Do Researchers Anchor their Beliefs on the Outcome of an Initial Study? Testing the Time-Reversal Heuristic

Ernst, Anja; Hoekstra, Rink; Wagenmakers, Eric-Jan; Gelman, Andrew; van Ravenzwaaij, Donald

Published in:
Experimental Psychology

DOI:
10.17605/OSF.IO/EHYM3

Do Researchers Anchor their Beliefs on the Outcome of an Initial Study? Testing the Time-Reversal Heuristic

Anja Franziska Ernst, MSc
Department of Psychometrics and Statistics
University of Groningen
Grote Kruisstraat 2/1
9712TS Groningen
The Netherlands
Phone: +31 50 36 34833
E-mail: a.f.ernst@rug.nl

Dr. Rink Hoekstra
Department Educational Science
University of Groningen

Prof. Dr. Eric-Jan Wagenmakers
Department of Psychological Methods
University of Amsterdam

Prof. Dr. Andrew Gelman
Department of Statistics
Columbia University

Dr. Don van Ravenzwaaij
Department of Psychometrics and Statistics
University of Groningen

Abstract

As a research field expands, scientists have to update their knowledge and integrate the outcomes of a sequence of studies. However, such integrative judgments are generally known to fall victim to a primacy bias where people anchor their judgments on the initial information. In this preregistered study we tested the hypothesis that people anchor on the outcome of a small initial study, reducing the impact of a larger subsequent study that contradicts the initial result. Contrary to our expectation, undergraduates and academics displayed a recency bias, anchoring their judgment on the research outcome presented last. This recency bias is due to the fact that unsuccessful replications decreased trust in an effect more than did unsuccessful initial experiments. We recommend the time-reversal heuristic to account for temporal order effects during integration of research results.

> The human understanding, when any proposition has been once laid down [...], forces everything else to add fresh support and confirmation; and although most cogent and abundant instances may exist to the contrary, yet either does not observe or despises them, or gets rid of and rejects them by some distinction, with violent and injurious prejudice rather than sacrifice the authority of its first conclusions.
>
> – Francis Bacon

In an ideal world, researchers are able to accurately assess the evidence from the published record and rationally update their knowledge as the literature expands. In the real world, however, the way in which researchers update their knowledge may be distorted by biases in human reasoning; specifically, researchers may exhibit a primacy or anchoring effect, overweighting the importance of the first study or set of studies. This possibility has recently been emphasized in the context of a thought experiment. Here we seek to examine this anchoring effect empirically.

It is well known that human judgments can be affected by the order in which information is processed (e.g., Landon Jr., 1971). Specifically, people tend to rely more heavily

on initial information, and interpret later evidence in light of this earlier information. For instance, clinical psychologists' initial belief in the effectiveness of a test has been shown to be relatively resistant to later evidence of the test's ineffectiveness (Chapman & Chapman, 1967, 1969). In other words, prior information can lead to bias that decreases the willingness to seek out contradictory evidence and entertain alternative hypotheses (Kuhn, Amsel, & O'Laughlin, 1980). This bias may arise from the difficulty to disregard one's initial beliefs when evaluating novel information (Baron, 2007). This primacy effect also expresses itself in the human tendency to seek confirmatory evidence for an initial belief, whereas evidence against an initial belief is only sought after explicit instruction (Hoch, 1985; Koriat, Lichtenstein, & Fischhoff, 1980; Martindale, 2005; Nickerson, 1998; Sinkey, 2015; Wason, 1960). People are prone to anchor on initial information even when the anchor is clearly irrelevant to the actual judgment, for instance when they are presented with random numbers before making an estimate (Tversky & Kahneman, 1974).

One of us (AG) suggested a similar anchoring bias is present in researchers when they interpret the evidence from an original study and a replication attempt (January 26 2016; `http://andrewgelman.com/2016/01/26/more-power-posing/`). Specifically, AG assumed that researchers attach more importance to their initial study than to the replication attempt. In other words, if an original small-N study reports a significant result, a large-N non-significant replication study does relatively little to reduce researchers' trust in their claim; in the same fashion, if an original large-N study reports a non-significant result, a small-N significant follow-up study does relatively little to reduce researchers' distrust in the effect. What this means is that the order in which the studies are presented can change the conclusions that researchers draw. To make researchers aware of their anchoring bias, AG proposed what he called the *time-reversal heuristic*. The heuristic invites researchers to imagine that the order of the studies were reversed, prompting these researchers to recalibrate their overall judgment. The researcher's anchoring effect (henceforth RAE) is

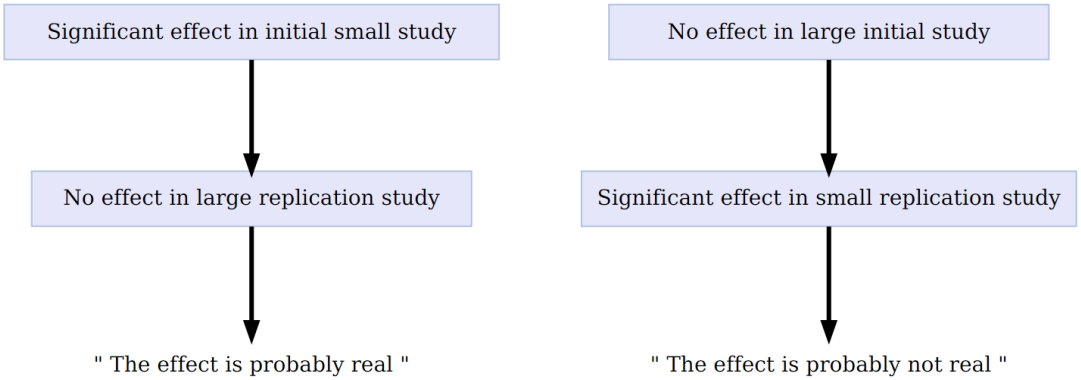


Figure 1. The researcher's anchoring effect: initial research findings have disproportionate impact on researchers' trust in the effect.

visualized in Figure 1.

The goal of the proposed study is to examine the RAE empirically, not in the perceptions of the researchers who produced the original study, but to outsiders, which is perhaps the more interesting population to be studying as they represent various aspects of the general scientific community. We will examine the RAE in a between-subjects experiment where each participant is presented with one of two hypothetical scenarios. We do not aim to assess researchers' aptitude to integrate research findings across different experiments; Instead, we wish to study a possible anchoring bias that, when brought to awareness, can benefit scientific judgments.

This study will be conducted both on a sample of academics and on a sample of undergraduates. In each sample, participants will be presented with a sequence of two research scenarios that is consistent either with the left or with the right panel of Figure 1.

This design allows us to assess the between-condition difference in the overall confidence for the presence of the effect. According to the RAE, overall trust should be higher in the left-panel sequence than in the right-panel sequence.

In order to prevent confirmation bias or hindsight bias from unwittingly skewing the interpretation of our findings we preregistered our analysis plan prior to data collection. The preregistration document can be found at `https://osf.io/j658h/`.

