Review of Economic Studies (2025) **00**, 1–41 doi:10.1093/restud/rdaf032

© The Author(s) 2025. Published by Oxford University Press on behalf of The Review of Economic Studies Limited. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Advance access publication 10 June 2025

Retractions: Updating from Complex Information

Duarte Gonçalves
Department of Economics, University College London, UK
Jonathan Libgober
Department of Economics, University of Southern California, USA
and
Jack Willis
Department of Economics, Columbia University, USA

First version received February 2023*; Editorial decision February* 2025*; Accepted May* 2025 *(Eds.)*

We modify a canonical experimental design to identify the effectiveness of retractions. Comparing beliefs after retractions to beliefs (1) without the retracted information and (2) after equivalent new information, we find that retractions result in diminished belief updating in both cases. We propose this reflects updating from retractions being more complex, and our analysis supports this: we find longer response times, lower accuracy, and higher variability. The results—robust across diverse participant groups and design variations—enhance our understanding of belief updating and offer insights into addressing misinformation.

Key words: Belief updating, Retractions, Information, Complexity

JEL codes: D83, D91, C91

1. INTRODUCTION

Retracted information often influences beliefs even once widely discredited. A notorious example is the enduring belief in a link between vaccines and autism, fuelled by a subsequently retracted study in *The Lancet*. The article's impact persists as the belief in such an association remains widespread, significantly harming public health (see Pluviano *et al*., 2017; Motta and Stecula, 2021; Pullan and Dey, 2021; Gabis *et al*., 2022). While this case is illustrative, retractions are pervasive, and retracted information is rarely erased entirely.[1]

[1] Focusing on tracking retractions of academic papers, the Retraction Watch Database lists over 45,000 articles, with error and failure to replicate constituting a significant fraction of the retraction notices, in addition to misconduct (Brainard and You, 2018). Of the 10 most cited retracted articles as of October 2023 in the Retraction Watch Database,

The editor in charge of this paper was Andrea Galeotti.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

Variations of this phenomenon arise in a wide range of situations, from groundless rumours to erroneous earnings reports and from fraudulent research to false political claims. Naturally, each case is different, and retraction effectiveness in particular cases can always be attributed to unique intervening factors—*e.g.* special media coverage, ulterior financial or political motives, or source reliability. However, while idiosyncratic factors may play important roles, issues in updating from retractions are documented too consistently and in too wide a range of settings for case-by-case explanations to be the whole story. This observation suggests moving beyond idiosyncratic factors to identify causes common to retractions generally.

In this article, we investigate if and why there is a fundamental friction in updating beliefs from retractions. To this end, we modify a canonical experimental design to identify and quantify diminished updating from retractions relative to direct evidence, absent a variety of idiosyncratic confounds. Our analysis reveals beliefs update significantly less from retractions than from direct evidence, a finding that challenges explanations unrelated to intrinsic characteristics of retractions. We propose a simple explanation: retractions convey more complex information than direct evidence. To support this hypothesis, we present evidence based on common empirical measures of complexity—specifically, accuracy, response time, and response variability. We document these basic patterns across numerous variations, show that they are robust to certain alternative implementation details, and argue against several natural competing explanations unrelated to the processing of retractions.

Identifying the diminished effectiveness of retractions requires a clear benchmark against which updating from retractions can be measured. The aforementioned confounding factors in particular settings complicate assessing how individuals should interpret any given retraction. Whether the retraction was prompted by negligence or malfeasance, casts doubt on other evidence, is politically motivated, or is disputed all influence a retraction's correct interpretation but may not be precisely quantifiable. At the same time, as individuals may err when interpreting *any* information, the mere presence of an error does not imply differential treatment of retractions compared to other pieces of new evidence. Indeed, previously documented updating biases may appear capable of explaining the diminished effectiveness of retractions. Perhaps most notable among these is *confirmation bias*—updating more when information confirms one's prior beliefs than when it does not (Rabin and Schrag, 1999)—as it suggests individuals resist disregarding information supporting their beliefs, such as a discredited study.

We develop a variation on the classic balls-and-urns experiment to identify and quantify updating from retractions absent a variety of idiosyncratic confounds. This canonical experimental design is widely used to study limitations in information processing, for example, in belief updating (Benjamin, 2019; Ba *et al.*, 2022; Augenblick *et al.*, 2023), social learning (Anderson and Holt, 1997; Weizsäcker, 2010; Angrisani *et al.*, 2021), and asset pricing (Halim *et al.*, 2019). Our version allows us to repeatedly provide retractions that are informationally equivalent to new observations to participants facing identical problems. At the same time, we, as analysts, have access to quantifiable information about the objective truth. These properties are essential to distinguish belief updating issues specific to retractions.

We briefly describe how we modify the canonical balls-and-urns design to accommodate retractions. As is standard, participants are presented with draws of balls from a box (with replacement), which are either blue or yellow. In our version, balls can be "noise balls," which are blue and yellow in equal proportion, or a "truth ball," which is either blue or yellow. Instead of asking if the box has a majority of blue or yellow balls, we elicit beliefs about whether the

seven had over a 100 citations since retraction, and two that had fewer had only been retracted in 2023 (Retraction Watch, 2023). We highlight that many papers that fail to be replicated are not retracted (Serra-Garcia and Gneezy, 2021).

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

truth ball is blue or yellow, an equivalent event. After a number of draws, in which participants are shown the colour but not the truth/noise status of the ball drawn, we then either present another such draw or inform participants whether a randomly chosen earlier ball draw was the truth ball or a noise ball. We refer to the disclosure that an earlier draw was noise as a *retraction*. In our formulation, retractions *only* provide information that a given signal was noise, analogous to a researcher having fabricated data or a news article relying on made-up claims.[2] In some practical instances, a retraction may also be coupled with additional information contradicting the initial subsequently retracted statement. Our design decouples these, as these are decoupled in several applications; however, our design also allows us to study updating from retractions with additional information.

We test for retraction effectiveness by comparing beliefs over the truth ball's colour after updating from retractions to (1) beliefs without having observed the retracted observation in the first place and to (2) beliefs updated from new draws with identical Bayes updates (*i.e.* a new draw of colour opposite the retracted observation). These comparisons identify whether participants update less from retractions than from either (1) the retracted observation or (2) a new informationally equivalent observation. We find participants update less from retractions in both comparisons. The magnitude of this diminished updating is significant: beliefs update on average about 50% less from retractions than new draws (see Section 3).

Why are retractions less effective? The minimality of our design strongly suggests that any explanation should be intrinsic to how retractions are processed. Consistent with this intuition, we consider a general class of *quasi-Bayesian* belief updating models that nests—but also accommodates usual deviations from—Bayesian updating. We show that results cannot be reconciled with any explanation that does not treat retractions as inherently different despite identical informational content (see Proposition 1 in Section 2.1). Notably, widely documented deviations, including confirmation bias, cannot rationalize our findings.

Our explanation is that retractions are more complex than direct information. Borrowing from Pearl (2009), we formally articulate a distinctive feature of retractions: Unlike the evidence they typically refer to, which is *directly informative* about the state—in our setting, the colour of a draw—retractions are only *indirectly informative* about the state, that is, they are only informative in light of the retracted evidence.[3] Consequently, retractions always require additional contingent reasoning relative to direct evidence. Indeed, recent literature has shown not only that considering more contingencies renders problems more complex and leads to inference errors in various domains (Esponda and Vespa, 2014; Martínez-Marquina *et al.*, 2019; Ali *et al.*, 2021; Esponda and Vespa, 2021) but also that complexity considerations can explain several well-documented behavioural biases (Oprea, 2020; Ba *et al.*, 2022; Oprea, 2022; Enke *et al.*, 2023a). This background motivates our hypothesis that the greater complexity inherent to retractions explains diminished updating.

