**Supplement to "Blinding Us to the Obvious?
The Effect of Statistical Training on the Evaluation of Evidence"**

Blakeley B. McShane[1] and David Gal[2]

[1] Kellogg School of Management, Northwestern University
[2] College of Business Administration, University of Illinois at Chicago

Section 1 of this supplement provides full details on all studies conducted including participants, procedures, and principal results. Section 2 provides a deeper examination of our choice data and Section 3 provides a deeper examination of our text response data. Finally, Section 4 provides statistical details. Our raw data can be found in Appendix A.

1 Study Details

Our first set of studies, labeled Study 1x, pertain to descriptive statements. To systematically examine whether researchers might be led by the notion of statistical significance to misconstrue evidence, we surveyed researchers across a wide variety of fields regarding their interpretation of data. In particular, we presented researchers with a study summary that showed a difference in an outcome variable associated with an intervention and a set of descriptions of that difference. We manipulated whether the difference in the outcome variable attained ($p = 0.01$) or failed to attain ($p = 0.27$) statistical significance. We posited that researchers would correctly identify that the outcome variable differed when the difference attained statistical significance but would fail to identify this difference when it failed to attain statistical significance. To test the robustness of any effect, we examined our hypothesis using different wordings for the response options, different question orders, and different subject populations including researchers in medicine, psychology, and business as well as undergraduates both trained and untrained in statistics.

In our second set of studies, labeled Study 2x, we examine whether any observed pattern of results would extend from descriptive statements to the evaluation of evidence via likelihood judgments. To do so, we presented additional groups of researchers with a study summary and a set of inferences that might be drawn from the data. As above, the p-value for the null hypothesis significance test of no difference between the two treatment groups was set above or below 0.05. While the p-value quantifies the strength of evidence against the null hypothesis, we hypothesized that participants would incorrectly judge (i.e., dismiss or undervalue in a relative sense) evidence that failed to attain statistical significance. Further, to examine whether participants' likelihood judgments would extend to decisions based on the data, we asked participants to report how they would choose to act in light of the data. To test the robustness of any effect, we examined our hypothesis using different wordings for the choice question and response options, different question orders, adding additional information (i.e., a posterior probability based on a Bayesian calculation), and different subject populations including researchers in cognitive science, psychology, and economics.

Studies labeled with a star appear in the main text; the discussion from the main text appears in this supplement along with additional material. Studies labeled with a letter were mentioned briefly in the Robustness subsections of the main text; a full discussion of them appears in this supplement.

In this section, we present data from the principal questions of each study graphically. For our raw data, see Appendix A.

1.1 Study 1*: *New England Journal of Medicine*

Objective:

The goal of Study 1 was to examine the hypothesis that a focus on statistical significance would lead researchers to misinterpret data. To systematically examine this question, we presented researchers with a study summary that showed a difference in an outcome variable associated with an intervention and a set of descriptions of that difference. We manipulated whether the difference in the outcome variable attained ($p = 0.01$) or failed to attain ($p = 0.27$) statistical significance. We posited that researchers would correctly identify that the outcome variable differed when the difference attained statistical significance but would fail to identify this difference when it failed to attain statistical significance.

Participants:

Participants were the authors of articles published in the 2013 volume of the *New England Journal of Medicine* (NEJM; issues 368.1-368.10). A link to our survey was sent via email to the 322 authors; about twenty email addresses were incorrect. Seventy-five authors completed the survey, yielding a completion rate of 25%.

Procedure:

Participants were asked to respond sequentially to two versions of a principal question followed by several follow-up questions. The principal question asked participants to choose the most accurate description of the results from a study summary that showed a difference in an outcome variable associated with an intervention. We manipulated whether this difference attained ($p = 0.01$) or failed to attain ($p = 0.27$) statistical significance within subjects, with participants first asked to answer the $p = 0.27$ version of the question and then, on the next screen, the $p = 0.01$ version of the question. To test for robustness to differences in the wording of the response options, participants were randomized to one of three variations. The first principal question using response wording one was presented in the main text of the paper and was the following:

> Below is a summary of a study from an academic paper.
>
> The study aimed to test how different interventions might affect terminal cancer patients' survival. Participants were randomly assigned to one of two groups.

Group A was instructed to write daily about positive things they were blessed with while Group B was instructed to write daily about misfortunes that others had to endure. Participants were then tracked until all had died. Participants in Group A lived, on average, 8.2 months post-diagnosis whereas participants in Group B lived, on average, 7.5 months post-diagnosis ($\boldsymbol{p = 0.27}$).

Which statement is the most accurate summary of the results?

A. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was **greater** than that lived by the participants who were in Group B.

B. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was **less** than that lived by the participants who were in Group B.

C. Speaking only of the subjects who took part in this particular study, the average number of post-diagnosis months lived by the participants who were in Group A was **no different** than that lived by the participants who were in Group B.

D. Speaking only of the subjects who took part in this particular study, it **cannot be determined** whether the average number of post-diagnosis months lived by the participants who were in Group A was greater/no different/less than that lived by the participants who were in Group B.

After seeing this question, each participant was asked the same question again but $p = 0.27$ was switched to $p = 0.01$.

Response wording two was identical to response wording one above except it omitted the phrase "Speaking only of the subjects who took part in this particular study" from each of the four response options. Response wording three was the following:

A. The participants who were in Group A tended to live **longer** post-diagnosis than the participants who were in Group B.

B. The participants who were in Group A tended to live **shorter** post-diagnosis than the participants who were in Group B.

C. Post-diagnosis lifespan **did not differ** between the participants who were in Group A and the participants who were in Group B.

D. It **cannot be determined** whether the participants who were in Group A tended to live longer/no different/shorter post-diagnosis than the participants who were in Group B.

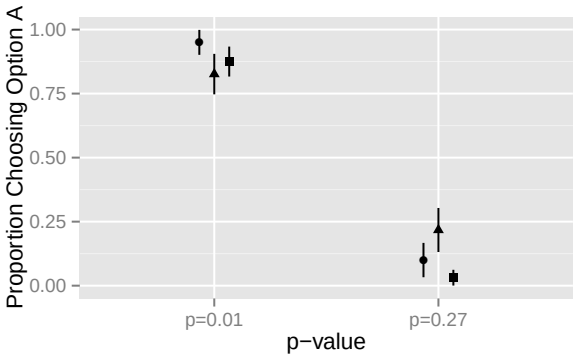


Figure 1: Data from Study 1⋆: *New England Journal of Medicine*. Points denote $\hat{p}_A$, the proportion of participants choosing option A, and lines denote $\hat{p}_A \pm \sqrt{\hat{p}_A(1-\hat{p}_A)/n}$. Response wording one is indicated by a circle, response wording two by a triangle, and response wording three by a square. Regardless of response wording, participants typically answered correctly when $p = 0.01$ but incorrectly when $p = 0.27$.

After these questions, participants were asked (i) a multiple choice question about their primary area of expertise (i.e., medicine, chemistry/biology, statistics/biostatistics, engineering, or other), (ii) a free response question asking at what p-value statistical significance is conventionally defined ($p < 0.05$; 97% of participants answered correctly), and (iii) a question about their statistical model for the data which read:

> Responses in the treatment and control group are often modeled as a parametric model, for example, as independent normal with two different means or independent binomial with two different proportions.
>
> An alternative model under the randomization assumption is a finite population model under which the permutation distribution of the conventional test statistic more or less coincides with the distribution given by the parametric model.
>
> Which of the following best describes your modeling assumption as you were considering the prior questions?
>
> A. I was using the parametric model.
>
> B. I was using the permutation model.
>
> C. I was using some other model.
>
> D. I was not using one specific model.

After this, the survey terminated.

Results:

Data from the principal question appears in Figure 1 (see Table A2 for data from the modeling question). We note only eleven of the seventy-five participants indicated an expertise in statistics or biostatistics with the majority (forty-eight) indicating an expertise in medicine.

For the principal question shown above, the correct answer is option A regardless of the p-value and the response wording: all four response options are descriptive statements and indeed the average number of post-diagnosis months lived by the participants who were in Group A was greater than that lived by the participants who were in Group B (i.e., $8.2 > 7.5$). However, participants were much more likely to answer the question correctly when the p-value in the question was set to 0.01 than to 0.27. For instance, among participants who saw response wording one, 95% correctly answered when the p-value was set to 0.01; on the other hand, only 10% correctly answered when the p-value was set to 0.27 with 55% choosing option C and 35% choosing option D. Responses in the other two response wording conditions were similar as can be seen in Figure 1.

These results are striking and suggest that, as hypothesized, a focus on statistical significance leads researchers to dichotomize evidence. In particular, participants failed to identify differences that were not statistically significant as different.

In attempting to explain these results, we considered a variety of explanations including wording, familiarity with the conventional level of statistical significance, modeling choice, and field of expertise. Given that our results were robust to three different wordings and that 97% of participants were familiar with the conventional level of statistical significance, we ruled the first two alternative explanations out. Further, while responses in treatment and control groups are often modeled using infinite population parametric models (e.g., independent normal with different means or independent binomial with different proportions), randomization only secures a finite population permutation model [Freedman et al., 2007, p. 511]. With this model, statements such as "the average for the treatment" can be ambiguous in terms of whether they refer to the average for those subjects who actually received the treatment in the trial versus the average for all subjects under the hypothetical that they all received the treatment; under the latter interpretation, one might be justified in giving a different response for $p = 0.01$ versus $p = 0.27$. As only 6.7% of participants reported using the permutation model, this explanation cannot hold in practice. Further, our wording generally precluded the latter interpretation (i.e., by asking about the average of "participants who were in Group A" it is unreasonable to assume we were asking about a hypothetical under which all participants were assigned to Group A). Finally, hypothesizing that a deep as opposed to cursory training in statistics and the concomitant exposure to forms of statistical reasoning outside the null hypothesis significance testing (NHST) paradigm would help participants focus on the descriptive nature of the question, we examined the performance of the eleven participants who indicated a primary expertise in statistics or biostatistics; indeed, statisticians performed somewhat better than other participants (see Section 4).

One potential criticism of our findings is that our question is essentially a trick question: researchers clearly know that 8.2 is greater than 7.5, but they might perceive that asking whether 8.2 is greater than 7.5 is too easy a question and hence they focus on whether the

difference is statistically significant. However, asking whether a p-value of 0.27 is statistically significant is also trivial, so this criticism does not resolve why researchers focus on the statistical significance of the difference rather than on the difference itself. A related potential criticism regards our question as a trick question for a different reason: by including a p-value, we naturally lead researchers to focus on statistical significance. However, this is essentially our point: researchers are so trained to focus on statistical significance that the mere presence of a p-value leads them to automatically view everything through the lens of the NHST paradigm even when it is not warranted. Moreover, in further response to such criticisms, we note that we stopped just short of explicitly telling participants that we were asking for a description of the observed data rather than asking them to make a statistical inference (e.g., response options read, "**Speaking only of the subjects who took part in this particular study**, the average number of post-diagnosis months lived by the **participants who were in Group A** was greater than that lived by the **participants who were in Group B**" and similarly; emphasis added).

While not directly relevant to our hypotheses, there are two additional points worth noting. First, even if we had asked participants to make a statistical inference under the NHST paradigm rather than to simply describe the data, option C (which stated that the average number of months lived by participants in the two groups did not differ) is never correct: failure to reject the null hypothesis does not imply or prove that the two treatments do not differ. Second, and again assuming we were asking an inferential question rather than a descriptive question, there is a sense in which option D (which states that it cannot be determined whether the average number of months lived by participants in the two groups differed) is the correct answer regardless of the p-value since at no p is the null definitively overturned. However, only a relatively small proportion of participants chose option D as their response to both versions of the question (i.e., the $p = 0.01$ version and the $p = 0.27$ version), with most choosing option A for the $p = 0.01$ version and option C or option D for the $p = 0.27$ version.

1.2 Study 1a: *Psychological Science*

Objective:

The main goal of Study 1a was to examine whether the effect identified in Study 1* would extend to another population. Two additional goals were to examine whether (i) a variation in the wording of the study description and the response options and (ii) the order in which the p-value in the principal questions was presented (i.e., $p = 0.01$ first versus $p = 0.27$ first) would materially affect the pattern of results.

Participants:

Participants were the members of the *Psychological Science* (PS) editorial board. A link to our survey was sent via email to the 116 board members who were not personal acquaintances or colleagues of the authors; two email addresses were incorrect. Fifty-four board members completed the survey, yielding a completion rate of 47%.

Procedure:

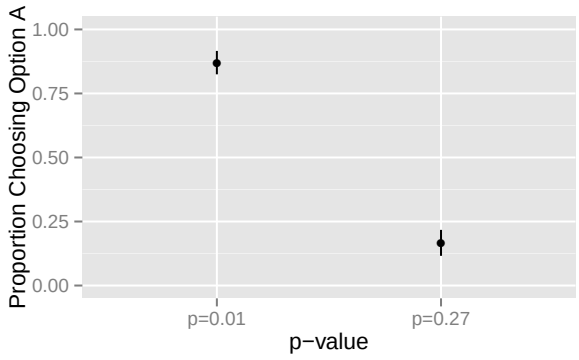
Participants were asked two principal questions very similar to those asked of authors from the *New England Journal of Medicine*. Participants were randomized to one of two conditions where the first condition saw the below question with $p = 0.27$ first and $p = 0.01$ second while the second condition saw the below question with $p = 0.01$ first and $p = 0.27$ second:

> Below is a summary of a study from an academic paper:
>
> The study aimed to test how different interventions might affect terminal cancer patients' survival. Participants were randomly assigned to either write daily about positive things they were blessed with or to write daily about misfortunes that others had to endure. Participants were then tracked until all had died. Participants who wrote about the positive things they were blessed with lived, on average, 8.2 months after diagnosis whereas participants who wrote about others' misfortunes lived, on average, 7.5 months after diagnosis ($\boldsymbol{p = 0.27}$).
>
> Which statement is the most accurate summary of the results?
>
> A. The results showed that participants who wrote about their blessings tended to live longer post-diagnosis than participants who wrote about others' misfortunes.
> B. The results showed that participants who wrote about others' misfortunes tended to live longer post-diagnosis than participants who wrote about their blessings.
> C. The results showed that participants' post-diagnosis lifespan did not differ depending on whether they wrote about their blessings or wrote about others' misfortunes.
> D. The results were inconclusive regarding whether participants' post-diagnosis lifespan was greater when they wrote about their blessings or when they wrote about others' misfortunes.