Method

Sampling Plan and Participants

Study 1: Researchers. The first study evaluated the RAE in academics. To this aim we emailed questionnaires to the first corresponding author of all articles published in 2014 and 2015 from the following journals: Journal of Experimental Psychology: General, Psychological Science, Journal of Abnormal Psychology, Journal of Consulting and Clinical Psychology, Journal of Experimental Social Psychology, and Journal of Personality and Social Psychology

These journals were chosen to provide a representative sample of researchers in the fields of experimental, social, and clinical psychology (see Cramer et al., 2016, for a similar approach). Selecting a single email address per article and removing duplicates left 640 unique email addresses for the 2015 articles and additional 706 unique addresses for the articles published in 2014. All participants selected in this manner were reminded one week after the initial invitation that their responses will be recorded only within the next seven days.[1]

Study 2: Undergraduates. In the second study we investigated the RAE in a naive student sample. All participants in the second study were first-year psychology students

[1]The initial email and the reminder email can be found in the supplementary material at `https://osf.io/j658h/`.

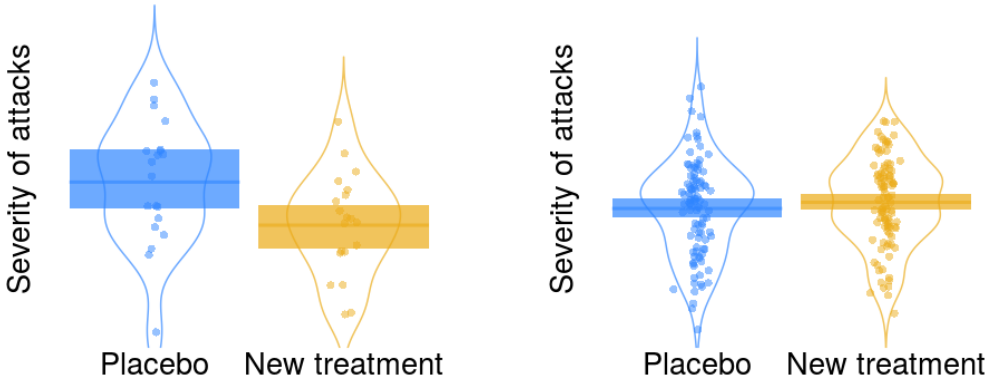
participating in this study as part of a large test battery at the University of Amsterdam in exchange for course credit. The students were asked to complete the same questionnaire as the academics in the first study. Additionally students received a short introduction on the purpose of the questionnaire, comparable to the invitation email that academics received. The questionnaire is outlined in detail below and can be viewed in the supplementary material.

Materials

Our between-subjects study employed two questionnaires, one for each condition.[2] Because we expected a smaller sample size in the study on researchers than in the study on undergraduates, researchers were semi-randomly assigned to either condition in alternating fashion to yield roughly equal numbers of participants in each condition while the undergraduate study employed random assignment to conditions. In the *initial-effect condition*, participants were first presented with a research scenario in which a research lab finds that a new treatment reduces the severity of panic attacks in a small sample of 40 individuals (i.e., p=.02, d=0.78). This research scenario was accompanied by a graphical depiction of the individual data points, as shown in Figure 2 (a). Participants were presented with effect size and sample size of the depicted effects; the scale on the y-axis was omitted deliberately to prevent participants from judging a given number of panic attacks as relatively low or high.

Subsequently participants were asked to indicate, on a nine-point Likert scale how strongly they believe in the effectiveness of the new treatment. The scale ranged from "There is definitely NO difference between the treatment and placebo" to "The treatment is definitely more effective than placebo"; participants were instructed that the middle value

[2]Duplicates of the questionnaires can be found in the supplementary material at `https://osf.io/j658h/`. These duplicates are pdf copies of the employed html questionnaires in which the second research scenario is presented only after participants submitted their response to the first research scenario.



(a) Based on these data, to what extent do you believe that the new treatment is more effective than placebo in reducing the severity of panic attacks?

(b) Overall, based on the data of both studies, to wha extent do you believe that the new treatment is more effective than placebo in reducing the severity of panic attacks?

Figure 2. Time-flow in the initial-effect condition. Participants are presented with research scenario (a) and are asked to rate their belief in the effect based on this scenario. Thereafter participants are presented with research scenario (b) and are asked to rate their belief in the effect based on both scenarios combined. In the initial-no-effect condition the scenarios are identical but are presented in reverse order.

of the scale represents perfect ambivalence about the treatment's effectiveness. Next participants were presented with a second research scenario in which a replication study, conducted by another research group, with a larger sample size of 200 individuals finds no evidence for the effectiveness of the new treatment (i.e., $p = .34$, $d = -0.14$). This research scenario was also illustrated in a graph, shown in Figure 2 (b). After having been presented with both research scenarios, participants were then asked to rate how strongly they believe in the effectiveness of the new treatment. Again, belief in the effectiveness of the treatment was assessed by a nine-point Likert scale ranging from "There is definitely NO difference between the treatment and placebo" to "The treatment is definitely more effective than placebo".

In the *initial-no-effect condition*, participants were presented with the identical two research scenarios as illustrated in Figure 2, but in reverse order. Hence, participants were first asked to indicate how strongly they believe in the effectiveness of the new treatment based on the scenario in which a large-sample study does not find the new treatment to be more effective than placebo (i.e., N= 200, $p = .34$, $d = -0.14$). After this, participants were presented with the second research scenario in which a smaller study finds the treatment to be more effective than placebo (i.e., N= 40, p=.02, d=0.78). Participants were then asked to rate their overall belief in the effectiveness of the new treatment, based on both studies they encountered. Thus, the only difference between our two questionnaires lies in the order of presentation for the two research scenarios. According to the RAE, participants in the initial-effect condition will have a stronger belief in the effect of the treatment than participants in the initial-no-effect condition, even though the presented information is identical.

Because it is possible that some of our academic participants have heard of the time-reversal heuristic, the academic questionnaire was concluded by an open-ended question that allowed participants to state their presumptions, if any, about the purpose of our study.

Participants who stated that they suspect our study to examine the time-reversal heuristic were excluded from the analyses. Our between-subjects design makes it difficult for an individual participant to discern the goal of the experiment, even if that participant happens to be familiar with the time-reversal heuristic.

Analysis Plan

There are three theories concerning the origin of the RAE, should it exist. The RAE might be a naturally occurring bias that is present in students and researchers alike. Alternatively, it is possible that the RAE is not a naturally occurring bias but is learned through repeated exposure to research results. In the latter case only researchers would exhibit a RAE. Lastly, the situation might be reversed with the RAE being a naturally occurring bias that is overridden by researchers through experience, while students might still be prone to anchor on prior information. We therefore investigate the presence of a RAE in researchers and students. Analyzing these two samples separately allows us to evaluate if either of the theories mentioned above might indeed explain the presence of a RAE, should it exist.

