provided probability of superiority estimates that exceeded 90%, while 35% of participants in the SE only condition did so ($t(142.47) = 5.11, p < .001$).

Despite these rather large differences in perceived effect size, we do not see a corresponding difference in average editorial opinion between conditions (Fig. 3B). Specifically, we did not find evidence that the SE + points format changed the average appeal of the work ($t(168.62) = -1.56, p = .120$), the average perceived sufficiency of sample size ($t(152.85) = -0.88, p = .380$), or the average overall recommendation ($t(171.26) = 0.52, p = .604$) compared to the SE only format. However, in a post hoc analysis we did find a systematic correlation between how large a participant perceived the effect presented in the study to be and their overall editorial recommendation (see Supporting Information Section S1 B).

As with our first experiment, participants showed a reasonable degree of confusion about both the type of error bars they saw and how to interpret them. Only 55% of people in the SE only condition and 51% of people in the SE + Points condition correctly responded that the error bars represented uncertainty in the estimation of the average, rather than variability in outcomes (see Supporting Information Section S1 B).

To check whether any differences between conditions was specific to the study in the extended abstract that we showed participants, or to the true underlying effect size of 59% for that study, we also showed each participant five additional figures with different (randomly generated) true underlying effect sizes ranging from 50–75%. In line with our previous findings, those who saw only SEs systematically overestimated the size of the effects they were shown, whereas those in the SE + points condition were, on average, well calibrated (Fig. 3C). A mixed effects model fit to predict absolute error in responses based on experimental condition and the true underlying effect size (both as fixed effects) and participant identity (as a random effect; see Materials and Methods) confirms this: participants in the SE + points condition made estimates that were on average 11 percentage points (95% CI: $[8.23, 13.7]$) closer to the true probability of superiorities compared to participants in the SE only condition. As with the extended abstract, participants who saw SEs only responded with a bimodal pattern, where a large cluster of extreme responses over 90% raised the overall average. In the SE + points condition, only 3.7% of responses were extreme, while 37% of responses in the SE only condition were extreme ($t(500.26) = 12.54$, $p < .001$). Further detailed analyses and participant feedback are available in the Supporting Information (see Section S1 B).

Experiment 3: Faculty

Our third and final experiment was identical to the previous experiment, but involved academic tenure-track faculty instead of professional data scientists. We recruited US tenure-track faculty from PhD-granting institutions in the fields of psychology, sociology, physics, biology, business, and computer science. Once again, all participation was voluntary and no direct payment was made, we instead donated personal protective equipment to the United Nations COVID-19 relief effort on behalf of each participant who completed the study.

A total of 368 participants completed Part 1 of the experiment, 339 participants completed Part 2, and 289 participants completed the optional background survey. Participants reported being highly experienced with the scientific process, with the modal participant indicating that they had performed over 100 peer reviews (Fig. S11). As per our pre-registration, we removed 63 participants who indicated that they were not currently tenure-track faculty, had no prior coursework in statistics, no experience conducting statistical analyses, or had never peer-reviewed a paper.

In line with our previous findings, participants in the SE only condition made substantially higher probability of superiority estimates ($\text{mean} = 76.4\%, \text{SD} = 19.9\%$) compared to those in the SE + points condition ($\text{mean} = 60.6\%, \text{SD} = 12.6\%$) on average ($t(159.36) = 6.34, p < .001$), and responses in the SE + points condition were well calibrated to the true value of 59% (Fig. 3D). Similarly, while 6% of participants in the SE + points condition provided probability of superiority estimates of 90% or greater, 35% of participants in the SE only condition did so, a statistically significant difference ($t(142.47) = 5.11, p < .001$).

Despite differences in perceived effect size by condition, we do not find a corresponding difference in average editorial opinion (Fig. 3E). Specifically, we did not find evidence that the SE + points format changed the average appeal of the work ($t(168.62) = -1.56, p = .120$), the average perceived sufficiency of sample size ($t(152.85) = -0.88, p = .380$), or the average overall editorial recommendation ($t(171.26) = 0.52, p = .604$). Mirroring the

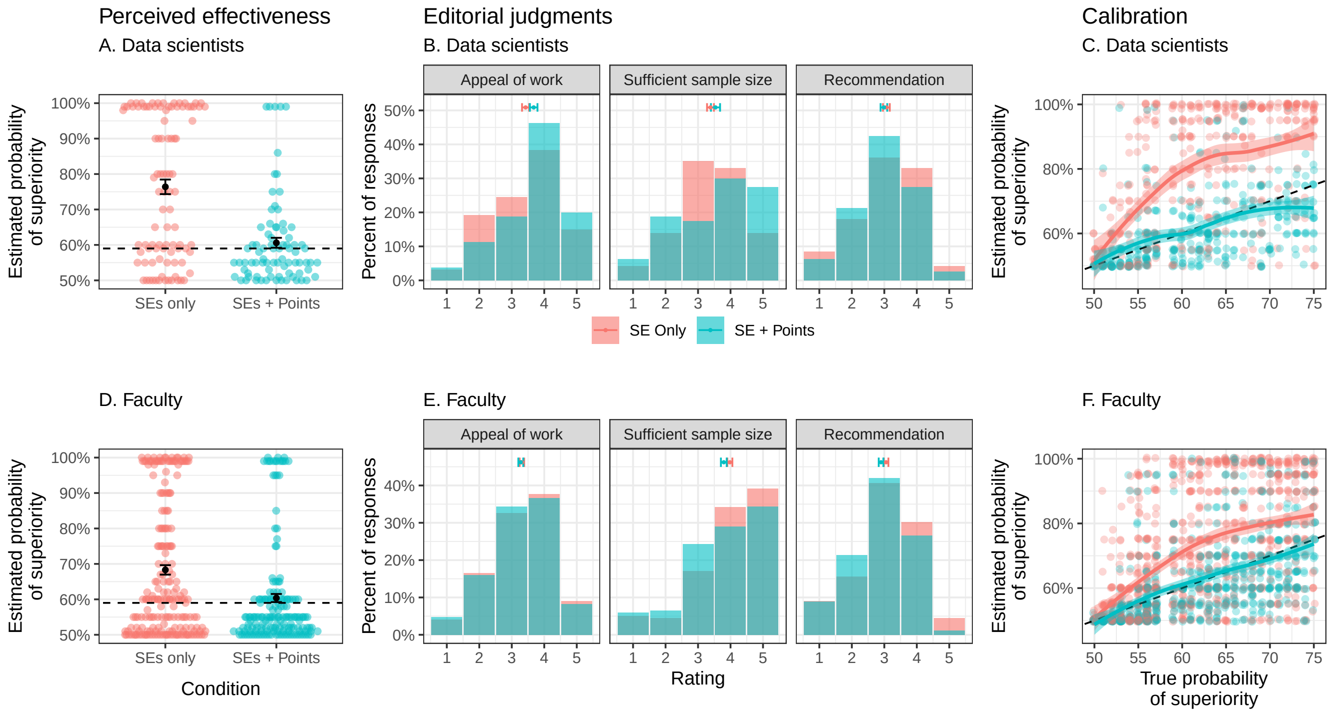


FIG. 3. **Results for data scientists and faculty** (A, D) Perceived probability of superiority of the experiment in the editorial judgment task between the conditions for data scientists and faculty, respectively. The black dot displays the mean, and the error bars are one standard error above and below the mean. The dotted line is the true probability of superiority of the underlying scenario. (B, E) Distributions of the editorial judgments between the two conditions for data scientists and faculty, respectively. The dot and error bars above the plots show the mean and one standard error above and below the mean. (C, F) For each of a series of hypothetical experiments with results generated from a random true probability of superiority, data scientists and faculty (respectively) estimated the true probability of superiority. The dotted line displays the correct answers. The colored line is a loess fit to the data, and the shaded region is a 95% confidence interval.

post hoc analysis from the previous experiment, we did find a systematic correlation between how large a participant perceived the effect presented in the study to be and their overall editorial recommendation (see Supporting Information Section S1 C).

As with our earlier experiments, participants showed a reasonable degree of confusion about the specific meaning of the error bars that they saw (see Supporting Information Section S1 C). In contrast to our previous experiment, however, we saw less confusion about the meaning of error bars for those in the SE + points condition compared to those in the SE only condition: 57.8% of participants in the SE only condition and 71% of participants in the SE + points condition recalling the correct meaning of the error bars ($t(363.70) = -2.67, p = .008$).

The second part of the experiment, which explored a wide range of true underlying effect sizes, showed a similar pattern to the previous experiment: participants in the SE + points condition made estimates that were on average 5.9 (95% CI: $[4.11, 8.61]$) percentage points closer to the true probability of superiorities compared to participants in the SE only condition, using the same mixed effects model as in the previous experiment (Fig. 3F). Extreme estimates drive this average difference: whereas only 5.4% of responses in the SE + points condition were above 90%, 22% of the responses were as extreme ($t(1, 362.43) = 9.68$, $p < .001$). Further detailed analyses and participant feedback are available in the Supporting Information (Section S1 C).

DISCUSSION

Taken together, the results of these three pre-registered experiments highlight a serious concern for the current state of scientific communication. Specifically, the pervasive focus on inferential uncertainty in scientific data visualizations can mislead even experts about the size and importance of scientific findings, leaving them with the impression that effects are larger than they actually are. This "illusion of predictability" is likely due to readers confusing the concepts of inferential uncertainty and outcome variability, and consequently mistaking precise statistical estimates for certain outcomes. Fortunately we have identified a straightforward solution to this problem: when possible, visually display both outcome variability and inferential uncertainty by plotting individual data points alongside statistical estimates.

There are, of course, several limitations to our work and to the accompanying recommendation of plotting individual outcomes. First, with regards to editorial judgments we tested only one extended abstract scenario. It could be the case that for another scenario, editorial opinions actually change along with the visual representation chosen to accompany the text. For example, perhaps there is a more polarizing setting for which people have weaker priors about the effect size and would be swayed more by visualizations of one type over the other. That said, if this were the case we would argue that the representation that results in the most veridical perceived effect size should be chosen, as this would lead reviewers to make the most well-informed decision possible about the merits of the work.[36] Likewise, these effects could be different in a "real stakes" settings (e.g., when actually reviewing for a high-stakes journal or making business decisions about the quality of a data analysis) compared to the hypothetical situation we presented our participants with. Another limitation of the settings we investigate concerns the ground truth effect sizes. While the values in our stimuli are similar in magnitude to those commonly found in medicine, neuroscience, psychology, and social sciences generally [27], we do not make claims of an illusion of predictability at considerably different effect sizes. However, investigating effect sizes that rarely occur in publications would have lower relevance for practice. In addition, there are cases where plotting individual outcomes is not as easy as it sounds. For instance, large datasets or extreme data skew can make it challenging to present all (or even a reasonable fraction of) the data in a way that allows one to see individual observations alongside statistical estimates. There are also complications when studying marginal effects while fixing or averaging over other factors, although techniques such as partial dependency plots could be adapted for these settings [37, 38]. Finally, there is the opportunity to study other visual encodings of uncertainty, including gradient and violin plots [14], hypothetical outcome plots [39], and quantile dot plots [40]. These limitations aside, we still endorse the idea that one should show outcome variability when possible, preferably by plotting individual outcomes alongside statistical estimates.

Our findings provide a clear and important opportunity to improve how statistical visualizations are presented to laypeople and experts alike. Such improvements should increase audience comprehension without sacrificing the details displayed in conventional plots. Having identified this problem and a solution to it, we might ask why it is has gone unnoticed for so long. Our conjecture is that this specific issue, while centered around data visualization, reflects a broader issue around how science is done and how scientific results are communicated.

Specifically, in many fields there has been a longstanding emphasis on inference (e.g., obtaining unbiased estimates of individual effects) over prediction (e.g., forecasting future outcomes), perhaps in part because prediction can be quite difficult, especially when compared to inference. It is surely easier to estimate an average effect across a large population, as is done in standard statistical inference, than it is to predict individual outcomes given all measurable factors that might be relevant to a given problem. But when the results of a study are communicated, they can often come across as having implied the latter when in fact they have only established the former. As a result there can be a great deal of confusion as to what we have actually learned about the world from a particular study, and as we have demonstrated, even experts mistake inferential visualizations as communicating information about prediction. Borrowing from Jacob Cohen's critique of hypothesis testing [41], we believe a similar logic applies to the display of inferential uncertainty:"Among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!"

To this end we believe that the solution of communicating both inferential uncertainty and outcome variability is merited. Rather than emphasizing inference over prediction (or vice versa), we should aim for integrative approaches that consider both aspects of scientific inquiry [42], and present them clearly alongside each other so that readers can themselves make accurate and appropriate inferences from them.

MATERIALS AND METHODS

The Institutional Review Board of Microsoft Corporation reviewed the protocol of these experiments and approved them for human subjects research under approval Ethics Review Portal #10159. Informed consent was obtained from all participants prior to starting any of the studies mentioned below. Full descriptions of the experimental protocol, including screenshots, are available in the Supporting Information (Section S2).