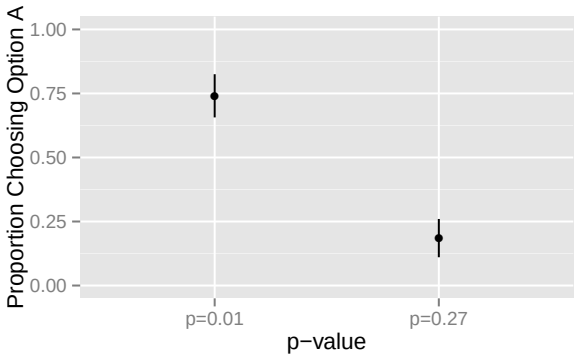
After seeing both the $p = 0.27$ and $p = 0.01$ versions of this question in random order, the survey terminated.

Results:

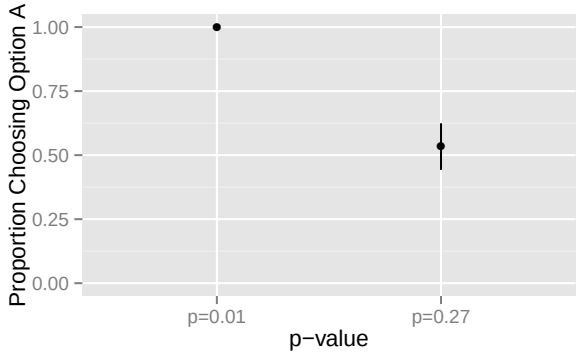
The pattern of results was not substantially affected by the order in which the p-value was presented (see Section 4). Consequently, we collapse across both order conditions and present our results in Figure 2(a). For the $p = 0.01$ version of the question, 87% of participants correctly answered option A. Conversely, for the $p = 0.27$ version of the question, only 17% correctly answered option A with 37% answering option C and 46% answering option D. In sum, the pattern of results was consistent with that obtained for the NEJM authors in Study 1*.



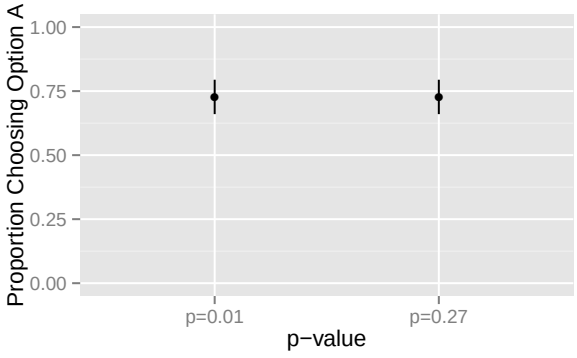
(a) *Psychological Science*



(b) Marketing Science Institute Young Scholars



(c) Statistically-Trained Undergraduates



(d) Statistically-Untrained Undergraduates

Figure 2: Data from Study 1a: *Psychological Science*, Study 1b: Marketing Science Institute Young Scholars, and Study 1c: Undergraduates. Points denote $\hat{p}_A$, the proportion of participants choosing option A, and lines denote $\hat{p}_A \pm \sqrt{\hat{p}_A(1-\hat{p}_A)/n}$. Participants with statistical training typically answered correctly when $p = 0.01$ but incorrectly when $p = 0.27$ thereby replicating the finding presented in Figure 1. Participants who lacked statistical training generally answered correctly regardless and outperformed the other groups when $p = 0.27$.

1.3 Study 1b: Marketing Science Institute Young Scholars

Objective:

The goal of Study 1b was to examine whether the effect identified in Studies 1* and 1a would extend to another population of researchers.

Participants:

Participants were the 2013 Marketing Science Institute Young Scholars (MSI)–individuals selected on the basis of their being "potential leaders of the 'next generation' of marketing academics." A link to our survey was sent via email to the thirty-five 2013 Young Scholars as well as the five Young Scholars advisors. Twenty-seven completed the survey, yielding a completion rate of 68%.

Procedure:

Participants were sent a survey identical to that sent to the editorial board members of PS however it contained one additional question presented last asking whether the participant identified as a quantitative, behavioral, or other researcher; nine answered quantitative, sixteen behavioral, and two other.

Results:

The pattern of results was not substantially affected by the order in which the p-value was presented (see Section 4). Consequently, we collapse across both order conditions and present our results in Figure 2(b). For the $p = 0.01$ version of the question, 74% of participants correctly answered option A. Conversely, for the $p = 0.27$ version of the question, only 19% of participants answered correctly with 15% answering option C and 67% answering option D. In sum, the pattern of results was consistent with that obtained for NEJM authors and PS editorial board members in Studies 1* and 1a respectively.

1.4 Study 1c: Undergraduates

Objective:

The goal of Study 1c was to examine the role of statistical training in leading individuals to misinterpret data. We have proposed that statistical training might lead individuals to focus on whether a result is statistically significant or not, thereby blinding individuals to differences that attain statistical significance. In order to examine this, we sought to compare how undergraduate students' interpretation of data differed from that of trained researchers and whether any difference was influenced by whether the undergraduates were or were not trained in statistics. While in most contexts we would expect even cursory training in statistics to improve the quality of statistical reasoning [Fong et al., 1986], we hypothesized that the focus placed on NHST in the training of professional researchers and in typical undergraduate courses would be associated with diminished performance on the

$p = 0.27$ version of the question among professional researchers and statistically-trained undergraduates vis-a-vis statistically-untrained undergraduates.

Participants:

Participants were undergraduates at a major U.S. university. Seventy-four were given our survey while in the lab participating in a series of unrelated studies.

Procedure:

Participants were given a survey identical to that sent to the editorial board members of PS however it contained several additional questions asked last. Of these, the first was a free response question asking at what p-value statistical significance is conventionally defined; this question served as a proxy for whether the participant possessed or lacked training in statistics and thirty of our seventy-four undergraduate participants answered it correctly. Participants were then asked about their age, sex, and Math SAT score percentile.

Results:

Again, the pattern of results was not substantially affected by the order in which the p-value was presented (see Section 4). Consequently, we collapse across both order conditions and present our results in Figures 2(c)-2(d). Thirty of our seventy-four undergraduate participants correctly identified the conventional level of statistical significance. Consistent with our prediction, on the $p = 0.27$ version of the question, 73% of statistically-untrained undergraduates answered correctly compared to 17% of PS editorial board members, 19% of MSI Young Scholars, and 53% of statistically-trained undergraduates who saw the identical question (see Figure 2). Further, and perhaps not surprisingly, statistically-untrained undergraduates answered the $p = 0.01$ and $p = 0.27$ versions of the question correctly at the same rate.

Across our first four studies, we found consistent support for the proposition that a focus on statistical significance can lead individuals to misinterpret data. As a result, it appears that professional researchers and undergraduates trained in statistics fail to recognize differences that are apparent to undergraduates untrained in statistics when the differences fail to attain statistical significance.

1.5 Study 2*: *American Journal of Epidemiology*

Objective:

Thus far, we have examined how differences in statistical significance affect researchers' descriptive statements about data. In Study 2*, we examine whether the observed pattern of results extends from descriptive statements to the evaluation of evidence via likelihood judgments. To do so, we presented researchers with a study summary and a set of inferences that might be drawn from the data. As above, the p-value for the null hypothesis significance test of no difference between the two treatment groups was set above or below 0.05. While

the p-value quantifies the strength of evidence against the null hypothesis, we hypothesized that participants would incorrectly judge (i.e., dismiss or undervalue in a relative sense) evidence that failed to attain statistical significance.

To examine whether participants' likelihood judgments would extend to decisions based on the data, we also asked participants to report how they would choose to act in light of the data. We hypothesized that when it comes to making a choice, researchers would, to some degree, shift their focus from whether a result is or is not statistically significant to which choice option represents the superior alternative (more precisely, here and hereafter, by the "superior alternative" we mean the alternative that is more likely to be superior which, in our setting, means the alternative that is more likely to be more effective in terms of the probability of recovery from a disease). As a result, we predicted that researchers would be more likely to select the superior alternative in the context of making a choice relative to the context of making a likelihood judgment. Moreover, we predicted that this effect would be more pronounced the more personally consequential the choice for the participant.

A further goal of Study 2* was to gain additional insight into researchers' reasoning when making likelihood judgments and choices by examining how varying (i) the degree to which the p-value is above the threshold for statistical significance and (ii) the magnitude of the treatment difference affects researchers' likelihood judgments and choices. We hypothesized that researchers would be substantially more likely to provide incorrect judgments when the p-value was set above 0.05 than when set below 0.05, but that (i) the degree to which the p-value exceeded 0.05 and (ii) the magnitude of the treatment difference would have little impact on the results as researchers' would focus almost solely on whether the difference between the treatments attained or failed to attain statistical significance.

Participants:

Participants were the authors of articles published in the 2013 volume of the *American Journal of Epidemiology* (AJE; issues 177.4 to 178.4). A link to our survey was sent via email to the 1,111 authors; about 110 email addresses were incorrect. 299 authors completed a survey, yielding a completion rate of 30%. Thirty-eight responses could not be used because they were inadvertently diverted to the wrong survey; consequently, we report results from the 261 participants who completed the correct survey.

Procedure:

Participants completed a likelihood judgment question followed by a choice question. Participants were randomly assigned to one of sixteeen conditions following a four by two by two design. The first level of the design varied whether the p-value was set to 0.025, 0.075, 0.125, or 0.175 and the second level of the design varied the magnitude of the treatment difference (52% and 44% versus 57% and 39%). The third level of the design applied only to the choice question and varied whether participants were asked to make a choice for a close versus distant other (see below). Participants saw the same p-value and magnitude of treatment difference in the choice question as they saw in the preceding judgment question.

The judgment question was:

Below is a summary of a study from an academic paper:

The study aimed to test how two different drugs impact whether a patient recovers from a certain disease. Subjects were randomly drawn from a fixed population and then randomly assigned to Drug A or Drug B. **Fifty-two percent (52%)** of subjects who took Drug A recovered from the disease while **forty-four percent (44%)** of subjects who took Drug B recovered from the disease.

A test of the null hypothesis that there is no difference between Drug A and Drug B in terms of probability of recovery from the disease yields a p-value of **0.175**.

Assuming no prior studies have been conducted with these drugs, which of the following statements is most accurate?

A. A person drawn randomly from the same population as the subjects in the study is **more likely** to recover from the disease if given Drug A than if given Drug B.

B. A person drawn randomly from the same population as the subjects in the study is **less likely** to recover from the disease if given Drug A than if given Drug B.

C. A person drawn randomly from the same population as the subjects in the study is **equally likely** to recover from the disease if given Drug A than if given Drug B.

D. It **cannot be determined** whether a person drawn randomly from the same population as the subjects in the study is more/less/equally likely to recover from the disease if given Drug A or if given Drug B.

For the choice question, participants were presented with the same study summary but were instead asked to make a hypothetical choice. Moreover, participants were randomized into one of two conditions: they were asked to choose a treatment for either a close other (i.e., a loved one) or a distant other (i.e., physicians treating patients). We predicted that participants would be more likely to choose a superior alternative for a close other than for a distant other when the superior alternative was not statistically significantly different from the inferior alternative. The basis for this prediction was our hypothesis that choice tends to shift the focus away from statistical significance and towards whether an option is superior combined with the logic that this shift would be greater the more consequential the choice for the individual. Participants in the close other condition saw the following wording:

If you were to advise a loved one who was a patient from the same population as those in the study, what drug would you advise him or her to take?

Participants in the distant other condition saw the following wording:

> If you were to advise physicians treating patients from the same population as those in the study, what drug would you advise these physicians prescribe for their patients?

All participants then saw the following response options:

A. I would advise Drug A.

B. I would advise Drug B.

C. I would advise that there is no difference between Drug A and Drug B.

In addition to asking participants to make judgments and choices, we also sought to gain insight into participants' reasoning by asking them to explain why they chose the option they chose in free response form both after the judgment question and after the choice question. Participants were provided with a text box to provide their response.

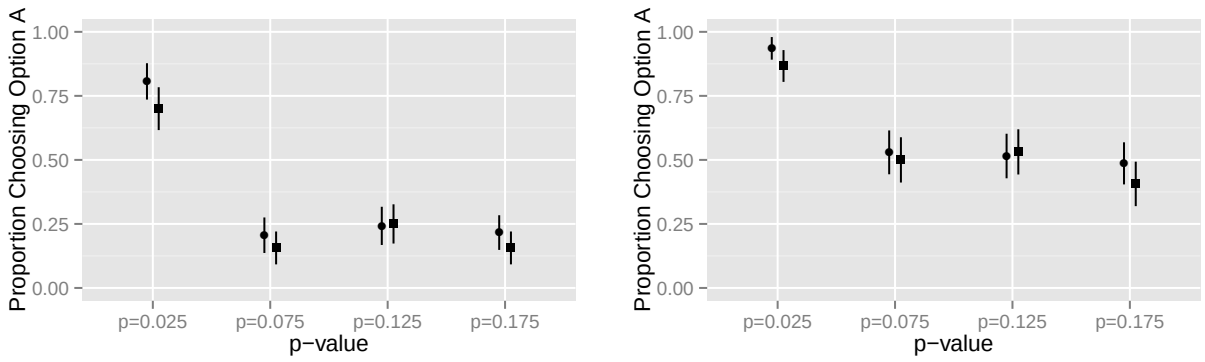
After these questions, participants were asked a multiple choice question about their primary area of expertise (epidemiology, medicine, statistics/biostatistics, or other) and a free response question asking at what p-value statistical significance is conventionally defined ($p < 0.05$; 99% of participants answered correctly).

Results:

As the effect of making a choice for a close versus distant other was a secondary hypothesis, we collapse over both "other" (i.e., close versus distant) conditions in the principal presentation of our results and return to the analysis of the effect of a close versus distant other below. We present our results in Figure 3.

The judgment question differs from the question seen by NEJM authors, PS editorial board members, the MSI Young Scholars, and the undergraduates in several ways. Most obviously, it asks about proportions as opposed to means; we did not expect this to impact results. More notably, it moves from descriptive statements to the evaluation of evidence via a likelihood judgment about what might happen to hypothetical new patients drawn from the same population as those in the study.

