Judgment and Decision Making, Vol. 14, No. 6, November 2019, pp. 671–682

How the public, and scientists, perceive advancement of knowledge from conflicting study results

Derek J. Koehler* Gordon Pennycook†

Abstract

Science often advances through disagreement among scientists and the studies they produce. For members of the public, however, conflicting results from scientific studies may trigger a sense of uncertainty that in turn leads to a feeling that nothing new has been learned from those studies. In several scenario studies, participants read about pairs of highly similar scientific studies with results that either agreed or disagreed, and were asked, "When we take the results of these two studies together, do we now know more, less, or the same as we did before about (the study topic)?" We find that over half of participants do not feel that "we know more" as the result of the two new studies when the second study fails to replicate the first. When the two study results strongly conflict (e.g., one finds a positive and the other a negative association between two variables), a non-trivial proportion of participants actually say that "we know less" than we did before. Such a sentiment arguably violates normative principles of statistical and scientific inference positing that new study findings can never reduce our level of knowledge (and that only completely uninformative studies can leave our level of knowledge unchanged). Drawing attention to possible moderating variables, or to sample size considerations, did not influence people's perceptions of knowledge advancement. Scientist members of the American Academy of Arts and Sciences, when presented with the same scenarios, were less inclined to say that nothing new is learned from conflicting study results.

Keywords: knowledge, science, inference, conflicting results

1 Introduction

In today's world it can be difficult to figure out what is true. Online it seems impossible to find any statement of fact that is not disputed somewhere, by somebody. Is the earth really (roughly) spherical, or is it flat? Did Apollo 11 really land humans on the moon? If you do not agree with the analysis of current events offered by one newspaper, you can often find another newspaper that offers a take more compatible with your own worldview. Politicians routinely make assertions of fact that directly contradict assertions of fact made by other politicians.

The remedy, when truth seems hard to find, is said to be evidence. Our conclusions about what is true should be guided by the best available evidence, and adhering to the evidence is prescribed as a means of reducing disagreement. Science is the ultimate manifestation of the use of evidence in drawing conclusions about what is true of the world. The scientific method offers a set of "best practices" for the systematic collection of evidence in the service of testing hypotheses, which are in effect claims about what is true or might be true (or, in some cases, what is *not* true).

Critics have pointed out many shortcomings in the actual practice of the scientific method that, arguably, reduce its stature as the ultimate arbiter of truth (e.g., Feyerabend, 1993; Latour & Woolgar, 1979). Nonetheless, there remains widespread agreement among practitioners and policymakers across a broad range of disciplines that important societal decisions ought to be evidence-based, and that science is a critical tool in the creation and adjudication of such evidence.

In many important societal decisions, however, politicians and policymakers are accountable to, and can only act with the approval of, the public. In some cases the public may not support certain actions even in the face of overwhelming supportive evidence. In other cases, though, the public can be highly responsive to perceived scientific consensus. For example, people's views on the need for action to mitigate climate change is correlated with the extent to which they perceive scientists as being in agreement that global warming is real and caused by humans (Ding et al. 2011), and belief in global warming is increased in response to a message describing the extremely high level of scientific consensus on the issue (van der Linden, Leiserowitz & Maibach, 2019). This might seem surprising in light of how highly politicized the topic of global warming has become, but political ideology is also predictive of the extent to which people see

This research was supported by grants to the first author from the Natural Sciences and Engineering Research Council of Canada and to the second author from the Social Sciences and Humanities Research Council of Canada and the Miami Foundation.

We thank John Randell of the American Academy of Arts and Sciences for coordinating the participation of Academy members in Experiment 6.

*Department of Psychology, University of Waterloo, Waterloo, Ontario, N2L 3G1 Canada. Email: dkoehler@uwaterloo.ca.

†Hill/Levene Schools of Business, University of Regina.

scientists as being in agreement on the topic. In the case of less polarized issues, the public is likely to be even more responsive to perceived scientific consensus.

What does it take for members of public to conclude that there is broad scientific consensus on an issue? For that matter, what determines whether the public views science as the best source of evidence on some pressing social issue? These questions are complicated not only because the public lacks (by definition) the subject-matter knowledge held by scientific experts on a given topic but also because they have a different understanding of, or different expectations for, how knowledge is advanced via the scientific method.

In the present research we investigated, specifically, how members of the public (and scientists) respond upon learning that two scientific studies on a particular topic have produced conflicting results. We compare this to how they respond when those same studies are said to have produced similar results. Our interest is in how agreement or disagreement in study outcomes impacts people's perception of whether, and to what extent, scientific knowledge has been advanced as a result of the studies being described. In other words, do people feel we know more, or are closer to the truth, as the result of conducting scientific studies even when those studies produce apparently discrepant results?

Understanding how the public perceives advancement of scientific knowledge is practically important for the obvious reason that taxpayers ultimately fund much of basic science. If conflicting results are a routine feature of productive science, but the public views them as a sign that knowledge is not being advanced, important research may not receive the public support it deserves. Indeed, perceptions of scientific dissent on a particular issue have been found to reduce public support for policy action on that issue (Aklin & Urpelainen, 2014).

The present research is also relevant to the development of evidence-based (e.g., based on behavioral science) public policy because scientific studies are often used in this context not only to develop effective policies but also as rhetorical support for their adoption. The public may be more skeptical than scientists of policy that emerges from a body of research that is anything less than completely in agreement.

Scientists recognize, by training or past experience, that different studies on a particular topic can sometimes produce conflicting results. Often we can learn something from such discrepancies as they can identify important moderators that determine, for example, the conditions under which one variable is, or is not, associated with another. Other times conflicting results are simply a consequence of statistical unreliability, such as that arising from sampling error. Either way, arguably, from the perspective of the scientific method, we have more data and therefore generally will be closer to the truth as the result of conducting the studies even if their results are not in agreement (Shiffrin, Borner & Stigler, 2018). Identification of, and competing attempts to explain, apparently discrepant results are, of course, a hallmark of the scientific method and the means by which better hypotheses eventually replace worse ones (Open Science Collaboration, 2015).