To test this hypothesis, we analyse three empirical complexity measures: (1) accuracy of belief reports, (2) speed of decision, and (3) variability of belief reports. All three of these data

[2] Retractions of scientific articles are often due to problems with experimental conduct, suggesting uninformative findings but leaving open the possibility that the tested hypotheses are true. One example illustrating this possibility is the retracted study on the impact of contact on opinion formation; despite the fabrication of evidence from an early study on this topic, Broockman and Kalla (2016) subsequently conducted an experiment that did indeed provide evidence for one of its key hypotheses.

[3] We say that x is directly informative about y if x and y are neither independent nor independent conditionally on some third variable z. In our setting, the colour of a draw is directly informative about the state, but learning about its noise status is indirectly informative: only by conditioning on its colour can it be informative, and it is otherwise independent.

types are borrowed from past work in which they were used as measures of complexity and cognitive noise—see, for example, Caplin *et al.* (2020) and Enke and Shubatt (2023) for the first; Wright and Ayton (1988), Krajbich *et al.* (2012), and Frydman and Jin (2022) for the second; and Khaw *et al.* (2021) and Enke and Graeber (2021) for the third. All proxies are larger when updating from retractions compared to equivalent new information, as well as compared to when the retracted signal had never been seen. These patterns suggest higher complexity for retractions, as proposed.

We leverage natural variation provided by our design, which suggests variation in the relative complexity of retractions and verify that these covary with updating strength. First, we compare updating from retractions of more or less recent observations. If the most recent observation is retracted, participants can simply "go back" to a past belief, making updating easier. This point suggests that retractions of more recent evidence are less complex, corroborated by our empirical complexity measures. In line with our mechanism, participants also update more when the most recent observation is retracted than when retractions refer to an earlier observation. Second, we examine updating from new observations after retractions. Beliefs respond less to new observations after retractions, and our empirical measures indicate inference is more complex.[4]

We further examine how updating patterns vary across histories. We use standard (Grether, 1980) log-odds regressions to compare biases from retractions to those typically documented in updating from new observations. While updating from new evidence exhibits confirmation bias, retractions entail both underinference and anticonfirmation bias. In line with this, confirmatory retractions are least effective (relative to equivalent new evidence) at histories inducing more extreme beliefs. These findings offer valuable insights into the unique influence of retractions on belief-updating behaviour.

We conducted a wide range of robustness checks to ensure the validity and generalizability of our results. First, we assessed whether our results simply reflect limited participant understanding and inattention despite our screening measures and attention checks. We consider removing participants who are "noisy" or prone to mistakes, as well as those who did not correctly answer unincentivized comprehension questions on the first try. We can also rule out misinterpreting that the draws are with replacement. A theme that emerges is that our results are maintained, if not strengthened, when restricting to participants who appear to have understood the task better.[5]

Second, we explored variations in participant characteristics. We find that our results on the diminished updating from retractions and its greater complexity are robust to whether participants perform better or worse in quantitative tasks, are more or less confident about their belief updating, are more or less experienced with the task, or more or less Bayesian in updating from observations. Although we are not powered for a fully fledged within-participant analysis, inspection of individual heterogeneity in our results indicates that the diminished effectiveness of retractions compared with new observations is a general phenomenon in our sample.

Third, we examined the impact of design variations, such as having shorter histories, omitting the history of past draws, garbling information so that the state is never perfectly learned, and different wording for retractions. These variations allowed us to assess whether our main findings were driven by specific features of the information process or details of the experimental. We

[4] This finding is relevant for situations where (1) some evidence is found inaccurate and (2) further contradictory evidence is subsequently revealed. The diminished updating from retractions under (1) and the diminished updating *following* retractions in (2) indicate that both elements contribute to a diminished updating from retractions.

[5] This finding is perhaps unsurprising since documenting any effect requires that participants act differently for retractions; if participants answered randomly or always answered 50–50, we would not document any difference. But it is worth emphasizing that most of our sample did very well on unincentivized comprehension questions, confirming our assertion that our design achieved its desired simplicity despite the richness it contains.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

found that our results remained robust across all these different experimental setups. Notably, our results are robust even when beliefs are only elicited at the end of each round—dispelling concerns that our findings are driven by information being hard to disregard once it has been "acted upon," as would be suggested by a cognitive dissonance explanation. Overall, our comprehensive analysis underscores the robustness and reliability of our findings across various conditions and contexts.

These observations support the claim that our work provides some of the first evidence that diminished retraction effectiveness could have origins (at least partially) in fundamental information processing properties. An advantage of showing this in a setting where beliefs can be elicited directly is that it suggests a unified and systematic approach to analysing patterns in belief updating from retractions. Of course, retractions in practice will differ from those we present to participants in our experiment. Indeed, we expect many elements deliberately precluded by design, such as memory frictions, salience, or motivated reasoning, to play a significant role in many settings where retractions appear less effective.

Our results are both of practical value and theoretical interest. We designed the experiment to connect the diminished effectiveness of retractions to information processing errors.[6] From a theoretical standpoint, our findings motivate the development of theoretical models of costly information processing that treat indirect information differently from direct information—even when their informational content is the same. From a practical standpoint, our analysis provides guidelines regarding how individuals respond to retractions, potentially relevant to campaigns targeting misinformation. The fact that retraction failures arise due to information processing errors suggests limits to the "this time is different" logic policy-makers may adopt—it is generally unreasonable to expect a retraction to be entirely successful in correcting beliefs. In many real-world cases, appreciating the inability to correct beliefs with retractions ex post may very well have changed the calculus regarding decisions to disseminate information ex ante.[7]

1.1. *Past work on causes and consequences of continued influence*

The closest precedent for the diminished effect of retractions comes from the literature on the *continued influence effect* in psychology. Reviewing this literature, Ecker *et al.* (2022) define this effect as the finding that "misinformation can often continue to influence people's thinking even after they receive a correction and accept it as true." Johnson and Seifert (1994) provided an early articulation of such a result, asking participants to recount the cause of the start of a fire and finding that they would still rely upon discredited information.[8] Chan *et al.* (2017) and Walter and Tukachinsky (2020) provide meta-analyses of the literature—across experiments that range from stories to advertising, scientific retractions, and beyond—and find that corrections fail to fully correct beliefs. These and similar patterns have been extensively documented in many settings; Appendix A discusses specific applications.

[6] In this sense, our article is part of a sizable literature that, while motivated by anecdotal or domain-specific evidence of biases, uses fundamental belief updating tasks to highlight a relevant theoretical mechanism; see, for example, Oprea and Yuksel (2022), Esponda *et al.* (2022), Hartzmark *et al.* (2021), and Agranov *et al.* (2022).

[7] We do not speak to issues of how these biases interplay with information *preferences*, although this might influence some of these decisions in practice; see Masatlioglu *et al.* (2021), Gul *et al.* (2021), Ambuehl and Li (2018), or Charness *et al.* (2021) for papers studying this element.

[8] A more extreme reaction is *backfiring*, in which participants believe more strongly in the retracted information. Nyhan and Reifler (2010) documented this pattern when providing participants with information about the presence of weapons of mass destruction in Iraq during the early 2000s (and subsequently providing corrections). But unlike continued influence, backfiring has not been replicated for the most part. See Nyhan (2021) for an authoritative discussion.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

Ecker *et al.* (2022) and Lewandowsky *et al.* (2012) discuss a number of channels for continued influence to emerge. These include biases related to memory storage (*e.g.* in terms of "mental models" individuals used) and retrieval,[9] as well as explanations based on the perceived credibility of a retraction and the extent to which it clashes with an individual's worldview.[10] A confounding factor, however, is that in many existing papers, the "continued influence effect" and related "biases" could actually be consistent with Bayesian updating, depending on the implementation of retractions (see Pennycook *et al.*, 2021; Guay *et al.*, 2023).