While the p-value quantifies the strength of the evidence regarding the likelihood that the efficacy of Drug A is higher than that of Drug B (and thus the likelihood of hypothetical new patients recovering under Drug A versus Drug B), again the level of p-value does not



(a) Judgment

(b) Choice

Figure 3: Data from Study 2⋆: *American Journal of Epidemiology*. Points denote $\hat{p}_A$, the proportion of participants choosing option A, and lines denote $\hat{p}_A \pm \sqrt{\hat{p}_A(1-\hat{p}_A)/n}$. The small (large) treatment difference condition is indicated by a square (circle). For both questions, the proportion choosing option A drops steeply once the p-value rises above 0.05 but it is stable thereafter. The drop is attenuated for the choice question relative to the judgment question. The magnitude of the treatment difference has no substantial impact on the proportion choosing option A for either question.

alter the correct response option. The correct answer is again option A as Drug A is more likely to have higher efficacy than Drug B regardless of the p-value. The share of participants who correctly answered the judgment question when the p-value was set to 0.025 was 70% and 81% in the small and large treatment difference conditions respectively. However, the share of participants who correctly answered the judgment question was substantially lower when the p-value was set to 0.075, 0.125, and 0.175, with no substantial variation in the share answering correctly across these three conditions (16%, 25%, and 16% respectively when the treatment difference was small and 21%, 24%, and 22% respectively when the treatment difference was large).

An argument might be made that there is a sense in which option D is the correct option for the judgment question because, as discussed above, at no p is the null definitively overturned. More specifically, under a classical frequentist interpretation of the question, which drug is 'more likely' to result in recovery depends upon the parameters governing the probability of recovery for each drug. As these parameters are unknown and unknowable, option D could be construed as the correct answer under this interpretation. We note that no such difficulty arises under a Bayesian interpretation of the question and for which option A is definitively the correct response.

Were participants approaching the judgment question with the classical interpretation, they would select option D both when presented with the $p < 0.05$ and the $p > 0.05$ versions of the question. However, participants overwhelmingly selected option A when presented with the $p < 0.05$ version of the question whereas they chose option D predominantly when presented with $p > 0.05$ versions. Thus, it seems highly improbable that participants approached the question classically.

Further, we note that both the judgment and choice questions ask the participant to evaluate evidence concerning what might happen to hypothetical new patients drawn from the same population as those in the study; thus, both are, in a sense, predictive questions. However, participants will not (and clearly do not) necessarily make the same evaluation in both contexts. As an additional point, we note that, by asking about hypothetical new patients, the questions prompt a parametric modeling approach as the permutation model is valid only for the finite population of patients who participated in the study. Finally, by asking about the likelihood that the efficacy of Drug A is higher than that of Drug B (via expected outcomes of hypothetical new patients), the questions also prompt a more Bayesian manner of reasoning as in classical statistical reasoning the efficacy of Drug A is either higher than that of Drug B or it is not; of course, as usual, one could adopt the classical position that the efficacy of Drug A is either higher than that of Drug B or it is not, that this fact is unknown, and that one's uncertainty about this fact is quantified using Bayesian reasoning.

We next examined participants' responses to the choice question. The share of participants choosing Drug A in the choice question when the p-value was set to 0.025 was 87% and 94% in the small and large treatment difference conditions respectively. This dropped substantially when the p-value was set to 0.075, 0.125, and 0.175, with no substantial variation across the three (50%, 53%, and 41% respectively when the treatment difference was small and 53%, 52%, and 49% respectively when the treatment difference was large).

In sum, the share of participants who correctly answer each question drops steeply once the p-value falls below 0.05 but is stable thereafter and the magnitude of the treatment difference has no substantial impact on the fraction answering correctly. This dichotomization of responses around the conventional threshold for statistical significance is consistent with the notion that dichotomization of evidence into statistically significant and not statistically significant biases researchers' judgments. Moreover, the lack of any substantial effect of the magnitude of the treatment difference suggests that, within the range of magnitudes we examined (i.e., a more than doubling of the magnitude), whether a result attains or fails to attain statistical significance has a far greater impact on the response than the magnitude of the treatment difference.

We further examined whether the choices made by participants varied by whether the choice was made on behalf of a close other or a distant other. When the p-value was set to 0.075, 0.125, and 0.175, the results showed that participants were more likely to choose Drug A when making a choice for a close other than when making a choice for a distant other (64%, 58%, and 53% respectively versus 39%, 47%, and 36% respectively). This finding supports our proposition that making a choice shifts participants' focus from whether a result attains or fails to attain statistical significance to the available evidence, and that this effect is greater the more consequential the choice for the participant. Nonetheless, we find it striking that–even when faced with a consequential choice–only 50% of participants across the two other conditions chose Drug A when the difference between the two drugs failed to attain statistical significance whereas 90% chose it when the difference attained statistical significance. For further analysis of how choices vary by the party on behalf of which the participant was making a choice, see Section 2.

We next examined participants' explanations for their answers. Of the 159 participants who incorrectly answered the judgment question when the p-value was above 0.05, 115 suggested that they chose the answer they did because the difference in treatment outcomes failed to attain statistical significance. Many of these responses alluded to the idea that they could not label as evidence differences that did not reach the threshold for statistical significance. Some representative responses were "test for statistical significance was 0.07, which is above the well-established standard of p-value < 0.05."; "H_0 is not rejected"; "p-value is > 0.05, indicating no statistical difference between groups."; and "because the p-value indicated that there was not a significant difference between groups and thus no detectable difference between drug A or B." Other responses that alluded to statistical significance indicated that the lack of statistical significance impacted participants' confidence: "Although the relative difference appears large, statistically the diff is not signif and not knowing more about the sample size and disease pathology or presumably drug mechanism I would not feel confident about 'prescribing' one drug over the other." A small minority of the responses among those assigned to the small treatment difference condition also expressed that the lack of statistical significance combined with the small magnitude of the treatment difference made any difference practically unimportant: "p-value > 0.05 plus from an intuitive standpoint both drugs essentially gave a 50-50 chance of recovery." Such explanations are consistent with our account that researchers' perceptions of evidence are dichotomized around the threshold for statistical significance and that this can manifest itself either as a total disregard for evidence for which $p > 0.05$ or a sharp change in confidence or perceptions of practical significance around $p = 0.05$. For further analysis of participants' explanations for their choices, see Section 3.

For further analysis of how choices vary by the party on behalf of which the participant was making a choice, see Section 2. For further analysis of participants' explanations for their choices, see Section 3. We note we also examined the performance of the thirty-four participants who indicated a primary expertise in statistics or biostatistics; indeed, statisticians performed somewhat better than other participants (see Section 4).

1.6 Study 2a: *Cognition*

Objective:

The main goal of Study 2a was to examine whether the effect identified in Study 2* would extend to another population of researchers. An additional goal of Study 2a was to examine whether presenting the same participant with versions of the judgment and choice questions that both attained and failed to attain statistical significance would alter the results and whether the order in which the two versions were presented would materially affect the pattern of results.

Participants:

Participants were the members of the *Cognition* editorial board. A link to our survey was sent via email to the seventy-four members of the *Cognition* editorial board who were

not personal acquaintances or colleagues of the authors and who were not also on the PS editorial board. Thirty-one members completed the survey, yielding a completion rate of 42%.

Procedure:

The editorial board members of *Cognition* were asked four principal questions, two judgment questions and two choice questions, followed by one follow-up question. Participants were randomized to one of two conditions where the first condition saw the two judgment questions with $p = 0.26$ first and $p = 0.01$ second and then the two choice questions with $p = 0.26$ first and $p = 0.01$ second while the second condition saw the two judgment questions with $p = 0.01$ first and $p = 0.26$ second and then the two choice questions with $p = 0.01$ first and $p = 0.26$ second. The judgment question was:

> Below is a summary of a study from an academic paper:
>
> The study aimed to test how two different drugs impact whether a patient recovers from a certain disease. Subjects were randomly assigned to Drug A or Drug B. Fifty-two percent (52%) of patients who took Drug A recovered from the disease while forty-four percent (44%) of patients who took Drug B recovered from the disease (***p* = 0.26**).
>
> Assuming no prior studies have been conducted with these drugs, which of the following statements is most accurate?
>
> A. A person drawn randomly from the same patient population as the patients in the study is **more likely** to recover from the disease if given Drug A than if given Drug B.
> B. A person drawn randomly from the same patient population as the patients in the study is **less likely** to recover from the disease if given Drug A than if given Drug B.
> C. A person drawn randomly from the same patient population as the patients in the study is **equally likely** to recover from the disease if given Drug A or if given Drug B.
> D. It **cannot be determined** whether a person drawn randomly from the same patient population as the patients in the study is more/less/equally likely to recover from the disease if given Drug A or if given Drug B.

After seeing this judgment question, each participant was asked it again but with $p = 0.26$ switched to $p = 0.01$ or vice versa.

After answering the judgment question with $p = 0.26$ and $p = 0.01$, participants were presented with the same summary information but asked to make a choice rather than a judgment:

If you were a patient from the same population as the patients in the study, what drug would you prefer to take to maximize your chance of recovery?

A. I prefer Drug A.

B. I prefer Drug B.

C. I am indifferent between Drug A and Drug B.

Again, after seeing this choice question, each participant was asked it again but with $p = 0.26$ switched to $p = 0.01$ or vice versa. Finally, as a follow-up question, participants were asked the modeling question asked of NEJM authors in Study 1*.

Results:

The pattern of results was not substantially affected by the order in which the p-value was presented (see Section 4). Consequently, we collapse across both order conditions and present our results in Figures 4(a)-4(b) (see Table A2 for data from the modeling question).

The judgment results are consistent with our prediction and the results from Study 2*: participants were substantially more likely to get the question correct when presented with $p = 0.01$ than with $p = 0.26$. In particular, 87% of participants chose option A for the $p = 0.01$ version of the question compared to 35% for the $p = 0.26$ version of the question with 10% choosing option C and 55% choosing option D.

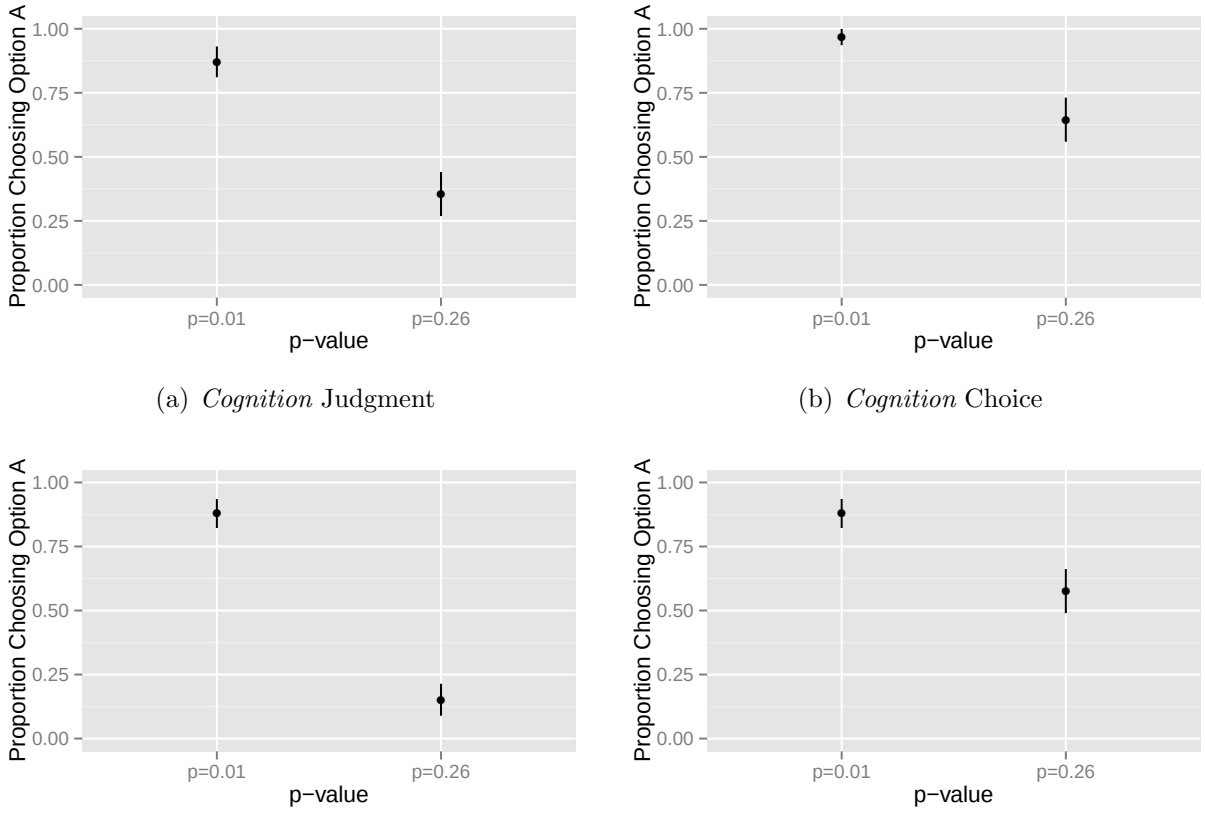
We next examined participants' responses to the choice question. There was no sharp difference between the judgment and choice questions in terms of the proportion of participants choosing option A when $p = 0.01$ (87% versus 97%) but there was when $p = 0.26$ (35% versus 65%). That is, even though participants typically claim no difference between the two drugs in terms of likelihood of recovery from the disease when $p = 0.26$, they still prefer (i.e., choose) Drug A. This finding appears consistent with the idea that making a choice shifts the focus from whether a difference is or is not statistically significant to which choice option is superior. Nonetheless, we find it striking that even when faced with a consequential choice only 65% of participants chose Drug A when the difference between the two drugs failed to attain statistical significance whereas 97% chose it when the difference attained statistical significance.

1.7 Study 2b: *Social Psychology and Personality Science*

Objective:

The main goal of Study 2b was to examine if the effects observed in Study 2* and Study 2a would extend to another population.

Participants:



(a) *Cognition* Judgment

(b) *Cognition* Choice

(c) *Social Psychology and Personality Science* Judgment

(d) *Social Psychology and Personality Science* Choice

Figure 4: Data from Study 2a: *Cognition* and Study 2b: *Social Psychology and Personality Science*. Points denote $\hat{p}_A$, the proportion of participants choosing option A, and lines denote $\hat{p}_A \pm \sqrt{\hat{p}_A(1-\hat{p}_A)/n}$. For both questions, the proportion choosing option A is much lower for the $p = 0.26$ version of the question as compared to the $p = 0.01$ version though this is attenuated for the choice question relative to the judgment question thereby replicating the findings presented in Figure 3.

Participants were the members of the *Social Psychological and Personality Science* (SPPS) editorial board. A link to our survey was sent via email to the ninety-one members of the SPPS editorial board who were not personal acquaintances or colleagues of the authors and who were not also on either the PS or *Cognition* editorial boards; two email addresses were incorrect. Thirty-three board members completed the survey, yielding a completion rate of 37%.