For both studies (i.e., academics and undergraduates), our main analysis concerned the impact of the order of individual-study results on the interpretation of the overall result. Specifically, the key dependent variable for our planned analysis was the nine-point Likert rating for treatment effectiveness based on the overall result (i.e., the second and final Likert rating). Thus, our crucial test contrasts the Likert ratings for the overall result in the initial-effect condition against those in the initial-no-effect condition. The RAE hypothesis predicts that the Likert scores will be higher in the initial-effect condition than in the initial-no-effect condition.

To quantify the statistical evidence for or against the RAE hypothesis we computed Bayes factors.[3] In the present context, Bayes factors compare the predictive adequacy of

[3]AG wishes to state that he hates Bayes factors. The reasons for his aversion are detailed in Gelman and

a null hypothesis versus the alternative "RAE" hypothesis. We compared the Likert scores in the two conditions using an independent-samples one-sided Bayesian t-test (Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009) with a default folded Cauchy effect size prior width of $r = \sqrt{2}/2$ (i.e., 0.707 or "medium"; for details see Rouder et al., 2009). In a secondary analysis we present the posterior distribution of the effect size as obtained from a two-sided Bayesian t-test with a default Cauchy effect size prior width of $r = \sqrt{2}/2$. In both the academic and the undergraduate studies, and in line with the classification scheme proposed by Jeffreys (1961), we regard a Bayes factor of 10 (whether in favor of the RAE hypothesis or in favor of the null hypothesis) as strong evidence.

Participants are assumed, on average, to have a neutral belief in the treatment's effectiveness before having been presented with any results. Consequently, we assume that the neutral rating of 5 will, on average, correspond to participants' belief in the effectiveness before having been presented with any results. Because of this, it is possible to compare the extent to which participants' belief in the effectiveness of the treatment was updated by the two scenarios, assuming linearity of our scale. We use this property of our scale to conduct a follow-up analysis in order to investigate the origin of the effect. We examine the following two origins of the effect: (1) the small-effect result is weighted more heavily when presented first; and, (2) the null result is weighted less when it is presented second. Specifically, the influence of a scenario in one condition will be compared to the degree by which participants' beliefs were updated by the same scenario in the other condition. Influence of a scenario is quantified as the difference between the participants' ratings before having been presented with the respective scenario and after having been presented with the scenario.

The RAE analysis and the follow-up analysis are illustrated in Figure 3 for fictional

---

Rubin (1995) and Gelman, Carlin, Stern, and Rubin (2004) (Chapter 6).

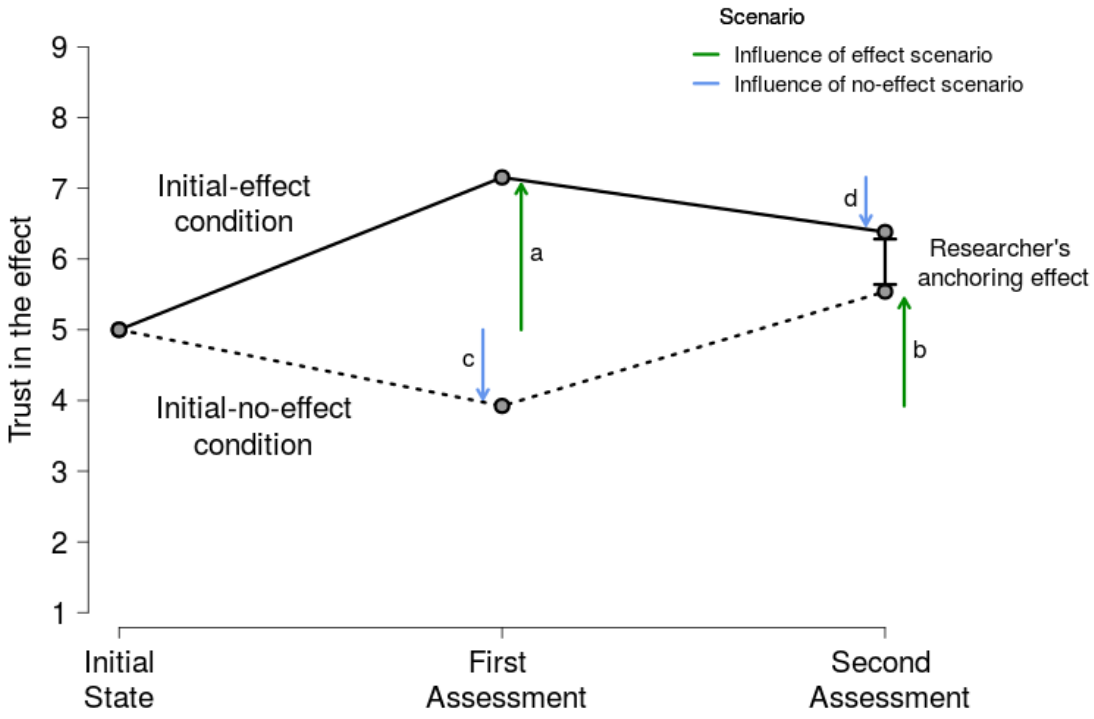


Figure 3. Visualization of the follow-up analysis based on fictional data that exhibits a researcher's anchoring effect.

data. The RAE is evaluated by comparing the second rating of our two conditions (i.e., the vertical black line labeled 'RAE'). The first potential origin of the effect, differential weighting of the *small-effect result*, is presented in Figure 3 as the difference between arrows *a* and *b*. According to hypothesis (1), this result should have a stronger influence in the initial-effect condition than in the initial-no-effect condition. As such, this scenario should increase participants' belief in the effect more in the initial-effect condition. This would be represented in Figure 3 with arrow *a* being greater than arrow *b*, $\mathscr{H}_1 : a > b$.

The second potential origin of the effect, differential weighting of the *null result*, is presented in Figure 3 as the difference in length between arrows *c* and *d*. According to