All t-tests are Welch's test for unequal variances unless otherwise noted [43], using the default settings in T.TEST in R. For median tests, we use the two-sample asymptotic Brown-Moody MEDIAN_TEST function from the COIN package in R. Bootstraps are performed with 10,000 resamples using the BOOT.CI function from the BOOT package in R, and the reverse percentile interval method for constructing confidence intervals. To analyze the calibration task for data scientists and faculty we fit the following pre-registered linear mixed effects model using the LME4 package in R:

$$|\text{error}| \sim (1|\text{participant}) + \text{psup} + \text{points} \tag{1}$$

where |error| is the absolute value between the true and guessed probability of superiority, psup refers to the true probability of superiority, and points is a binary indicator variable that is 1 if the participant was in the SE + points condition, 0 otherwise. Probability of superiorities aare expressed as a percentage between 50% and 100%.

To analyze the role of perceived probability of superiority on overall editorial opinion for the data scientists and the faculty, we used the following linear regression model:

$$\text{overall} \sim \text{psup} + \text{condition} + \text{controls} \qquad (2)$$

where overall is the overall editorial judgment of the extended abstract on a 5-point Likert scale, psup refers to the estimated probability of superiority, and controls refers to background variables reported in each experiment.

ACKNOWLEDGMENTS

We would like to thank Alex Blanton and Chris Shaffer, who helped with the experiment involving data scientists, as well as all of the data scientists who participated in that experiment. We would also like to thank the tenure-track faculty who participated in this research. We also wish to thank the medical providers who participated in our first experiment and those who provided feedback on the hypothetical blood pressure RCT results and experimental task we developed. Finally, we would like to thank Aaron Clauset and Nick LaBerge for their help with the studies we conducted. Funding: This work was supported in part by an NSF Graduate Research Fellowship Award DGE 2040434 (SZ). The views expressed here are those of the authors and do not represent the views of the Consumer Financial Protection Bureau or the United States.

[1] F. Fidler and G. Cumming, Teaching confidence intervals: Problems and potential solutions, Proceedings of the 55th International Statistics Institute Session **50** (2005).

[2] C. A. Anderson and K. E. Dill, Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life., Journal of Personality and Social Psychology **78**, 772 (2000).

[3] G. Cumming and S. Finch, Inference by eye: Confidence intervals and how to read pictures of data., American Psychologist **60**, 170 (2005).

[4] G. Cumming, F. Fidler, P. Kalinowski, and J. Lai, The statistical recommendations of the american psychological association publication manual: Effect sizes, confidence intervals, and meta-analysis, Australian Journal of Psychology **64**, 138 (2012).

[5] G. Cumming, *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis* (Routledge, 2013).

[6] `http://news.bbc.co.uk/2/hi/health/720707.stm`.

[7] M. J. Gardner and D. G. Altman, Confidence intervals rather than p values: Estimation rather than hypothesis testing, BMJ **292**, 746 (1986), https://www.bmj.com/content/292/6522/746.full.pdf.

[8] M. Krzywinski and N. Altman, Visualizing samples with box plots, Nature Methods **11**, 119 (2014).

[9] T. L. Weissgerber, N. M. Milic, S. J. Winham, and V. D. Garovic, Beyond bar and line graphs: Time for a new data presentation paradigm, PLOS Biology **13**, e1002128 (2015).

[10] Show the dots in plots, Nature Biomedical Engineering **1**, 0079 (2017).

[11] J. Ho, T. Tumkaya, S. Aryal, H. Choi, and A. Claridge-Chang, Moving beyond p values: Data analysis with estimation graphics, Nature Methods **16**, 565 (2019).

[12] M. Allen, D. Poggiali, K. Whitaker, T. R. Marshall, and R. A. Kievit, Raincloud plots: A multi-platform tool for robust data visualization, Wellcome Open Research **4** (2019).

[13] G. E. Newman and B. J. Scholl, Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias, Psychonomic Bulletin & Review **19**, 601 (2012).

[14] M. Correll and M. Gleicher, Error bars considered harmful: Exploring alternate encodings for mean and error, IEEE Transactions on Visualization and Computer Graphics **20**, 2142 (2014).

[15] A. Kale, M. Kay, and J. Hullman, Visual reasoning strategies for effect size judgments and decisions, IEEE Transactions on Visualization and Computer Graphics **27**, 272 (2020).

[16] A. P. Association, *Publication Manual of the American Psychological Association*, 7th ed. (American Psychological Association, 2019).

[17] JAMA: Instructions for Authors, `https://web.archive.org/web/20220412040639/https://jamanetwork.com/journals/jama/pages/instructions-for-authors` (2022).

[18] New England Journal of Medicine: Statistical Reporting Guidelines, `https://web.archive.org/web/20220405233315/https://www.nejm.org/author-center/new-manuscripts` (2022).

[19] G. Cumming, J. Williams, and F. Fidler, Replication and researchers' understanding of confidence intervals and standard error bars, Understanding Statistics **3**, 299 (2004).

[20] S. Belia, F. Fidler, J. Williams, and G. Cumming, Researchers misunderstand confidence intervals and standard error bars, Psychological Methods **10**, 389 (2005).

[21] R. Hoekstra, R. D. Morey, J. N. Rouder, and E.-J. Wagenmakers, Robust misinterpretation of confidence intervals, Psychonomic Bulletin & Review **21**, 1157 (2014).

[22] E. Soyer and R. M. Hogarth, The illusion of predictability: How regression statistics mislead experts, International Journal of Forecasting **28**, 695 (2012).

[23] G. Cumming, F. Fidler, and D. L. Vaux, Error bars in experimental biology, The Journal of Cell Biology **177**, 7 (2007).

[24] M. Krzywinski and N. Altman, Points of significance: Error bars (2013).

[25] P. Nagele, Misuse of standard error of the mean (SEM) when reporting variability of a sample. a critical evaluation of four anaesthesia journals, British Journal of Anaesthesia **90**, 514 (2003).

[26] J. M. Hofman, D. G. Goldstein, and J. Hullman, How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results, in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020) pp. 1–12.

[27] Y. Kim, J. M. Hofman, and D. G. Goldstein, Putting scientific results in perspective: Improving the communication of standardized effect sizes, in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).

[28] K. O. McGraw and S. Wong, A common language effect size statistic, Psychological Bulletin **111**, 361 (1992).

[29] J. Ruscio, A probability-based measure of effect size: Robustness to base rates and other factors., Psychological Methods **13**, 19 (2008).

[30] https://osf.io/9gxva/?view_only=644d7ab3aaaa48d8bf6b9a521a55b363, https://aspredicted.org/BZ3_37S, and https://aspredicted.org/B3N_9K2.

[31] https://github.com/jhofman/illusion-of-predictability.

[32] W. F. Sharpe, D. G. Goldstein, and P. W. Blythe, The distribution builder: A tool for inferring investor preferences (2000), preprint.

[33] D. G. Goldstein, E. J. Johnson, and W. F. Sharpe, Choosing outcomes versus choosing products: Consumer-focused retirement investment advice, Journal of Consumer Research **35**, 440 (2008).

[34] distBuilder: a Javascript library to facilitate the elicitation of subjective probability distributions, https://doi.org/10.5281/zenodo.166736 (2016).

[35] Results are qualitatively similar if we do not remove these participants.

[36] Ideally reviewers would judge work based on the quality of the questions it asks and the methods used to answer that question, but in the absence of registered reports [**?**] the results themselves likely might impact editorial judgments.

[37] J. H. Friedman, Greedy function approximation: A gradient boosting machine, The Annals of Statistics **29**, 1189 (2001).

[38] W. S. Cleveland, Coplots, nonparametric regression, and conditionally parametric fits, in *Multivariate Analysis and Its Applications* (1994) pp. 21–36.

[39] J. Hullman, P. Resnick, and E. Adar, Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering, PLOS ONE **10**, e0142444 (2015).

[40] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson, When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems, in *Proceedings of the 2016 chi conference on human factors in computing systems* (2016) pp. 5092–5103.

[41] J. Cohen, The earth is round (p ¡ .05), American Psychologist **49**, 997 (1994).

[42] J. M. Hofman, D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire, *et al.*, Integrating explanation and prediction in computational social science, Nature **595**, 181 (2021).

[43] M. Delacre, D. Lakens, and C. Leys, Why psychologists should by default use Welch's t-test instead of Student's t-test, International Review of Social Psychology **30** (2017).

Supplementary Material to An Illusion of Predictability in Scientific Results

Sam Zhang

Department of Applied Mathematics,

University of Colorado, Boulder, CO

Patrick R. Heck

Office of Research, Consumer Financial Protection Bureau, Washington, DC

Michelle N. Meyer and Christopher F. Chabris

Department of Bioethics & Decision Sciences,

Geisinger Health System, Danville, PA

Daniel G. Goldstein and Jake M. Hofman

Microsoft Research, New York, NY

S1. SUPPLEMENTARY RESULTS

A. Experiment 1: Medical providers

Distribution estimates In addition to asking for explicit probability of superiority estimates, we also asked participants to provide a full distribution of treatment and control outcomes in each of the scenarios they saw. Each participant provided 100 data points for each of four distribution builders, generating a total of 65,200 data points (Fig. 2C). Comparing the top figure to the bottom figure within each scenario (column) shows that those who saw SEs first generated narrower distributions overall compared to those who saw SDs first. Using these distributions, we calculated the implied probability of superiority between the treatment and control distributions for each participant. Participants who saw the results of the hypothetical experiment presented using SEs generated distributions with estimated probabilities of superiority that were 9.1 percentage points higher than participants who saw SEs ($t(160) = -4.92$, $p < .001$) (Fig. S8A).

Participant feedback After completing the experiment, participants were given the chance to provide open-ended comments about what they had seen. These responses revealed widespread confusion about standard errors. For instance, one participant who was shown standard errors but mistook them for standard deviations wrote, "'Standard error' is incorrectly applied," and estimated a 95% probability of superiority. Even when participants understood the correct meaning of the error bars, they often made large overestimates: one participant who saw SEs in the blood pressure medication scenario first wrote, "On a bell curve 2 standard deviations from the mean should show 95% of all values whereas standard error with a 95% confidence interval should give you a 95% confidence that the mean lies within that range," which is an accurate understanding of the different meaning of the error bars, yet they still estimated the probability of superiority to be 95%.

Participant feedback also revealed the complex factors that influenced estimates of drug pricing, where one participant wrote, "Providers would decide best treatments and costs would be usually by a pharmacy committee". Other participants mentioned using priors to estimate drug pricing, with one participant sharing, "New drugs seem to cost more than older or even popular therapies." The length of hospitalization factored into participants' decision around pricing the COVID-19 medicine, with one participant writing, "My 'willingness to pay' cost estimate was based on a guess at how many days a patient might reduce their admission." and another writing, "I didn't know the cost per day in the hospital so the dollar estimate may be way off."

B. Experiment 2: Data scientists

Perceived effect size and overall editorial recommendation We modeled each participant's overall editorial recommendation based on their effect size estimate for the study, along with the condition as a fixed effect and their responses to the background questions as random effects. This analysis reveals that, all else equal, an increase of one percentage point in estimated probability of superiority was associated with an average increase of 0.016 points in the overall editorial judgment ($p < 0.001$, using robust standard errors). In other words, the average difference in editorial recommendation between participants who estimated the probability of superiority to be 50 and those who estimated the probability of superiority to be 100 was 0.8 points on a 5-point Likert scale (Fig. S2B).

Recalled meaning of error bars All participants saw error bars depicting 1 standard error. In the SE + points condition only 46% (95% CI: [30%, 63%]) of participants correctly recalled that this was the case, and in the SE only condition 61% (95% CI: [48%, 74%]) did so. In the SE only condition, 65% (95% CI: [53%, 77%]) recalled seeing standard errors (as opposed to standard deviations), and 51% (95% CI: [36%, 67%]) in the SE + points condition recalled seeing standard errors. Regardless of what type of error bars they recalled, participants frequently misinterpreted the meaning of the error bars at a rate indistinguishable from random guessing.

Participant feedback After completing the experiment, participants provided open-ended comments in response to a question asking for feedback on the task. A common reaction among the data scientists was to confuse the meaning of the error bars, thinking that they referred to a measure of outcome variability (such as standard deviations) rather than inferential uncertainty of estimating the average. One participant wrote, "I think the first graph was incorrect. The boxplot lines should have overlapped (mean of 6.8 + s.d. of 0.5 should have spanned 7.3 to 6.3 which included the mean of the other group)." However it wasn't the standard deviations that were being shown, but rather the standard errors, which were an order of magnitude smaller.

In the SE + points condition, a participant revealed to us their thought process for estimating the probability of superiority: "Interesting study to guess probabilities - felt like I should revise [sp] my big book of stats to calculate it exactly, but then just did it visually based on relative population dispersions." The presence of points allowed that participant to simply visually estimate the effect size, avoiding any confusion about standard errors and standard deviations.

C. Experiment 3: Faculty

Perceived effect size and overall editorial recommendation We find a similar correlation between perceived effect size and overall editorial recommendation. All else equal, we find that an increase of one percentage point in estimated probability of superiority is associated with an average increase of 0.013 points in the overall editorial judgment ($p < 0.001$, using robust standard errors). To put this in perspective, this translates to a 0.7 difference on a 5-point Likert scale between participants who perceived the probability of superiority to be 50% compared to 100%.