Procedure:

Participants were sent a survey identical to that sent to the editorial board members of *Cognition*.

Results:

Again, the pattern of results was not substantially affected by the order in which the p-value was presented (see Section 4). Consequently, we collapse across both order conditions and present our results in Figures 4(c)-4(d) (see Table A2 for data from the modeling question). As shown in the figures, the pattern of results matched those from Study 2a for the *Cognition* editorial board: only 15% of participants correctly answered the $p = 0.26$ version of the likelihood judgment question compared to 88% for the $p = 0.01$ version while only 58% of participants correctly answered the $p = 0.26$ version of the choice question compared to 88% for the $p = 0.01$ version.

1.8 Study 2c: Economists

Objective:

The main goal of Study 2c was to examine whether the findings of Studies 2⋆-2b would extend to another researcher population, namely economists. An additional goal of study Study 2c was to examine whether adding a posterior probability for the likelihood that one treatment is more effective than the other would moderate our results. We hypothesized that this information would decrease the share of participants selecting the incorrect option when the difference fails to attain statistical significance as it would shift the focus away from statistical significance and towards the posterior probability. A final goal was to examine whether the order in which the judgment and choice questions were presented would materially affect the pattern of results.

Participants:

Participants were the authors of articles published in the 2012 volume of the *American Economic Review* (AER; issues 102.1-102.7) and 2012 volume of the *Journal of Political Economy* (JPE; issues 120.1-120.6). A link to our survey was sent via email to the 553 authors; about fifteen email addresses were incorrect. 176 authors completed the survey, yielding a completion rate of 33%.

Procedures:

Participants were asked two principal questions followed by several follow-up questions. Participants were randomized to one of four conditions following a two by two design. The first factor of the design varied whether the participant was asked a judgment question or a choice question first while the second factor varied the absence or presence of a posterior probability based on a Bayesian calculation. In all cases, the p-value presented was $p = 0.26$.

The judgment and choice questions asked of and response wordings given to the economists were identical to those presented to the editorial board members of *Cognition* and *Social Psychology and Personality Science* with one exception. Rather than presenting the p-value in parentheses, a sentence was inserted on the next line giving the p-value. Thus, the judgment question was:

> Below is a summary of a study from an academic paper:
>
> The study aimed to test how two different drugs impact whether a patient recovers from a certain disease. Subjects were randomly assigned to Drug A or Drug B. Fifty-two percent (52%) of patients who took Drug A recovered from the disease while forty-four percent (44%) of patients who took Drug B recovered from the disease.
>
> A test of the null hypothesis that there is no difference between Drug A and Drug B in terms of probability of recovery from the disease yields a p-value of **0.26**.
>
> Assuming no prior studies have been conducted with these drugs, which of the following statements is most accurate?
>
> A. A person drawn randomly from the same patient population as the patients in the study is **more likely** to recover from the disease if given Drug A than if given Drug B.
> B. A person drawn randomly from the same patient population as the patients in the study is **less likely** to recover from the disease if given Drug A than if given Drug B.
> C. A person drawn randomly from the same patient population as the patients in the study is **equally likely** to recover from the disease if given Drug A or if given Drug B.
> D. It **cannot be determined** whether a person drawn randomly from the same patient population as the patients in the study is more/less/equally likely to recover from the disease if given Drug A or if given Drug B.

After answering the judgment question, participants were presented with the same summary information but asked to make a choice rather than a judgment:

> If you were a patient from the same population as the patients in the study, what drug would you prefer to take to maximize your chance of recovery?

A. I prefer Drug A.

B. I prefer Drug B.

C. I am indifferent between Drug A and Drug B.

Participants randomized to the conditions indicating the presence of the posterior probability saw an additional sentence in the study summary. After the sentence that provided the p-value, they also saw a sentence that read:

> Given the numbers presented above and under mild assumptions, a simple calculation shows that there is an 87% chance that Drug A is more effective than Drug B in terms of probability of recovery from the disease.

The 87% probability is a posterior probability based on a Bayesian calculation that holds for a variety of non-informative priors (see below). As it is calculated based on the proportions and p-value reported in the study summary plus some mild assumptions, it is objectively redundant information (i.e., it adds no information to that present in the original version of the study summary; for a discussion of whether it is perceived as redundant, see below).

After the two principal questions presented above, participants were asked a free response question about their principal research method (e.g., mathematics, econometrics, experiments), a free response question asking at what p-value statistical significance is conventionally defined (94% of participants answered correctly), and the modeling question asked of NEJM authors in Study 1*.

Posterior Probability:

The logic underlying the statement "that there is an 87% chance that Drug A is more effective than Drug B in terms of probability of recovery from the disease" requires a Bayesian calculation. In particular, given the p-value presented in the text of the question (i.e., $p = 0.26$) and assuming the test conducted was the standard test for comparing two proportions, one knows $z = |\Phi(\frac{p}{2})^{-1}| = |\Phi(0.13)^{-1}| = 1.13$ where Φ is the standard normal distribution function and thus

$$z = 1.13 = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\bar{p}\,(1-\bar{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \tag{1}$$

where $\hat{p}_A$ and $\hat{p}_B$ are the observed proportion of individuals who recovered from the disease under Drug A and Drug B respectively (i.e., 52% and 44%), n_A and n_B are the number of subjects assigned to Drug A and Drug B respectively, and $\bar{p} = \frac{\hat{p}_A n_A + \hat{p}_B n_B}{n_A + n_B}$. If one makes an assumption about the ratio of n_A and n_B (i.e., $n_B = c \cdot n_A$), one can then solve the above equation for n_A and thus n_B. For instance, assuming they are equal yields $n_A = n_B = 100$.

If, in addition to the equality assumption, one also takes a Bayesian approach using independent non-informative priors on p_A and p_B (i.e., the parameters governing the proportion of individuals who recover from the disease under Drug A and Drug B respectively), a simple calculation yields a 0.87 posterior probability that $p_A > p_B$ (i.e., that Drug A is more effective that Drug B). This holds under a wide variety of non-informative priors including the Haldane, Jeffreys, and Bayes-Laplace priors (i.e., $p \sim \text{Beta}(a,a)$ with $a = 0.0$, 0.5, and 1.0 respectively) as well as the Zellner prior (i.e., $p \propto p^p(1-p)^{(1-p)}$).

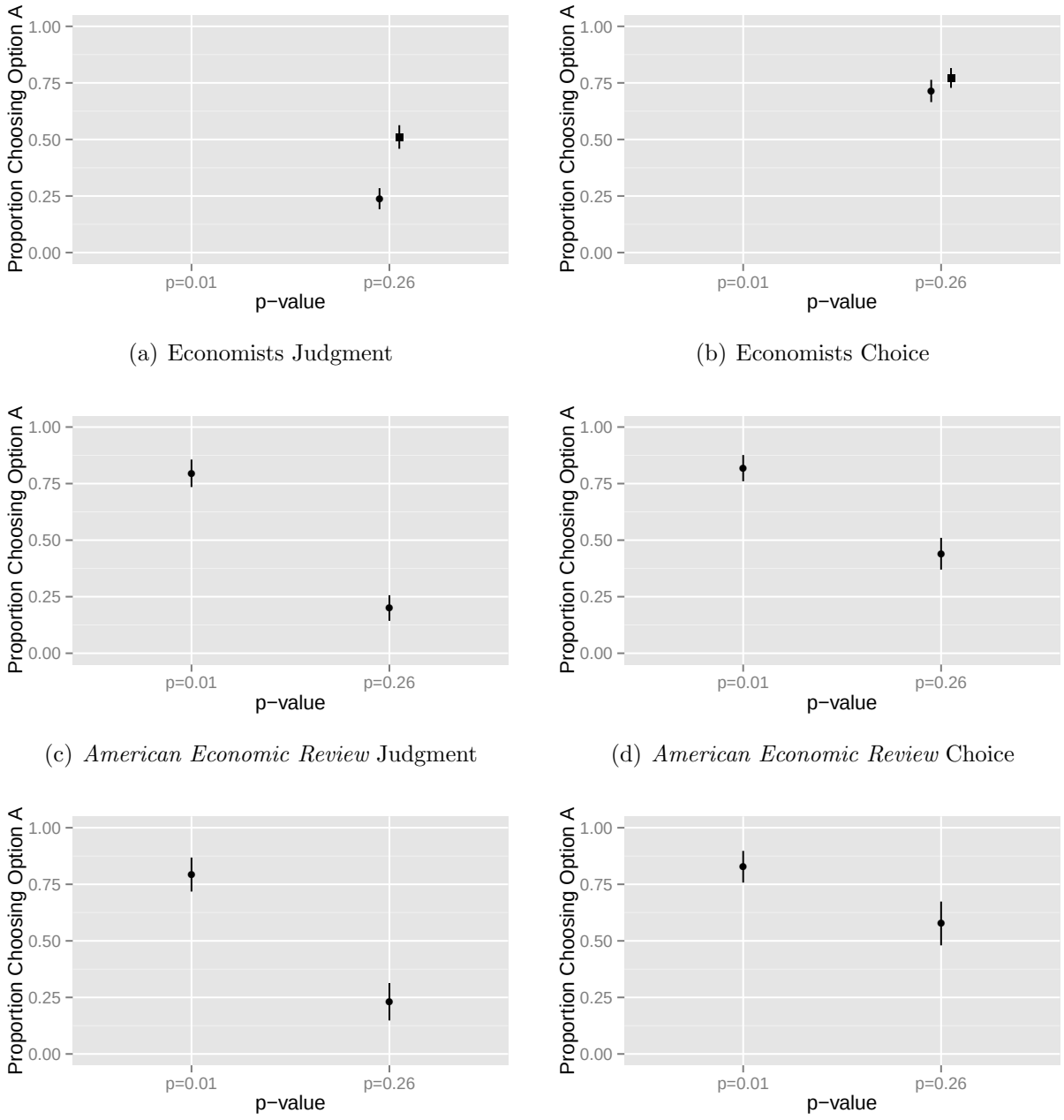
As the prior calculations are based on the information presented in the study summary (i.e., the observed proportion of individuals who recovered from the disease under Drug A and Drug B respectively and the p-value) plus the mild assumption that the sample size in each condition is the same, the result "that there is an 87% chance that Drug A is more effective than Drug B in terms of probability of recovery from the disease" is thus, objectively speaking, redundant information (i.e., it adds no information to that present in the original version of the study summary).

Results:

The pattern of results was not substantially affected by the order in which the judgment and choice questions were asked (see Section 4). Consequently, we collapse across both order conditions and present our results in Figures 5(a)-5(b) (see Table A2 for data from the modeling question). The results are similar to those from the prior studies. That is, in the absence of the posterior probability, 24% of participants answered the judgment question correctly while 71% answered the choice question correctly. Inclusion of the posterior probability increased the share of participants correctly answering the judgment question (24% versus 51%) but did not substantially affect the response to the choice question (71% versus 77%).

These findings suggest that a focus on statistical significance can bias judgments and choices. Further, the finding that the inclusion of a posterior probability strongly attenuates the proportion of participants answering the likelihood judgment question incorrectly is consistent with our hypothesis that the inclusion of a posterior probability can shift attention from a focus on whether a finding is statistically significant or not. That said, we wish to add one qualification to our interpretation of the effect of the posterior probability on participants' responses. Although as noted above the posterior probability is calculated from information already contained in the scenario and is thus objectively redundant information, it may not be perceived as such by participants. Instead, the explicit provision of the posterior probability might effectively be new information for participants (i.e., participants were not previously aware of this information even though it could be derived from other information in the problem). Regardless, as in the case of the choice question, we are struck by the high share of participants answering incorrectly even when being provided with the posterior probability.

As an aside, we note that it has been suggested that posterior probabilities calculated based on non-informative priors as here can be too high in practice [Gelman, 2013].



(a) Economists Judgment

(b) Economists Choice

(c) *American Economic Review* Judgment

(d) *American Economic Review* Choice

(e) *Quarterly Journal of Economics* Judgment

(f) *Quarterly Journal of Economics* Choice

Figure 5: Data from Study 2c: Economists, Study 2d: *American Economic Review* and Study 2e: *Quarterly Journal of Economics*. Points denote $\hat{p}_A$, the proportion of participants choosing option A, and lines denote $\hat{p}_A \pm \sqrt{\hat{p}_A(1-\hat{p}_A)/n}$. The absence (presence) of the posterior probability is indicated by a circle (square). For both questions, the proportion choosing option A is much lower for the $p = 0.26$ version of the question as compared to the $p = 0.01$ version though this is attenuated for the choice question relative to the judgment question thereby replicating the findings presented in Figures 3-4. The proportion choosing option A was larger for the $p = 0.26$ version of the judgment question when the posterior probability was present.

1.9 Study 2d: *American Economic Review*

Objective:

The main goal of Study 2d was to gain additional insight into participants' reasoning when evaluating differences that fail to attain statistical significance. We did this by asking participants to explain their answers in free response form. A secondary goal was to gain additional insight into why the share of participants selecting the inferior option is reduced in the context of making a choice relative to the context of making likelihood judgment. We did this by examining whether participants are more likely to choose a superior alternative when they are choosing for a close other (i.e., a loved one) versus a distant other (i.e., physicians treating patients). As noted above, we predicted that participants would be more likely to choose a superior alternative for a close other than for a distant other when the superior alternative was not statistically significantly different from the inferior alternative. The basis for this prediction was our hypothesis that choice tends to shift the focus away from statistical significance and towards whether an option is superior combined with the logic that this shift would be greater the more consequential the choice for the individual. As a final matter, we wanted to test the effect of a minor wording variant for the third response option of the choice question.

Participants:

Participants were the authors of articles published in the 2013 volume of the AER (issues 103.1-103.5). A link to our survey was sent via email to the 315 authors; thirteen emails addresses were incorrect. Ninety-four researchers completed our survey for a response rate of 31%.

Procedure:

Participants were randomly assigned to one of four conditions following a two by two design. The first level of the design varied whether they saw the $p = 0.01$ versions of the judgment and choice questions or the $p = 0.26$ versions and the second level of the design varied whether the choice question asked them to advise a close or distant other. Participants first encountered the judgment question and then a choice question, both with the same p-value (i.e., either $p = 0.01$ or $p = 0.26$ depending on the condition to which they were assigned).