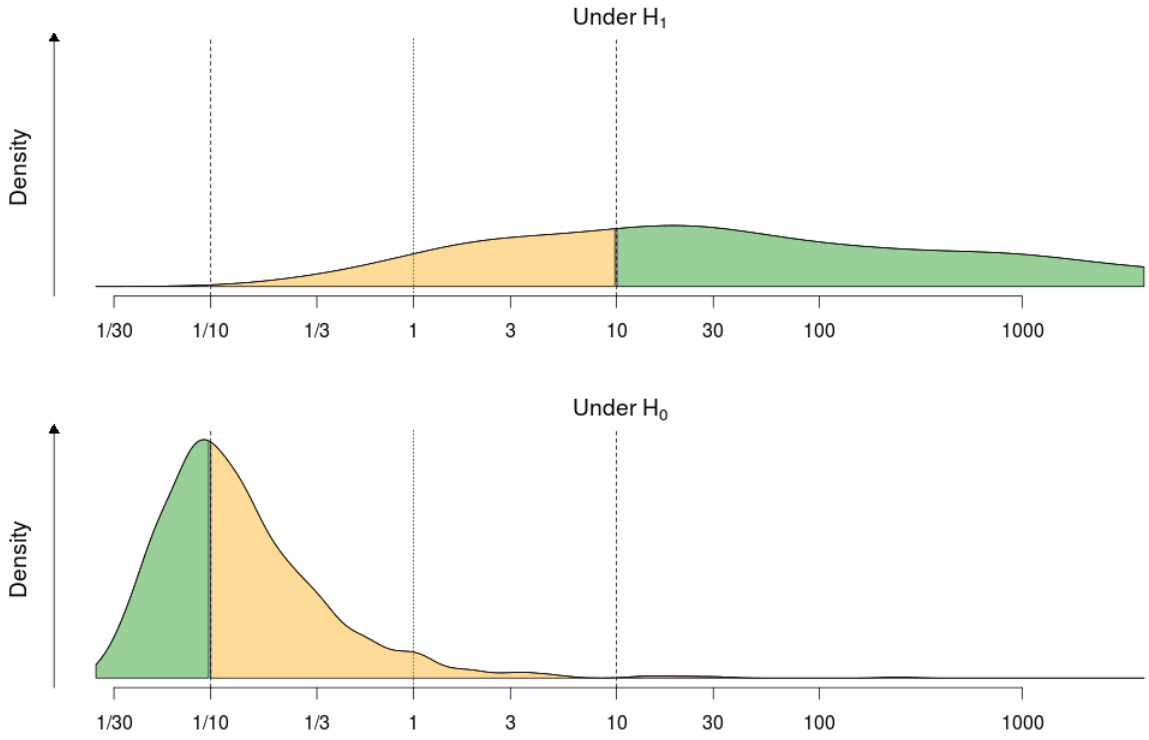


Figure 4. Distribution of Bayes factors produced by our design analysis. Vertical lines indicate regions of compelling evidence against the RAE (BF $\leq 1/10$) or in favor of the RAE (BF ≥ 10). The vertical line around BF$= 1$ corresponds to neutral evidence.

hypothesis (2), this result should have a weaker influence in the initial-effect condition than in the initial-no-effect condition. As such, this scenario should decrease participants' belief in the effect less in the initial-effect condition. In the figure, this would be represented by arrow d being shorter than arrow c, $\mathscr{H}_2 : d < c$.

Formally, our follow-up analysis consists of two additional one-sided t-tests, both with a folded Cauchy effect size prior width of $r = \sqrt{2}/2$. One of the t-tests evaluates hypothesis (1) by comparing the degree to which the small-effect result increased belief across the two conditions (cf. Figure 3, arrows a and b), while the other t-test evaluates hypothesis (2) by comparing the degree to which the null result decreased trust across the two conditions (cf. Figure 3, arrows c and d). Additionally, we present the posterior

distribution of the effect sizes associated with our two follow-up hypotheses obtained from two one-sided Bayesian t-test with a default Cauchy effect size prior with width of $r = \sqrt{2}/2$. The influence of the first research scenario was calculated as the difference from 5 (the neutral rating on the Likert scale), the influence of the second research scenario will be calculated as the rating difference from the first result.

Given our available sample size of about 350 undergraduate students and a fixed modest effect size of Cohen's $d = .35$, we undertook a Bayesian design analysis in order to determine the expected strength of evidence (Schönbrodt & Wagenmakers, 2016). The design analysis was performed using the R package BFDA (Schönbrodt, 2016). The resulting distribution of Bayes factors is displayed in Figure 4. This distribution reveals that with $N = 350$ and $d = .35$, a one-sided t-test has a 68% chance to reach a Bayes factor of 10 or higher in favor of the RAE hypothesis. If, on the other hand, no RAE effect exists (i.e., $d = 0$), our study on undergraduates would yield strong evidence against the RAE hypothesis in 42% of the cases, with associated Bayes factor values lower than 1/10. We conclude that our study on undergraduates has a considerable probability of yielding compelling evidence under either hypothesis.

Results

All analyses were carried out in R (R Core Team, 2017) using the BayesFactor package (Morey, 2017) and in JASP (JASP Team, 2017). Figures 6, 8, 10 and 12 were created with JASP. Figure 2 was produced with the yarrr R package (Phillips, 2017). All data and analysis scripts are available at OSF.[4]

[4] `https://osf.io/j658h/`

Academic Sample

Of all contacted researchers, 320 (23.8%) filled out the entire questionnaire. The responses of 22 researchers were excluded: 18 presumed we studied temporal order effects,[5] and 4 admitted to having misinterpreted the graphs. The final sample size retained for analysis was $N = 298$ with 158 researchers in the initial-effect condition and 140 in the initial-no-effect condition. The distributions of Likert ratings are displayed in Figure 5. Black lines indicate mean ratings of overall trust in the effect, based on both research scenarios.

Contrary to our expectation, Figure 5 shows that overall trust in the effect is *lower* in the initial-effect condition than in the initial-no-effect condition. The one-sided Bayesian t-test evaluating the RAE hypothesis against a point null indicates strong relative evidence in favor of the null, $BF_{RAE/\mathscr{H}_0} = .004$ (or $BF_{\mathscr{H}_0/RAE} = 232.4$). A two-sided Bayesian t-test shows strong support against the null hypothesis $BF_{two-sided\ RAE/\mathscr{H}_0}) = 2.41 \times 10^{15}$.

Figure 6 displays samples from the posterior distribution of the effect size obtained by the two-sided t-test. The figure allows to contrast support for the RAE hypothesis to support for a reversed RAE hypothesis. A line above the posterior distribution indicates the 95% credible interval of the effect size, $[-1.31, -.82]$, around a median value of -1.07. Negative effect sizes correspond to a reversed RAE. Thus, Figure 6 shows support for a reversed RAE (i.e., researchers' overall trust is higher in the initial-effect condition).

Figure 7 shows the impact of both research scenarios on the researchers' trust, split by condition. The influence of the effect scenario, depicted as lines a and b, appears roughly equal across the two conditions. We formally test this pattern with the first pre-specified follow-up analysis. The analysis indicates that the significant result scenario had roughly equal effects on researchers' trust across the two conditions; the one-sided Bayesian t-test

[5] Of the 18 researchers who suspected temporal order effects, 6 were assigned to the initial-effect condition and 12 were assigned to the initial-no-effect condition.