Recalled meaning of error bars When asked to recall the meaning of the error bars, (Fig. S5). only 55% (95% CI: [45%, 65%]) in the SE + points condition correctly recalled them as representing one standard error above and below the mean (as opposed to two standard errors, or one or two standard deviations), and similarly only 63% (95% CI: [54%, 71%]) in the SE only condition did so. In the SE only condition, 68% (95% CI: [60%, 76%]) recalled seeing standard errors (as opposed to standard deviations), and 64% (95% CI: [55%, 73%]) in the SE + points condition recalled seeing standard errors.

Participant feedback Even tenure-track faculty who teach statistics fell prey to the illusion of predictability, but sometimes they realized this during the feedback stage. One participant reflected, "Clever... I realize in retrospect I was initially not answering right question. In years of teaching stat, I learned what you are learning now ... that even learned colleagues have a difficult time keeping straight the distinction between the distribution of a variable and the distribution of estimates of the mean of that variable." We also saw a similar confusion between standard deviations and standard errors of the estimated means, just like with the data scientists. One participant writes, "I didn't notice the 1 sd (rather than 2 sd) confidence interval info right away," when in fact the confidence interval did not show standard deviations at all, but rather standard errors.

S2. MATERIALS AND METHODS

A. Medical providers

1. *Participants*

The first experiment was conducted from a sampling frame of around 3,100 medical providers with prescribing privileges. All participation was voluntary and no direct payment was made; instead we donated Thanksgiving meals to a local food bank for each completion of the experiment.

2. *Procedures*

We sought to design a scenario that would be easy to understand and appear realistic and familiar to medical experts. After surveying peer-reviewed randomized controlled trials (RCTs) published in medical journals [1], we created a vignette describing and visualizing the results of a double-blind, placebo controlled RCT conducted on a new, hypothetical medication designed to lower blood pressure in high-risk patients. To ensure that our scenario was reasonably realistic, familiar, and easy to understand, we conducted eleven 30-minute structured interviews with physicians including domain experts in blood pressure treatment (e.g., cardiologists, including an editor of a cardiology journal) and other physicians who are highly educated but not necessarily experts in blood pressure treatment (e.g., residents; specialists outside of cardiology). We slightly modified our scenario to reflect some experts' recommendations, but the results of these interviews generally indicated that the blood pressure treatment scenario we constructed was realistic and would be presentable to our target population. We also constructed a COVID-19 medication scenario to appear similar in structure and style to the blood pressure treatment scenario, but for a medication designed to reduce time to recovery for patients hospitalized for COVID-19. We tested our final design on Amazon's Mechanical Turk platform before recruiting medical professionals (see the Supplement for results). Our sampling method, interview materials, respondent characteristics, and the results of these interviews are available in the

Supplementary Materials.

We used both a between-subjects and within-subjects design for this experiment. For the between-subjects portion of the design, participants were randomly assigned to see the results of hypothetical trials for one of two types of medications: blood pressure medication or COVID-19 medication. For each type of medication we constructed two versions of the results: both had identical text, but one showed a figure depicting inferential uncertainty (with error bars encoding standard errors) while the other showed a figure depicting outcome variability (with error bars encoding standard deviations). In the within-subjects part of the design we randomly varied which of these two versions of the results participants saw first (SEs first or SDs first).

In all cases, the hypothetical RCTs involved 300 patients, with 150 patients in the treatment group and 150 in the control group. The text reported the average outcome in each group and described the type of error bars shown in the accompanying figure: either 95% confidence intervals (approximately 2 standard errors above and below the mean) or 95% predictive intervals (2 standard deviations above and below the mean). We explicitly clarified the meaning of the error bars, using bold text to assert for the SD condition that "**the error bars in the figure error bars in the figure show two standard deviations above and below the average in each group**. Predictive intervals like these are constructed such that they should, in the long run, contain outcomes for 95% of similar patients in future studies." Similarly, for the SE condition, we assert, "**The error bars in the figure show two standard errors above and below the average in each group**. Confidence intervals like these are constructed such that 95% percent of them should, in the long run, contain the true average for similar patients in future studies."

After reading these results, participants were asked to estimate the probability that a randomly selected patient in the treatment group had a better outcome than a randomly selected patient in the control group. For the blood pressure scenario this read as follows: "What is your best estimate of the probability that, after the experiment, a randomly selected patient in the treatment group had a lower systolic blood pressure than a randomly selected patient in the control group?" For the COVID-19 scenario participants were instead asked: "What is your best estimate of the probability that a randomly selected patient in the treatment group recovered more quickly than a randomly selected patient in the control group?" In both cases participants were informed that "A 50% probability would indicate no difference in outcomes between the treatment and control groups." Input validation was performed to ensure that the estimate was between 50-100%.

Following this, participants were asked to estimate how much the treatment would worth to patients. In the blood pressure scenario we asked: “Based on its effectiveness and the price of other medications, how much do you think a 30 day supply of this new medication would be worth to patients with high blood pressure? For reference, a 30 day supply of generic ACE inhibitors, also used to lower blood pressure, costs about $20.” For the COVID-19 scenario this was phrase as follows: “Based on its effectiveness and the price of other medications, how much do you think a 5 day supply of this new medication would be worth to patients hospitalized with COVID-19? For reference, a 5 day supply of the most popular drug used for this purpose costs about $2500.”

After making these effect size and value estimates, the details of the RCT and the accompanying figure were hidden from participants and they were asked to recall what kind of error bars were shown in the plot on the previous page: “standard errors, showing uncertainty in estimating the average in each condition” or “standard deviations, showing the variation in individual outcomes in each condition”. Next, on a separate page, each participant was asked to specify a full distribution for what they thought outcomes were for 100 patients in each of the treatment and control conditions of the study using the Distribution Builder tool.

This concluded the first half of the experiment, after which participants were informed that they would see “another scenario” for the same type of medication and repeat the entire process. Although not revealed to them, the second scenario was identical to the first except for the accompanying figure—those who saw SEs first saw the exact same results but with the figure depicting SDs, and vice versa. They repeated all of the steps specified above, from reading about the next RCT to providing Distribution Builder estimates.

With this completed, participants were given an optional background survey, which we used to assess their medical and statistical training. We asked for sources of informal and formal training in research and/or statistics, comfort with understanding the results of RCTs, experience producing or consuming results of RCTs, their role in the medical field, and how long they have worked in medicine. Full details of these questions are provided in the supplement. This concluded the experiment.

B. Data scientists

1. *Participants*

The second experiment was conducted from a sampling frame of around 1,600 data scientists at a large software company. All participation was voluntary and no direct payment was made; instead we donated one set of personal protective equipment to the United Nations COVID-19 relief effort on behalf of each participant who completed the experiment.

2. *Procedures*

Participants were randomly assigned to one of two conditions in a between subjects design: the "SE only" condition where figures depicted inferential uncertainty only (showing standard errors around an estimated mean) or the "SE + points" condition where figures depicted inferential uncertainty and outcome variability (showing standard errors around an estimated mean plus the individual data points from which the mean is estimated). Then participants were asked to complete two tasks: first they were asked to read and review an extended abstract describing a psychology experiment and second they were asked to make a series of effect size estimates for simulated data.

For the first task we created an extended abstract based on a highly-cited psychology study testing the impacts of playing violent video games on aggressive behaviors in the laboratory [2]. In the study, undergraduates from a Midwestern introduction to psychology course were assigned to play either non-violent video games (control), or violent video games (treatment). Then in a subsequent task, their aggressiveness was measured as a continuous score. We created an extended abstract using the framing and results from the actual paper, including the means, standard deviation, and effect size of their experiment. We increased the sample size within their study from 210 participants to 800 participants, to reflect changing standards within experimental psychology around statistical power and sample sizes.

In our experiment all participants saw the same text for this extended abstract, with the accompanying figure and caption varying by experimental condition. Specifically, for participants in the SE only condition, we displayed the outcomes using only the means and one standard error above and one standard error below the means. For the SE + points condition, we also showed the individual outcomes (Fig. S12). However, the real outcomes were not available, so we generated synthetic data to match the summary statistics from the paper. The paper did not provide standard deviations for the data, but it did provide means and Cohen's d, which is sufficient for recovering the pooled standard deviation. The violent video game condition had a mean of 6.81 and the control condition had a mean of 6.65, and Cohen's d was 0.31, yielding a pooled standard deviation of 0.52. To avoid making the two distributions identical, we rounded the standard deviations for the treatment and control to the nearest hundredth place, subtracted 0.01 for the treatment standard deviation, and added 0.01 for the control standard deviation. This yielded standard deviations of 0.51 for treatment and 0.53 for control. We generated standard normal data for each condition, and recentered and scaled to match the means and standard deviations above.

We collected two sets of dependent variables for participants' reviews of the extended abstract. The first were editorial judgments where participants rated the "Appeal of this work to a broad interdisciplinary audience" (from 1=Not at all appealing to 5=Very appealing), "How sufficient is the experiment's sample size?" (from 1=Completely insufficient to 5=Completely sufficient), and "Overall recommendation of abstract for publication" (reverse coded, from 1=Strong accept to 5=Strong reject, where 2=Accept, 3=Possible reject, and 4=Reject).

The second type of dependent variable was participants' estimates of the probability of superiority for the effect described in the abstract. Specifically, we asked: "What is the probability that a randomly selected member of the treatment group (someone who played a violent video game) had a higher aggressiveness score than a randomly selected member of a control group (someone who played a non-violent video game)? (A 50% probability indicates no difference in outcomes between the treatment and control groups, on average.)" Input validation was performed to ensure that the estimate was between 50-100%.

Following this we tested participants' understanding of what they had seen by hiding the figure and abstract and asking whether the error bars on the figure on the previous page showed 1 standard error, 2 standard errors, 1 standard deviation, or 2 standard deviations above and below the mean. We also asked participants whether the error bars conceptually captured "uncertainty in estimating the average in each condition" or "variation in individual outcomes in each condition".

After reviewing the extended abstract, we presented participants with the second phase of the experiment: a series of five hypothetical outcomes from studies with a treatment and a control group. We generated the means of the two outcomes as uniform random variables between 0 and 3, picking the larger of the two as the treatment. Then we generated the variance of each outcome as a uniform random variable between 0.5 and 2. Lastly, we picked sample sizes for the two experiments by drawing from a binomial distribution with true parameter $p = 1/2$ and n uniformly random between 100 and 400. We computed the probability of superiority and performed rejection sampling until we found a scenario with the true probability of superiority between 50% and 75%. These parameters produced effect sizes roughly uniformly in that range. We presented the sample sizes, the sample means, and the standard errors on these estimates for each experiment [3], and manipulated whether participants saw visualizations depicting only inferential uncertainty or inferential uncertainty and outcome variability depending on their experimental condition. Then we asked participants for their evaluation of the probability of superiority of the treatment over the control. For participants who saw outcome variability (points) in the visualization, we drew the points from the distributions above, but to ensure that the points aligned with the true probability of superiority, we centered and scaled the points so that their mean and variance matched the mean and variance of the hidden parameters (Fig. S13).

To conclude the experiment we presented participants with several questions about their scientific and statistical background, which we used in our preregistered exclusion criteria to filter out participants without some experience in science and statistics. In particular, we excluded any participants whose responses indicate that a) they have had no formal or informal training in research, research methods, or statistics; b) they are “not at all” comfortable understanding the results of randomized experiments; or c) that they have never done any of a set of 6 activities related to expertise in this area (reading scientific results, publishing a scientific paper, working on a randomized experiment, took a course in statistics or a related field, analyzed data outside of a course requirement, or used statistical software). As per our preregistration, if a participant did not fully complete a part of the experiment (Part 1 or Part 2), we eliminated their incomplete responses from the relevant analysis.

C. Faculty

1. Participants

The third experiment was conducted using a stratified sample of 9,000 tenure-track psychologists, sociologists, physicists, biologists, business faculty, and computer scientists from PhD-granting institutions in the US. All participation was voluntary and no direct payment was made; instead we donated one set of personal protective equipment to the United Nations COVID-19 relief effort on behalf of each participant who completed the experiment.

2. Procedures

This experiment different from the previous one only in the participant population and the relevant background/experience questions. We still randomly assigned participants to one of two display conditions: inferential uncertainty ("SE only") or inferential uncertainty with outcome variability ("SE + points"). Participants completed the same editorial judgment task and statistical estimation task.

Background questions. After the statistical estimation task, we presented participants with background questions. This time, we preregistered different questions and exclusion criteria. In particular, we excluded any participants whose responses indicate that a) they are not tenure-track faculty at a PhD-granting institution, b) they are ”not at all” comfortable understanding inferential statistics and hypothesis tests, c) they are ”not at all” comfortable analyzing data and using statistical software, or d) that they have never reviewed any papers during their academic career. As in Experiment 2, if a participant did not complete a part of the experiment (Part 1 or Part 2), we eliminated their incomplete responses from the relevant analysis.