The judgment question was identical to that encountered by AJE participants in Study $2^{\star}$ (and thus was very similar to that encountered by *Cognition* and SPPS participants in Studies 2a and 2b respectively and to that encountered by AER and JPE participants in Study 2c). In particular, the judgment question was:

> Below is a summary of a study from an academic paper:
>
> The study aimed to test how two different drugs impact whether a patient recovers from a certain disease. Subjects were randomly drawn from a fixed population

and then randomly assigned to Drug A or Drug B. Fifty-two percent (52%) of subjects who took Drug A recovered from the disease while forty-four percent (44%) of subjects who took Drug B recovered from the disease.

A test of the null hypothesis that there is no difference between Drug A and Drug B in terms of probability of recovery from the disease yields a p-value of **0.26**.

Assuming no prior studies have been conducted with these drugs, which of the following statements is most accurate?

A. A person drawn randomly from the same patient population as the patients in the study is **more likely** to recover from the disease if given Drug A than if given Drug B.

B. A person drawn randomly from the same patient population as the patients in the study is **less likely** to recover from the disease if given Drug A than if given Drug B.

C. A person drawn randomly from the same patient population as the patients in the study is **equally likely** to recover from the disease if given Drug A or if given Drug B.

D. It **cannot be determined** whether a person drawn randomly from the same patient population as the patients in the study is more/less/equally likely to recover from the disease if given Drug A or if given Drug B.

After completing the judgment question, participants were asked to explain why they chose the option they chose and were provided with a text box to provide their response. Next, participants were presented with a choice question. As in prior studies, the choice question contained the same study summary as the judgment question. However, as in Study 2*, participants were randomized into one of two conditions: they were asked to choose a treatment either for a close other (i.e., a loved one) or a distant other (i.e., physicians treating patients). The choice question in the close other condition was:

If you were to advise a loved one who was a patient from the same population as those in the study, what drug would you advise him or her to take?

Participants in the distant other condition saw the following wording:

If you were to advise physicians treating patients from the same population as those in the study, what drug would you advise these physicians prescribe for their patients?

All participants then saw the following response options:

A. I would advise Drug A.

B. I would advise Drug B.

C. I would advise that there is no evidence of a difference between Drug A and Drug B.

We note that this is identical to the response wording in Study 2* except that option *C* reads "I would advise that there is *no evidence of a difference* between Drug A and Drug B" in this study as compared to the rather stronger "I would advise that there is *no difference* between Drug A and Drug B" in Study 2*. After completing the choice question, participants were asked to explain why they chose the option they chose and were provided with a text box to provide their response. Finally, participants were asked a free response question about their principal research method (e.g., mathematics, econometrics, experiments) and a free response question asking at what p-value statistical significance is conventionally defined (90% of participants answered correctly).

Results:

As the effect of making a choice for a close versus distant other was a secondary hypothesis, we collapse over both "other" (i.e., close versus distant) conditions in the principal presentation of our results and return to analysis of the effect of a close versus distant other below (see Sections 2 and 4). Data from the principal questions appear in Figures 5(c)-5(d).

Consistent with the pattern observed in our prior studies, 80% of participants answered the judgment question correctly when $p = 0.01$ as compared to 20% when $p = 0.26$ while 82% answered the choice question correctly when $p = 0.01$ as compared to 44% when $p = 0.26$.

For analysis of how choices vary by the party on behalf of which the participant was making a choice, see Section 2. For analysis of participants' explanations for their choices, see Section 3.

1.10 Study 2e: *Quarterly Journal of Economics*

Objective:

In Study 2e, we sought to replicate the findings of Study 2d in another sample of economists. As a secondary matter, we wanted to test the effect of the order in which the questions were asked as well as a minor wording variant for the third response option of the choice question.

Participants:

Participants were the authors of articles published in the 2012 and 2013 volumes of the *Quarterly Journal of Economics* (QJE; issues 127.1-128.3). A link to our survey was sent via email to the 147 authors who were not personal acquaintances or colleagues of the authors; five emails addresses were incorrect. Fifty-five authors completed the survey for a completion rate of 39%.

Procedure:

The design of Study 2e was identical to that of Study 2d, with two exceptions. First, response option C of the choice question was the stronger statement that was also given in Study 2* (i.e., it read "I would advise that there is *no difference* between Drug A and Drug B" in Studies 2* and 2e as opposed to "I would advise that there is *no evidence of a difference* between Drug A and Drug B" in Study 2d). Second, the order in which participants saw the judgment and choice questions was randomized (i.e., some participants saw the judgment question first and then the choice question whereas others saw the choice question first and then the judgment question). Consequently, there were eight conditions in total (i.e., two p-values, two "others" (i.e., close versus distant), and two orders). 93% of participants correctly answered the question asking at what p-value statistical significance is conventionally defined.

Results:

As the effect of making a choice for a close versus distant other was a secondary hypothesis and as the pattern of results was not substantially affected by the order in which the judgment and choice questions were asked, we collapse over both "other" (i.e., close versus distant) and order conditions in the principal presentation of our results (see Sections 2 and 4). Data from the principal questions appear in Figures 5(e)-5(f).

Consistent with our prior studies, a much larger share of the participants answered the judgment question correctly when the p-value was set to 0.01 as opposed to 0.26 (79% versus 23%). The same pattern held for the choice question, although, as before, the difference was attenuated (83% versus 58%).

For analysis of how choices vary by the party on behalf of which the participant was making a choice, see Section 2. For analysis of participants' explanations for their choices, see Section 3.

2 Choice Data

One goal of our study designs was to gain additional insight into why the share of participants selecting the inferior option is reduced in the context of making a choice relative to the context of making a likelihood judgment. We did this by examining whether participants were more likely to choose a superior alternative when they were choosing for the self or for a close other

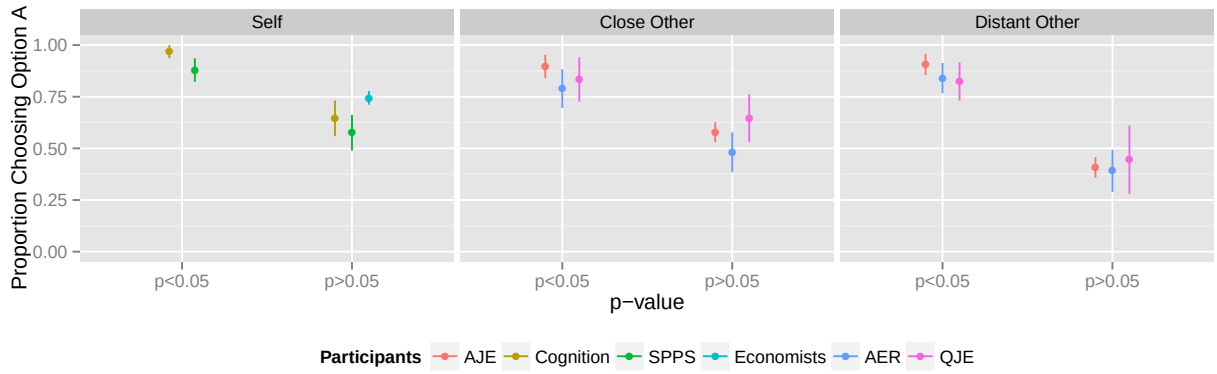


Figure 6: Data from Studies 2 for Choices Made on Behalf of Self versus Close Others versus Distant Others. Points denote $\hat{p}_A$, the proportion of participants choosing option A, and lines denote $\hat{p}_A \pm \sqrt{\hat{p}_A(1-\hat{p}_A)/n}$. The proportion choosing option A is similar for self, close others, and distant others when $p < 0.05$ but it is higher for self and close others as compared to distant others when $p > 0.05$.

(i.e., a loved one) versus when they were choosing for a distant other (i.e., physicians treating patients). Our hypothesis was that there would be no difference in the fraction of participants choosing the superior alternative among the various conditions when the difference presented in the study summary attained statistical significance but that participants in the self and close other conditions would be more likely than those in the distant other condition to choose the superior alternative when the difference presented in the study summary failed to attain statistical significance. The basis for this prediction was our hypothesis that choice tends to shift the focus away from statistical significance and towards whether an option is superior combined with the logic that this shift would be greater the more consequential the choice for the individual. We report our data in Figure 6.

As can be seen, when the difference presented in the study summary attained statistical significance, the vast majority of participants (between 79% and 97% depending on the participant population) chose the superior alternative regardless of whether they were choosing for the self, a close other, or a distant other. Participants were much less likely to choose the superior alternative when the difference presented in the study summary failed to attain statistical significance; nonetheless, they were more likely to do so for the self and close others (between 48% and 74% depending on the participant population) than for distant others (between 39% and 44% depending on the participant population). This finding supports our proposition that making a choice shifts participants' focus from whether a result attains or fails to attain statistical significance to the available evidence, and that this effect is greater the more consequential the choice for the participant.

For further statistical treatment of this issue, see Section 4.

3 Text Response Data

Schema:

We have proposed that statistically trained researchers fail to select the correct alternative (option A) because they are focused on whether a difference attains or fails to attain statistical significance and are thereby blinded to differences that fail to attain statistical significance (we refer to this as the "statistical significance" account). However, at least two alternative explanations are possible. Although in our scenario, Drug A is more likely to be more effective than Drug B, it may be that when responding to the likelihood judgment question, people interpret option D (and possibly also option C) to mean "It has not yet been clearly established that Drug A is (more likely to be) more effective than Drug B" or "There is no very strong reason to conclude that Drug A is (more likely to be) more effective than Drug B." In other words, participants select option C or option D because they are insufficiently certain that Drug A is more likely to be more effective than Drug B on account of the high p-value (we refer to this as the "insufficient certainty" explanation). However, they might nonetheless select Drug A when given a choice because of their correct understanding that, based on the little information available, Drug A is more likely to be more effective than Drug B and hence that choosing Drug A is the optimal strategy.

A related but distinct explanation arises from the fact that some of scenarios studied involved a relatively small difference in the observed proportions of recovery between the two treatments (i.e, 52% versus 44%). Given this small treatment difference and a relatively high p-value, participants might decide to describe a difference as not different based on the reasoning that the proportions are not substantively different. That is, they describe the proportions as the same even though they recognize that they are different. When it comes to acting on the data, however, people select the treatment with the higher likelihood of greater efficacy because they recognize it is superior, even if only slightly. Stated differently, a participant might recognize that Drug A is more likely to be more effective than Drug B in a literal sense, but might believe that, pragmatically, the difference is "not a big deal." However, in a choice context, even a difference that is "not a big deal" might be dispositive if there is no other basis for making a decision. We refer to this explanation as the "practical importance" explanation and note it does not apply to data from the large treatment difference condition of Study $2^{\star}$.

While each of these two alternative explanations lead to answers that are technically incorrect, the answers may or may not reflect an error in reasoning. To the extent that individuals correctly identify evidence as weak and choose to describe it as the absence of evidence because they either lack sufficient certainty in the evidence or believe it to be practically unimportant, it is not necessarily an error in reasoning. However, if either their level of certainty or their belief that a difference is practically important is impacted by whether a p-value has crossed the arbitrary 0.05 threshold or not, this suggests an error in reasoning involving the dichotomization of evidence. Although similar to the error in reasoning that we have posited in that they also arise from dichotomizing evidence, these errors involve the dichotomization of certainty in and practical importance of evidence respectively, rather

than the *per se* existence of evidence. Of note, interpretations of the data based on these explanations imply that the bias might actually be in the $p = 0.01$ condition. That is, a small difference between proportions, which might otherwise be perceived as too small to be practically important or to give one certainty in the evidence, is perceived as practically important or sufficient to give one certainty in the evidence in the $p = 0.01$ condition because the difference attains statistical significance. We examine these explanations by studying the text responses provided by participants in Study $2^{\star}$ and Studies 2d-2e; as the magnitude of the treatment difference was always small in Studies 2d-2e, we begin with those.

Study 2d: American Economic Review

We analyzed the explanations of the forty participants who answered the $p = 0.26$ version of the judgment question incorrectly. Three of explanations could not be categorized, none matched the practical importance account, twenty-eight matched our statistical significance account, three matched the insufficient certainty account (all of these responses belonged to participants who selected option D), and six focused on the idea that power and sample size might have been insufficient to draw a conclusion.

Representative explanations matching our statistical significance account included, "With a p-value of 0.26, we cannot reject the null hypothesis of equality," "Cannot reject the null (at conventional levels of confidence)," "The null that the statistical effect of Drug A is no different than that of Drug B cannot be rejected at conventional levels of statistical significance," "Random assignment and failure to reject null of equal recovery rates," and "p=0.26 for no diff."

A representative explanation matching the insufficient certainty account was "the significance value is not high enough to give me confidence that I had enough information to say one way or the other."

Of the explanations focusing on power and sample size, all belonged to participants who selected option D and appeared to be a justification for choosing option D over option C (i.e., not for choosing option D over option A). For example, one participant stated, "No information on the sample size was provided / Given the p-value and the size of the difference, it seems like the study was underpowered. It is possible that with a larger sample the research would have led to significant differences between the two drugs but it is not possible to know necessarily which way it would have gone." Such a statement suggests that the participant is unwilling to label data as evidence when it does not reach the threshold for statistical significance, which is consistent with our account.

We next analyzed the choice explanations of a subset of the forty participants who failed to choose option A for the judgment question, namely the the fourteen among them who "switched" to Drug A in the choice question. Ten of the fourteen had previously responded with option D to the judgment question and the remaining four had answered with option C to the judgment question. In addition, all three participants whose explanation for the judgment question matched the insufficient certainty account were represented in this group, as were two of the six who gave the power and sample size explanation and two whose

explanation could not be categorized. Most participants expressed either that (i) there was weak evidence for Drug A being more likely to be more effective than Drug B or (ii) Drug A was more likely to be more effective than Drug B despite no statistical evidence of a difference. A response representative of the former included, "There is weak evidence in favour of one of the drugs, so it seems slightly better than picking at random." Responses representative of the latter included, "though insignificant, seems more likely that A is better than B," and "No statistical evidence of difference. Since 52 is non trivially larger than 44, might as well go with A." One participant responded, "I like these questions :-) / Even though it is inconsistent with my previous answer, I would probably not disregards the point estimates at this point..."

In sum, the preponderance of participants' own explanations for their incorrect judgments appear consistent with our account that individuals fail to identify treatment differences because a focus on statistical significance blinds them to the differences. Of those that switch in the choice question, some of the explanations appear consistent with our account for why they switch (i.e., making a consequential choice diminishes the focus on statistical significance).