Members of the public, however, may not share the same perspective on how the scientific method advances knowledge. Indeed, it has been argued that scientists and the public may hold different mental models of the scientific method (Rabinovich & Morton, 2012). According to this account, the public views the scientific method as forging a direct path to the truth; scientists, by contrast, view it as a productive debate in which conflict – between study results and between hypotheses held by different scientists – plays a central role. Consequently, when confronted with conflicting results from scientific studies, members of the public may be more likely than scientists to conclude that nothing new has been learned from those studies, and that we are no closer to the truth than we were before they were conducted. It is even possible that conflicting results lead some people to feel that we know *less* than we did before.

Such a sentiment arguably violates normative principles of statistical and scientific inference which posit that new study findings cannot, generally[1], reduce our level of knowledge (and that only completely uninformative studies can leave our level of knowledge unchanged). Scientists may be more likely to subscribe to such principles than members of the public, who instead may rely on judgmental heuristics and intuitions about uncertainty in making their determinations of whether knowledge has been advanced by a particular set of study results.

Relevant findings from research on the psychology of judgment demonstrate how intuitive responses to conflicting evidence can violate normative principles of statistical inference. Kahneman and Tversky (1973) described an illusion of validity, in which people feel less uncertain making predictions based on redundant (and therefore less informative) cues, which necessarily agree with one another, than they do based on independent (and therefore more informative) cues, which are more likely to conflict with one another. One illustration comes from a study in which people predicted the outcome of a jury trial based on seeing only the arguments made by one side (either the plaintiff or the defendant) or on both sets of arguments (Brenner, Koehler & Tversky, 1996). The latter condition necessarily entailed more conflict, and, although it led to more accurate predictions than did the conditions exposed to only one side of the case, those predictions were actually made with lower confidence. This research suggests that a sense of uncertainty can be generated even as a consequence of an evidence-collection process that in fact advances our knowledge and thereby our ability

[1]If a misleading finding causes you to give up (or question) a true belief, you are now, in a sense, farther from the truth as a result. In our experiments, we assume that participants do not think of this sort of example.

to make accurate predictions, when the evidence collected is conflicting or less than entirely consistent in its implications.

Based on these ideas, we hypothesized that, when presented with the results of a pair of scientific studies on a given topic, members of the public would be more likely to say that nothing new had been learned, and that we are no closer to the truth, when those studies produced conflicting rather than consistent results. We hypothesized that scientists, who are more familiar with the conflict inherent in the practice of science, would be less likely than members of the public to conclude that we learn nothing new, and move no closer to the truth, when scientific studies produce conflicting results.

The experiments reported here share a common design: Participants read about two scientific studies on a common topic. The first of the two studies described to participants was said to have had the same result in all conditions, showing a significant correlation between the key variables or a significant effect of a treatment. For some participants, the second study was reported to have found a similar result to the first. For other participants, the second study was said to have produced either a null result or an effect in the opposite direction of that found in the first study. Participants were asked, after reading about the two study results, to make judgments of the extent to which the studies advance scientific knowledge. Specifically, they were asked whether "we know more" and "are closer to the truth" as a result of those studies. Our primary hypothesis is that people will be more likely to say we do not know more, and are no closer to the truth, when the two studies are said to have produced conflicting, rather than consistent, results.

From a normative perspective, is it always the case that we know more as the result of new studies even when their findings conflict? A strong position, which we have implied in this section, is yes, at least assuming that those studies were competently conducted. We have more data than we did before the studies were conducted, so we know more. A weaker position might state that whether we know more or not should not depend on the results of the studies, so if we conclude that we do not know more when the study results conflict, we also should conclude that we do not know more when they agree. Our main hypothesis implies that people will violate this weaker normative position as well as the stronger one. A potential problem, in either case, is that respondents to our scenarios may confuse knowing with believing, and it might be possible, normatively, to defend the position that we "knew" (believed) something based on one study that was then contradicted by the results of a second study, so now we "know" less. (However, the first few experiments we report actually ask about how much we know from both studies taken together, not what we know from the second study after already knowing something from the first.) One way we try to address this concern is by also asking, in later experiments, about closeness to truth rather than knowing, which could be less susceptible to an interpretation in terms of belief. But we acknowledge that there are possible criticisms of the normative position that we always know more even from conflicting study results.[2] Setting aside the normative question, of course, it is still of interest to consider the descriptive question of how members of the public perceive knowledge advancement from conflicting (vs. consistent) study results, and whether their perceptions differ from those of scientists.

Experiments 1–4 involved scenarios describing studies from a variety of scientific disciplines and reporting results in terms of an accessible effect-size measure (percentage points). Participants judged how much was learned from both studies taken together. Experiment 5 used a different scenario and reported results in terms of presence or absence of statistically significant differences rather than effect size. Participants judged how much was learned from a follow-up study relative to what was already known from an initial study. Experiment 6 presented the scenarios used in the earlier experiments to renowned scientists as well as to a comparison group of laypeople.

2 Experiment 1

University students read scenarios involving studies conducted in three different research areas. For each area, a pair of studies was described. The first study in the pair always produced a positive result. The second study result varied between subjects, and produced either a similar positive result, a null result, or a negative result (i.e., opposite in direction to the first study). We also varied whether or not an additional paragraph was provided noting methodological differences between studies; it is possible that members of the public do not spontaneously consider possible mediating variables but do so when they are drawn to their attention, possibly making them more inclined to say that something new has been learned even from conflicting studies. Individual difference measures of science beliefs and trust in scientists were also collected to test if they moderated how participants responded to the scenarios.

2.1 Method

Participants. Canadian undergraduate students were recruited from the University of Waterloo participant pool to complete an online study in early 2016. In total, 176 students began the survey but 8 did not complete it. A further 35 participants reported responding randomly at some point during the survey and 1 responded negatively to our question about English proficiency (2 did not respond) – these participants were removed from the data set. The remaining

[2] See note 1.

130 participants (*Mean* age = 20.6) consisted of 81 females and 49 males.

Materials and Procedure. Participants were randomly assigned to 1 of 6 possible conditions. In each condition, participants were presented with three science-based scenarios that concerned three different categories of content (education, health, psychology). Full materials are provided in the supplemental materials. As an illustration, here is the education scenario:

> Educational researchers in a school district in California tested the impact of a new "experiential" (hands-on learning) mathematics curriculum. In three high schools where the new curriculum was introduced, scores on a standardized math test administered at the end of the year were **18% higher** than scores had been the year before.
>
> An independent team of researchers working in a school district in Texas conducted a similar test, and found that standardized math test scores [*same effect* condition: **increased by 15%]** [*no effect* condition: **increased by only 3%**] [*opposite effect* condition: **actually decreased by 3%**] compared to scores from the year before the new curriculum was introduced.