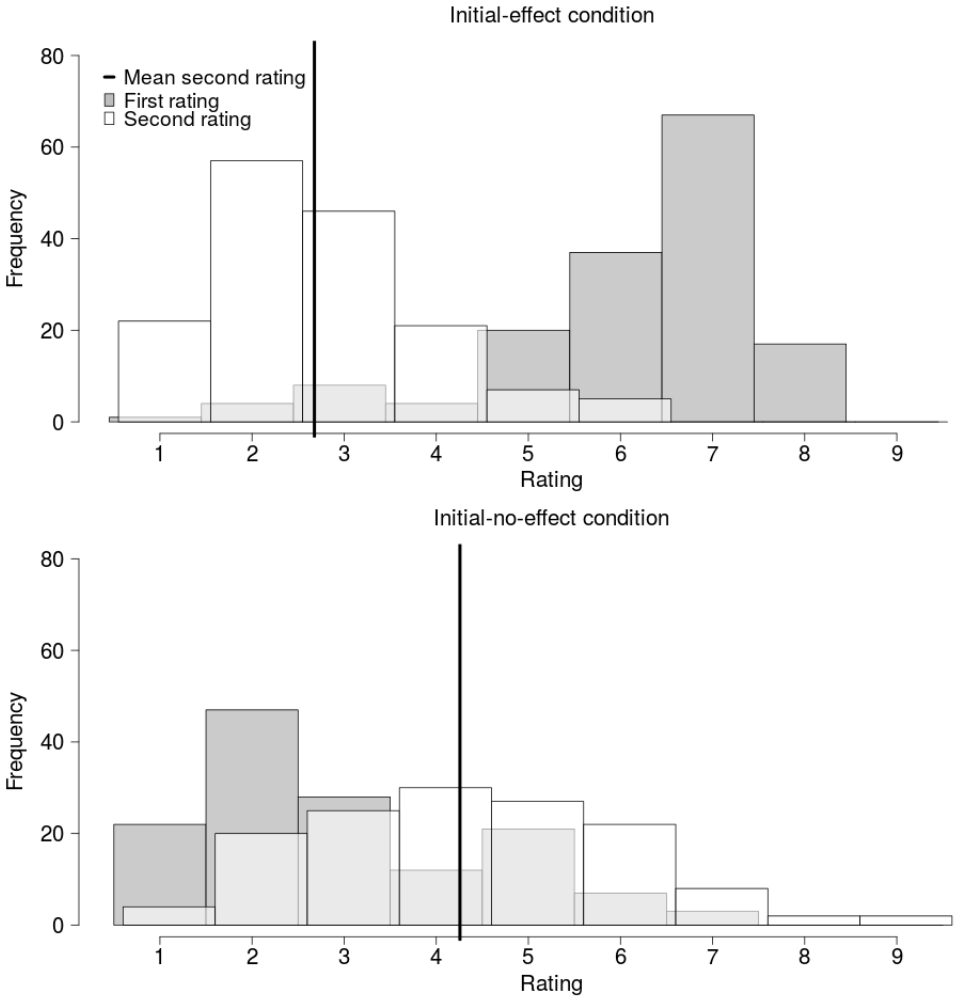


Figure 5. Distribution of Likert ratings in the academic sample. The focal ratings, group means on the second question, are shown by vertical lines.

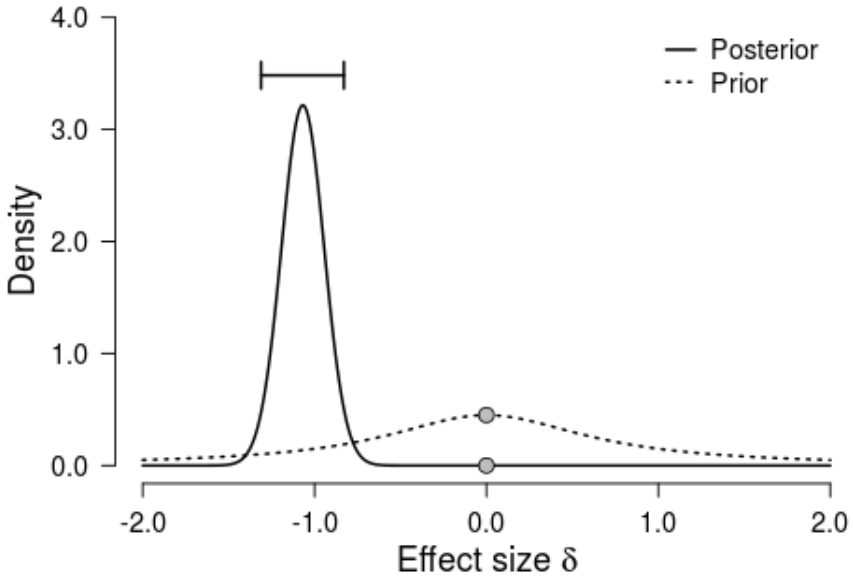


Figure 6. Samples from the posterior distribution of the effect size obtained by a two-sided Bayesian *t*-test in the academic sample. A line above the posterior spans the 95% credible interval of the effect size.

evaluating the hypothesis that the small effect scenario had a stronger impact on researchers in the initial-effect condition (i.e.,$\mathscr{H}_1 : a > b$) shows strong relative support for the null hypothesis $BF_{\mathscr{H}_1/\mathscr{H}_0} = .08$ (or $BF_{\mathscr{H}_0/\mathscr{H}_1} = 12.06$).

A two-sided Bayesian *t*-test yields relative support for the null-hypothesis as well, $BF_{two-sided\ \mathscr{H}_1/\mathscr{H}_0} = 0.16$ (or $BF_{\mathscr{H}_0/two-sided\ \mathscr{H}_1} = 6.46$). The posterior distribution of effect size for the two-sided follow-up analysis are displayed in Figure 8 (a). The figure shows samples from the posterior distribution of effect sizes evaluating the difference in influence of the effect scenario across the two conditions; the 95% credible interval is centered around $-.07$ with $[-.30, .15]$, indicating that there is no evidence for a difference in influence across the two conditions.

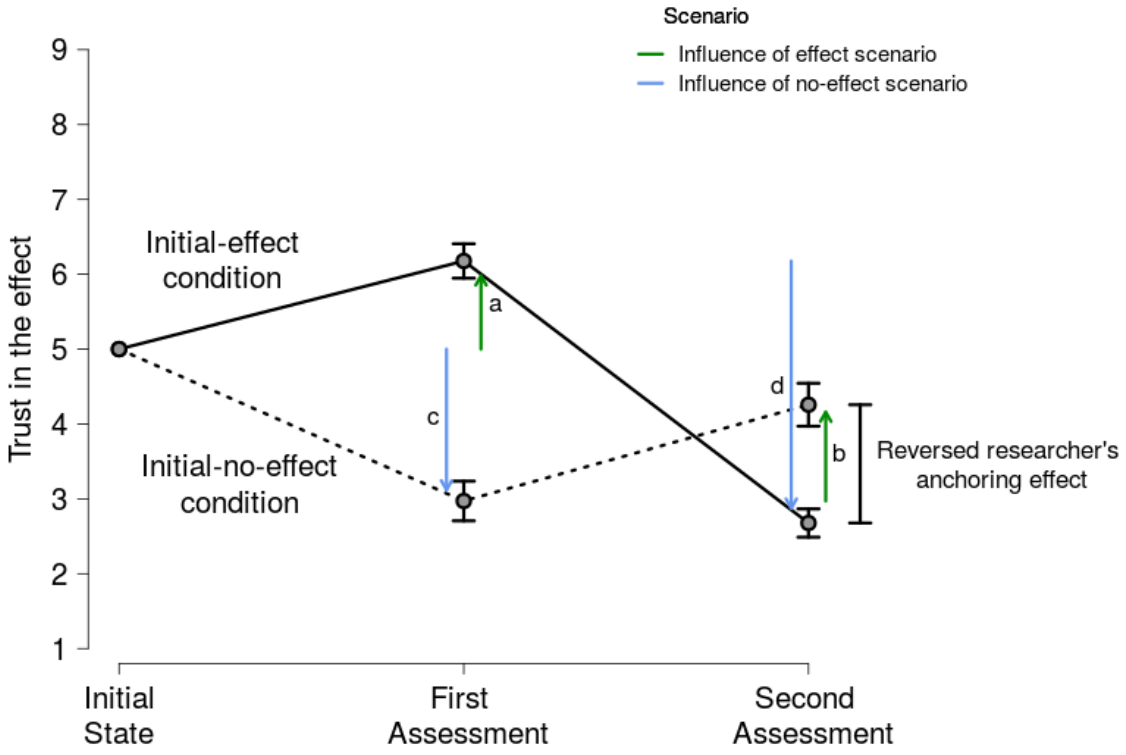
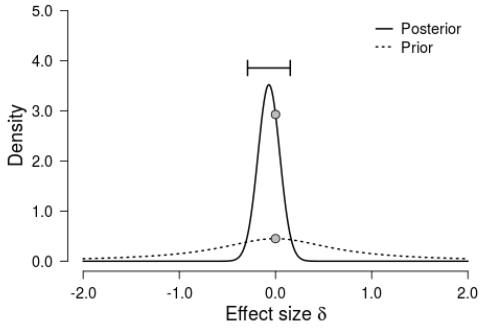