Study 2e: Quarterly Journal of Economics

We analyzed the explanations of the twenty participants who answered the $p = 0.26$ version of the judgment question incorrectly. Two of explanations could not be categorized, none matched the practical importance account, ten matched our statistical significance account, two matched the insufficient certainty account (all of these responses belonged to participants who selected option D), and three focused on the idea that power and sample size might have been insufficient to draw a conclusion. In addition, two responses had elements of both the statistical significance and insufficient certainty accounts, and one response had elements of both the statistical significance account and a focus on power and sample size considerations. In sum, the explanations largely replicate the findings of Study 2d.

Study 2*: American Journal of Epidemiology

Our account suggests that errors in judgment surrounding evidence that fails to attain statistical significance stems from the dichotomization of evidence. In this context, we note that versions of the insufficient certainty and practical importance explanations that do not imply an error in judgment due to the dichotomization of evidence imply that the degree to which the p-value is above the threshold for statistical significance should affect participants' judgments. In particular, these explanations suggest that the share of participants answering incorrectly should attenuate as the degree to which the p-value is above the threshold for statistical significance shrinks (i.e., share answering incorrectly should shrink as the p-value shrinks). Further, the practical importance explanation suggests the share of participants answering incorrectly should attenuate as the magnitude of the difference between treatments increases. The data from Study 2* can speak to this (see also the Results of Study 2*).

The share of participants in Study 2* who correctly answer each question drops steeply once the p-value falls below 0.05 but is stable thereafter and the magnitude of the treatment

difference has no substantial impact on the fraction answering correctly (see Figure 3). This dichotomization of responses around the conventional threshold for statistical significance is consistent with the notion that dichotomization of evidence into statistically significant and not statistically significant biases researchers' judgments. While such biased judgments are consistent with the statistical significance account, the insufficient certainty account, and the practical importance account, they are are not consistent with versions of the insufficient certainty or practical importance account which suggest that a sense of certainty or practical importance is independent of whether the p-value attains or fails to attain statistical significance. Moreover, the lack of any substantial effect of the magnitude of the treatment difference suggests that, within the range of magnitudes we examined (i.e., a more than a doubling of the magnitude), whether a result attains or fails to attain statistical significance has a far greater impact on the response than the magnitude of the treatment difference.

Of the 159 participants who incorrectly answered the judgment question when the p-value was above 0.05, eighteen could not be categorized, nine most closely matched the insufficient certainty explanation, zero most closely matched the practical importance explanation, sixteen focused predominantly on whether the study sample size gave sufficient power to test the hypothesis, and seventy-six most closely matched the statistical significance explanation. In addition, several explanations matched two or more categories: two responses contained elements of both the insufficient certainty explanation and a focus on power and sample size considerations, twenty-eight contained elements of both the statistical significance explanation and a focus on power and sample size considerations, seven contained elements of both the statistical significance explanation and the insufficient certainty explanation, one had elements of both the statistical significance explanation and the practical importance explanation, and two responses had elements of three different explanations.

Participants who focused on power and sample size considerations in their explanations often appeared to be explaining why they chose option D rather than option C (or, to a lesser extent, vice versa). In particular, these responses tended to note that the results in the scenario might have attained significance if the study had been sufficiently powered. Thus, these explanations still tended to reflect a focus on whether the result was statistically significant, consistent with our theorizing. For example, one representative response was, "The n for the study population was not given. Therefore, if the sample size for the study was small, it may be that the study lacked statistical power to detect an association. If the sample size was sufficiently large, I would have chosen C, instead." Another was, "Given the proximity to 0.50 the drug really made no difference in the two arms. So, my thinking was that about half recover and half do not, and given the p-value, I would not draw an inference of any difference. Third option was tempting. What I would have like to have seen was the sample size..."

In sum, as in Studies 2d-2e, the weight of the evidence from Study 2* appears to support our account that a focus on statistical significance blinds researchers to differences. At the same time, among the participants were some whose certainty or sense of whether a result is practically important appears to be influenced by whether a result attains or fails to attain statistical significance thereby also contributing to the share that answer the judgment

question incorrectly. Conversely, the weight of the evidence does not suggest that a high proportion of researchers are answering incorrectly independent of being influenced by whether a result attains or fails to attain statistical significance.

4 Methods and Results

4.1 Methods

Comparisons in our studies generally come in two forms, (i) those that compare the proportion of a given group of participants who answered a given question correctly to the proportion of another group who answered that same question correctly and (ii) those that compare the proportion of a given group of participants who answered a given question correctly to the proportion of that same group who answered some other question correctly. Statistical tests for such comparisons can be conducted using the standard test for comparing two proportions in combination with the data appearing in the tables in Appendix A.

A limitation of the standard test is that it assumes the two proportions are independent. This is unlikely to be the case for the second kind of comparison discussed above which involves repeated measures data. While there are tests that account for such data (e.g., McNemar's test), we instead model our data in full using versions of the Rasch model [Rasch, 1961, Linacre, 1999] that account for individual-specific variance. In particular, if Y_{ij} is a Bernoulli-distributed random variable indicating whether participant i answered question j correctly, the Rasch model assumes

$$\mathbb{P}(Y_{ij} = 1) = \text{logit}^{-1}(a_i - b_j)$$

where a_i denotes the "ability" of participant i and b_j denotes the "difficulty" of question j. In all of the models that follow, we let $a_i = \beta_i + \theta_i$ where β_i is a fixed effect describing either participant i (e.g., expert in statistics or not) or the version of the survey seen by participant i (e.g., various response wordings) and θ_i is an individual-specific random effect and let b_j be a question-specific fixed effect (although for directional consistency with β_i we parameterize our models in terms of $-b_j$ rather than b_j). The parameters of interest will typically be the β_i and b_j.

Except where otherwise noted, all models were fit in `R` [R Core Team, 2012] using the functions contained in the `lme4` library [Bates et al., 2014]. In particular, estimates and model fit statistics were obtained from `glmer` and (parametric bootstrap) standard errors were obtained from `bootMer`; bootstrap-based standard errors are typically more reliable than the standard asymptotic ones for generalized linear mixed models such as ours. To compensate the computational burden associated with the bootstrap, the faster `nAGQ=0` option was utilized.

Finally, while we agree with Gelman and Hill [2006] that the fixed effects and random effects terminology is problematic and often unclear, we use it here because it is both

	Model 1		Model 2	
Coefficient Estimates & Standard Errors				
Intercept	2.41	0.69	2.63	0.77
p-value=0.27	-4.29	0.64	-4.81	0.77
Wording 2	-0.04	0.71	-0.13	0.85
Wording 3	-0.75	0.76	-0.86	0.81
Self-Identified Statistician			-0.66	3.88
Self-Identified Statistician & p-value=0.27			2.18	4.38
Variance Component Estimates & Standard Errors				
Participant	0.76	0.39	1.01	0.61
Model Fit Statistics				
n	150		150	
Deviance	103.59		99.91	

Table 1: Models fit to Data from Study 1*: *New England Journal of Medicine*.

widespread and clear in the context of the `lme4` library.

4.2 Study 1*: *New England Journal of Medicine*

The models fit to data from *New England Journal of Medicine* authors appear in Table 1. In the first model, the $p = 0.01$ version of the question with wording version one serves as the baseline version and it is typically answered correctly as indicated by the relatively large intercept. On the other hand, participants are very unlikely to get the $p = 0.27$ version of the question correct as indicated by the large negative coefficient. Finally, response wording does not appear to play a major role.

Participant-level heterogeneity is estimated to have a variance of 0.76 indicating that there is considerable variation in participant ability, a feature that holds across all surveys we conducted. The sample size $n = 150$ in tandem with the knowledge that we asked each participant two questions indicates that we had seventy-five participants while the deviance gives minus two times the log-likelihood.

The second model builds on the first model by adding a fixed effect for the eleven participants who indicated an expertise in statistics or biostatistics as well as its interaction with the p-value presented in the question. While the improvement in model fit is small as indicated by the change in deviance, this new model specification allows us to estimate certain quantities of interest. As in the first model, the $p = 0.01$ version of the question with wording version one serves as the baseline version. Further, in this model, those not self-identifying as statisticians or biostatisticians serve as the baseline participants. The results are consistent with the first model. In particular, the large positive estimate for the intercept and the large negative coefficient estimate for the $p = 0.27$ version of the question show that those

who did not indicate an expertise in statistics or biostatistics typically answer the $p = 0.01$ version of the question correctly and the $p = 0.27$ version of the question incorrectly while the wording coefficients suggest that wording does not appear to play a major role. Those who self-identified as statisticians or biostatisticians did not appear to perform substantially better on either question as indicated by the two coefficients. However, unlike those who do not self-identify as statisticians or biostatisticians, those who do get both the $p = 0.01$ and $p = 0.27$ versions of the question correct at a *statistically* similar rate (0.82 and 0.27 respectively for statisticians versus 0.08 and 0.89 respectively for non-statisticians); indeed, the coefficient that compares the performance for statisticians on the $p = 0.27$ version of the question versus the $p = 0.01$ version of the question is estimated at -2.64 (i.e., $-4.81 + 2.18$) with a standard error of 4.36 while the corresponding estimate for non-statisticians is -4.81 with a standard error of 0.77. Nonetheless, even self-identified statisticians did not perform particularly well on the $p = 0.27$ version of the question.

While the second model suggests that statisticians perform somewhat worse than non-statisticians on the $p = 0.01$ version of the question and somewhat better on the $p = 0.27$ version of the question, the paucity of self-identified statisticians as well as the presence of the individual-specific random effects make it difficult to estimate these fixed effects as indicated by the considerable standard errors of the two relevant coefficients. In order to address these comparisons in a different manner, we can make recourse to the standard test for comparing two proportions which is appropriate here since we are focusing on how frequently two different groups of participants answered a given question correctly. As expected, there was no major difference between statisticians and non-statisticians on the $p = 0.01$ version of the question: the proportion of participants who self-identified as statisticians or biostatisticians and answered the $p = 0.01$ version of the question correctly was estimated to be modestly lower than the proportion of participants who self-identified as something other than statisticians and answered the $p = 0.01$ version of the question correctly ($d = 0.82 - 0.89 = -0.07$; $se = 0.11$; here and hereafter d denotes a difference in two proportions and se is given by the denominator of Equation 1). On the other hand, there was a larger difference for the $p = 0.27$ version of the question with those self-identifying as statisticians performing somewhat better ($d = 0.27 - 0.08 = 0.19$, $se = 0.10$). Nonetheless, the general pattern held: both groups of participants performed quite well on the $p = 0.01$ version of the question and quite poorly on the $p = 0.27$ version of the question.

There are three final points worth noting. First, among both statisticians and non-statisticians, anyone who answered the $p = 0.27$ version of the question correctly also answered the $p = 0.01$ version of the question correctly. Second, self-identified statisticians appeared less swayed by the p-value in an absolute sense: they were more likely to give the same answer to both questions ($d = 0.45 - 0.14 = 0.31$; $se = 0.13$). Finally, the results of this paragraph and the prior paragraphs regarding the performance of statisticians versus non-statisticians are necessarily preliminary and tentative given the paucity of self-identified statisticians among our participants; however, it is interesting and hope-inducing (if not entirely surprising) that advanced training in statistics, beyond the rote and recipe-like training typical in introductory undergraduate and graduate courses, may yield improved performance even if the effects are somewhat modest in our data.

	Psychological Science		Marketing Science Institute Young Scholars	
Coefficient Estimates & Standard Errors				
Intercept	1.69	0.49	1.02	0.62
p-value=0.27	-3.79	0.83	-2.92	1.03
Order (p=0.01 First)	0.60	0.86	0.44	1.04
Variance Component Estimates & Standard Errors				
Participant	0.99	0.52	2.09	1.79
Model Fit Statistics				
n	108		54	
Deviance	88.36		55.10	

Table 2: Models fit to Data from Study 1a: *Psychological Science* and Study 1b: Marketing Science Young Scholars.

4.3 Study 1a: *Psychological Science*, Study 1b: Marketing Science Institute Young Scholars, and Study 1c: Undergraduates

The models fit to data from the *Psychological Science* editorial board members and Marketing Science Young Scholars appear in Table 2. The $p = 0.01$ version of the question serves as the baseline version of the question and, for comparison with results from Study 1*: *New England Journal of Medicine*, the survey with $p = 0.27$ first serves as the baseline order. The results are remarkably consistent: researchers typically answer the $p = 0.01$ version of the question correctly and the $p = 0.27$ version of the question incorrectly and order is comparably unimportant.

The models fit to data from undergraduates appear in Table 3. Again, the $p = 0.01$ version of the question serves as the baseline version of the question and, for comparison with results from Study 1*: *New England Journal of Medicine*, the survey with $p = 0.27$ first serves as the baseline order. Further, undergraduates who lack statistical training serve as the baseline participants. The results show that the p-value appearing in the question has no impact on undergraduates who lack statistical training and that order is relatively unimportant. On the other hand, statistical training is associated with an increased likelihood of answering the $p = 0.01$ version of the question correctly but a decreased likelihood of answering the $p = 0.27$ version of the question correctly. Our primary interest in this model is comparing the performance of statistically-trained versus untrained undergraduates on the $p = 0.27$ version of the question; we note our estimate is -1.07 (i.e., $19.64 + -20.71$) with a standard error of 0.57 suggesting that statistical training is associated with decreased performance on this question.

Given the extremely large coefficient estimates on statistical training and its interaction with the p-value of the question, we are reluctant to trust the estimates and believe them to be imprecise despite the relatively small bootstrap standard errors (we note the usual asymptotic

	lme4		Bayesian	
Coefficient Estimates & Standard Errors				
Intercept	1.32	0.48	3.07	1.41
p-value=0.27	0.00	0.44	0.01	0.78
Order (p=0.01 First)	-0.14	0.51	-0.21	1.15
Statistical Training	19.64	0.74	6.76	3.82
Statistical Training & p-value=0.27	-20.71	0.85	-9.32	4.39
Variance Component Estimates & Standard Errors				
Participant	4.02	2.09	20.39	18.83
Model Fit Statistics				
n	148		148	
Deviance	137.82		70.17	

Table 3: Models fit to Data from Study 1c: Undergraduates.

standard errors for these two coefficients are both 2955.98 indicating imprecision). The imprecision in the estimates is an artifact of the data: all statistically-trained undergraduates answered the $p = 0.01$ version of the question correctly and therefore this covariate (i) serves as a perfect predictor for this question for these undergraduates and (ii) induces a negative correlation between the coefficient estimates for statistical training and its interaction with $p = 0.27$.