(The condition labels are used for consistency across studies even though they only approximately apply to the numerical result values given, i.e., "same effect" was not exactly the same change in percentage points as in the first study, "no effect" was not exactly 0 percentage points change, and "opposite effect" was in the opposite direction of the first study but generally not as extreme as in the first study.) Participants read three scenarios with evidence/result condition varied within subjects so that each level of that variable was experienced once by each participant across the three scenarios. We counterbalanced the scenarios such that the second scenario either produced the same effect, no effect, or the opposite effect equally frequently across participants (example: Education had the same effect for condition 1, no effect for condition 2, and the opposite effect for condition 3).

A second (between-subjects) factor was the provision (or not) of an explanation following the description of the study results that identified methodological differences between the studies, for example:

> The two teams of researchers identified several potentially important differences in how the new curriculum was implemented and tested in the two studies, such as the amount of teacher training with the new curriculum and the particular standardized test that was administered, as well as differences in student demographics (e.g., family income, cultural background) in the two school districts.

Following each scenario, participants were asked four questions. First, they were asked "When we take the results of these two studies together, do we now know more, less, or the same as we did before about the [description of study content]?" (We know less, We know the same amount, We know more). This was the key dependent variable. We also asked (in the following order): "Do these studies advance our knowledge about the impact of [description of study content]" (yes, no); "The body of knowledge built from this area of research seems:" (Very weak, Weak, Neither weak nor strong, Strong, Very strong); "Should research on this topic receive more, less, or the same continued level of funding" (Less funding, The same amount of funding, More funding).

After the full set of scenarios, participants were asked a number of pro-science belief questions (Pennycook, Cheyne, Koehler & Fugelsang, 2019) and a trust in science questionnaire (Nadelson et al., 2014). We then asked participants if they responded randomly at any point during the survey (and noted that they will get their compensatory credit regardless of their response). The survey closed with a short demographics questionnaire (age, gender, English fluency, and social/fiscal conservatism).

2.2 Results

We averaged across the 3 scenarios and report the within-subject difference between the extent to which people report knowing more, the same, versus less (1 = 'we know less', 2 = 'we know the same', 3 = 'we know more'; i.e., a higher score = reporting more knowledge) based on whether the second study produced the same effect, no effect, or the opposite effect. We therefore analyzed the data using a 3 (evidence: same effect, no effect, opposite effect) x 2 (explanation, no explanation) mixed design ANOVA. This produced a significant main effect of main effect of evidence type ($F(2, 254) = 18.57$, $MSE = .40$, $p < .001$, $\eta^2 = .13$), such that reported knowledge was lower given an opposite effect ($M = 2.22$ [2.08, 2.37]) than given no effect ($M = 2.50$ [2.39, 2.62]), which was lower than when the same effect was present ($M = 2.70$ [2.60, 2.8]) – see Figure 1. There was no main effect of having an explanation or interaction between evidence type and explanation, F's < 1. This pattern of results is identical when people's perceptions of whether knowledge has advanced given the two studies (main effect $F = 18.10$, $p < .001$), whether the body of knowledge is strong versus weak (main effect $F = 36.40$, $p < .001$), and (more weakly) whether research on the topic should receive more funding (main effect $F = 3.21$, $p = .042$). There were no reliable correlations between pro-science beliefs *or* trust in scientists and people's knowledge ratings for any evidence type, all r's < .11, p's > .235).

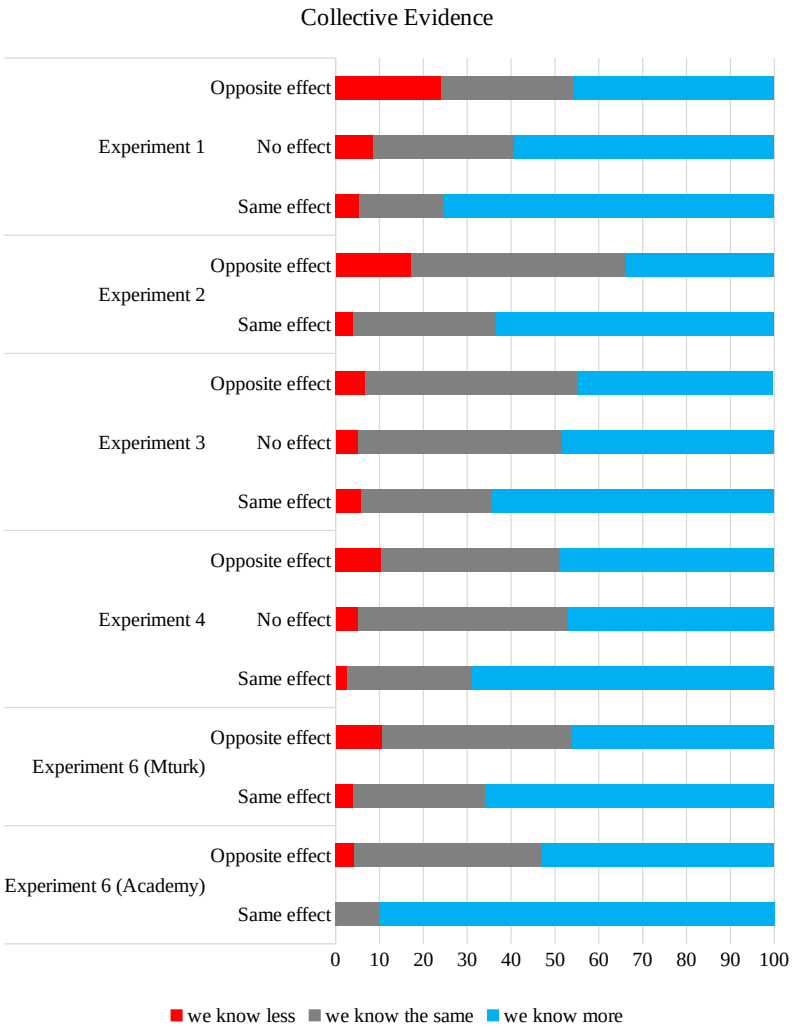

FIGURE 1: Proportion of individuals indicated that, based on the collective evidence of two studies, we know less (red), the same (grey), or more (blue) than before.