Figure 7. Effect of the two scenarios on trust across conditions in the academic sample. Dots represent condition mean values of trust, arrows indicate the region within 2 standard errors.

Blue lines in Figure 7 suggest that the no-effect condition had a large influence on researchers' ratings especially in the initial-effect condition. We formally test this pattern with the second follow-up analysis. The one-sided Bayesian *t*-test comparing the hypothesis that the no-effect scenario had a weaker impact on researchers in the initial-effect condition (i.e., $\mathscr{H}_2 : d < c$, cf. Figure 7) provides strong relative support for $\mathscr{H}_0$ with $BF_{\mathscr{H}_2/\mathscr{H}_0} = 0.006$ (or $BF_{\mathscr{H}_0/\mathscr{H}_2} = 176.5$). The two-sided Bayesian *t*-test provides strong support against the null hypothesis, $BF_{two-sided\ \mathscr{H}_2/\mathscr{H}_0} = 4.48 \times 10^{11}$. The posterior distribution of effect size for the two-sided t-test is displayed in Figure 8 (b). The 95% credible

(a) Influence of effect scenario



(b) Influence of no-effect scenario

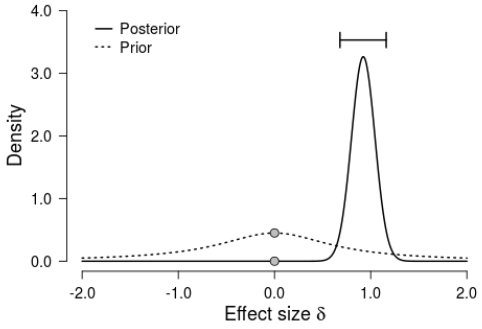


Figure 8. Posterior distributions of the effect size yielded by the two-sided Bayesian t-tests in the follow-up analysis on the academic data. A line above the posterior spans the 95% credible interval of the effect size.

interval shown in the figure spans $[.67, 1.17]$, the median of the distribution equals .92. Positive effect sizes indicate that the impact of the effect scenario was larger in the initial-effect condition than in the initial-no-effect condition. It appears, thus, researchers' trust in an effect was influenced particularly by a failed replication attempt resulting in a reversed RAE. The reversed RAE is visualized as a black line in Figure 7.

Student Sample

Our student sample consists of 365 introductory psychology students; 191 of which were randomly assigned to the initial-effect condition, 174 to the initial-no effect condition. No exclusion of participants took place. Figure 9 displays the distribution of rating scores in the student sample. The mean overall trust rating of the two groups is indicated by vertical black lines. The figure shows that similar to academics, students display a reversed RAE. The Bayes factor contrasting the RAE to the null is $BF_{RAE/\mathscr{H}_0} = .005$ (or $BF_{\mathscr{H}_0/RAE} = 183.3$), indicating strong relative support for the null hypothesis. The Bayes

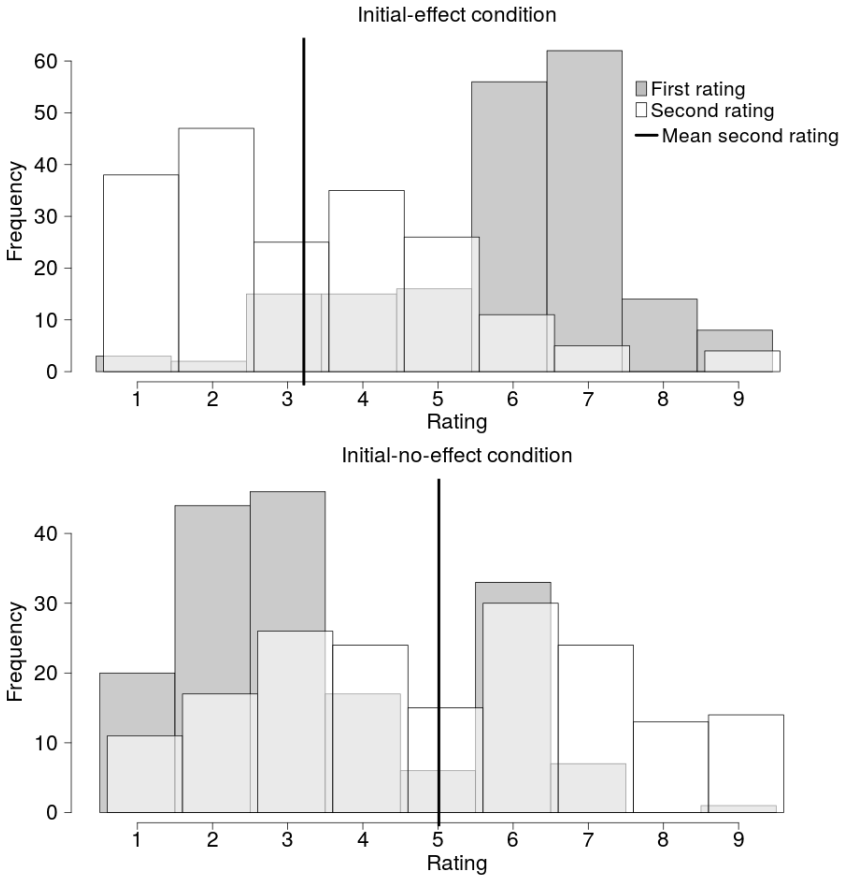


Figure 9. Distribution of Likert ratings in the student sample. The focal ratings, group means on the second question, are shown by vertical lines.

factor evaluating the two-sided hypothesis provides strong support for the presence of a reversed RAE, $BF_{two-sided\ RAE/\mathscr{H}_0} = 1.58 \times 10^{12}$. Figure 10 shows samples from the posterior distribution of the effect size obtained by a two-sided Bayesian t-test contrasting the overall trust of both groups. The figure also displays the 95% credible interval of the effect size, the interval spans $[-1.06, -.64]$ around median value $-.85$, indicating a large reversed RAE.