S3. PRELIMINARY STUDY OF PHYSICIANS USING IN-DEPTH-INTERVIEWS

A. Recruitment and Participants

During May and June 2020, we conducted ten individual, 30-minute in-depth-interviews (IDIs) with physicians who belonged to the target population, and one additional interview with a physician outside of the target population. We recruited participants using the snowball method, where at the end of each interview, an interviewee was asked to recommend other physicians who have relevant expertise or who might otherwise be interested in participating in an interview. These recommended physicians were then invited to participate via personalized emails. All interviews were conducted using video calling over the internet by the same interviewer (PRH). Participation was voluntary and respondents were not paid. All physicians who were interviewed were excluded from the list of eligible participants for the full experiment.

Our final sample of eleven participants comprised practicing physicians across a variety of specialties, career stages, and levels of involvement in research. Specialties included cardiology (3), developmental pediatrics (2), psychiatry (2), internal medicine (hospitalist) (1), pediatric gastroenterology (1), geriatric medicine and neurology (1), and rheumatology (1). Two respondents were in their residency training at the time the interviews were conducted and the remaining nine were distributed across career stages. All respondents had some experience with research and three respondents reported substantial current involvement in research, including one editor of a cardiology journal and another physician with over 100 peer-reviewed publications in cardiology and related fields.

B. Procedure

At the start of each interview, each participant was given a brief overview of the research topic and the purpose of the interview. After introducing himself, the interviewer used the following script to orient the participant: “We are interested in studying how healthcare and medical professionals judge the effectiveness of treatments after viewing some example visualizations of research conducted on those treatments. Our aim is to study treatments that are modern, realistic, and familiar to healthcare professionals.”

After asking the interviewee for permission to share materials via screenshare, the interviewer displayed a single full-screen image of the primary task, including a description of the results of a hypothetical RCT, a visualization displaying the results of this RCT using error bars that represented either confidence intervals or standard deviations, and the primary dependent measure that requested participants to provide an estimate for the probability of superiority. The participant was given unlimited time to verbally respond to this measure. After recording the participant's response, the interviewer then presented the task again, noting that the visualization had changed—it now displayed whichever set of error bars was not displayed before. Participants then provided a probability of superiority estimate for this new scenario.

After completing the task, the participant was asked to respond freely to the prompt, "Did anything in the graph, the description, or the question stand out as confusing or unclear?" The interviewer took notes on responses to this question, occasionally asking the participant to give more details to relevant observations. Participants were allowed as much time as they needed to react and respond to the scenario and task.

For the remainder of the interview, the interviewer selected targeted questions from a pre-constructed list (see instrument in Appendix). These primarily comprised questions about the hypothetical blood pressure medication RCT scenario that we had designed. These questions were designed to solicit participants' advice regarding the setup, language, experimental approach, figure, and numbers we had used to develop the scenario. In some cases, participants were asked whether specific alternative scenarios (e.g., a design using a comparative effectiveness trial; an outcome measure using an average change score from pre- to post-treatment) would be more realistic, familiar, or easier to comprehend than the scenario we had designed (using, e.g., a double-blind, placebo-controlled trial, displaying post-treatment systolic blood pressures as outcomes). Other targeted questions asked participants to comment on our dependent measures, including probability of superiority and willingness to pay for or prescribe the experimental medication.

When the session was nearly over, the interviewer asked the participant if they had any questions about the research or task, requested recommendations for other physicians who might be interested in being interviewed, offered to send the participant a copy of the published report, and thanked them for their time. After the participant had left the session, the interviewer summarized and synthesized their responses into a tracking spreadsheet.

C. Results

We present the results of these interviews organized according to what we sought to learn.

First, we sought to expose a small sample of our target population to the experimental task to observe whether interviewees' responses were consistent with our hypothesized pattern. To this end, we asked each interviewee to provide a probability of superiority estimate immediately after viewing each scenario. Nearly all participants (10/11) reported that the probability of superiority was lower when viewing the figure displaying confidence intervals than when viewing the figure displaying error bars representing standard deviations. We note that some respondents preferred not to give specific numeric estimates, instead preferring to use descriptive terms (e.g., "high;" "very high;" "almost 100%;" "lower than the last one," etc.). This prevented us from computing descriptive statistics, but it was nevertheless striking that most of our sample made estimates that were consistent with the hypothesized pattern of results. The participant who did not show this pattern said that she would assign the same probability of superiority estimate in both conditions.

Second, we sought expert feedback on the level of familiarity, believability, and accuracy of the scenario we had constructed: the results of an RCT comparing an experimental blood pressure medication against a placebo control. Responses to the open-ended question immediately following the experimental task, and to more targeted questions about our design choices, indicated that participants generally found our scenario to be realistic, believable, and familiar. All eleven participants recommended that the hypothetical RCT we present should compare a treatment to a placebo control (as opposed to another drug or type of treatment). Of the eight participants who were asked the targeted question about which outcome measure to present, six agreed that absolute post-treatment blood pressure was the most appropriate outcome measure to display (one preferred lines indicating pre-to-post-treatment change; one simply stated that multi-time point treatment displays would be interesting for future study). Most participants preferred that we use generic terms like "treatment" and "control," rather than naming a specific drug in our hypothetical RCT; one participant preferred a specific drug name, and one did not express a preference.

Third, we solicited reactions and points of confusion in response to our dependent measures. Specifically, we asked participants to comment on how we could make the probability of superiority measure clearer, and how we might phrase a measure of willingness to pay for (or prescribe) the experimental treatment. Most participants (8/11) recommended that we do nothing to change the probability of superiority measure—that it is unlikely we would be able to improve it. Some of these participants (2) noted that it was difficult to reason in this way but nevertheless concluded that it was acceptable as written. One participant reported not liking the measure but did not recommend any alternatives. Several participants (4) explicitly mentioned that the descriptive sentence clarifying that "a 50% probability would indicate no difference in outcomes between the treatment and control group" was helpful. Participants' responses to a willingness to pay for (or prescribe) measure were mixed. Participants did not complete such a measure as part of the interview; the interviewer described what this measure might look like and asked participants to comment on whether it would be a useful measure to collect. There was little consensus about the usefulness or feasibility of this measure; proponents described a measure of this kind as valuable and interesting, while detractors commented that there are too many external factors to consider when thinking about paying for or prescribing a medication.

D. Appendix: Interview Protocol and Questions for Physicians

Give brief overview: we are interested in studying how healthcare and medical professionals judge the effectiveness of treatments after viewing some example visualizations of research conducted on those treatments. Our aim is to study treatments that are modern, realistic, and familiar to healthcare professionals.

[Ask interviewee to complete the task using an example visualization.]

SD version probability of superiority answer:

SEM version probability of superiority answer:

Free-response question: Did anything in the graph, the description, or the question stand out as confusing or unclear?

Targeted questions about setup:

- When thinking about the effectiveness of blood pressure treatments, is it easier to think about how effective a treatment is relative to:

  1. A nondescript control (i.e., the standard of care);
  2. Another drug or behavioral treatment regimen;

3. Nothing (i.e., to what would happen if no drug or behavioral therapy was introduced)

- When thinking about what improvement means after undergoing a treatment for high blood pressure, is it easier to conceptualize clinical improvement as:

  1. An average change score. (For example, those who received the treatment lowered their blood pressure by an average of 50 points. Those who did not receive the treatment lowered their blood pressure by an average of 10 points.)
  2. An average pre- and post-measure. (o For example, those who received the treatment began with an average measure of 160 and ended with an average measure of 110 points. Those who received the treatment began with an average measure of 160 and ended with an average measure of 150 points.)

- As part of this research, we are interested in learning how effective a new treatment would have to be for a provider to seriously consider prescribing it instead of what they typically prescribe. In your opinion, how likely are most healthcare providers to begin prescribing a new drug or treatment (relative to what they are used to prescribing)? In other words, how much improvement, and how much certainty about that level of improvement, would be required for most providers to implement new prescribing going forward? You can assume that the new drug is more effective than the old one, and all external considerations (like adverse side effects, cost, etc.) are roughly equivalent between the new and old treatment.

- Consider your understanding of blood pressure treatments. Can you list a few common drug or treatment names that other healthcare providers will likely be familiar with? Can you think of any good baseline or experimental drugs or treatments we could use or adapt?

Questions about dependent measures:

- Did the Probability of Superiority item make sense? Is there any way we can make this clearer?

- Are there any other items, or language that you are used to seeing, that can help us measure people's perception of how effective one treatment is over another?

- Willingness to Pay: do you have any ideas for a way to make this kind of measure feasible?
- Willingness to Prescribe: do you have any ideas for a way to make this kind of measure feasible?

S4. STIMULI FOR MEDICAL PROVIDER EXPERIMENT

See Fig S1 for the different figures used in the experiment. Participants were randomly assigned to one of the scenarios (rows) and then randomly assigned to see either the SE or the SD version of the figure first, followed by the other figure next. See Fig S15 for a screenshot of the entire page presenting the COVID-19 medication scenario, and Fig S14 for a screenshot of the page presenting the blood pressure medication scenario.

S5. BACKGROUND QUESTIONS FOR MEDICAL PROVIDERS

A. Risk of re-identification

We removed the background questions from the publicly released data to avoid the risk of re-identification of medical providers. The background questions were unnecessary for any of the main analyses, and only used to perform analyses of the aggregate analyses.

B. Background questions

1. What sources of formal or informal training or education have you had in research, research methods, or statistics? Select all that apply.
    - Undergraduate coursework
    - Master's program in a data science related field
    - PhD program in a data science related field
    - Postgraduate coursework (e.g., during internship, fellowship, etc...)
    - Continuing education courses
    - Self-instruction via textbooks or peer-reviewed literature

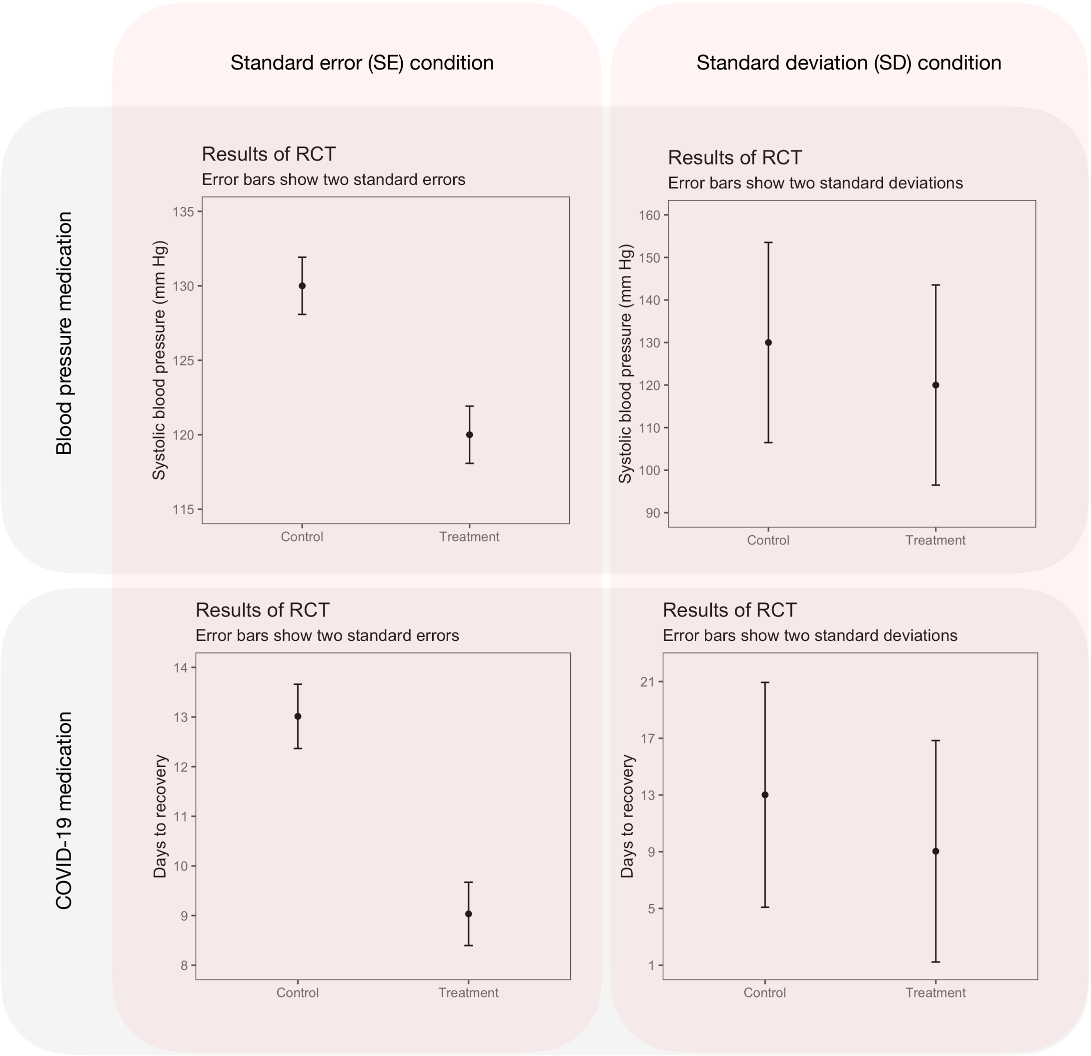


FIG. S1. **Experimental stimuli for medical providers** The statistical graphic showing (left column) standard errors or (right column) standard deviations in either (top row) the blood pressure medication scenario or (bottom row) the COVID-19 medication scenario for the medical provider participants.