3 Experiment 2

The psychology (video games and aggression) scenario from Experiment 1 was presented to a larger and more demographically diverse set of participants recruited via Mechanical Turk. Only the consistent (same effect) and conflicting (opposite effect) versions of the second study result were used.

3.1 Method

Participants. Americans were recruited from Amazon's Mechanical Turk to complete an online study in spring 2016. In total, 406 participants began the survey but 9 did not complete it. A further 5 participants reported responding randomly at some point during the survey and 1 responded negatively to our question about English proficiency (2 did not respond) – these participants were removed from the data set. The remaining 395 participants (*Mean* age = 36.1) consisted of 183 females and 206 males (3 individuals did not indicate their gender).

Materials and Procedure. Participants were randomly assigned to 1 of 4 possible conditions. All participants received a variant of the psychology scenario from Experiment 1. However, in this case, participants were given only a single scenario. Moreover, we employed only the same-effect and opposite-effect conditions. As in Study 1, half of the participants were given an explanation as to how the replication differed from the initial study. Thus, our design is a 2 (same effect, opposite effect) x 2 (explanation, no explanation). Participants were given the same questions about the scenario as in Experiment 1.

In lieu of the pro-science beliefs questionnaire, we also administered the 15-item Need for Closure scale (Roets & Van Hiel, 2011). Otherwise, the materials and procedure were identical to Experiment 1.

3.2 Results

We analyzed the data using a 2 (evidence: same effect, opposite effect) x 2 (explanation, no explanation) univariate ANOVA with reported knowledge as the DV. This produced a significant main effect of evidence ($F(1, 388) = 43.14$, $MSE = .41$, $p < .001$, $\eta^2 = .10$), such that knowledge was judged to be lower in the opposite effect ($M = 2.17$ [2.08, 2.26]) than the same effect ($M = 2.59$ [2.50, 2.68]) condition – see Figure 1. As in Experiment 1, there was no main effect of having an explanation or interaction between explanation and evidence (F's < 1). Furthermore, the same pattern of results was found for whether knowledge has been advanced (main effect $F = 46.85$, $p < .001$) and strength of the body of knowledge (main effect $F = 56.44$, $p < .001$). However, beliefs in whether the research should continue to be funded was only marginally lower in the inconsistent evidence condition ($F(1, 388) = 3.37$, $MSE = .48$, $p = .067$, $\eta^2 = .01$).

In the opposite effect condition, reported knowledge did not correlate with trust in scientists ($r = .07$, $p = .320$) or Need for Closure ($r = .01$, $p = .924$). However, there were significant correlations in the same effect condition such that people who trusted science more were more confident about how much knowledge was gained by the replication study ($r = .18$, $p = .010$). In contrast, people who were higher in Need for Closure were *less* confident about how much knowledge was gained ($r = -.17$, $p = .016$).

4 Experiment 3

The three scenarios (different research disciplines) from Experiment 1 were presented to Mechanical Turk participants; this time each participant read only one scenario rather than all three as in Experiment 1. The numerical (percentage change) results that were described from each study were now varied in a more consistent fashion across scenarios, and the conflict (opposite effect) condition described a generally stronger effect in the second study that was almost exactly as large as the positive result from the first study (but in the opposite direction). It was also explicitly stated that an independent reviewer had found both studies to have been well executed, to reduce perceptions that the conflicting results signaled incompetence in how the studies were conducted.

4.1 Method

Participants. Americans were recruited from Amazon's Mechanical Turk to complete an online study in fall 2016. In total, 936 participants began the survey but 35 did not complete it. A further 27 participants reported responding randomly at some point during the survey and 3 responded negatively to our question about English proficiency (1 did not respond) – these participants were removed from the data set. The remaining 873 participants (*Mean* age = 36.6) consisted of 437 females and 436 males.

Materials and Procedure. Participants were randomly assigned to 1 of 9 possible conditions. All participants received only a single scenario. Following Experiment 1, the scenarios either presented a replication study that produced the same effect, no effect, or the opposite effect as the original study. However, we used all three different types of scenarios across participants (hence, 9 conditions). Moreover, every participant was given an explanation for why the replication study might differ from the initial study and the explanation was strengthened such that it began with a note that both studies were determined to be well-executed by an independent researcher.

We also revised scenarios in a number of ways. First, we adjusted the described results in each scenario across the three content domains to use the same numerical values (and, specifically, effects of 20% for the initial study and 21% for the replication). The "no effect" scenarios were revised so that the results of the replication where framed as providing absolutely no effect at all. Furthermore, the "same effect" scenarios were made to be more consistent (e.g., Study 1: 20% increase in aggressive behaviors; Study 2: 21% increase in aggressive behaviors). Finally, the "opposite effect" scenarios were made to be more strongly inconsistent (e.g., Study 1: 20% increase in aggressive behaviors; Study 2: 21% decrease in aggressive behaviors). See supplemental materials for full details. Participants were asked only two questions about the presented scenario: Whether we know more/the same/less (taking the results of both studies together) and whether the studies advance our knowledge of the topic (yes/no).

As individual difference measures, we also gave participants the Scientific Reasoning Scale (Drummond & Fischhoff, 2017) and a 4-item non-numeric Cognitive Reflection Test (CRT) (Thomson & Oppenheimer, 2016).

4.2 Results

We analyzed the data using a 3 (evidence: same effect, no effect, opposite effect) x 3 (domain: education, psychology, health) univariate ANOVA with reported knowledge as the DV. This produced a significant main effect of main effect of evidence type ($F(2, 864) = 9.83$, $MSE = .34$, $p < .001$, $\eta^2 = .02$), such that reported knowledge was lower for opposite effects ($M = 2.38$ [2.31, 2.44]) than same effects ($M = 2.58$ [2.52, 2.65]) – see Figure 1. No effect in the follow-up produced judgments of knowledge ($M = 2.43$ [2.37, 2.50]) that were lower than the same effect ($t(572) = 3.08$, $p = .002$, $d = .25$), but similar to the opposite effect ($t(585) = 1.06$, $p = .289$, $d = .08$). There was no interaction between evidence type and domain of study, $F < 1$, indicating that the decrease in judged knowledge for opposite effect and no effect conditions relative to same effect condition was

equivalent across the three domains. This pattern of results is identical when people's perceptions of whether knowledge has advanced given the two studies (main effect $F = 29.22$, $p < .001$). Judgments of knowledge did not reliably correlate with cognitive reflection (r's < .11, p's > .085) or scientific reasoning (r's < .03, p's > .620) regardless of evidence type.