The issues which cause this imprecision, namely lack of identifiability, collinearity, and instability arising from separation due to perfect predictors, are well known [Albert and Anderson, 1984, Lesaffre and Albert, 1989, Gelman et al., 2008]. Indeed, Gelman et al. [2008] notes "separation is surprisingly common in applied logistic regression, especially with binary predictors, and, as noted by Zorn [2005], is often handled inappropriately." Gelman et al. [2008] suggest (i) standardizing binary input variables to have mean zero and to differ by one (i.e., between their high and low values) and standardizing continuous variables to have mean zero and standard deviation 0.5 [Gelman and Pardoe, 2007, Gelman, 2008] and (ii) performing a Bayesian analysis with an independent Cauchy prior with center zero and scale 2.5 on each predictor (except the intercept which has a Cauchy prior with center zero and scale 10). While Gelman et al. [2008] provide an approximate EM algorithm to obtain point estimates and standard errors, they note that their approach "can be used as part of a fully Bayesian computation in more complex settings such as hierarchical models." As this reflects our setting, it is the approach we take in order to more fully address the problem posed by the imprecise estimates presented in Table 3. In particular, we (i) standardize the input variables and let the "fixed effects" β_i and b_j follow independent, scaled Cauchy priors, (ii) let the "random effects" θ_i be independently and identically distributed Normal$(0, \sigma_\theta^2)$, and (iii) let the participant-level standard deviation σ_θ be distributed Uniform$(0, 100)$ following the recommendation of Gelman [2006] for variance parameters in hierarchical models; this in combination with our Bernoulli likelihood fully specifies our Bayesian model.

Estimates (in particular, posterior means and standard deviations) obtained from our

fully Bayesian model are presented in Table 3. As can be seen, the two estimates of the intercept, the main effect of the p-value appearing in the question, and the order of the question are reasonably similar. On the other hand, the prior serves to regularize the two rather extreme estimates obtained using maximum likelihood techniques, and, thus, the corresponding Bayesian estimates, while still large in absolute value, are both more reasonable and are estimated with reasonable precision. In terms of implications for our research question, the results are qualitatively the same. Again, the p-value appearing in the question has no impact on undergraduates who lack statistical training and order is relatively unimportant. On the other hand, statistical training is associated with increased likelihood of answering the $p = 0.01$ version of the question correctly but decreased likelihood of answering the $p = 0.27$ version of the question correctly; in particular, we estimate the sum of the two relevant coefficient at -2.56 (i.e., $6.67 + -9.32$) with a standard deviation of 1.70 and a 0.97 posterior probability of being less than zero; this estimate is substantially larger in magnitude than the classical estimate of -1.07.

In sum, our model fits show that (i) statistically-untrained undergraduates perform equally well on $p = 0.01$ version of the question and the $p = 0.27$ version of the question, (ii) statistically-trained undergraduates perform better on the $p = 0.01$ version of the question than the $p = 0.27$ version of the question, and that (iii) statistically-trained undergraduates perform better (worse) on the $p = 0.01$ ($p = 0.27$) question as compared to statistically-untrained undergraduates. As the last set of comparisons compares the performance of the two different groups on a given question, we can also turn to the standard test for comparing two proportions. On the $p = 0.01$ version of the question, the proportion of statistically-trained undergraduates who answered correctly was estimated to be substantially larger ($d = 1.00 - 0.73 = 0.27$; $se = .09$) while on the $p = 0.27$ version of the question it was estimated to be substantially smaller ($d = 0.53 - 0.73 = -0.19$; $se = 0.11$). Also of interest is comparing our statistically-untrained undergraduates to the academic researchers who saw the same response wording. On the $p = 0.01$ version of the question, the proportion of statistically-untrained undergraduates who answered correctly was estimated to be lower as compared to *Psychological Science* editorial board members ($d = 0.73 - 0.87 = -0.14$; $se = 0.08$) and Marketing Science Institute Young Scholars ($d = 0.73 - 0.74 = -0.01$; $se = 0.11$). On the other hand, on the $p = 0.27$ version of the question, the proportion of statistically-untrained undergraduates who answered correctly was estimated to be higher as compared to *Psychological Science* editorial board members ($d = 0.73 - 0.17 = 0.56$; $se = 0.10$) and Marketing Science Institute Young Scholars ($d = 0.73 - 0.19 = 0.54$; $se = 0.12$). Thus, statistically-untrained undergraduates perform the best of the four groups on the $p = 0.27$ version of the question.

There are three additional facts that are potentially of note. First, across both questions asked of the four groups of participants who saw the same response wording, it appears statistically-trained undergraduates performed best. Second, even if we had asked participants to make a statistical inference from the data under the NHST paradigm rather than to simply describe the data, option C is never correct: failure to reject the null hypothesis does not imply or prove that the two treatments do not differ. No statistically-untrained undergraduate and only one statistically-trained undergraduate made this mistake for the

$p = 0.27$ version of the question whilst a considerable number of *Psychological Science* editorial board members and Marketing Science Young Scholars (and, for that matter, *New England Journal of Medicine* authors) did. Third, and again assuming we were asking an inferential question rather than a descriptive question, there is a sense in which option D is the correct answer regardless of the p-value; indeed "God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p" [Rosnow and Rosenthal, 1989] and thus at no p is the null definitively overturned. However, only a relatively small proportion of participants chose option D as their response to both versions of the question (i.e., the $p = 0.01$ version and the $p = 0.27$ version), with most choosing option A for the $p = 0.01$ version and option C or option D for the $p = 0.27$ version (except for the statistically-untrained undergraduates who mostly chose option A for both versions).

4.4 Study 2*: *American Journal of Epidemiology*

The models fit to data from *American Journal of Epidemiology* authors appear in Table 7. The $p = 0.025$ version of the judgment question with a small treatment difference serves as the baseline version of the question. Unlike the prior surveys which asked descriptive statements about what happened in the actual study, this survey asked participants to make a judgment about which drug is likely to be more effective based on the data and to make a choice as to what drug they would choose. As the close versus distant other condition was applicable only to the choice question, we include a coefficient for this in the model so that the close other condition serves as the baseline version of the choice question; we note our hypothesis regarding the effect of a close versus distant other on choice is for a nonzero effect only when $p > 0.05$.

The second model builds on the first model by adding a fixed effect for the thirty-four participants who indicated an expertise in statistics or biostatistics as well as relevant interactions. As with the model fit to data from Study 1*: *New England Journal of Medicine*, the improvement in model fit resulting from these additional effects is small as indicated by the change in deviance; however, this second model specification allows us to estimate certain quantities of interest.

Given the similarity in coefficient estimates between the two models, we discuss only the second model. The coefficients for the first three batches of coefficients (i.e., those pertaining to the judgment question, those pertaining to the choice question, and those pertaining to the magnitude of the treatment difference) are, unsurprisingly, entirely consistent with the data displayed in Figure 3. In particular, (i) participants perform much worse on both the judgment and choice questions when $p > 0.05$ versus when $p < 0.05$, (ii) participants perform better on the choice question relative to the judgment question when $p > 0.05$, and (iii) the magnitude of the treatment difference has no substantial impact.

In sum, participants appear to make judgments about data in the same manner that they assess statements about descriptive statistics: they perform poorly when $p > 0.05$ and well when $p < 0.05$. Further, when $p > 0.05$, they perform much better when asked to make

	Model 1		Model 2	
Coefficient Estimates & Standard Errors				
Intercept	0.98	0.46	1.33	0.51
p-value=0.075 Judgment	-2.95	0.60	-3.25	0.65
p-value=0.125 Judgment	-2.50	0.59	-2.85	0.64
p-value=0.175 Judgment	-2.92	0.61	-3.24	0.64
p-value=0.025 Choice	1.42	2.11	0.90	2.16
p-value=0.075 Choice	-0.43	0.58	-0.94	0.65
p-value=0.125 Choice	-0.42	0.58	-0.83	0.64
p-value=0.175 Choice	-0.80	0.59	-1.23	0.64
Large Difference	0.65	0.65	0.50	0.67
Large Difference Choice	0.30	2.62	0.55	2.73
Large Difference & p-value > 0.05	-0.41	0.76	-0.24	0.77
Large Difference Choice & p-value > 0.05	-0.35	2.66	-0.63	2.76
Distant Choice	-0.32	2.17	-0.21	2.16
Distant Choice & p-value > 0.05	-0.76	2.20	-0.84	2.18
Self-Identified Statistician			-1.13	0.92
Self-Identified Statistician Choice			1.64	5.23
Self-Identified Statistician & p-value > 0.05			0.71	1.54
Self-Identified Statistician Choice & p-value > 0.05			-0.30	5.36
Variance Component Estimates & Standard Errors				
Participant	3.11	0.70	3.25	0.74
Model Fit Statistics				
n	522		522	
Deviance	553.80		547.40	

Table 4: Models fit to Data from Study 2*: *American Journal of Epidemiology.*

a choice based on the data.

We note that there is also strong support for our hypothesis concerning the effect of a close versus distant other on choice when $p > 0.05$: participants are more likely to answer the choice question correctly when $p > 0.05$ when asked to make a choice for a close versus distant other ($d = 0.58 - 0.41 = 0.17$; $se = 0.07$; the model coefficient estimate is $-0.21 + -0.84 = -1.06$ with a standard error of 0.31) while there is no substantial difference when $p < 0.05$ ($d = 0.90 - 0.91 = -0.01$; $se = 0.08$; the model coefficient estimate is -0.21 with a standard error of 2.16).

As noted above, we added a fixed effect for the thirty-four participants who indicated an expertise in statistics or biostatistics as well as relevant interactions just as we did for Study 1*: *New England Journal of Medicine*. Such fixed effects are particularly important for making comparisons of the performance of a single group (i.e., those who did versus

did not indicate an expertise in statistics or biostatistics) across different questions. For this study, unlike in the case of Study 1*: *New England Journal of Medicine*, we are not as interested in such comparisons (for completeness, however, we note that those who did versus did not indicate an expertise in statistics or biostatistics performed similarly in terms of the relevant comparisons; in particular, those who received the $p = 0.025$ versions of the judgment and choice questions performed similarly on the two questions regardless of whether they had or not not indicated an expertise in statistics or biostatistics while those who received the $p > 0.05$ versions of the questions performed comparably better on the choice question as compared to the judgment question regardless of their expertise). Of more interest in this case is the absolute performance of those indicating an expertise in statistics or biostatistics as compared to those who did not noting that the hypothesis would be that those who indicated an expertise would perform similarly to those who did not on the $p < 0.05$ versions of the questions but better on the $p > 0.05$ versions of the questions. The two groups do indeed perform relatively similarly on the $p < 0.05$ versions of the questions ($d = 0.58 - 0.80 = -0.21$; $se = 0.13$ for the judgment question; $d = 0.92 - 0.90 = 0.02$; $se = 0.10$ for the choice question; model coefficient estimates have relatively large standard errors), and, further, they perform similarly on the $p > 0.05$ versions of the judgment question ($d = 0.14 - 0.21 = -0.08$; $se = 0.09$; model coefficient estimates have relatively large standard errors). On the $p > 0.05$ versions of the choice question, there is weak evidence that those who indicated an expertise in statistics or biostatistics perform better ($d = 0.68 - 0.47 = 0.21$; $se = 0.11$; the model coefficient estimate is $-1.13 + 1.64 + 0.71 + -0.30 = 0.92$ with a standard error of 0.58). We emphasize the results of this paragraph are necessarily preliminary and tentative given the paucity of self-identified statisticians among our participants; however, it is interesting and hope-inducing (if not entirely surprising) that advanced training in statistics, beyond the rote and recipe-like training typical in introductory undergraduate and graduate courses, again may yield improved performance however modest.

4.5 Study 2a: *Cognition* and Study 2b: *Social Psychology and Personality Science*

The models fit to data from the *Cognition* and *Social Psychology and Personality Science* editorial board members appear in Table 5. For comparison with results from Study 2*: *American Journal of Epidemiology*, the $p = 0.01$ version of the judgment question serves as the baseline version of the question as this most closely matches the baseline version of the question used in the models fit to data from that study. For comparison with results from Study 1*: *New England Journal of Medicine* where the $p = 0.27$ version of the question was presented before the $p = 0.01$ version of the question, the survey with $p = 0.26$ first serves as the baseline order.

The same pattern of results observed before emerges. First, for the judgment questions, participants typically answer the $p = 0.01$ version correctly and the $p = 0.26$ version incorrectly. However, when participants are asked to make a choice, the results change substantially. When participants are asked to make a choice with the $p = 0.01$ version of the

	Cognition		*Social Psychology and Personality Science*	
Coefficient Estimates & Standard Errors				
Intercept	2.91	1.42	3.57	1.57
p-value=0.26 Judgment	-3.37	1.28	-6.33	1.94
p-value=0.01 Choice	1.81	5.47	0.00	1.48
p-value=0.26 Choice	-1.71	1.15	-2.94	1.42
Order (p=0.01 First)	-0.54	1.00	-0.42	1.08
Variance Component Estimates & Standard Errors				
Participant	3.81	4.65	9.58	0.73
Model Fit Statistics				
n	124		132	
Deviance	103.91		104.18	

Table 5: Models fit to Data from Study 2a: *Cognition* and Study 2b: *Social Psychology and Personality Science*.

choice question, there is not much difference relative to the $p = 0.01$ version of the judgment question (the model coefficient estimates are 1.81 with a standard error of 5.47 and 0.00 with a standard error of 1.60 respectively); this is not surprising as a large majority of participants give the correct response to the judgment question with $p = 0.01$. Second, when participants are asked to make a choice with the $p = 0.26$ version of the choice question, they do much better than on the $p = 0.26$ version of the judgment question; we estimate the model coefficient at 1.66 (i.e., $-1.71 - -3.37$) with a standard error of 0.69 for the *Cognition* editorial board members and 3.39 (i.e., $-2.94 - -6.33$) with a standard error of 0.98 for the *Social Psychology and Personality Science* editorial board members. Third, as in prior surveys, order does not play a major role.