- Other

2. How comfortable are you with understanding the results of randomized controlled trials, including reading graphs and plots?

- Not at all
- Somewhat

   - Moderately
   - Very
   - Extremely

3. Have you ever done any of the following? Select all that apply.
   - Read the results of a randomized controlled trial in a peer-reviewed journal article
   - Changed what you typically prescribed or recommend after personally reading the results of a randomized controlled trial in a peer-reviewed journal article
   - Published a scientific paper in a peer-reviewed journal
   - Conducted or worked on a team conducting a randomized controlled trial
   - Took a course or class in statistics, biostatistics, or research methods (online, in-person, CME, etc.)
   - Analyzed data for statistical significance outside of a course requirement
   - Used SPSS, R, Python, Stata, SAS, or any other statistical software
   - None of the above

4. Are you currently involved in research?
   - Yes
   - No

5. Please select the option that best describes you below.
   - Doctor (MD or DO)
   - Physician Assistant
   - Nurse Practitioner
   - Non-prescribing clinician or staff without clinical credential
   - Other

6. How long have you been working in the medical field?
   - Less than 1 year

- 1–2 years
- 3–5 years
- 6–10 years
- 11–20 years
- 21–30 years
- 30+ years

S6. BACKGROUND QUESTIONS FOR DATA SCIENTISTS

1. What sources of formal or informal training or education have you had in research, research methods, or statistics? Select all that apply.
    - Undergraduate coursework
    - Master's program in a data science related field
    - PhD program in a data science related field
    - Postgraduate coursework (e.g., during internship, fellowship, etc...)
    - Continuing education courses
    - Self-instruction via textbooks or peer-reviewed literature
    - Other
2. How comfortable are you with understanding the results of randomized controlled trials, including reading graphs and plots?
    - Not at all
    - Somewhat
    - Moderately
    - Very
    - Extremely
3. Have you ever done any of the following? Select all that apply.
    - Read the results of a randomized controlled trial in a peer-reviewed journal article

- Changed what you typically prescribed or recommend after personally reading the results of a randomized controlled trial in a peer-reviewed journal article
- Published a scientific paper in a peer-reviewed journal
- Conducted or worked on a team conducting a randomized controlled trial
- Took a course or class in statistics, biostatistics, or research methods (online, in-person, CME, etc.)
- Analyzed data for statistical significance outside of a course requirement
- Used SPSS, R, Python, Stata, SAS, or any other statistical software
- None of the above

4. Are you currently involved in research?
   - Yes
   - No

5. How long have you been working in the data science field?
   - Less than 1 year
   - 1–2 years
   - 3–5 years
   - 6–10 years
   - 11–20 years
   - 20+ years

S7. BACKGROUND QUESTIONS FOR FACULTY

1. Are you currently a tenure-track faculty at a PhD granting institution?
   - Yes
   - No

2. What best describes the department you're primarily affiliated with? (Choose "Other" if none are at all appropriate)

- Business
- Computer science
- Physical sciences
- Life sciences
- Social sciences
- Mathematics or Statistics
- Other

3. How comfortable are you with understanding inferential statistics and hypothesis tests?
    - Not at all
    - Somewhat
    - Moderately
    - Very
    - Extremely

4. How comfortable are you with analyzing data and using statistical software?
    - Not at all
    - Somewhat
    - Moderately
    - Very
    - Extremely

5. Have you ever taught statistics or a related course at the undergraduate or graduate level?
    - Yes
    - No

6. How many papers have you reviewed during your academic career?
    - 0

- 1–10
- 11–20
- 21–50
- 51–100
- More than 100

S8. REFERENCES AND FOOTNOTES

[1] S. R. Group, “A randomized trial of intensive versus standard blood-pressure control,” *New England Journal of Medicine*, vol. 373, no. 22, pp. 2103–2116, 2015.

[2] C. A. Anderson and K. E. Dill, “Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life.,” *Journal of Personality and Social Psychology*, vol. 78, no. 4, p. 772, 2000.

[3] This alone constitutes enough information to compute the implied standard deviations and hence effect sizes for each stimulus participants were shown, although doing so might be relatively difficult.

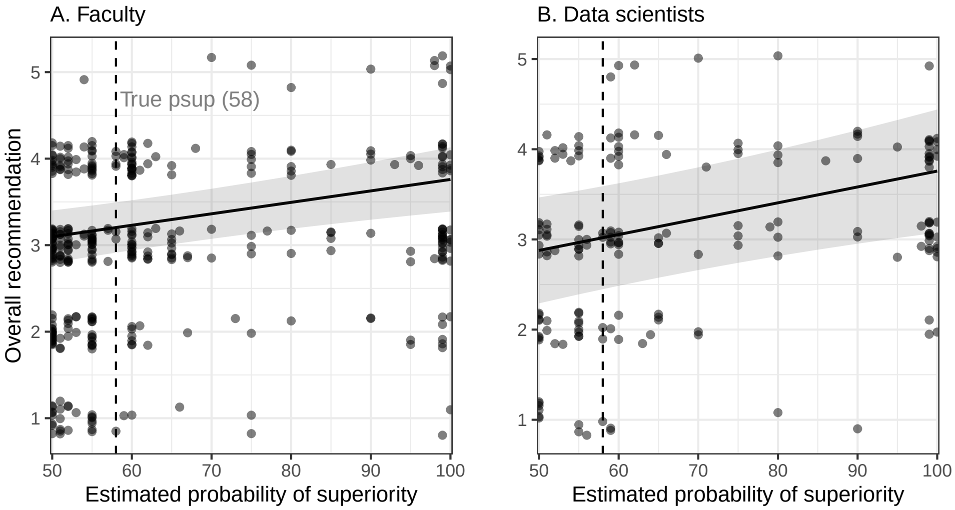


FIG. S2. **Probability of superiority vs. overall editorial judgment** Overall editorial judgment of (A) faculty and (B) data science participants as a function of their estimated probability of superiority. Regression line shows estimated marginal means after controlling for background variables, with shaded interval representing the 95% confidence interval of the estimated marginal mean.

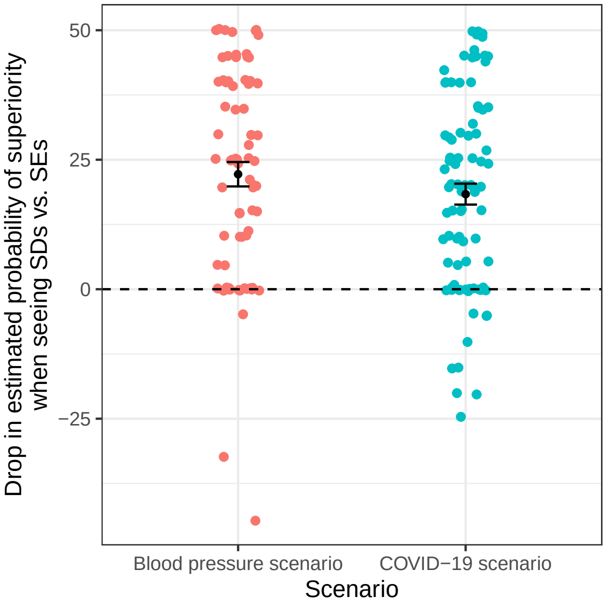


FIG. S3. **Within-subjects results for medical providers** Difference in individual medical providers' estimated probability of superiority when seeing SEs vs. SDs. Black dot is the mean, with error bars depicting one standard error above and one standard error below the mean.

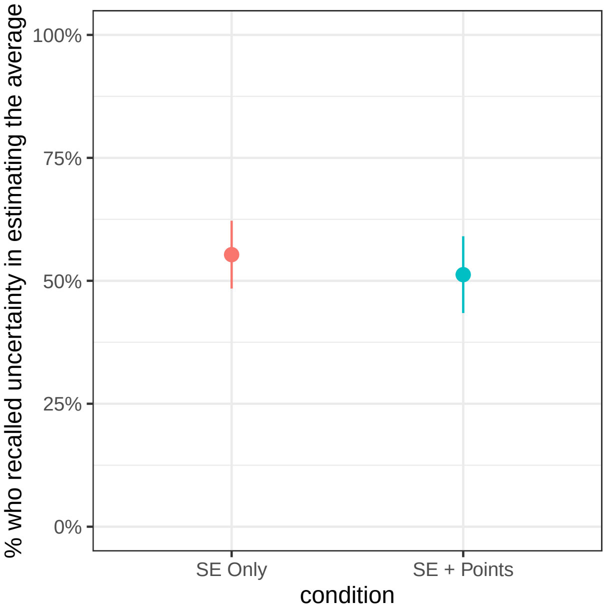


FIG. S4. **Recall of data scientist participants** Displayed is the sample mean and one standard error above and one standard error below the mean of the binary outcome of whether the data scientist participant correctly recalled that the error bars in the hypothetical study depicted uncertainty in estimating the average. In particular, we asked “What did the error bars on the previous page capture conceptually?” with possible answers “Uncertainty in estimating the average in each condition” and “Variation in individual outcomes in each condition”.

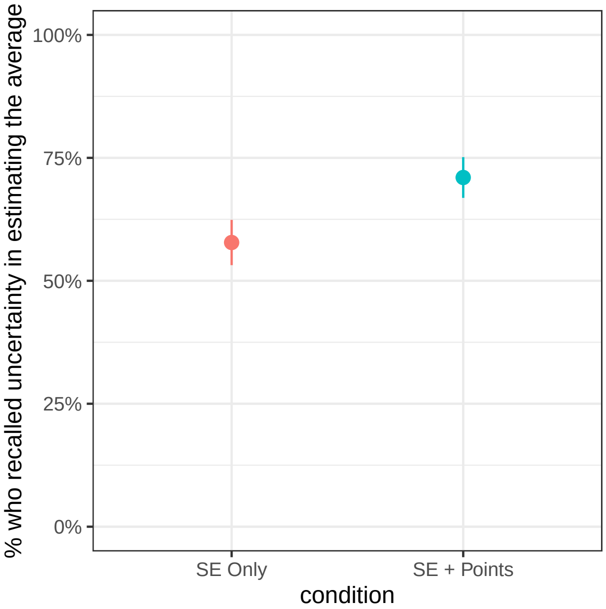


FIG. S5. **Recall of faculty participants** For both conditions, we depict the proportion of faculty participants who correctly recalled that the error bars in the hypothetical study depicted uncertainty in estimating the average (solid dot) as well as one standard error above and one standard error below the mean. In particular, we asked “What did the error bars on the previous page capture conceptually?” with possible answers “Uncertainty in estimating the average in each condition” and “Variation in individual outcomes in each condition”.

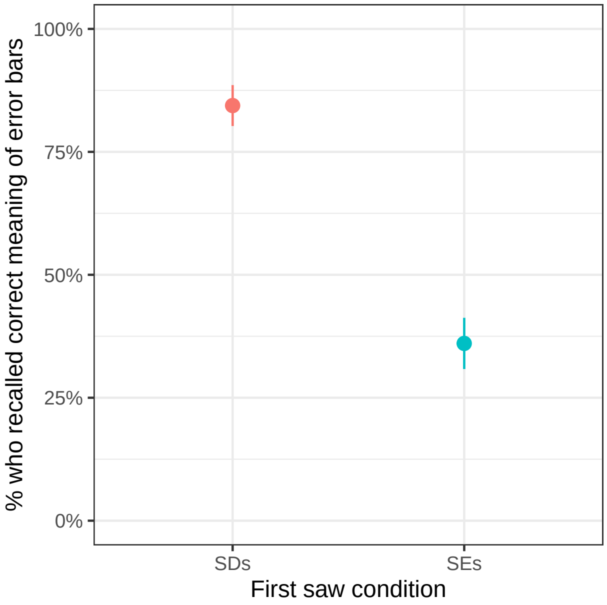


FIG. S6. **Recall of medical provider participants** For the first scenario, we depict the proportion of medical provider participants who recalled the correct meaning of the error bars in the hypothetical study they had seen by condition. We show the average (solid dot) as well as one standard error above and one standard error below the mean. In particular, we asked “What kind of error bars were shown in the plot on the previous page?” with possible answers “Standard errors, showing uncertainty in estimating the average in each condition” and “Standard deviations, showing the variation in individual outcomes in each condition”.