The Scientific Reasoning Scale and CRT were not significantly correlated with reported knowledge (more/less/same) or advances in knowledge (yes/no) for any type of content (same effect, no effect, opposite effect), all r's < .1, p's > .08, with one exception: CRT was positively correlated with judgments about whether knowledge has been advanced for opposite effect studies, $r(305) = .15$, $p = .008$.

5 Experiment 4

The general design followed that of Experiment 3 but with a few changes to materials and measures. It was emphasized that the two studies had been conducted independently, and participants were explicitly asked to evaluate whether knowledge was gained relative to what was known before *either* study was conducted. An additional measure was included that asked whether, as a result of the two studies, we are now closer to the truth than we were before they were conducted. It is possible that people interpret the "we know more" item as concerning beliefs, justifying a sense (e.g., in the opposite effect condition) that we initially "knew" (believed) something that was later refuted such that we now "know" less. Because the new item about closeness to truth concerns the actual state of the world rather than beliefs, we thought it would be less susceptible to this interpretation.

5.1 Method

Participants. Americans were recruited from Amazon's Mechanical Turk to complete an online study in winter 2016. In total, 941 participants began the survey but 37 did not complete it. A further 17 participants reported responding randomly at some point during the survey and 2 responded negatively to our question about English proficiency – these participants were removed from the data set. The remaining 885 participants (*Mean* age = 33.5) consisted of 373 females and 512 males.

Materials and Procedure. Participants were randomly assigned to 1 of 9 possible conditions, as in Experiment 3. However, we again revised scenarios in a number of ways in service of undermining the impetus for participants to say that we know less given the conflicting studies. First, we added emphasis to the fact that the initial study and replication were completed independently. Moreover, we gave more context about the purpose of the studies at the outset of the scenario. Most importantly, the studies were introduced together to emphasize their equivalence in terms of evidence.

We also made a few changes to the dependent variables. First, we removed the broad (secondary) question about whether the research improves our knowledge. Second, we emphasized that we were asking about knowing more (or less, or the same) about the topic of study relative to *before* both studies were run. The question read: "When we take the results of these two studies together, do we now know more, less, or the same as we did **before the two studies were conducted** about [description of study content]". Third, we added a new question (which came first after the scenarios) about closeness to truth. The question read: "In your opinion, do the results of these two studies, taken together, move us closer toward the truth about [description of study content]? Taken together, relative to where we were before either study was completed, these two studies:" (Move us further from the truth, Do not move us any closer to the truth, Move us closer to the truth). Otherwise, the procedure was identical to Experiment 3, except that we did not include individual differences measures. See supplemental materials for details.

5.2 Results

We analyzed the data using a 3 (evidence: same effect, no effect, opposite effect) x 3 (domain: education, psychology, health) univariate ANOVA with reported knowledge as the DV. This produced a significant main effect of evidence type ($F(2, 876) = 20.86$, $MSE = .33$, $p < .001$, $\eta^2 = .05$), such that reported knowledge was lower when presented with opposite effects ($M = 2.38$ [2.32, 2.45]) than presented with the same effects ($M = 2.66$ [2.60, 2.73]) – see Figure 1. No effect in the follow-up study produced judgments of knowledge ($M = 2.42$ [2.35, 2.48]) that were lower than the same effect condition, $t(586) = 4.88$, $p < .001$, $d = .43$, but similar to the opposite effect condition ($t(585) = 1.31$, $p = .192$, $d = .05$). There was no interaction between evidence type and domain of study ($F(2, 876) = 1.66$, $MSE = .33$, $p = .156$, $\eta^2 = .01$), indicating that the decrease in judged knowledge for opposite- and no-effect conditions relative to the same-effect condition was equivalent across the three domains.

The pattern of results for judgments about whether the studies collectively move us closer (or further) from the truth was identical. Specifically, there a significant main effect of evidence type, $F(2, 876) = 23.76$ ($MSE = .30$, $p < .001$, $\eta^2 = .05$), such that movement toward truth was lower in the opposite-effect condition ($M = 2.33$ [2.27, 2.39]) than in the same-effect condition ($M = 2.62$ [2.56, 2.69]), but was similar to the no-effect condition ($M = 2.39$ [2.33, 2.46]). There was no interaction between evidence type and domain ($F < 1$).

6 Experiment 5

A different scenario, involving genetic expression of blood pressure, was used. This topic might be viewed by the public as more rigorously "scientific" than were those described in the previous experiments. Further, by use of this topic and accompanying abstract labels ("Gene X"), we reduced the likelihood that participants had any prior beliefs about the hypothesis being tested by the described study. Results of the study were described not in terms of effect size (percentages, as in the previous experiments) but rather in terms of the presence of absence of a statistically significant difference. Attention was drawn instead to sample size, which had not been explicitly reported in the scenarios used in the earlier experiments. It is possible that drawing attention to considerations of sample size might make people more appreciative of the cumulative impact on knowledge of conducting multiple studies, even when their results conflict. Finally, in contrast to the earlier experiments, which focused on how much knowledge was gained from two independent studies taken together, in this experiment the first study was described as the "initial" study and the second as a "follow-up" study and participants were asked to judge how much knowledge was advanced by the follow-up study beyond what was already known from the initial study. It is possible that participants are more likely to agree that something new has been learned from the follow-up study when its sample size is larger than that of the initial study, which we tested by varying the reported sample size of the follow-up study.

6.1 Method

Participants. Americans were recruited from Amazon's Mechanical Turk to complete an online study in spring 2017. In total, 417 participants began the survey but 15 did not complete it. A further 3 participants reported responding randomly at some point during the survey (1 did not respond) and 1 responded negatively to our question about English proficiency (1 did not respond) – these participants were removed from the data set. The remaining 397 participants (*Mean* age = 34.1) consisted of 168 females and 228 males (and 1 person who did not answer the gender question).