To determine how the reversed RAE arose, we carried out the follow-up analysis that had been specified in the analysis plan above. A one-sided Bayesian t-test evaluating the hypothesis that the small effect scenario had a stronger impact on students in the initial-

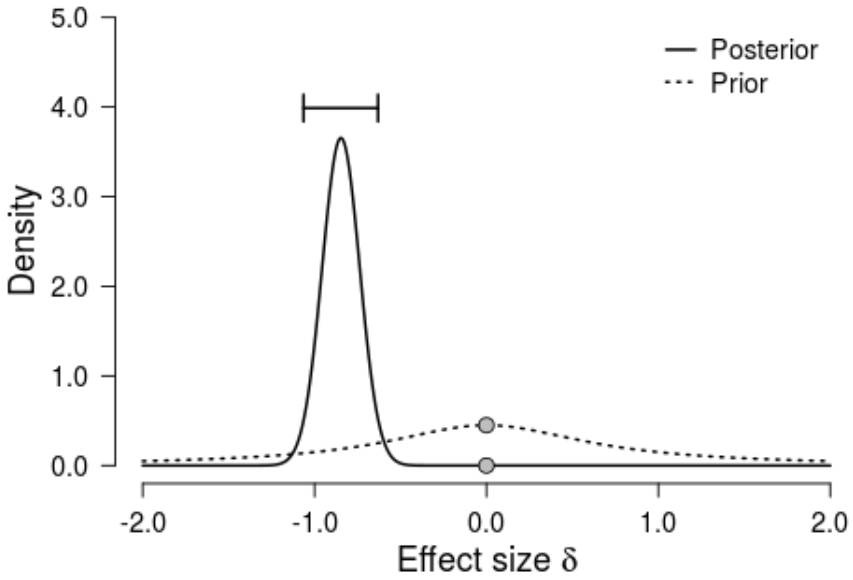


Figure 10. The figure shows samples form the posterior distribution of the effect size obtained by a two-sided Bayesian *t*-tests in the student sample. A line above the posterior spans the 95% credible interval of the effect size.

effect condition ($\mathscr{H}_1$) shows strong relative support for the null hypothesis $BF_{\mathscr{H}_1/\mathscr{H}_0} = .03$ (or $BF_{\mathscr{H}_0/\mathscr{H}_1} = 31.2$). The Bayes factor evaluating the two-sided hypothesis provides only anecdotal evidence against the null hypothesis, $BF_{two-sided\ \mathscr{H}_1/\mathscr{H}_0} = 3.29$. Figure 12 (a) shows posterior distribution and the 95% credible interval of effect sizes for the two-sided test: $[-.47, -.07]$ centered on $-.27$. Negative effect sizes indicate that the influence of the effect scenario was weaker in the initial-effect condition.

The Bayes Factor associated with $\mathscr{H}_2$, the hypothesis that the no-effect scenario would have a weaker impact on students in the initial-effect condition, shows strong support in favor of the null: $BF_{\mathscr{H}_2/\mathscr{H}_0} = .01$ (or $BF_{\mathscr{H}_0/\mathscr{H}_2} = 97.54$). The two-sided *t*-test is $BF_{two-sided\ \mathscr{H}_2/\mathscr{H}_0} = 4.51 \times 10^5$ showing strong support for the two-sided alternative. Fig-

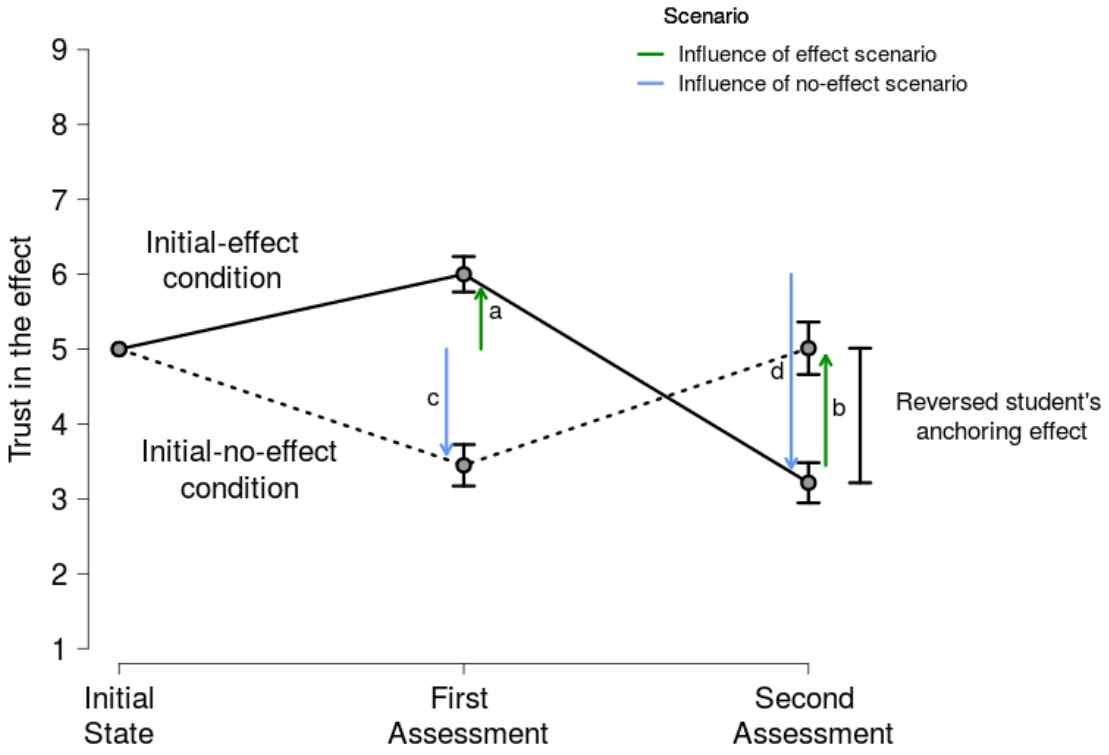
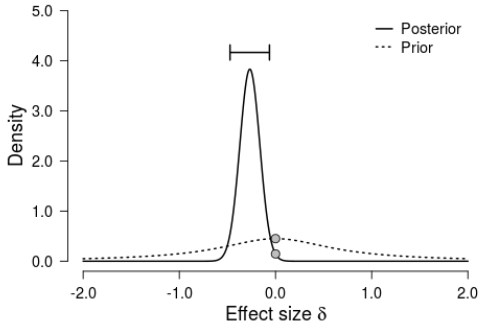


Figure 11. Effect of both research scenarios on trust across conditions in the student sample. Dots represent condition mean values of trust, arrows indicate the region within 2 standard errors

ure 12 (b) shows samples from the posterior distribution of effect sizes for this test. The positive effect sizes included in the 95% credible interval, $[.37, .79]$ around median value .58, suggest that the impact of the no-effect scenario was stronger in the initial-effect condition.