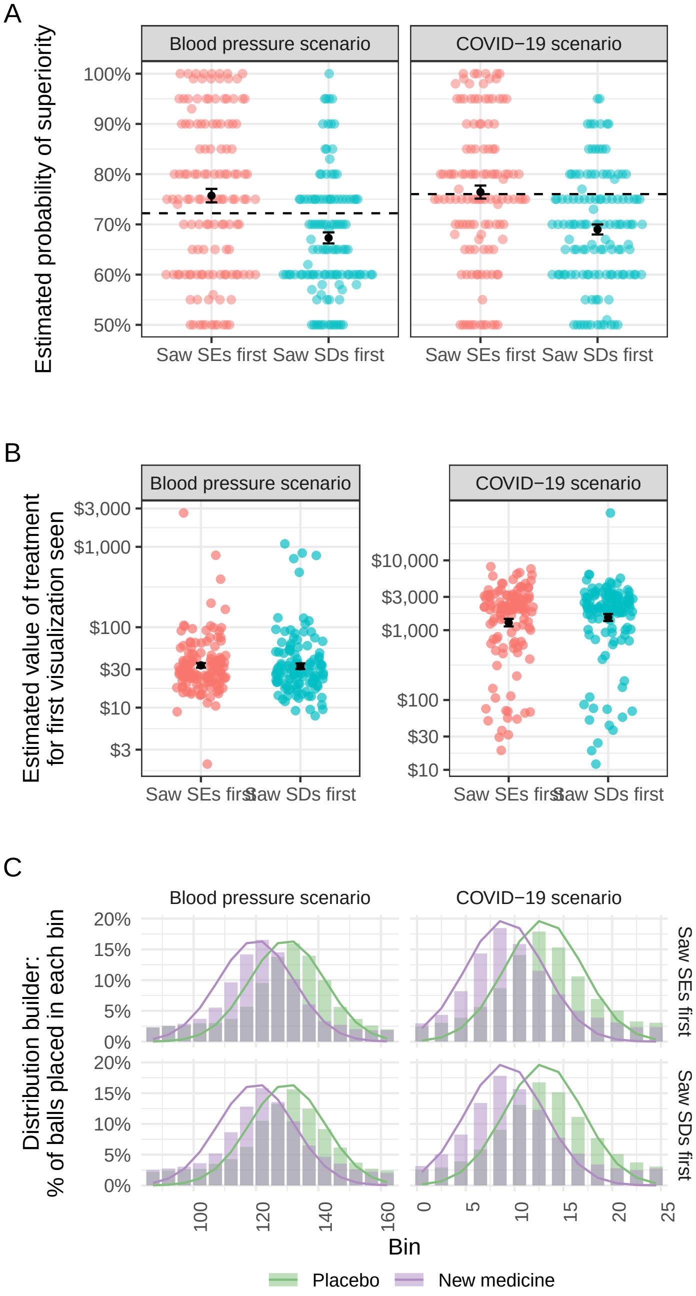


FIG. S7. **Results for mechanical turkers** (A) The estimated probability of superiority of each treatment is depicted for each turker, with black dot and error bars signifying the mean and one standard error of the mean, above and below the mean, as well as dashed lines representing the true underlying effect size. (B) The perceived value of treatment is shown with a logarithmic y axis, and the black dot and error bars depict the median with 95% bootstrapped confidence intervals. (C) Perceived distributions (bars) versus actual distributions (line) of the effectiveness of treatment.

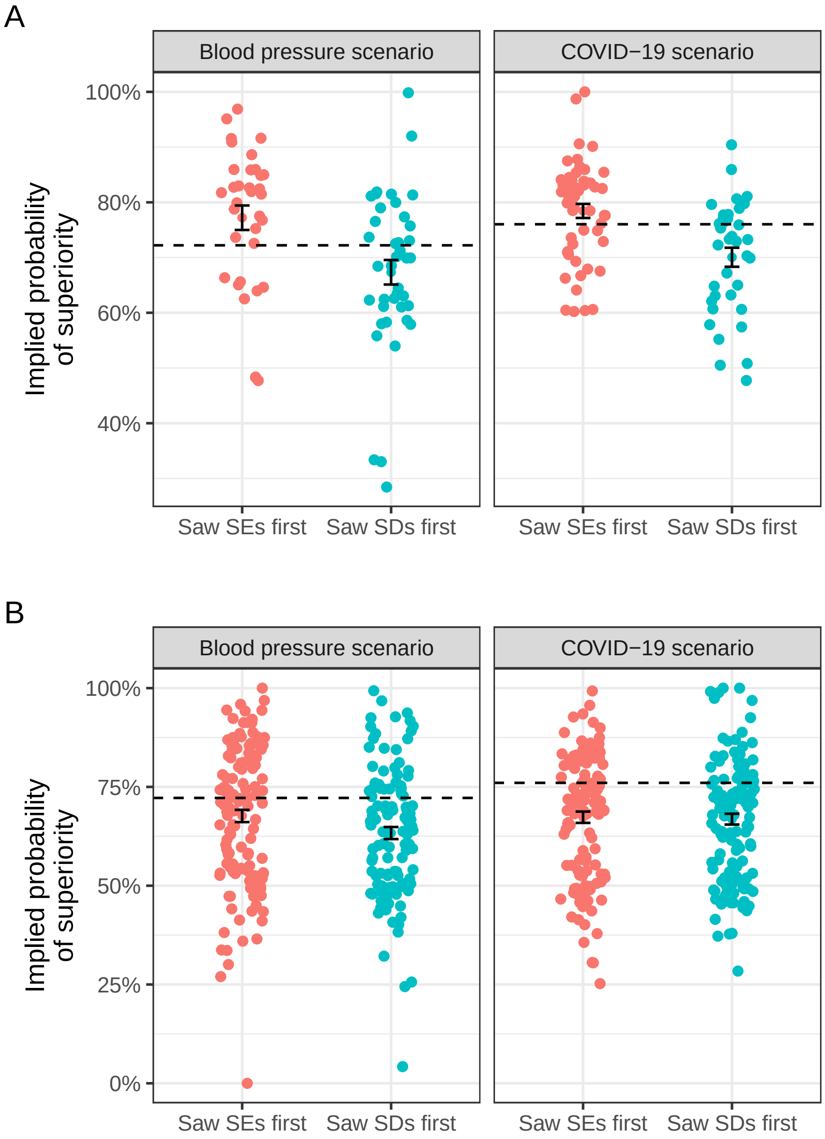


FIG. S8. **Implied probability of superiority from distribution builders** Implied probability of superiority for (A) medical provider and (B) mechanical turk participants. For each participant, we calculate the implied probability of superiority between the treatment and control conditions for the first hypothetical medication they see. The average probability of superiority is the black dot, and the error bar shows one standard error of the mean above and below the mean for each condition.

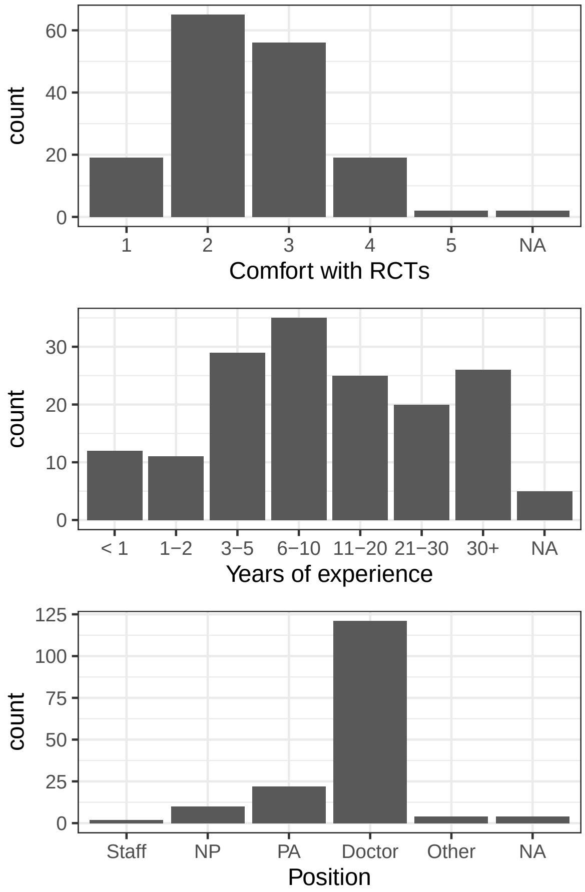


FIG. S9. **Background of the medical practitioner participants** Responses to the background questions at the end of the study for the medical practitioners. NP=Nurse Practitioner, PA=Physician Assistant.

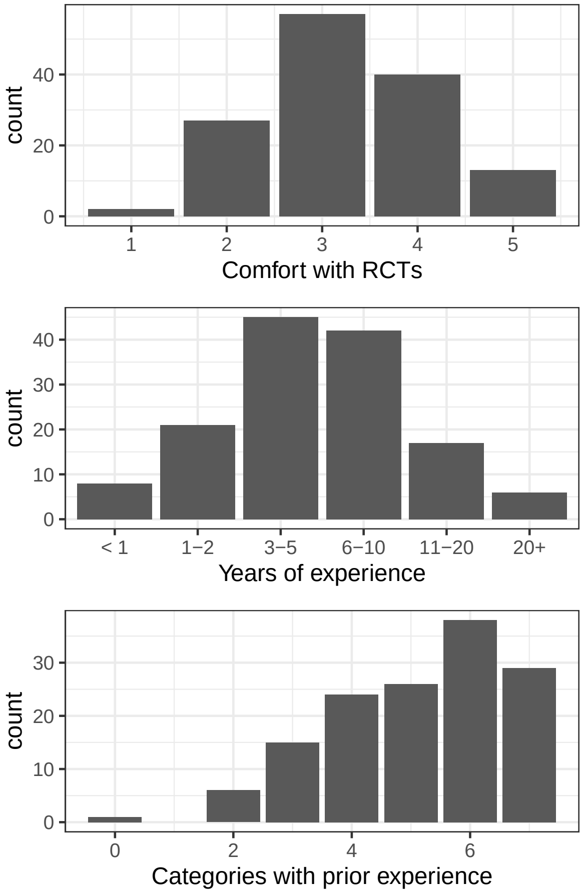


FIG. S10. **Background of the data scientist participants**. We asked participants for their comfort between 1 (Not at all comfortable) to 5 (Extremely comfortable) with "understanding the results of randomized experiments, including reading graphs and plots" (top). We also asked "How long have you been working in the data science field?" (middle). We asked participants "Have you ever done any of the following?" ("Read the results of a randomized experiment in a peer-reviewed journal", "Published a scientific paper in a peer-reviewed journal", "Conducted or worked on a team conducting a randomized controlled trial such as an A/B test", "Took a course or class in statistics, data science, or research methods (online, in-person, etc)", "Analyzed data for statistical significance outside of a course requirement", or "Used SPSS, R, Stata, SAS, or any other statistical software.") and we display the number of selected checkboxes (bottom).

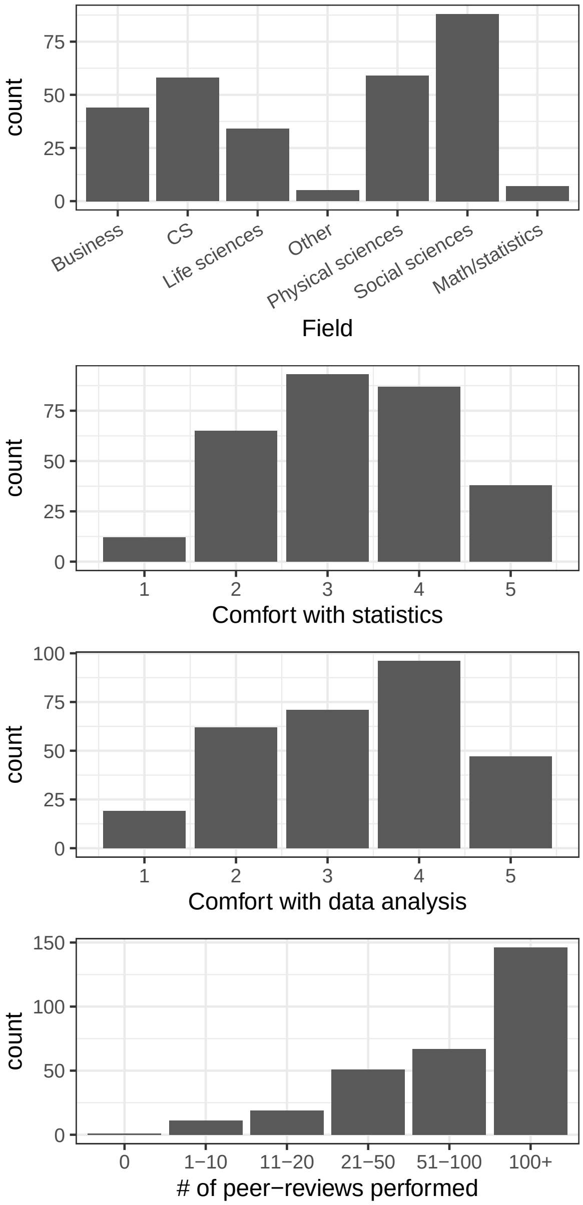
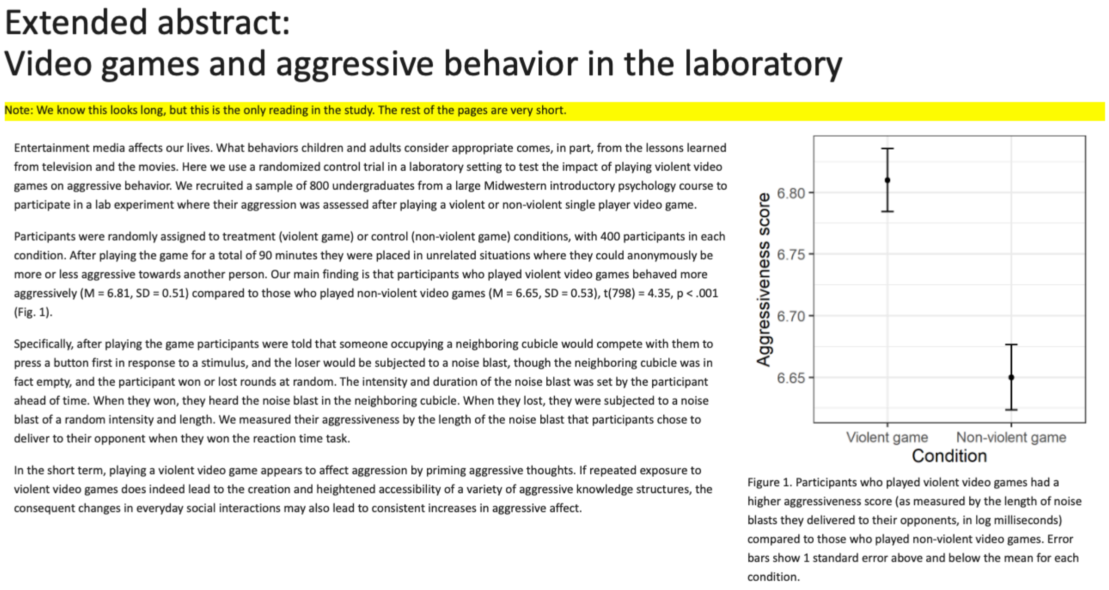