Materials and Procedure. Participants were randomly assigned to 1 of 4 possible conditions. The replication study was either consistent with the initial study (same effect) or inconsistent (no effect). Furthermore, we reported the sample size for both studies: in one condition both the initial and follow-up studies had equal samples sizes; in the other the follow-up study had twice the sample size of the initial study. We also made a number of changes to the scenarios. As noted, we set out to test perceptions about how the results of a follow-up study change people's understanding *relative to* an initial study. Thus, our questions asked specifically about the follow-up study as opposed to how much we know about the topic of interest given the collective evidence presented in the two studies taken together. Furthermore, we revised the scenario to be more focused on basic science as opposed to social science – and on a topic for which people would not have prior beliefs. In particular, we asked about a hypothetical association between a particular gene ("Gene X") and blood pressure. See supplemental materials for full details. We also included the 4-item Cognitive Reflection Test from Study 3 as an individual difference measure.

6.2 Results

We analyzed the data using a 2 (evidence: same effect, no effect) x 2 (same sample size, larger sample size) univariate ANOVA with reported knowledge as the DV. This produced a significant main effect of evidence ($F(1, 392) = 46.34$, $MSE = .39$, $p < .001$, $\eta^2 = .11$), such that knowledge gained from the replication was judged to be lower in the no-effect condition ($M = 2.27$ [2.18, 2.36]) than in the same-effect condition ($M = 2.70$ [2.61, 2.78]) – see Figure 2. Doubling the sample size for the replication had no impact on people's perceptions of how much knowledge was gained from the replication (nor was there an interaction between sample size and evidence) (F's < 1).

The pattern of results for judgments about whether the "follow-up study" (replication) move us closer (or further) from the truth was identical. There was a significant main effect of evidence ($F(1, 392) = 87.06$, $MSE = .34$, $p < .001$, $\eta^2 = .18$), such that the replication was judged to bring us *less* close to the truth in the no-effect condition ($M = 2.29$ [2.21, 2.37]) than in the same-effect condition ($M = 2.84$ [2.76, 2.92]). There was again no main effect or interaction of sample size (F's < 1).

Performance on the Cognitive Reflection Test did not correlate with judgments of knowledge ($r = .06$, $p = .375$) or truth ($r = .04$, $p = .612$) in the no effect (failed replication) condition. Individual differences results were also not significant for knowledge ($r = .06$, $p = .366$) or truth ($r = -.13$, $p = .062$) in the same effect condition.

7 Experiment 6

We had the opportunity to present some of the scenarios from the previous experiments (slightly modified as described below) to scientist members of the American Academy of Arts and Sciences, a distinguished honorary organization recognizing exceptional professional accomplishments. We hypothesized that, because they hold a different mental model in which conflict is a feature of the scientific process, scientists would be less likely than members of the public to say that nothing new has been learned when two scientific studies produce conflicting results. Judgments from this group were compared to new data collected from members of the

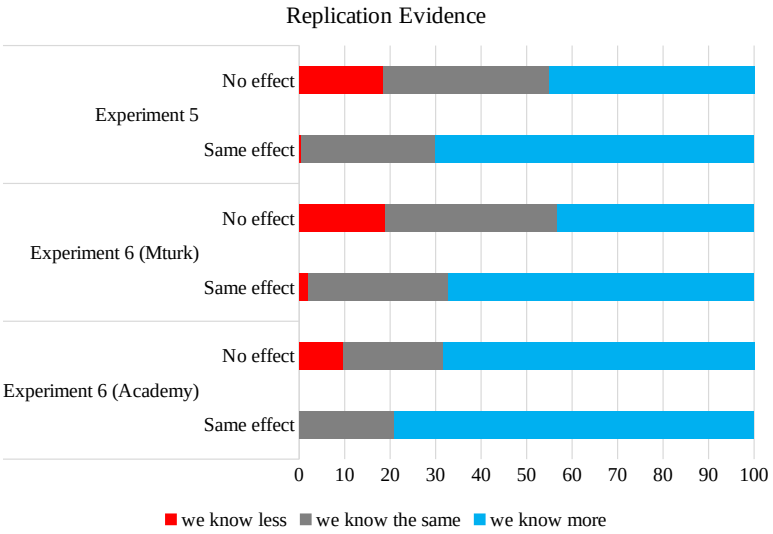

FIGURE 2: Proportion of individuals indicated that, based on replication study, we know less (red), the same (grey), or more (blue) than before.

public via Mechanical Turk. To test the mental-model idea, a new scale was introduced that asked respondents explicitly about their views and expectations regarding outcomes of scientific studies.

7.1 Method

Participants. We recruited from two sources: Mechanical Turk and the American Academy of Arts and Sciences. Both studies were completed in the summer of 2017. In total, 541 participants from Mechanical Turk began the survey but 40 did not complete it. Twelve Academy members began the survey but did no completed it. We did not ask Academy members if they responded randomly or were proficient in English and therefore did not remove any additional participants from the Mechanical Turk data set on this basis either. Academy members came from all five membership classes: Class I: Mathematical and Physical Sciences (N = 121), Class II: Biological Sciences (N = 86), Class III: Social Sciences (N = 87), Class IV: Humanities and Arts (N = 63), and Class V: Public Affairs, Business, and Administration (N = 42). For purposes of analysis, we only analyzed data from Classes I-III, which are based more strongly in quantitative scientific research (N = 294). See the dataset in the supplemental materials for data from the other Academy Classes. The Mechanical Turk sample consisted of 501 participants (*Mean* age = 34.9) consisted of 185 females and 316 males. We did not ask Academy members for their age or gender.

Materials and Procedure. Apart from removing the demographic questions for the Academy members (and including the question about membership class), the studies were identical for the two samples. We presented two scenarios to all participants: The Gene X and blood pressure scenario from Experiment 5 and the health (arthritis and exercise) scenario from Experiment 4. These were judged to be the strongest cases where participants *should* indicate that knowledge has been gained (and that we have moved closer to truth). Thus, in the gene scenario, the replication ("follow-up") study always had twice the sample as the initial study. Again, participants were asked about whether the follow-up study moves us closer to the truth and helps us know more about the association between Gene X and high blood pressure. Each participant was given either the same-effect or no-effect version of this scenario. For the health scenario, participants were always given an explanation for why the two studies differ and were asked about the collective knowledge gained from both studies. Each participant was given either the same-effect or opposite-effect version of this scenario. Participants were randomly assigned to the two between-subject conditions in a full factorial design. We also counterbalanced the order of the scenarios such that half of the participants saw the health scenario first and the other saw the gene scenario first. See supplemental materials for details.