The effect of the two research scenarios on students' ratings is displayed in Figure 11. Blue lines show the impact of the small-effect research scenario, green lines give indication of the no-effect scenario's influence. In line with the results outlined above, comparison between blue lines suggest the reversed RAE to arise mainly through the no-effect scenario's

(a) Influence of effect scenario



(b) Influence of no-effect scenario

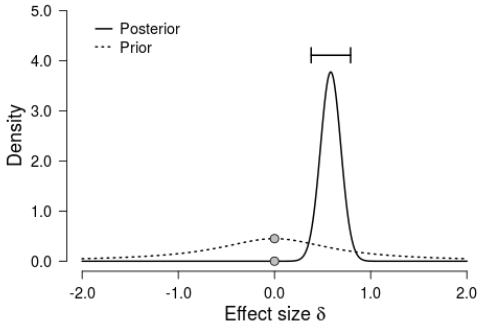


Figure 12. Posterior distributions of the effect size yielded by the two-sided Bayesian *t*-test in the follow-up analysis on the student data. A line above the posterior spans the 95% credible interval of the effect size.

impact that is more pronounced in the initial-effect condition.

Robustness Analysis

To ensure that results are not distorted by inattentive participants or responses based on possible misinterpretation, a secondary analysis was carried out. For this analysis we excluded all participants whose second rating changed in the direction opposite to what could be expected (i.e., who showed decreased trust after having been presented with a positive research outcome or vice versa). In the academic sample 7 participants showed such a pattern: 1 in the initial-effect condition, and 6 in the initial-no effect condition. This number was higher in the student sample with 38 participants updating their belief in an unexpected direction; 13 in the initial-effect, 25 in the initial-no-effect condition.[6]

None of the results changed direction. Changes in magnitude were negligible, except for the influence of the small-effect condition on students' ratings. Before exclusion only

[6]The result of the secondary analysis can be found on OSF: `https://osf.io/j658h/`.

moderate support was found for an increased weighting of the effect scenario in the initial-no-effect condition $BF_{two-sided\ \mathscr{H}_1/\mathscr{H}_0} = 3.29$; once students with unexpected changes in trust were excluded, the support for this hypothesis was strong with $BF_{two-sided\ \mathscr{H}_1/\mathscr{H}_0} = 2.3 \times 10^4$.

Discussion

In this paper we have investigated the presence of a primacy bias, the researcher's anchoring effect (RAE), in students and academics. According to the RAE hypothesis, subjects anchor on the most prior study outcome when integrating the results of several studies. The findings of our study contradicted our expectation and revealed a reverse anchoring effect for both researchers and students.

We found strong support for the existence of a *recency RAE* with researchers and students being most severely influenced by study outcomes presented last. This effect was caused mostly by failed replications having a stronger discounting impact on overall trust than failed initial experiments.

The higher importance researchers and students placed on unsuccessful replications might be indicative of the crisis of confidence (Pashler & Wagenmakers, 2012), caused by the replicability problems encountered in psychology. After trust in many prominent research outcomes has been undermined by unsuccessful replications or convictions of fraud (Open Science Collaboration, 2015; Schweinsberg et al., 2016; Yong, 2012), researchers might now show decreased trust in their colleagues' work after a result does not successfully replicate. A recency bias was only present for negative, not for positive study outcomes in psychological researchers. There is mild evidence, however, that undergraduates might show a recency bias for positive study outcomes as well: once inattentive participants were excluded from the student sample, the follow-up analysis provided strong support for the existence of a recency bias for positive study outcomes.

In this study we have evaluated only superficial trust. As one of our academic participants remarked: “Normally, we take more time to understand our results, not 90 seconds”. Another caveat is that in our study researchers were presented with research outcomes in immediate succession. It is possible that a RAE might arise over time once the belief in a research outcome has had time to sink in and manifest itself before contradictory evidence is encountered. This study was not intended to provide a holistic account of the way researchers update their knowledge and beliefs; rather, we investigated the presence of implicit anchoring biases during the intuitive integration of research results. These situations are prone to influence trust in scientific results, although they do not prescribe to all efforts of research integration.

Ideally, the integration of study outcomes should not depend on the temporal order in which results are presented. Although we found a recency REA, not a primacy REA as expected, the time-reversal heuristic proposed by AG may still help adjust for implicit bias when evaluating the credibility of an effect. When integrating findings researchers should mentally reverse the temporal order of results, according to the heuristic, in order to account for the presence of temporal order effects. By evaluating trust in an effect for different temporal sequences of the results at hand researchers can adjust their evaluations to account for temporal order effects, such as the recency RAE.

References

Baron, J. (2007). *Thinking and Deciding* (4th ed.). New York, NY: Cambridge University Press.

Chapman, L. J., & Chapman, J. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, *72*, 193-204.

Chapman, L. J., & Chapman, J. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, *74*, 271-280.

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman,

R. P. P. P., ... Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*, 640-647.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis (2nd ed.)*. Boca Raton (FL): Chapman & Hall/CRC.

Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology*, *25*, 165–173.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *11*, 719-731.

JASP Team. (2017). *JASP (Version 0.8.3.1)[Computer software].* Retrieved from `https://jasp-stats.org/`

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107-118.

Kuhn, D., Amsel, E., & O'Laughlin, M. (1980). *The development of scientific thinking skills*. New York, NY: New York: Academic Press.

Landon Jr., E. L. (1971). Order bias, the ideal rating, and the semantic differential. *Journal of Marketing Research*, *8*, 375-378.

Martindale, D. A. (2005). Confirmatory bias and confirmatory distortion. *Journal of Child Custody: Research, Issues, and Practices*, *2*, 31-48.

Morey, R. D. (2017). *BayesFactor package for R.* `https://cran.r-project.org/web/packages/BayesFactor/index.html`.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175-220.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 1 - 8. doi: 10.1126/science.aac4716

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on Psychological Science*,

7(6), 528 - 530. doi: 10.1177/1745691612465253

Phillips, N. (2017). yarrr: A Companion to the e-Book "YaRrr!: The Pirate's Guide to R" [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=yarrr (R package version 0.1.5)

R Core Team. (2017). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225-237.

Schönbrodt, F. D. (2016). *BFDA: Bayes factor design analysis package for R.* https://github.com/nicebread/BFDA.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2016). *Bayes Factor Design Analysis: Planning for Compelling Evidence.* Available at SSRN: https://ssrn.com/abstract=2722435 or http://dx.doi.org/10.2139/ssrn.2722435.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., ... Ly, A. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55 - 67. doi: 10.1016/j.jesp.2015.10.001

Sinkey, M. (2015). How do experts update beliefs? Lessons from a non-market environment. *Journal of Behavioral and Experimental Economics*, *57*, 55-63.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1131.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology*, *12*, 129-140.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t-test. *Psychonomic Bulletin & Review*, *16*, 752-760.

Yong, E. (2012). Replication studies: Bad copy. *Nature*, *485*(7398), 298 - 300. doi: 10.1038/

485298a