FIG. S11. **Background of faculty participants** Faculty self-reported their discipline, and even though we restricted our survey to CS, business, biology, physics, and psychology, we had some faculty self-identify as mathematics/statistics or other. Most faculty were moderately comfortable with statistics and data analysis. Our participants tended to be strongly familiar with the peer-review system, with over half of the participants reviewing over 100 papers.



Extended abstract: Video games and aggressive behavior in the laboratory

Note: We know this looks long, but this is the only reading in the study. The rest of the pages are very short.

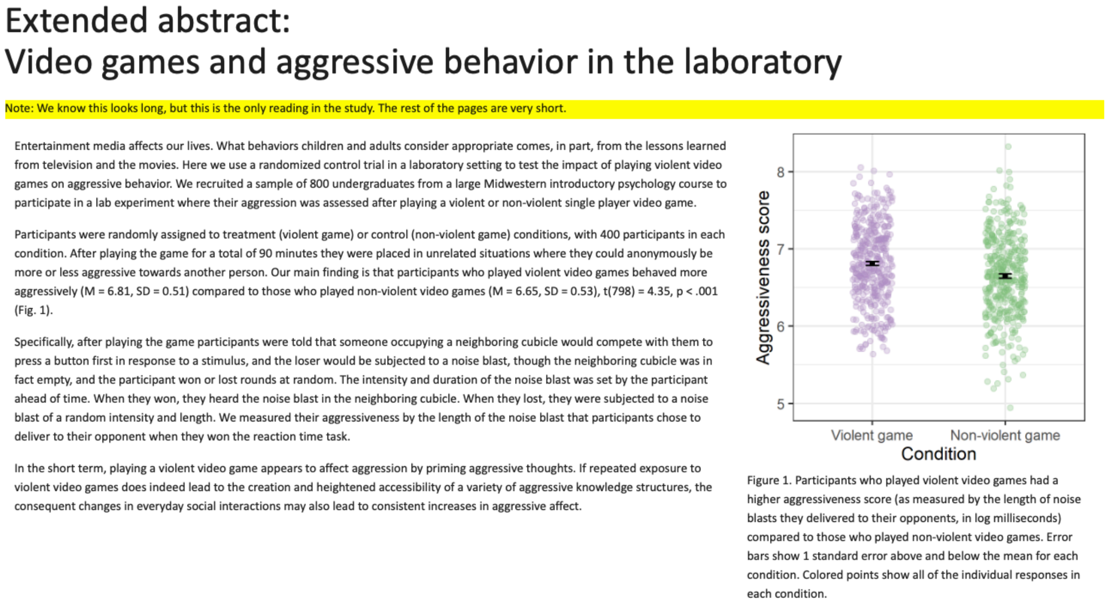
Entertainment media affects our lives. What behaviors children and adults consider appropriate comes, in part, from the lessons learned from television and the movies. Here we use a randomized control trial in a laboratory setting to test the impact of playing violent video games on aggressive behavior. We recruited a sample of 800 undergraduates from a large Midwestern introductory psychology course to participate in a lab experiment where their aggression was assessed after playing a violent or non-violent single player video game.

Participants were randomly assigned to treatment (violent game) or control (non-violent game) conditions, with 400 participants in each condition. After playing the game for a total of 90 minutes they were placed in unrelated situations where they could anonymously be more or less aggressive towards another person. Our main finding is that participants who played violent video games behaved more aggressively (M = 6.81, SD = 0.51) compared to those who played non-violent video games (M = 6.65, SD = 0.53), t(798) = 4.35, p < .001 (Fig. 1).

Specifically, after playing the game participants were told that someone occupying a neighboring cubicle would compete with them to press a button first in response to a stimulus, and the loser would be subjected to a noise blast, though the neighboring cubicle was in fact empty, and the participant won or lost rounds at random. The intensity and duration of the noise blast was set by the participant ahead of time. When they won, they heard the noise blast in the neighboring cubicle. When they lost, they were subjected to a noise blast of a random intensity and length. We measured their aggressiveness by the length of the noise blast that participants chose to deliver to their opponent when they won the reaction time task.

In the short term, playing a violent video game appears to affect aggression by priming aggressive thoughts. If repeated exposure to violent video games does indeed lead to the creation and heightened accessibility of a variety of aggressive knowledge structures, the consequent changes in everyday social interactions may also lead to consistent increases in aggressive affect.

Figure 1. Participants who played violent video games had a higher aggressiveness score (as measured by the length of noise blasts they delivered to their opponents, in log milliseconds) compared to those who played non-violent video games. Error bars show 1 standard error above and below the mean for each condition.



Extended abstract: Video games and aggressive behavior in the laboratory

Note: We know this looks long, but this is the only reading in the study. The rest of the pages are very short.

Entertainment media affects our lives. What behaviors children and adults consider appropriate comes, in part, from the lessons learned from television and the movies. Here we use a randomized control trial in a laboratory setting to test the impact of playing violent video games on aggressive behavior. We recruited a sample of 800 undergraduates from a large Midwestern introductory psychology course to participate in a lab experiment where their aggression was assessed after playing a violent or non-violent single player video game.

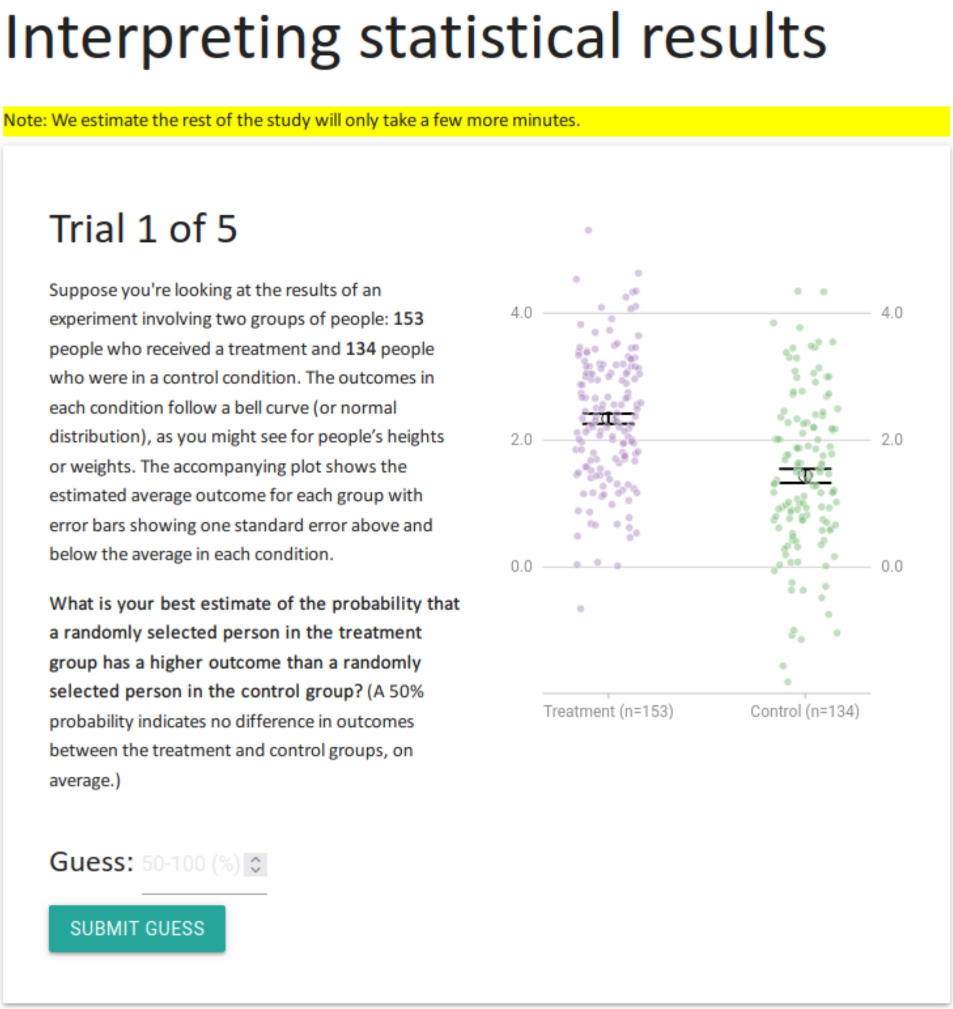
Participants were randomly assigned to treatment (violent game) or control (non-violent game) conditions, with 400 participants in each condition. After playing the game for a total of 90 minutes they were placed in unrelated situations where they could anonymously be more or less aggressive towards another person. Our main finding is that participants who played violent video games behaved more aggressively (M = 6.81, SD = 0.51) compared to those who played non-violent video games (M = 6.65, SD = 0.53), t(798) = 4.35, p < .001 (Fig. 1).

Specifically, after playing the game participants were told that someone occupying a neighboring cubicle would compete with them to press a button first in response to a stimulus, and the loser would be subjected to a noise blast, though the neighboring cubicle was in fact empty, and the participant won or lost rounds at random. The intensity and duration of the noise blast was set by the participant ahead of time. When they won, they heard the noise blast in the neighboring cubicle. When they lost, they were subjected to a noise blast of a random intensity and length. We measured their aggressiveness by the length of the noise blast that participants chose to deliver to their opponent when they won the reaction time task.

In the short term, playing a violent video game appears to affect aggression by priming aggressive thoughts. If repeated exposure to violent video games does indeed lead to the creation and heightened accessibility of a variety of aggressive knowledge structures, the consequent changes in everyday social interactions may also lead to consistent increases in aggressive affect.

Figure 1. Participants who played violent video games had a higher aggressiveness score (as measured by the length of noise blasts they delivered to their opponents, in log milliseconds) compared to those who played non-violent video games. Error bars show 1 standard error above and below the mean for each condition. Colored points show all of the individual responses in each condition.

FIG. S12. **Editorial study screenshot** The extended abstract shown to data scientist and faculty participants in the SE + Points condition (top) and the SE Only (bottom) condition. The text is exactly identical except for the presentation of the figure and one sentence in the figure caption.



Interpreting statistical results

Note: We estimate the rest of the study will only take a few more minutes.