Our implementation of retractions within a balls-and-urns design differs from the existing literature in that we, as analysts, know a retraction's objective informational content. This advantage facilitates the identification of differences in information processing *due to information being a retraction*, and our proposed mechanism is intrinsically tied to how retractions generate information. Further, while each explanation above is undoubtedly important in some circumstances and less relevant in others, our design allows us to differentiate our proposed mechanism from these setting-specific explanations—issues discussed in more detail in Section 6.

1.2. *Other work on belief updating biases*

Our article builds on the experimental literature studying belief updating. Benjamin (2019) provides a comprehensive survey; of independent interest, we replicate many of its key findings.[11]

We aim to identify and distinguish the updating from retractions and other well-known biases. For instance, we document *base-rate neglect* (whereby agents underweight the prior when updating; see, *e.g.* Esponda *et al.*, 2024), as well as *confirmation bias*, discussed above (see also Rabin and Schrag, 1999).[12] We show in our theoretical framework that the diminished effectiveness of retractions is *distinct from these biases* and cannot be explained by models that do not treat retractions inherently differently.

Our analysis suggests that "indirect information" is more complex to process than "direct information." Though our focus on retracting information is new, the idea that contingent reasoning entails higher cognitive effort has been illustrated in different settings. One of the first documented difficulties of contingent reasoning was Charness and Levin (2005) for the winner's curse.[13] Closer to our study is Enke (2020), which documents in a pure prediction setting that many participants consistently fail to account for the informational content from the absence of observations, suggesting a failure of contingent reasoning. One microfoundation driving greater complexity for "indirect information" than "direct information" is that participants face *higher cognitive imprecision* in their understanding of the informativeness of a retraction than of an observation—see Woodford (2020) for a survey, and Enke and Graeber (2022) and Augenblick *et al.* (2023) for recent applications to belief updating.

[9] In particular, the mere passage of time may affect the perception of evidence (Jacoby *et al.*, 1989).

[10] As illustrated by Susmann and Wegener (2022), a possible reason underlying this belief-updating pattern is that it reflects an implied cognitive dissonance (Harmon-Jones and Mills, 2019), owing to the psychological discomfort following from holding two contradicting ideas that retractions induce.

[11] For recent papers studying these patterns in belief updating, see, for instance, Ambuehl and Li (2018), Coutts (2019), and Augenblick *et al.* (2023).

[12] To avoid confounding factors, our design features exogenous information; Charness *et al.* (2021) study how biases may influence participants' *choice* of sources of information.

[13] See Esponda and Vespa (2014) and Martínez-Marquina *et al.* (2019) for more on difficulties in contingent reasoning in particular games.

2. FRAMEWORK AND DESIGN

2.1. *Information arrival: draws and retractions*

Our experiments consider a simple belief updating problem. Participants form beliefs over a state θ, which takes one of two values with equal probability, say $\theta \in \{yellow,\ blue\}$. We write $\hat{p}_t$ to denote a participant's belief that $\theta = yellow$, given all the information observed by period t. We use the term "signal" as a generic term for information throughout. Our interest is in two kinds of information participants may have access to: *draws* and *retractions*.

Draws. In a given period t, participant i may observe $s_t \in \{yellow,\ blue\}$, a signal informative about θ and drawn independently conditional on θ. We refer to this kind of information as an "observation" or "draw." In our baseline experiment, each observation s_t can correspond either to the *truth*, in which case $s_t = \theta$, or to *noise*, in which case it is given by an independent $\epsilon_t \in \{yellow,\ blue\}$. Denoting the former event by $\{n_t = T\}$ and the latter by $\{n_t = N\}$, we focus on cases where these events are independent of θ. To summarize, we have $s_t = \theta$ if $n_t = T$, and $s_t = \epsilon_t$ if $n_t = N$, where $n_t \in \{T, N\}$ and n_t, ϵ_t, and θ are independent. For simplicity, we write $S_t = \{s_1, \ldots, s_t\}$. In this setup, if $n_t = T$, then s_t reveals the state. In one variant, we additionally allow s_t to be *im*perfectly informative even when $n_t = T$, but we defer our discussion of this possibility.

Retractions. The second kind of signal a participant may receive in period t is a *retraction*. Formally:

Definition 1. A *retraction* of the ρth observation informs that it was noise, *i.e.* $n_\rho = N$.

Retractions provide information about *past signals*. The process by which retractions are determined—for example, how observation ρ is chosen to be retracted—matters for how they should be interpreted, a theme we return to later. Important for identification in our experimental paradigm, we focus on the following type of retraction:

Definition 2. A *verifying retraction* of the ρth observation is a retraction in which ρ (the period that the retraction refers to) is chosen independently of that or other observations' truth value.

Our experiment implements verifying retractions by selecting ρ uniformly at random from all past observations and subsequently revealing n_ρ, that is, whether this observation was noise.[14] We refer to the signal that informs the participant of n_ρ as a *verification*, noting that a verification is a retraction when $n_\rho = N$. The indicator variable r denotes the occurrence of a retraction, whereby $r_t = 1$ if a retraction occurs in period t and $r_t = 0$ otherwise.

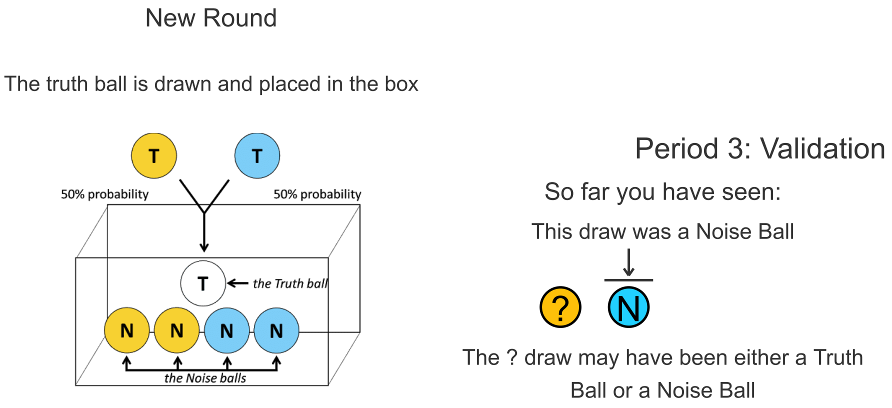
2.2. *Experimental design*

We turn to how we operationalized this information arrival process in our experiments. Here, we focus on our baseline setup and subsequently discuss how we modified it in our variants (Figure 1).

In each *round* of the experiment, we provided information about a state across up to four *periods*:

[14] This implementation implies one learns that past information was not noise when $n_\rho = T$, which, in the current setting, perfectly reveals θ in turn.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025



New Round

The truth ball is drawn and placed in the box

Period 3: Validation

So far you have seen:

This draw was a Noise Ball

The ? draw may have been either a Truth Ball or a Noise Ball

FIGURE 1

Screenshots of the experimental implementation

Notes: This figure provides screenshots of the visuals provided to the participants corresponding to the operationalization of the information structure.

(1) At the start of the round, a *truth ball* (corresponding to the state θ) is chosen at random to be either yellow or blue, with equal probability. The truth ball is then placed into a box with four *noise balls*, two yellow and two blue (corresponding to $P(n_t = N) = 4/5$ and $P(\epsilon_t = yellow) = 1/2$).

(2) In periods one and two, participants obtain a *new observation*: a draw from the box with replacement. They see the ball's colour (s_t) but not whether it is the truth ball or a noise ball (n_t).