As an individual difference measure of people's mental models of the scientific process, we adapted a 3-item beliefs about science (BAS) scale from Rabinovich and Morton (2012), which indexes the extent to which people believe that science produces a single correct answer (high score) or multiple possible answers (low score). Our adapted version of this scale includes three items (participants rated their level of agreement on a 7-point scale: 1) "There may be more than one possible explanation for the results of a scientific study" (reverse scored), 2) "If a scientific hypothesis is correct, every study that tests it will produce supporting results," and 3) "Scientists' knowledge should be called into question

when scientific studies produce contradictory results." However, the first item was not associated with the other two and was therefore removed from analysis. Laypeople (MTurkers) scored much higher on the scale (M = 4.87) than Academy members (science Classes I-III M = 2.74; arts Classes IV-V M = 2.98), reflecting a greater likelihood of holding a mental model of the scientific process as a "straight march to truth" that does not tolerate conflicting data or explanations.

7.2 Results

We first analyzed the health-scenario data (where participants judged the collective evidence from the two studies) using a 2 (evidence: same effect, opposite effect) x 3 (expertise: layperson, Academy member) univariate ANOVA with reported knowledge as the DV. This produced a significant main effect of evidence type ($F(1, 791) = 25.01$, $MSE = .32$, $p < .001$, $\eta^2 = .03$), such that reported (collective) knowledge was lower when the follow-up produced an opposite effect (M = 2.42 [2.36, 2.48]) than when it produced the same effect (M = 2.76 [2.70, 2.82]) – see Figure 1. There was also a main effect of expertise ($F(1, 791) = 66.62$, $MSE = .32$, $p < .001$, $\eta^2 = .08$), such that Academy members rated collective knowledge to be higher (M = 2.69 [2.63, 2.76]) than did laypeople (M = 2.49 [2.44, 2.54]). However, there was not a significant interaction between expertise and evidence (although it was marginal) ($F(1, 791) = 3.46$, $MSE = .32$, $p = .063$, $\eta^2 = .004$). The closeness to truth measure (i.e., whether the two studies bring us closer to the truth) also did not produce a significant interaction between expertise and evidence ($F < 1$). There were nonetheless significant main effects of evidence ($F(1, 791) = 12.28$, $MSE = .30$, $p < .001$, $\eta^2 = .02$), and expertise ($F(1, 791) = 60.36$, $MSE = .30$, $p < .001$, $\eta^2 = .07$), for judgments of truth.

The results for the gene-scenario (where participants judged the evidence from the replication only) produced more reliable differences between experts and laypeople. Specifically, there was again a main effect of evidence ($F(1, 791) = 28.44$, $MSE = .38$, $p < .001$, $\eta^2 = .04$), such that knowledge gained from the replication was judged to be lower in the no-effect condition (M = 2.42 [2.35, 2.48]) than in the same-effect condition (M = 2.72 [2.66, 2.79]) – see Figure 2. There was also a main effect of expertise ($F(1, 791) = 45.74$, $MSE = .38$, $p < .001$, $\eta^2 = .06$), such that Academy members rated collective knowledge to be higher (M = 2.69 [2.62, 2.76]) than did laypeople (M = 2.45 [2.39, 2.50]). Finally, there was an interaction between evidence and expertise ($F(1, 791) = 5.06$, $MSE = .38$, $p = .025$, $\eta^2 = .01$), such that the difference between the same effect and no effect (or failed replication) conditions was smaller among Academy members than among laypeople (see Figure 2). The same pattern was evident for judgments of closeness to truth: a main effect of evidence ($F(1, 791) = 37.03$, $MSE = .29$, $p < .001$, $\eta^2 = .05$), a main effect of expertise ($F(1, 791) = 32.65$, $MSE = .29$, $p < .001$, $\eta^2 = .04$), and an interaction between evidence and expertise ($F(1, 791) = 6.61$, $MSE = .29$, $p = .010$, $\eta^2 = .01$).

Notably, however, there was a significant effect of evidence in every single case. That is, for both laypeople and experts, presenting two studies with conflicting results patterns of evidence produced lower judgments of (collective) gained truth and knowledge (t's > 4.7, p's < .001), and presenting a failed replication (relative to a successful replication) led to lower judgments of gained truth and knowledge from the replication (t's > 2.7, p's < .007). However, as is evident from Figures 1 and 2, this occurred largely by Academy members moving from "we know more" to "we know the same", with relatively few (relative to laypeople) indicating that "we know less".

The individual difference measure of science mental models was uncorrelated with judgments of knowledge advancement and closeness to truth for either the gene (collective) or health (replication) scenarios, regardless of whether the results conflicted or not (r's < .081, p's > .20). Among Academy members, the beliefs about science scale correlated with judgments about collective evidence for the conflicting results gene scenario ($r(154) = .20$, $p = .013$), but not for judgments of truth ($r < .02$), or for the same results gene scenario (for both judgments) (r's < .08, p's > .40). There was, however, consistent evidence for an association between the BAS scale and both knowledge and truth judgments among Academy members for both the same and no effect versions of the health (replication) scenario (r's > .17, p's < .04). Although the evidence is not overwhelmingly strong, taken together with the large mean difference between scientists and laypeople on this scale, it is worth exploring further how different mental models of the scientific process between scientists and the public related to differing perceptions of scientific progress, particularly from conflicting study results.

8 General Discussion

In six experiments, people consistently perceived less advancement of knowledge from scientific studies when their results conflicted rather than agreed. This finding held across scenarios involving different scientific disciplines (health, education, psychology, genetics) and across different measures of perceived scientific progress (e.g., whether we know more, or are closer to truth). Ratings of perceived advancement of knowledge in turn correlated with perceived strength of the research area and support for further research funding. Reassuring participants that the studies were competently executed, drawing attention to methodological differences that might help explain discrepant results, and explicitly noting the sample sizes of the studies did not affect the general disinclination to say "we know more" when studies produce

conflicting results. The same pattern of results held whether participants evaluated how much knowledge was gained from the two studies taken together (against an unspecified baseline level of knowledge) or whether instead they evaluated how much knowledge was gained from the second (follow-up) study beyond what was already known from the first (initial) study.

Scientists presented with these same scenarios responded somewhat differently from members of the public. They were, in general, more inclined to say "we know more" and are closer to the truth from the two studies whether their results agreed or disagreed. They were also somewhat less responsive to the absence or presence of conflicting results when making their evaluations. One factor that differed between scientists and members of the public and also correlated, across the entire sample, with judgments of knowledge advancement following replications related to mental models of the scientific process: Members of the public were more likely than scientists to expect that all studies testing a correct hypothesis should consistently produce supportive results. That said, on average, scientists – like members of the public – did generally tend to see less advancement of knowledge from conflicting results compared to congruent results.