4.6 Study 2c: Economists

The model fit to data from *American Economic Review* and *Journal of Political Economy* authors appears in Table 6. As all questions in this survey used only $p = 0.26$, the $p = 0.26$ version of the judgment question serves as the baseline version of the question, and, for comparison with results from Study 2*: *American Journal of Epidemiology* in which the judgment question was presented first and the choice question second, the survey with the judgment question first serves as the baseline order. In line with prior results, participants presented with a p-value above 0.05 were more likely to answer correctly when they were asked to make a choice rather than a judgment. Further, the presence of the 0.87 posterior probability that Drug A is more effective than Drug B increased the likelihood that participants would respond correctly; however, the impact was comparatively minor for the choice question while dramatic for the judgment question (i.e., the model coefficient is estimated to be $2.83 - 2.49 = 0.33$ with a standard error of 0.37 for the choice question and 1.39

Coefficient Estimates & SEs		
Intercept	-1.12	0.30
Judgment & Posterior	1.39	0.36
Choice	2.49	0.34
Choice & Posterior	2.83	0.39
Order (Choice First)	-0.51	0.28
Variance Component Estimates & SEs		
Participant	2.39	0.64
Model Fit Statistics		
n	352	
Deviance	402.20	

Table 6: Model fit to Data from Study 2c: Economists.

with a standard error of 0.36 for the judgment question). Consequently, it appears that the presence of a posterior probability can at least partially override the tendency of participants perform poorly when $p = 0.26$. This suggests that objectively redundant information like the posterior probability can, if framed differently, substantially impact participants' judgments (though we reiterate the qualification made previously that the explicit provision of the posterior probability might effectively be new information for participants). Finally, order did not play a large role.

As before, we can also use the standard test for comparing two proportions to compare the performance of those who saw the posterior probability versus those who did not on each question. On the judgment question, the proportion of those who saw the posterior probability who answered correctly was substantially larger ($d = 0.51 - 0.24 = 0.27$; $se = 0.07$) while on the choice question it was similar ($d = 0.77 - 0.71 = 0.07$; $se = .07$).

4.7 Study 2d: *American Economic Review* and Study 2e: *Quarterly Journal of Economics*

The models fit to data from *American Economic Review* and *Quarterly Journal of Economics* authors appear in Table 7. For comparison with results from Study 2*: *American Journal of Epidemiology*, Study 2b: *Cognition* and Study 2c: *Social Psychology and Personality Science*, the $p = 0.01$ version of the judgment question serves as the baseline version of the question for both sets of authors; for comparison with results from Study 2*: *American Journal of Epidemiology*, Study 2a: *Cognition*, Study 2b: *Social Psychology and Personality Science*, Study 2c: Economists, and Study 2d: *American Economic Review*, the survey with the judgment question first serves as the baseline order for Study 2e: *Quarterly Journal of Economics* authors.

The patterns observed in prior surveys hold here. Participants typically answer the $p = 0.01$ version of the judgment correctly and the $p = 0.26$ version incorrectly. When

	American Economic Review		*Quarterly Journal of Economics*	
Coefficient Estimates & Standard Errors				
Intercept	1.66	0.43	1.67	0.64
p-value=0.26 Judgment	-3.43	0.60	-2.92	0.85
p-value=0.01 Choice	0.34	1.63	0.19	3.38
p-value=0.26 Choice	-1.60	0.63	-0.90	0.79
Distant Choice	-0.27	1.78	0.12	3.96
Distant Choice & p-value=0.26	-0.45	1.89	-0.77	4.18
Order (Choice First)			-0.28	0.58
Coefficient Estimates & Standard Errors				
Participant	3.73	1.81	2.17	2.03
Model Fit Statistics				
n	188		110	
Deviance	194.09		114.84	

Table 7: Models fit to Data from Study 2d: *American Economic Review* and Study 2e: *Quarterly Journal of Economics*.

asked to make a choice with the $p = 0.01$ version of the choice question, there is not much difference relative to the $p = 0.01$ version of the judgment question. On the other hand, when participants are asked to make a choice with the $p = 0.26$ version of the choice question, they do better than on the $p = 0.26$ version of the judgment question; we note this effect is substantially greater for those asked to make a choice on behalf of a close versus distant other (because those asked to make a choice on behalf of a close other are observed to choose the superior alternative more frequently than those asked to make a choice on behalf of a distant other when $p = 0.26$; see Table A9). There appears to be rather weak support for our hypothesis that participants are more likely to choose the superior alternative when making choices on behalf of a close as opposed to distant other when $p = 0.26$ ($d = 0.48 - 0.39 = 0.09$; $se = 0.14$; the model coefficient estimate is $-0.27 + -0.45 = -0.72$ with a standard error of 0.66 for *American Economic Review* authors; $d = 0.65 - 0.44 = 0.20$; $se = 0.20$; the model coefficient estimate is $0.12 + -0.77 = -0.65$ with a standard error of 1.38 for *Quarterly Journal of Economics* authors). Finally, as in prior surveys, order does not play a major role. In sum, we replicate our findings that participants appear to make judgments about data in the same manner that they assess statements about descriptive statistics and that they appear to perform much better when asked to make choices based on the data when $p = 0.26$.

4.8 Extensions

We note that, in addition to the models presented here, we considered several generalizations. For instance, we considered interactions of, for example, the p-value presented in the question with the wording variations (for the survey sent to *New England Journal of Medicine* authors)

or with the survey question order (for several of the other surveys). Our findings were robust to these variations in our model specifications.

References

A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71:1–10, 1984.

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. *lme4: Linear mixed-effects models using Eigen and S4*, 2014. URL `http://CRAN.R-project.org/package=lme4`. R package version 1.1-7.

Geoffrey T. Fong, David H. Krantz, and Richard E. Nisbett. The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, 18:253–292, 1986.

David Freedman, Robert Pisani, and Roger Purves. *Statistics*. W. W. Norton and Company, New York, 4 edition, 2007.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:515–533, 2006.

Andrew Gelman. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27:2865–2873, 2008.

Andrew Gelman. p values and statistical practice. *Epidemiology*, 24(1):69–72, 2013.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, NY, 2006.

Andrew Gelman and Iain Pardoe. Average predictive comparisons for models with nonlinearity, interactions, and variance components. *Sociological Methodology*, 37(1):23–51, 2007.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2(4):1360–1383, 2008.

E. Lesaffre and A. Albert. Partial separation in logistic discrimination. *Journal of the Royal Statistical Society, Series B*, 51:109–116, 1989.

John M. Linacre. Understanding rasch measurement: Estimation methods for rasch measures. *Journal of Outcome Measurement*, 3(4):382–405, 1999.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Georg Rasch. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. IV*, pages 321–333. University of California Press, 1961.

Ralph L. Rosnow and Robert Rosenthal. Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44(10):1276–1284, 1989.

C. Zorn. A solution to separation in binary response models. *Political Analysis*, 13:157–170, 2005.

A Data Tables

Option	A	B	C	D
A	2	0	0	0
B	0	0	0	0
C	10	0	0	1
D	7	0	0	0

(a) Wording 1

Option	A	B	C	D
A	5	0	0	0
B	0	0	0	0
C	7	0	0	1
D	7	0	0	3

(b) Wording 2

Option	A	B	C	D
A	1	0	0	0
B	0	0	0	0
C	18	0	2	0
D	9	1	0	1

(c) Wording 3

Table A1: Data from Study 1⋆: *New England Journal of Medicine*. Each cell gives the number of participants who gave the response indicated by the row to the $p = 0.27$ version of the question and the response indicated by the column to the $p = 0.01$ version of the question.

Study	Option			
	A	B	C	D
Study 1⋆: *New England Journal of Medicine*	35	5	2	36
Study 2a: *Cognition*	14	3	2	12
Study 2b: *Social Psychology and Personality Science*	16	3	1	13
Study 2c: Economists	81	5	6	84

Table A2: Data from Modeling Question. Each cell gives the number of participants from the study indicated by the row who who gave the response indicated by the column.

Option	A	B	C	D
A	9	0	0	0
B	0	0	0	0
C	18	0	2	0
D	20	0	0	5

(a) *Psychological Science*

Option	A	B	C	D
A	5	0	0	0
B	0	0	0	0
C	3	0	0	1
D	12	0	0	6

(b) Marketing Science Institute Young Scholars

Option	A	B	C	D
A	16	0	0	0
B	0	0	0	0
C	1	0	0	0
D	13	0	0	0

(c) Statistically-Trained Undergraduates

Option	A	B	C	D
A	28	2	0	2
B	0	0	1	0
C	0	0	0	0
D	4	0	0	7

(d) Statistically-Untrained Undergraduates

Table A3: Data from Study 1a: *Psychological Science*, Study 1b: Marketing Science Institute Young Scholars, and Study 1c: Undergraduates. Each cell gives the number of participants who gave the response indicated by the row to the $p = 0.27$ version of the question and the response indicated by the column to the $p = 0.01$ version of the question.

Option	A	B	C
A	7	0	0
B	0	0	0
C	1	0	1
D	4	0	1

(a) Close Other, Small Difference, $p = 0.025$

Option	A	B	C
A	2	0	0
B	0	0	0
C	2	0	3
D	8	0	2

(b) Close Other, Small Difference, $p = 0.075$

Option	A	B	C
A	0	0	1
B	0	0	0
C	4	0	4
D	6	0	2

(c) Close Other, Small Difference, $p = 0.125$

Option	A	B	C
A	2	0	0
B	0	0	0
C	2	0	2
D	4	0	5

(d) Close Other, Small Difference, $p = 0.175$

Option	A	B	C
A	13	0	0
B	0	0	0
C	0	0	0
D	1	0	1

(e) Close Other, Large Difference, $p = 0.025$

Option	A	B	C
A	3	0	0
B	0	0	0
C	2	0	3
D	4	0	4

(f) Close Other, Large Difference, $p = 0.075$

Option	A	B	C
A	4	0	0
B	1	0	0
C	0	0	2
D	4	0	5

(g) Close Other, Large Difference, $p = 0.125$

Option	A	B	C
A	2	0	1
B	0	0	0
C	0	0	3
D	9	0	6

(h) Close Other, Large Difference, $p = 0.175$

Option	A	B	C
A	13	0	1
B	0	0	0
C	0	0	1
D	1	0	0

(i) Distant Other, Small Difference, $p = 0.025$

Option	A	B	C
A	2	0	1
B	0	0	0
C	0	0	2
D	2	0	8

(j) Distant Other, Small Difference, $p = 0.075$

Option	A	B	C
A	6	0	1
B	0	0	0
C	1	0	2
D	0	0	5

(k) Distant Other, Small Difference, $p = 0.125$

Option	A	B	C
A	3	0	0
B	0	0	0
C	0	0	8
D	2	0	4

(l) Distant Other, Small Difference, $p = 0.175$

Option	A	B	C
A	12	0	0
B	0	0	0
C	1	0	0
D	2	0	1

(m) Distant Other, Large Difference, $p = 0.025$

Option	A	B	C
A	3	0	1
B	0	0	0
C	2	0	5
D	4	0	3

(n) Distant Other, Large Difference, $p = 0.075$

Option	A	B	C
A	3	0	1
B	0	0	0
C	0	0	3
D	5	0	5

(o) Distant Other, Large Difference, $p = 0.125$

Option	A	B	C
A	4	0	1
B	0	0	0
C	0	0	3
D	3	0	5

(p) Distant Other, Large Difference, $p = 0.175$

Table A4: Data from Study 2*: *American Journal of Epidemiology*. Each cell gives the number of participants who gave the response indicated by the row to the judgment question and the response indicated by the column to the choice question. The close versus distant other distinction applies only to the choice question.

Option	A	B	C	D
A	11	0	0	0
B	0	0	0	0
C	3	0	0	0
D	13	0	0	4

(a) *Cognition* Judgment

Option	A	B	C
A	20	0	0
B	0	0	0
C	10	0	1

(b) *Cognition* Choice

Option	A	B	C	D
A	5	0	0	0
B	0	0	0	0
C	10	0	0	1
D	14	0	0	3

(c) *Social Psychology and Personality Science* Judgment

Option	A	B	C
A	19	0	0
B	0	0	0
C	10	0	4

(d) *Social Psychology and Personality Science* Choice

Table A5: Data from Study 2a: *Cognition* and Study 2b: *Social Psychology and Personality Science*. Each cell gives the number of participants who gave the response indicated by the row to the $p = 0.27$ version of the question and the response indicated by the column to the $p = 0.01$ version of the question.

Option	A	B	C
A	19	0	1
B	0	0	0
C	5	0	10
D	36	0	13

(a) Posterior Probability Absent

Option	A	B	C
A	44	0	3
B	0	0	0
C	6	0	8
D	21	0	10

(b) Posterior Probability Present

Table A6: Data from Study 2c: Economists. Each cell gives the number of participants who gave the response indicated by the row to the judgment question and the response indicated by the column to the choice question.

Option	A	B	C
A	13	0	0
B	0	0	0
C	0	0	3
D	2	0	1

(a) Close Other, $p = 0.01$

Option	A	B	C
A	3	0	1
B	0	0	0
C	4	0	6
D	6	0	7

(b) Close Other, $p = 0.26$

Option	A	B	C
A	19	0	3
B	0	0	0
C	0	0	0
D	2	0	1

(c) Distant Other, $p = 0.01$

Option	A	B	C
A	5	0	1
B	0	0	0
C	0	0	3
D	4	0	10

(d) Distant Other, $p = 0.26$

Table A7: Data from Study 2d: *American Economic Review*. Each cell gives the number of participants who gave the response indicated by the row to the judgment question and the response indicated by the column to the choice question. The close versus distant other distinction applies only to the choice question.

Option	A	B	C
A	9	0	1
B	0	0	0
C	0	0	0
D	1	0	1

(a) Close Other, $p = 0.01$

Option	A	B	C
A	4	0	1
B	0	0	0
C	2	0	1
D	5	0	4

(b) Close Other, $p = 0.26$

Option	A	B	C
A	12	0	1
B	0	0	0
C	0	0	0
D	2	0	2

(c) Distant Other, $p = 0.01$

Option	A	B	C
A	1	0	0
B	0	0	0
C	0	0	2
D	3	0	3

(d) Distant Other, $p = 0.26$

Table A8: Data from Study 2e: *Quarterly Journal of Economics*. Each cell gives the number of participants who gave the response indicated by the row to the judgment question and the response indicated by the column to the choice question. The close versus distant other distinction applies only to the choice question.

Other	Study	$p < 0.05$		$p > 0.05$	
		$\hat{p}_A$	n	$\hat{p}_A$	n
Self	Study 2a: *Cognition*	97	31	65	31
	Study 2b: SPPS	88	33	58	33
	Study 2c: Economists			74	176
Close	Study 2*: AJE	90	29	58	102
	Study 2d: AER	79	19	48	27
	Study 2e: QJE	83	12	65	17
Distant	Study 2*: AJE	91	32	41	98
	Study 2d: AER	84	25	39	23
	Study 2e: QJE	82	17	44	9

Table A9: Data from Studies 2 for Choices Made on Behalf of Self versus Close Others versus Distant Others. Each cell the percentage of participants who chose option A or the sample size. The proportion choosing option A is similar for self, close others, and distant others when $p < 0.05$ but it is higher for self and close others as compared to distant others when $p > 0.05$.