(3) In periods three and four, and independently across periods, participants either obtain a new observation (as above) or observe a verification of an earlier observation (ρ) from the same round, with equal probability. Under a verification, one of the previous draws is chosen uniformly at random, and it is revealed whether that draw was a noise ball ($n_\rho = N$)—a *retraction*—or the truth ball ($n_\rho = T$). If the draw turns out to have been the truth ball, the round ends, as at that point, the state (the colour of the truth ball) is fully revealed.

Participants report their belief regarding the probability that the truth ball is blue or yellow ($\hat{p}_t$) at the end of each period, that is, after each new signal (observation or retraction). These reports are incentivized, as detailed in Section 2.3. Each participant plays a total of 32 rounds, and no feedback on performance is provided until performance-based payouts are made at the end of the experiment.

Variants. Sections 3, 4, and 5 present results using the described implementation. However, in total we ran four experiments with six main, across-participant treatments (including the baseline). Table 1 summarizes these treatments and details where the article discusses them. These variants aimed to demonstrate the robustness of our findings and to investigate the underlying mechanisms. Table 5 presents sample characteristics for each treatment. We defer detailed descriptions of each variant until Section 6.

2.3. *Implementation details*

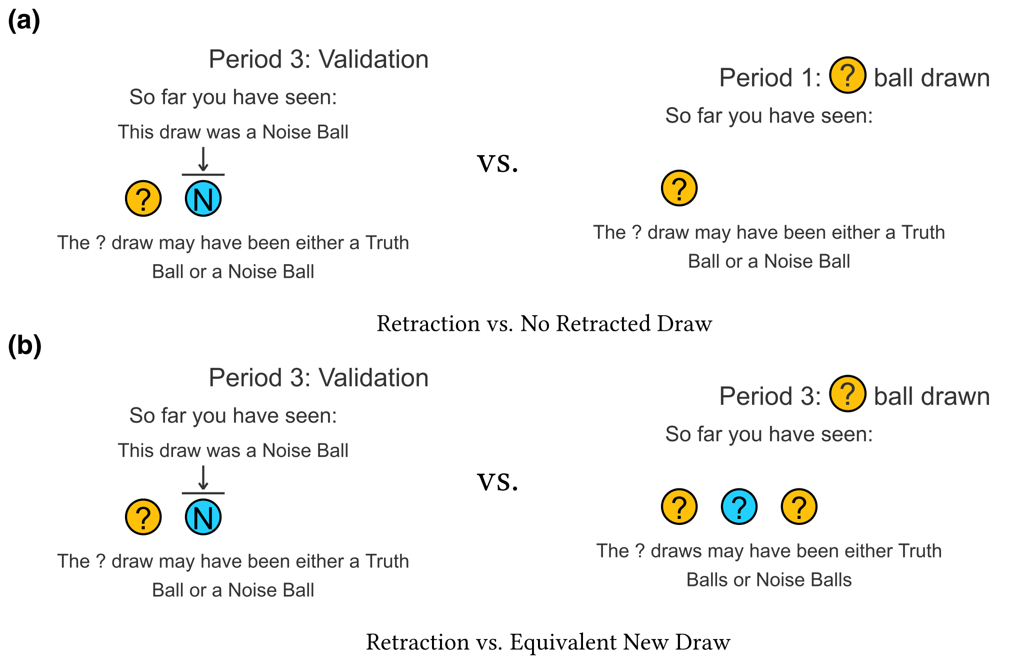
This section discusses implementation details for all experiments discussed in the article.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

TABLE 1
Summary of treatments

Experiment	Treatment	Venue	No. of participants	Duration (min)	Payment	Sections
A	Baseline	MTurk	211	31	$11.96	Throughout
A	Elicit at end	MTurk	204	24	$8.14	6.3
B	Garbled information	MTurk	164	40	$11.03	6.3
C	Baseline	Prolific	155	49	$11.64	Throughout
C	Retraction information	Prolific	164	52	$11.76	6.1
C	No history	Prolific	164	51	$11.80	6.3
D	Short histories	Prolific	150	26	$12.02	6.3

Notes: This table summarizes the four experiments, their respective treatments, and the sections of the article where they are discussed. "Duration" and "Payment" refer to the average time spent in the experiment in minutes and to the average payment in USD, respectively.


(a)
Period 3: Validation
So far you have seen:
This draw was a Noise Ball
?
N
The ? draw may have been either a Truth Ball or a Noise Ball
VS.
Period 1: ? ball drawn
So far you have seen:
?
The ? draw may have been either a Truth Ball or a Noise Ball

Retraction vs. No Retracted Draw

(b)
Period 3: Validation
So far you have seen:
This draw was a Noise Ball
?
N
The ? draw may have been either a Truth Ball or a Noise Ball
VS.
Period 3: ? ball drawn
So far you have seen:
? ? ?
The ? draws may have been either Truth Balls or Noise Balls

Retraction vs. Equivalent New Draw

FIGURE 2
Illustrative examples to explain the empirical strategy (a) retraction versus no retracted draw and (b) retraction versus equivalent new draw

Notes: This figure provides an illustrative example of the empirical strategy for testing Hypothesis 1. According to Proposition 1, for any quasi-Bayesian, beliefs should be identical for each of the displayed histories. (a) Comparing beliefs following (yellow, blue, *retraction of the* blue) to those following (yellow), tests if beliefs are the same when evidence gets retracted and when such evidence was never observed. (b) comparing beliefs following (yellow, blue, *retraction of the* blue) to those following (yellow, blue, yellow), tests whether updating from retractions is the same as from otherwise equivalent direct information.

Experimental interface. Figure 2 summarizes the explanatory visuals shown to participants in our baseline treatment, and Supplementary Appendix L contains the experiment's instructions. Participants reported beliefs using a slider, which displayed the stated probability assigned to the truth ball being yellow and the complementary probability assigned to it being blue. After the instructions, participants were given two rounds of unincentivized "practice" to familiarize themselves with the interface.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

Participant pool. We ran four experiments (labelled A–D) comprising different treatments as described in Table 1. The first (A and B) were on Amazon Mechanical Turk (MTurk), and the remaining two (C and D) were on Prolific, with different experiments corresponding to different requests for participants.[15] The latter two experiments were run to address the mechanisms underlying our results, in response to reviewer feedback. To ensure that the choice of venue did not influence our main findings, we ran our baseline treatment on both platforms. We recruited 1,212 participants in total; Appendix B presents sample characteristics for all experiments. The assignment of participants to treatments was randomized within each experiment. We took several steps to ensure that our participant pool was of high quality; Section 6.1 describes these steps in greater detail.

Payments. We incentivized participants to report their beliefs truthfully using a binarized scoring rule (Hossain and Okui, 2013; Mobius *et al.*, 2022). By reporting $\hat{p}_t \in [0, 1]$, a participant would receive $\$High$ with probability $(1 - (\mathbf{1}\{\theta = yellow\} - \hat{p}_t)^2)$ and $\$Low$ with complementary probability, where $\$High$ and $\$Low$ correspond to \$12.00 and \$6.00 for experiments A and B (ran in 2020 and 2021, respectively) and to \$13.00 and \$7.00 for experiments C and D (ran in 2024). To determine payments, we used a report from a single randomly selected period of a randomly selected round.[16]

In the instructions—but not in the main interface—we provided information on the elicitation procedure, phrased as eliciting the probability that the truth ball was either yellow or blue. The instructions explained that the procedure was meant to ensure they were incentivized to answer truthfully. As the elicitation scheme we used may appear complicated, we sought to limit the extent to which participants were required to focus on it while maintaining transparency. Danz *et al.* (2022) show that the binarized scoring rule can introduce noise and "pull beliefs toward the centre," although the magnitude appears to vary across participant pools and might be lower for online platforms (see Healy and Kagel, 2023). As our primary focus is on how updating from retractions compares to direct information, any difference is still meaningful. Moreover, any potential underreaction would make it *harder* to detect an effect of retractions, not easier.