What does it mean to say two studies produce conflicting results? How do people code the magnitude of conflict and then translate it into a judgment of how much, or little, knowledge has been advanced as a result? In a potentially informative comparison, Experiments 1, 3, and 4 included two levels of conflict. Relative to a first study that produced a positive effect of a given size (e.g., 20% higher scores under a new math curriculum compared to the old one), the second study could produce a conflicting result either through a null effect (scores under new curriculum identical to those under old curriculum, i.e., 0% change) or an effect in the opposite direction (scores 20% *lower* under the new curriculum). The results, particularly in Experiments 3 and 4, look quite similar for these two levels of conflict. In a between-subjects comparison, it may be that the quantitative difference in effect size does not matter much as it is not highly evaluable in isolation (Hsee, 1996); instead, compared to the first study which found "an effect", the second may be coded similarly as having failed to find that effect (and thus is "conflicting"), regardless of whether it produced a null result or an effect in the opposite direction of the first.

A few previous studies have also examined the effects of calling attention to disagreement among scientists and scientific studies. One study (Chang, 2015) presented pairs of news articles reporting health studies that either agreed or conflicted in their findings (e.g., one study indicates milk consumption reduces cancer risk while another indicates it increases cancer risk). Those presented with conflicting findings found the news articles less credible, had lower intentions to adopt the prescribed behavior, and expressed greater uncertainty about the research conclusions and rated the research field as less helpful and useful. In a different domain, Nagler (2014) found that people who reported having been exposed via the news media to higher levels of conflicting information from nutrition experts on health risks and benefits of certain foods (e.g. coffee) were more likely to say they were confused about which foods are best to eat, which in turn predicted nutrition "backlash" (belief that nutritionists keep changing their minds and therefore can be ignored) and reduced intentions to eat indisputably healthy foods. Both of these studies are consistent with our finding that conflicting study results lead people to see a scientific discipline as less strong and less deserving of research support compared to when those same studies instead produced congruent results.

Research on how people respond to conflicting scientific study results is also relevant to the ongoing "replication crisis" in psychology and other fields. Considerable attention and effort have been devoted to identifying sources of replication failure, from insufficient sample sizes to questionable research practices. Researchers in psychology have also become more alert to the presence of methodological variation in potential moderator variables. Large-scale collaborative studies, such as those conducted by the Open Science Collaboration, and more generally the synthesis of multiple studies through meta-analysis, are giving researchers an unprecedented vantage point for examining variation in study methods and results. An optimistic view of the replication crisis is that it has offered an opportunity to harness conflicting study results in the service of advancing knowledge, following the "productive debate" mental model of the scientific process: "Accumulating evidence is the scientific community's method of self-correction and is the best available option for achieving that ultimate goal: truth." (Open Science Collaboration, 2015). If the public does not hold this mental model, however, our research suggests that they may not share the perception that scientific knowledge is being advanced when studies produce conflicting results. Public confidence in science may be affected as a consequence (e.g., Wingen, Berkessel & Englich, in press). A pressing challenge for science communication (Jamieson, 2018), then, is to foster appreciation for advances in knowledge that accumulate even from an imperfect science in which conflicting results are a feature rather than a bug.

References

Aklin, M., & Urpelainen, J. (2014). Perceptions of scientific dissent undermine public support for environmental policy. *Environmental Science & Policy, 38*, 173–177.

Brenner, L. A., Koehler, D. J., & Tversky, A. (1996). On the evaluation of one-sided evidence. *Journal of Behavioral Decision Making, 9*(1), 59–70.

Chang, C. (2015). Motivated processing: How people perceive news covering novel or contradictory health research findings. *Science Communication*, *37*(5), 602–634.

Ding, D., Maibach, E. W., Zhao, X., Roser-Renouf, C., & Leiserowitz, A. (2011). Support for climate policy and societal action are linked to perceptions about scientific agreement. *Nature Climate Change, 1*(9), 462–466.

Drummond, C., & Fischhoff, B. (2017). Development and Validation of the Scientific Reasoning Scale. *Journal of Behavioral Decision Making*, *30*(1), 26–38. https://doi.org/10.1002/bdm.1906.

Feyerabend, P. (1993). *Against method*. New York: Verso.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational behavior and human decision processes*, *67*(3), 247–257.

Jamieson, K. H. (2018). Crisis or self-correction: Rethinking media narratives about the well-being of science. *Proceedings of the National Academy of Sciences, 115*(11), 2620–2627.

Kahneman, D., & Tversky, A. (1973). On the Psychology of Prediction. *Psychological Review, 80*(4), 237–251.

Latour, B., & Woolgar, S. (1979). *Laboratory Life*. Beverly Hills: Sage.

Nadelson, L., Jorcyk, C., Yang, D., Jarratt Smith, M., Matson, S., Cornell, K., & Husting, V. (2014). I Just Don't Trust Them: The Development and Validation of an Assessment Instrument to Measure Trust in Science and Scientists. *School Science and Mathematics*, *114*(2), 76–86. https://doi.org/10.1111/ssm.12051.

Nagler, R. H. (2014). Adverse outcomes associated with media exposure to contradictory nutrition messages. *Journal of health communication*, *19*(1), 24–40.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2019). On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs. *[Working Paper]*. https://doi.org/10.31234/OSF.IO/A7K96.

Rabinovich, A., & Morton, T. A. (2012). Unquestioned answers or unanswered questions: Beliefs about science guide responses to uncertainty in climate change risk communication. *Risk Analysis, 32*(6), 992–1002.

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences*, *50*(1), 90–94. https://doi.org/10.1016/j.paid.2010.09.004.

Shiffrin, R. M., Börner, K., & Stigler, S. M. (2018). Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences*, *115*(11), 2632–2639.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113.

van der Linden, S., Leiserowitz, A., & Maibach, E. (2019). The gateway belief model: A large-scale replication. *Journal of Environmental Psychology, 62*, 49–58.

Wingen, T., Berkessel, J., & Englich, B. (in press). No Replication, no Trust? How Low Replicability Influences Trust in Psychology. *Social Psychological and Personality Science*.