Trial 1 of 5

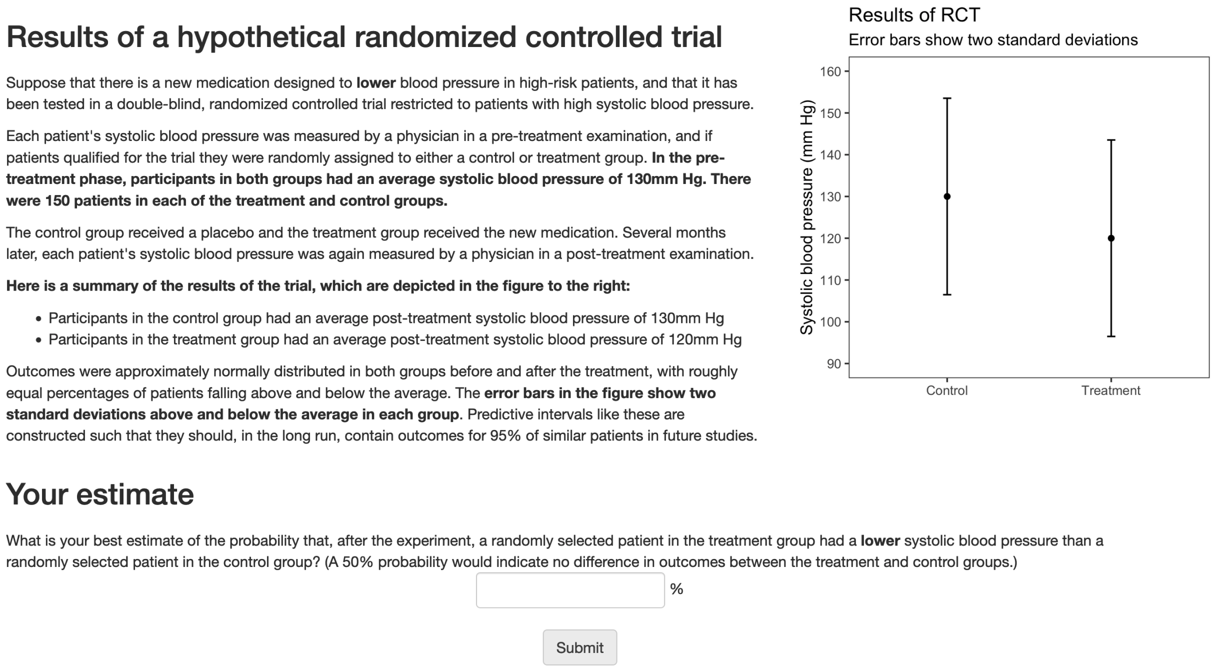
Suppose you're looking at the results of an experiment involving two groups of people: **153** people who received a treatment and **134** people who were in a control condition. The outcomes in each condition follow a bell curve (or normal distribution), as you might see for people's heights or weights. The accompanying plot shows the estimated average outcome for each group with error bars showing one standard error above and below the average in each condition.

What is your best estimate of the probability that a randomly selected person in the treatment group has a higher outcome than a randomly selected person in the control group? (A 50% probability indicates no difference in outcomes between the treatment and control groups, on average.)

Guess: 50-100 (%)

SUBMIT GUESS

FIG. S13. **Statistical estimation task screenshot** Screenshot of the statistical estimation task (Part 2) in the experiment for data scientists and faculty in the SE + points condition. Participants are asked to read a hypothetical outcome of an experiment. The experiment has randomly generated outcomes from a known distribution, so we can compute the probability of superiority from those distributions. The participant is asked to guess the probability of superiority.



Results of a hypothetical randomized controlled trial

Suppose that there is a new medication designed to **lower** blood pressure in high-risk patients, and that it has been tested in a double-blind, randomized controlled trial restricted to patients with high systolic blood pressure.

Each patient's systolic blood pressure was measured by a physician in a pre-treatment examination, and if patients qualified for the trial they were randomly assigned to either a control or treatment group. **In the pre-treatment phase, participants in both groups had an average systolic blood pressure of 130mm Hg. There were 150 patients in each of the treatment and control groups.**

The control group received a placebo and the treatment group received the new medication. Several months later, each patient's systolic blood pressure was again measured by a physician in a post-treatment examination.

Here is a summary of the results of the trial, which are depicted in the figure to the right:

- Participants in the control group had an average post-treatment systolic blood pressure of 130mm Hg
- Participants in the treatment group had an average post-treatment systolic blood pressure of 120mm Hg

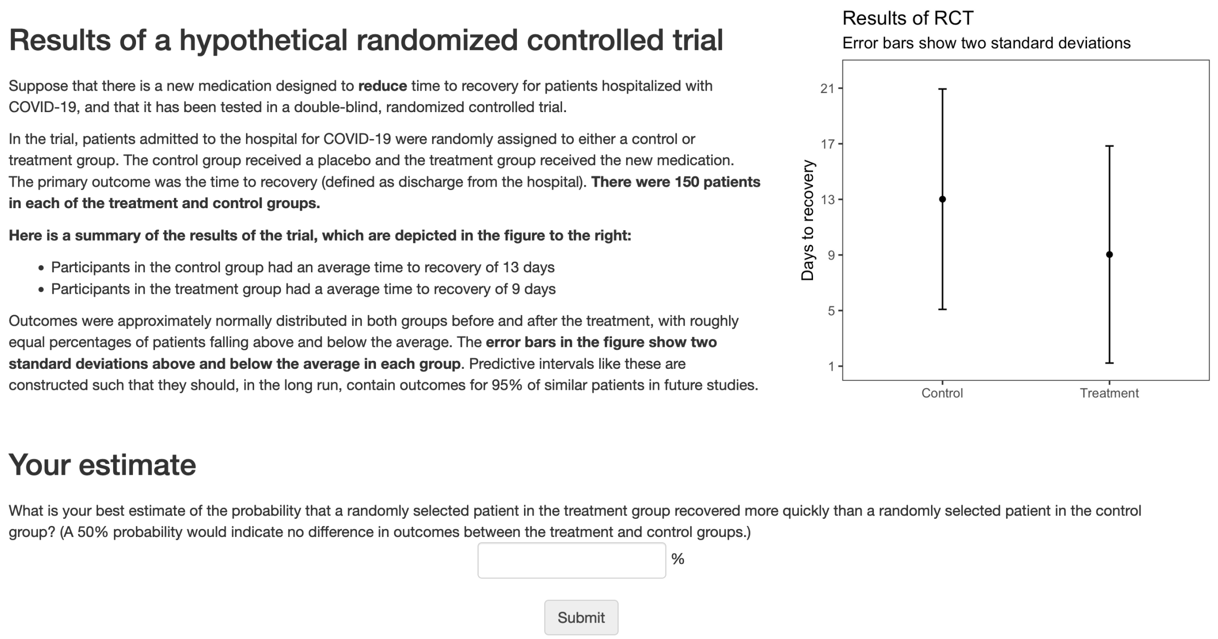
Outcomes were approximately normally distributed in both groups before and after the treatment, with roughly equal percentages of patients falling above and below the average. The **error bars in the figure show two standard deviations above and below the average in each group**. Predictive intervals like these are constructed such that they should, in the long run, contain outcomes for 95% of similar patients in future studies.

Results of RCT
Error bars show two standard deviations
Systolic blood pressure (mm Hg)
160
150
140
130
120
110
100
90
Control
Treatment

Your estimate

What is your best estimate of the probability that, after the experiment, a randomly selected patient in the treatment group had a **lower** systolic blood pressure than a randomly selected patient in the control group? (A 50% probability would indicate no difference in outcomes between the treatment and control groups.)

FIG. S14. **Medical providers experiment screenshot: blood pressure medication scenario** Screenshot of the experiment for medical provider and Mechanical Turk participants in the blood pressure medication scenario and the outcome variability (SD) condition. Participants are asked to read a hypothetical outcome of an RCT. The participant then estimates the perceived probability of superiority of that RCT.



Results of a hypothetical randomized controlled trial

Suppose that there is a new medication designed to **reduce** time to recovery for patients hospitalized with COVID-19, and that it has been tested in a double-blind, randomized controlled trial.

In the trial, patients admitted to the hospital for COVID-19 were randomly assigned to either a control or treatment group. The control group received a placebo and the treatment group received the new medication. The primary outcome was the time to recovery (defined as discharge from the hospital). **There were 150 patients in each of the treatment and control groups.**

Here is a summary of the results of the trial, which are depicted in the figure to the right:

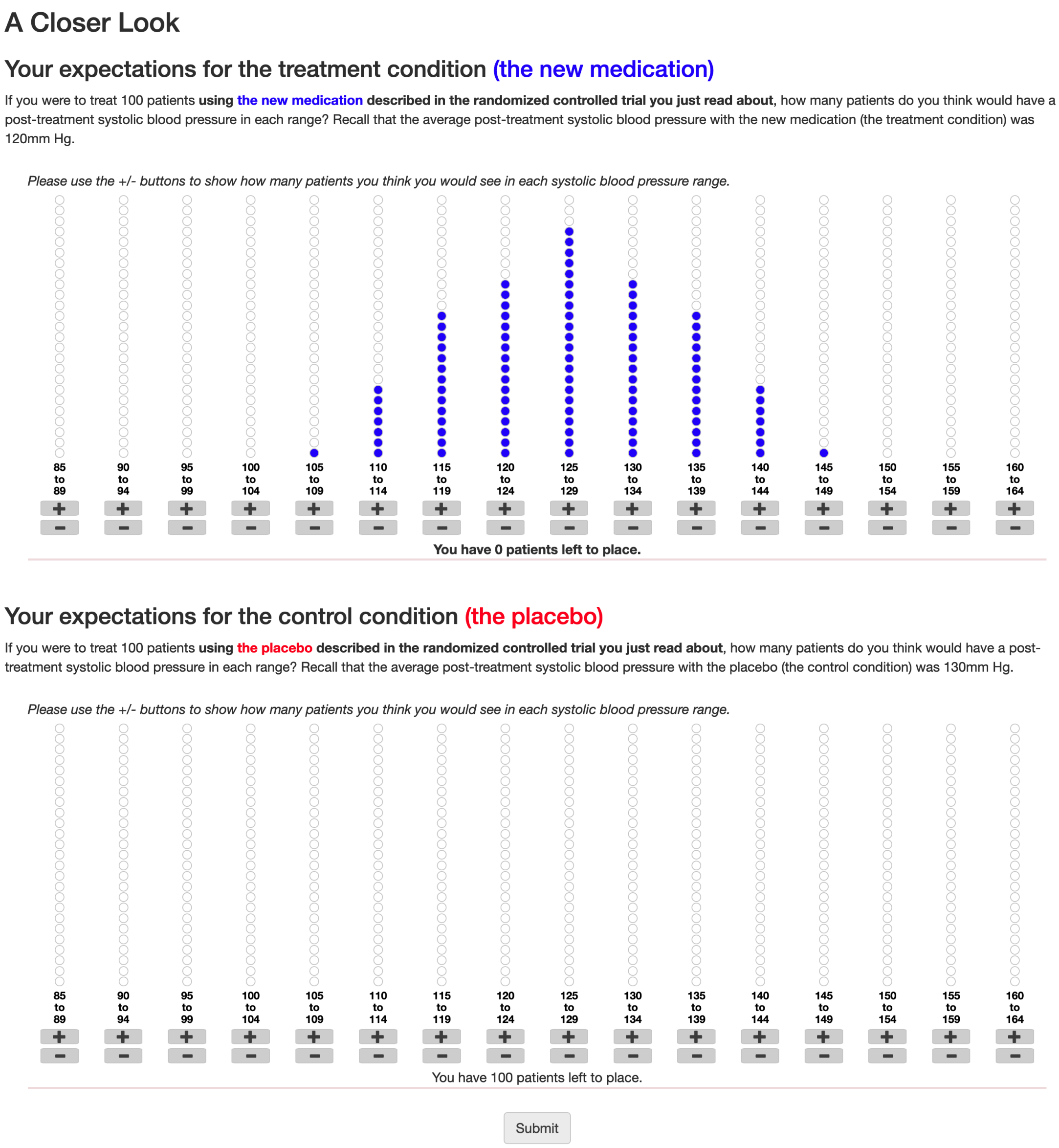
- Participants in the control group had an average time to recovery of 13 days
- Participants in the treatment group had a average time to recovery of 9 days

Outcomes were approximately normally distributed in both groups before and after the treatment, with roughly equal percentages of patients falling above and below the average. The **error bars in the figure show two standard deviations above and below the average in each group**. Predictive intervals like these are constructed such that they should, in the long run, contain outcomes for 95% of similar patients in future studies.

Your estimate

What is your best estimate of the probability that a randomly selected patient in the treatment group recovered more quickly than a randomly selected patient in the control group? (A 50% probability would indicate no difference in outcomes between the treatment and control groups.)

FIG. S15. **Medical providers experiment screenshot: COVID-19 medication scenario** Screenshot of the experiment for medical provider and Mechanical Turk participants in the COVID-19 medication scenario and the outcome variability (SD) condition. Participants are asked to read a hypothetical outcome of an RCT. The participant then estimates the perceived probability of superiority of that RCT.



A Closer Look

Your expectations for the treatment condition (the new medication)

If you were to treat 100 patients **using the new medication described in the randomized controlled trial you just read about**, how many patients do you think would have a post-treatment systolic blood pressure in each range? Recall that the average post-treatment systolic blood pressure with the new medication (the treatment condition) was 120mm Hg.

Please use the +/- buttons to show how many patients you think you would see in each systolic blood pressure range.

You have 0 patients left to place.

Your expectations for the control condition (the placebo)

If you were to treat 100 patients **using the placebo described in the randomized controlled trial you just read about**, how many patients do you think would have a post-treatment systolic blood pressure in each range? Recall that the average post-treatment systolic blood pressure with the placebo (the control condition) was 130mm Hg.

Please use the +/- buttons to show how many patients you think you would see in each systolic blood pressure range.

You have 100 patients left to place.

FIG. S16. **Distribution builder screenshot** Screenshot of the distribution builder for medical provider participants, where the treatment condition has been filled in, and the control condition has not been filled in yet. The distribution builder allows us to elicit participants' estimate of the entire outcome distribution for both conditions.