We also asked additional questions on mathematical ability, which were incentivized via a \$0.50 reward if a randomly chosen question was answered correctly. The average duration and compensation were 31 min and \$9.98 (\$24.36/h) for experiments A and B and 45 min and \$11.81 (\$20.52/h) for C and D. For comparison, this rate is higher than the MTurk experiment of Enke and Graeber (2022) and four times the MTurk average of \$5.00.

Preregistration. Experiment A was registered using the AEA RCT Registry under RCT ID AEARCTR-0003820, while Experiment B was registered using the AEA RCT Registry under RCT ID AEARCTR-0006106. The experimental design and recruitment targets for these experiments were pre-registered. Of our four main hypotheses presented below, Experiment A's preregistration formulated Hypotheses 1, 3, and 4, in addition to the analysis in Section 5. While our registration discusses difficulty of updating from retractions as a mechanism, our formal hypothesis on complexity, Hypothesis 2, was introduced subsequently, as feedback we received convinced us they provided evidence for our proposed mechanism. Experiments C and D were run to test hypotheses suggested by reviewers at this journal and not preregistered.

[15] On each platform, we excluded participants who participated in the earlier experiment on that platform.

[16] Azrieli *et al.* (2018) show that random selection is essentially the unique problem-selection mechanism inducing incentive compatibility when preferences satisfy state-wise monotonicity, namely that participants prefer higher payments given any realization of uncertainty (selected problem/underlying states).

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

3. DIMINISHED UPDATING FROM RETRACTIONS

3.1. *Theoretical predictions*

We start by clarifying how our design enables us to identify if and how updating differs between retractions and direct information. The core of our identification strategy comes from our result that, in our setting, any difference in updating would be inconsistent with any explanation that does not treat retractions differently from direct information—including the general class of frameworks used to explain many known deviations from Bayesian updating. In the process, we clarify why seemingly similar paradigms fail to do so and the extent to which continued influence could be consistent with rational belief updating.

Let $P(\cdot)$ denote *objective* probabilities associated with the data generating process, and $\hat{P}(\cdot)$ denote i's *subjective* beliefs. For a Bayesian decision-maker, subjective beliefs about θ, $\hat{p}_t := \hat{P}(\theta \mid \mathcal{H}_t)$ coincide with the objective probability that $p_t := P(\theta \mid \mathcal{H}_t)$, where $\mathcal{H}_t$ represents the entire history at period t, that is, the set of all the draws observed as well as any retractions, fixing the order. Past work has routinely rejected this hypothesis. A common alternative is to assume there is a strictly increasing f_i such that $\hat{p}_t = f_i(p_t)$. It follows that, upon observing some event E at t, updating of beliefs $\hat{p}_{t-1}$ is given by the following identity:

$$\mathcal{L}(f_i^{-1}(\hat{p}_t)) = \mathcal{L}(f_i^{-1}(\hat{p}_{t-1})) + K(E), \tag{1}$$

where $\mathcal{L}(p) := \ln(\frac{p}{1-p})$ denotes the log-odds of $p \in (0, 1)$, and $K(E) := \ln(\frac{P(E|\theta=yellow)}{P(E|\theta=blue)})$ the log-likelihood of E, with the understanding that $\mathcal{H}_t = \mathcal{H}_{t-1} \cup E$. As long as $\hat{p}_t = f_i(p_t)$, this relationship holds for all histories $\mathcal{H}_t$; this point will be useful in our analysis.

Inspired by Cripps's (2021) axiomatic work, we call a decision-maker who updates according to (1) "quasi-Bayesian":

Definition 3. We say that a decision-maker is *quasi-Bayesian* if there exists a strictly increasing f_i such that, for any information $\mathcal{H}_{t-1}$ and event E, $\hat{p}_t = \hat{P}(\theta \mid \mathcal{H}_{t-1}, E)$ can be derived from $\hat{p}_{t-1} = \hat{P}(\theta \mid \mathcal{H}_{t-1})$ according to (1).

Note that, to accommodate some forms of confirmation bias, it may be necessary to allow the function f_i to depend on the prior belief from which participants update; we strive to be as agnostic as possible and our comparisons will hold across a number of possible assumptions. We return to a discussion of possible microfoundations for distortions under quasi-Bayesianism in our discussion of mechanisms in Section 4.

Our main comparisons in the article relate to the following subjective beliefs:

(1) $\hat{P}(\theta|S_t, n_\rho = N)$: the participant's belief after observing the retraction $n_\rho = N$ in period $t+1$;
(2) $\hat{P}(\theta|S_t \setminus s_\rho)$: the participant's belief had the retracted observation s_ρ never been observed; and
(3) $\hat{P}(\theta|S_t \cup s_{t+1})$: the participant's belief following a new observation s_{t+1} instead of the retraction.

Proposition 1. *Suppose retractions are verifying. For any quasi-Bayesian,*

(a) *their belief after observing the retraction $n_\rho = N$ in period $t+1$ is the same as their belief had the retracted observation s_ρ never been observed, i.e. $\hat{P}(\theta|S_t, n_\rho = N) = \hat{P}(\theta|S_t \setminus s_\rho)$;*

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

(b) *their belief after observing the retraction* $n_\rho = N$ *in period* $t+1$ *is the same as their belief following a new draw* s_{t+1} *instead of the retraction, i.e.* $\hat{P}(\theta|S_t, n_\rho = N) = \hat{P}(\theta|S_t \cup s_{t+1})$, *if and only if the log-likelihood of the new draw is negative of the retracted observation,* $K(s_{t+1}) = -K(s_\rho)$.

The proof of this proposition essentially follows from an application of Bayes rule and the observation that quasi-Bayesian updating rules still satisfy this identity under the transformation f_i^{-1}. An identical argument could be used to introduce additional history dependence into the updating rule; our identification strategy below would remain valid. More generally, while our framework allows decision-makers to exhibit a plethora of biases, any differences between (1) and (2) or (3) in our experimental setup will require retractions to be treated as intrinsically different.

We focus on verifying retractions to ensure equivalence to signal histories with only new draws and that updating is equivalent to simply never having observed the retracted evidence, and nothing more. In particular, the log-likelihood of retracting an observation exactly offsets the log-likelihood of the retracted observation, *i.e.* $K(n_\rho = N) = -K(s_\rho)$. This property contrasts with setups where participants consider restricted information structures, a factor (Miller and Sanjurjo, 2019) argue leads to mistakes in probabilistic reasoning, such as those in the Monty Hall Problem.[17] Here, the *selection* of a signal is independent of its and other observations' truth value, making our implementation of retractions *unrestricted*. In fact, Proposition 1 no longer generally holds if retractions are not verifying and unrestricted.

A provocative implication of this observation is that sometimes "continued influence" or related "biases" could simply reflect Bayesian updating (Pennycook *et al.*, 2021). If, for instance, only uninformative signals are selected ($\rho = t$ implies $n_t = N$), retracting an observation gives more credence to *nonretracted* evidence, which can lead to updating patterns resembling "continued influence" (Johnson and Seifert, 1994)—as well as patterns resembling backfiring (discussed in Nyhan, 2021).[18] While in many important settings, disclosure is targeted and retractions are restricted, verifying retractions allow direct comparisons and serve as a natural starting point—thus implying that a (quasi-)Bayesian agent would *not* exhibit continued influence.[19]

3.2. *Hypothesis and identification*

This article aims to study updating from retractions and, in particular, compare it to updating from direct information. Our first hypothesis concerns our two basic approaches to doing so:

Hypothesis 1 (Retractions are less effective). *Participants (1) fail to internalize retractions fully and (2) treat retractions as less informative than an otherwise equivalent piece of new information.*

[17] In the Monty Hall Problem, a participant selects one of three doors, one of which hides a prize. After making a choice, an *un*selected door that does *not* hide the prize is opened. The participant can then switch choices. Since only unselected doors *without a prize* are opened, the other unselected door is more likely to hide a prize, making switching optimal. Friedman (1998) finds participants err with striking consistency, often choosing to keep their choices.

[18] Related to this point, Guay *et al.* (2023) mentions that studies often obtain different results depending on whether they vary the extent to which participants are shown exclusively fake news versus a mix.

[19] In ongoing research, we examine a version of this experiment using targeted (*i.e.* non-verifying) retractions; the results are largely consistent, although direct comparisons between the two are unwarranted. These results are available from the authors upon request.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

We emphasize that our usage of "retractions" reflects the meaning in Definition 1, with "otherwise equivalent" reflecting the last case of Proposition 1. In our experimental setting, the log-likelihood of a *blue* draw exactly offsets that of a *yellow* draw ($K(blue) = -K(yellow)$), so a retraction of a *blue* draw is informationally equivalent to a new *yellow* draw, and vice versa.

We will refer to retractions having *diminished effectiveness* as the finding that belief updates are diminished when generated by retractions, reflecting either part of this hypothesis. Proposition 1 shows retractions should be as effective as new direct information unless participants treat these two types of information differently. Therefore, we identify the diminished effectiveness of retractions as a phenomenon distinct from belief-updating biases that are not intrinsically related to retractions.

In our context, parts (a) and (b) of Hypothesis 1 correspond to the following comparisons, which we will make repeatedly in the article, explained visually in Figure 2:

(a) *Comparing beliefs with retractions and without the retracted observation*: Are participants' beliefs after seeing a retraction the same as if the retracted observation had never been observed in the first place?

(b) *Comparing beliefs with retractions and with equivalent new observation*: Are participants' beliefs following a retraction of a *yellow* signal the same as when observing a new *blue* draw?

While (a) and (b) can both be used to assess whether retractions are less effective, and although one conclusion may be *suggestive* of the other, they are ultimately distinct. In principle, both new observations and retractions could be treated as equivalent and less informative than an earlier observation, leading to (a) without (b)—diminished updating from retractions could be driven by a feature of belief updating common to both retractions and new information. Conversely, new observations and retractions could be treated differently, but with retracted evidence treated as if it had never been seen, and with over-reaction to new observations driven by some other channel—leading to (b) without (a).

3.3. *Estimation strategy*

We start by noting that belief updates in log-odds should be $\pm K(yellow)$, no matter the signal (a draw or a retraction) and no matter the prior (moderate or extreme).[20] This is because a Bayesian would have constant log-odds updates for any prior. Therefore, since using log-odds beliefs allows us to more easily compare and interpret our results, and in keeping with standard practice in the literature on belief updating (as in Benjamin, 2019), we will specify all our regressions using log-odds of beliefs, defined as $\hat{\ell}_t := \mathcal{L}(\hat{p}_t)$. Our conclusions are, however, robust to relying either on log-odds or level beliefs, as shown below.

The key element of our estimation strategy relies on precisely defining fixed effects based on the comparisons described in Proposition 1 (and illustrated in Figure 2) to identify diminished effectiveness of retractions. For this, we will pair histories $\mathcal{H}_t$ with and without retractions. Recall that $\mathcal{H}_t$ denotes the history up to and including period t: the set of all the draws observed and any retractions, fixing the order. Except for Section 4, where we explicitly consider updating *after* retractions, we do not include histories in which there was previously a retraction or where the truth ball was revealed so as to avoid any confounding factors.

For *comparing beliefs with retractions and without the retracted evidence*, test (a), we define the *compressed history*, $C(\mathcal{H}_t)$: the history with the retracted observations removed, as if they

[20] In contrast, the change in levels is lower the farther away from 1/2 the prior belief is.

had never occurred to begin with. Taking as an example the top panel of Figure 2, the compressed history of *(yellow, blue, retraction of the blue)* is simply *(yellow)*.[21] According to Hypothesis 1a—based on Proposition 1(a)—histories sharing a common compressed history should also share common beliefs and, therefore, the same log-odds beliefs.

We then test Hypothesis 1a with the following regression,

$$\hat{\ell}_{i,t} = \beta_0 \cdot r_{i,t} + \beta_1 \cdot r_{i,t} \cdot K(s_{i,\rho_{i,t}}) + \gamma_{C(\mathcal{H}_{i,t})} + \varepsilon_{i,t}, \tag{2}$$

where i denotes the participant, $r_{i,t}$ denotes a dummy variable indicating in period t there is a retraction ($r_{i,t} = 1$) or a new observation ($r_{i,t} = 0$), $s_{\rho_{i,t}}$ denotes the colour of the retracted observation, $\gamma_{C(\mathcal{H}_{i,t})}$ are fixed effects for compressed history, and $\varepsilon_{i,t}$ is a noise term. Note that we do not include $K(s_{i,\rho_{i,t}})$ as a term in this regression, since this term is the same for every observation with the same compressed history.

The coefficient of interest is β_1. In the context of the illustrative example, our compressed-history fixed effects allow us to take differences in beliefs across histories that induce the same compressed history, *(yellow)*, such as *(yellow)* and *(yellow, blue, retraction of blue)*. As $K(blue) < 0$, retracting *blue* should increase the belief that $\theta =$*yellow*, and so β_1 captures how much less beliefs update from a retraction compared to how much they update from the retracted observation when it was first observed. Hypothesis 1a corresponds to $\beta_1 > 0$.[22]

For *comparing retractions to new evidence*, test (b), we define *sign history*, $S(\mathcal{H}_t)$, which is the history without distinguishing whether signals were new observations or retractions. For example, as illustrated in the bottom panel of Figure 2, *(yellow, blue, retraction of blue)* and *(yellow, blue, yellow)* both have the same sign history. We then run the same regression as before, equation (2), except with sign-history fixed effects, $\gamma_{S(\mathcal{H}_t)}$, instead of compressed-history fixed effects, $\gamma_{S(\mathcal{H}_t)}$. β_1 again is the coefficient of interest, measuring how much less beliefs update from retractions than from (informationally) equivalent new observations.

3.4. *Updating from new observations*

As a first step in our analysis, and in part as a test of the validity of our experimental setting, we examined participants' belief updating from (nonretracted) new observations using a standard empirical approach in this literature. Here, we simply note that our findings are consistent with existing literature—we present the results more in-depth in Section 5, where we investigate how retractions affect belief-updating patterns.

In the absence of a retraction, the design is similar to many others surveyed by Benjamin (2019). Participants appear to correctly understand the setting, with reported beliefs tracking Bayesian posteriors closely.[23] We consider Grether-style (Grether, 1980) regressions—a workhorse model of analysis in this literature—enabling a direct comparison to existing experimental results on belief updating. Specifically, we replicate common patterns in belief updating, such as base-rate neglect and confirmation bias. While participants depart from Bayesian updating, our theoretical framework implies that any additional departure due to retractions cannot

[21] Note that compressed histories do not distinguish between the retracted observation having been drawn in Period 1 or Period 2. For example, both *(yellow, blue, retraction of the blue)* and *(blue, yellow, retraction of the blue)* have the same compressed history, *(yellow)*.

[22] The scaling by $K(s_{\rho_t})$ will prove useful when discussing how much participants infer from observations in the same log-likelihood scale to enable a comparison to Bayesian updating. Note that, upon observing s_t, Bayesian updating implies that $\ell_t = \ell_{t-1} + K(s_t)$, where $\ell_t := \mathcal{L}(p_t)$.

[23] Supplementary Appendix F.1 presents beliefs and Bayesian posteriors disaggregated by history; in Supplementary Appendix E, we report the difference and the distance between beliefs and Bayesian posteriors.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

TABLE 2
Updating from retractions (Hypothesis 1)

Retraction versus	No retracted draw (1) $\hat{\ell}_t$	Equivalent new draw (2) $\hat{\ell}_t$
Retraction (r_t)	0.011	−0.019
	(0.018)	(0.025)
Retracted draw ($r_t \cdot K(s_{\rho_t})$)	0.586^{***}	0.603^{***}
	(0.067)	(0.087)
Compressed history FEs	Yes	No
Sign history FEs	No	Yes
R^2	0.26	0.27
N	39,162	39,162

Notes: Column (1) tests Hypothesis 1a by estimating equation (2). Column (2) tests Hypothesis 1b by estimating a variant of equation (2), in which compressed-history fixed effects are replaced with sign-history fixed effects. The sample includes all observations of participants in the baseline treatment, excluding periods in which the truth ball is disclosed or in which there was a retraction in an earlier period.
Clustered standard errors at the subject level in parentheses.
$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

be attributed to explanations that are not specific to the nature of the information source. We first focus on how belief updating from retractions differs from updating from new observations, deferring the detailed reporting and discussion of general departures from Bayesian updating to Section 5.

3.5. *Updating from retractions*

We now present our first central finding: empirical support of Hypothesis 1. We estimate the differences in beliefs specific to retractions using equation (2) on our baseline treatments.

We find a diminished effectiveness of retractions: participants update beliefs less from retractions than from both the retracted observation (retraction versus no retracted draw) and an equivalent new observation (retraction versus equivalent new draw). Table 2 presents our estimates for our baseline treatments. Belief updates are significantly lower for retractions than new information: by 0.586 for retractions compared to belief updates had the retracted evidence never been observed and by 0.603 compared to equivalent new draws.

In order to contextualize this number, we compare it to the estimate of how much beliefs update following a new observation. This estimate is given by the coefficient β_1 from the regression specification $\Delta\hat{\ell}_{i,t} = \beta_0 + \beta_1 \cdot K(s_{i,t}) + \gamma_{S(\mathcal{H}_{i,t})} + \varepsilon_{i,t}$, where $\Delta\hat{\ell}_{i,t} = \hat{\ell}_{i,t} - \hat{\ell}_{i,t-1}$, restricted to histories $\mathcal{H}_{i,t}$ consisting only of new draws.[24] Since the left-hand side is $\Delta\hat{\ell}_{i,t}$, Bayesian updating corresponds to $\beta_0 = 0$ an $\beta_1 = 1$. We find that a new draw moves beliefs by 1.081 times the log-likelihood of a new draw, providing a rough estimate of how much *less* beliefs update from retractions relative to new draws: 0.603/1.081, approximately 55%. Figure 3(a) provides a visualization of these estimates. Figure 3(b) provides analogous estimates using levels ($\hat{p}_t$), instead of log-odds, and exhibits consistent results. Specifically, we find that following retractions (1) beliefs update insufficiently and remain on average 3.2 percentage points away from the beliefs held absent the retracted evidence and (2) participants update

[24] Note that when we restrict to histories without retractions, compressed and sign histories are the same: $C(\mathcal{H}_t) = S(\mathcal{H}_t)$; hence, this normalization is appropriate for both comparisons.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

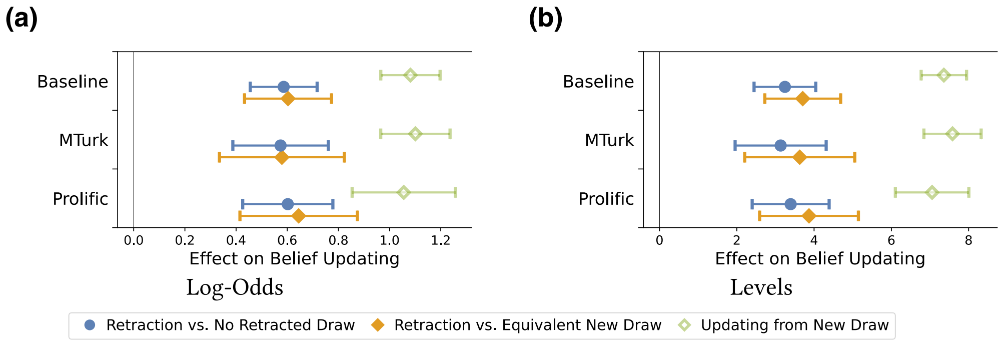


FIGURE 3

Retractions are less effective (Hypothesis 1) (a) log-odds and (b) levels

Notes: This figure depicts the effects of retractions on belief updating, showing how much less participants update beliefs from retractions than from the retracted evidence (retractions versus no retracted draw; blue solid circle) and from new direct evidence (retractions versus equivalent new draw; orange solid diamond). The green hollow diamond depicts how much beliefs update on average from new draws, for comparison. (a) These estimates for beliefs in log-odds ($\hat{\ell}_t$) as per equation (2) are shown, while panel (b) provides the analogous estimates for beliefs in levels ($\hat{p}_t$), measured in percentage points (0–100%). The sample includes all observations of participants in the baseline treatment, excluding periods in which the truth ball is disclosed or in which there was a retraction in an earlier period. The figure displays results both pooled (Baseline) and separated by recruitment platform. Plot whiskers represent 95% confidence intervals.

beliefs on average 3.7 percentage points less than from new draws—about 50% of the average belief updates from observations of 7.4 percentage points.

We conclude that participants infer substantially less from retractions than direct evidence. Furthermore, this difference does not depend on whether test (a) or test (b) is considered.

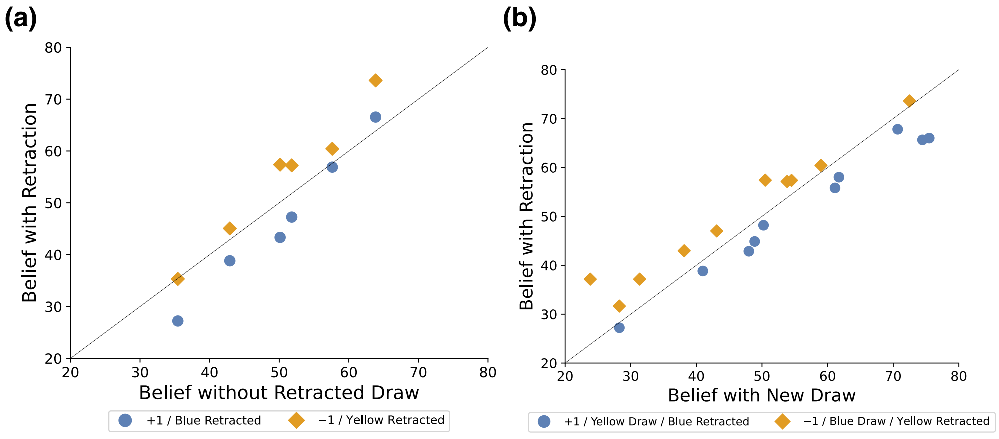
These findings represent average estimates, and a natural question is the extent to which there is heterogeneity in the effects across histories. Throughout, we will discuss different meaningful dimensions of heterogeneity, namely with respect to how recent retracted observations are and the number of draws observed (Section 4.4), as well as if the retraction is confirmatory (reinforces the prior belief) or not (Section 5). While we lack statistical power at the most disaggregated level, Figure 4 provides indicative evidence that our results are robust across histories, and we also report results fully disaggregated by history, with consistent conclusions across histories (see Supplementary Appendix F.2).

We note that we collected data for our baseline design twice, on Amazon Mechanical Turk in 2020 and again on Prolific in 2024. We obtained remarkably similar estimates of the effect across both platforms, as seen in Figure 3. In Appendix C, we show there are no significant differences between the two recruitment platforms across all our main specifications. We discuss the robustness of our results further in Section 6, only mentioning for now that restricting to particular rounds or to participants that appear to perform better does not affect our conclusions.

4. INFORMATIONAL COMPLEXITY AND DIMINISHED UPDATING FROM RETRACTIONS

Having documented differences in beliefs in updating from retractions, we now turn to a discussion of mechanisms. We divide our analysis of possible mechanisms into two parts. In this section, we propose and analyse the hypothesis that retractions are less effective because they entail greater informational complexity. We defer to the following sections the discussion of alternative explanations that could plausibly generate our results—and show that they do not.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025



Retractions vs. No Retracted Draw

Retractions vs. Equivalent New Draw

FIGURE 4

Retractions are less effective (Hypothesis 1) (a) Retractions versus No Retracted Draw and (b) Retractions versus Equivalent New Draw

Notes: This figure exhibits the effect of retractions on belief updating across the fixed effects used in our baseline specifications, reported in Table 2. Each marker in panel (a) represents average beliefs with a retraction (y-axis) and without the retracted draw (x-axis) for a specific compressed history. Analogously, each marker in panel (b) represents average beliefs with a retraction (y-axis) and with an equivalent new draw (x-axis) for a specific sign history. Blue dots correspond to cases in which a blue draw is retracted and orange diamonds to those in which the retraction refers to a yellow draw. Retractions being less effective corresponds to blue dots being below the 45-degree line and orange diamonds above. The sample includes all observations of participants in the baseline treatments, excluding periods in which the truth ball is disclosed or in which there was a retraction in an earlier period.

4.1. *Retractions provide more complex information*

While the informational content (as captured by the log-likelihood) of a retraction is the same as that of a new observation, we argue that properties inherent to retractions render it more complex and lead to the observed diminished belief updates.

One such property refers to the kind of information retractions provide. In contrast to observations that provide direct evidence about the state (*e.g.* statements, trials, data), retractions provide only indirect information. To see this, note that retractions' meaning is obtained by informing about the quality or properties of direct evidence and are hence "one step removed" from the state relative to observations. Inference from retractions, therefore, necessitates an additional layer of contingent reasoning compared to observations, which renders them more complex. Indeed, there is abundant evidence that contingent reasoning renders problems more complex and explains deviations from optimality. These include failure to incorporate pivotality considerations in voting (Esponda and Vespa, 2021), neglecting correlation in information sources (Enke and Zimmermann, 2019), or in common value auctions (Eyster *et al.*, 2019). Even in very simple environments, an added layer of contingent reasoning entails a significantly greater propensity for suboptimal choices (Martínez-Marquina *et al.*, 2019).

In our setup, this additional layer of contingent reasoning can be precisely seen using a simple causal model as given by a directed acyclical graph (Pearl, 2009). Figure 5 represents how θ, s_t, ϵ_t, and n_t are related, whereby an arrow from variable x to variable y means that x determines (in part) the value of y. We say that an observation s_t provides *direct information* about the state θ, since s_t is directly connected to θ, with θ directly influencing the distribution over the observation's realization. However, information obtained from a retraction—disclosing n_t—is only indirectly informative about θ, as θ and n_t are independent. Dependence emerges

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

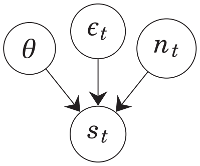


FIGURE 5

Graphical model representation of a new draw.

only through conditioning on s_t: information that an observation is or is not noise (n_t) is only informative about θ *contingent on* s_t. Pearl (2009) refers to such connections as *indirect*. Pearl and Mackenzie (2018) argue that this phenomenon—that is, that independent variables can become correlated conditional on another variable—is responsible for several apparent logical paradoxes.[25]

In our subsequent analyses, we turn to measuring complexity and identifying its prominent association with updating from retractions.

4.2. *Tracing retraction complexity*

We now turn to our empirical measures of complexity. A common microfoundation for deviations from Bayesian updating is the hypothesis that the agent faces cognitive imprecision, as posited by models of cognitive uncertainty, efficient coding, and sequential sampling.[26] Our hypothesis is that this cognitive imprecision is higher for retractions. We provide evidence for this using two broad strategies. First, we consider different empirical measures of complexity borrowed from the literature and show that these generally are larger for retractions. Second, we consider treatments of and variation in our baseline design where retraction complexity would appear to increase, showing that this correspondingly strengthens the effect.

Before presenting our evidence for such a mechanism, we briefly sketch a model in the spirit of this literature, which ties complexity to empirical measures that we can infer from the data. Suppose decision-maker i faces uncertainty about how to interpret the likelihood of evidence E and update beliefs. In particular, for tractability, we assume the decision-maker's prior about θ is Gaussian, with $K(E) \sim \mathcal{N}(0, \sigma^2)$, and that they obtain T noisy estimates $K(s_t) + \sigma_\zeta \cdot \zeta$, where $\zeta \sim \mathcal{N}(0, 1)$ denotes (Gaussian) noise. Using the Bayesian updating formulas for normal distributions, this yields posterior log-odds updates as

$$\hat{\ell}_t = \hat{\ell}_{t-1} + \beta K(E) + \beta \frac{\sigma_\zeta}{\sqrt{T}} \zeta,$$

[25] For instance, the Monty Hall problem is central among the paradoxes described by Pearl and Mackenzie (2018), connecting this observation to our discussion of Miller and Sanjurjo (2019) from Section 3.1. Other related phenomena are the observed difficulty people have in thinking through problems involving higher-order reasoning, expressed in aversion to compound lotteries (Abdellaoui *et al.*, 2015; Dean and Ortoleva, 2019) and in mistaken higher-order beliefs in strategic settings (Crawford *et al.*, 2013; Kneeland, 2015; Alaoui and Penta, 2016; Alaoui *et al.*, 2020).

[26] While distinct, the literatures are closely related. Efficient coding (Wei and Stocker, 2015) and cognitive uncertainty models have been increasingly popular in economics; *e.g.* Khaw *et al.* (2021), Frydman and Jin (2022), Enke and Graeber (2022), and Augenblick *et al.* (2023). Models of sequential sampling provide a relationship between cognitive uncertainty and time through evidence accumulation (Krajbich *et al.*, 2010; Bhui and Gershman, 2018). See Ratcliff *et al.* (2016) for a survey of sequential sampling models in psychology and neuroscience, and Fudenberg *et al.* (2018), Alós-Ferrer *et al.* (2021), and Gonçalves (2023) for recent applications in economics.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

with $\beta = (1+\sigma_\zeta^2/(\sigma^2 T))^{-1}$. Section 3 shows that β is lower for retractions. The hypothesis that retractions increase complexity is reflected in an increase of σ_ζ.

We test falsifiable predictions from this setup that could explain our results. For that, we use three behavioural markers of complexity: (1) *accuracy*, *i.e.* how close belief reports are to Bayesian posteriors; (2) *speed*, *i.e.* decision times; and (3) *variability* in belief reports.

Accuracy. Our first indicator measures the distance between belief reports and the Bayes posterior. This variable captures accuracy since, based on our incentivization, the optimal report given the provided information coincides with the Bayesian posterior, and the expected payoff is decreasing in the absolute error of beliefs, that is, the distance between the belief reported and the Bayes posterior, $|\hat{p}_t - p_t|$.

Speed. Our second indicator captures how much effort individuals exert. A standard approach in the literature associates T with decision time, the idea being that the decision-maker obtains one such signal per unit of time spent deliberating (see footnote 26). In line with the general finding that decision-makers take more time and do less well on simple tasks when these tasks become less immediately apparent, we will interpret longer decision times, together with lower accuracy, as suggestive evidence that complexity is higher in the updating problem.[27]

Variability. Our third measure is the variability in the belief reports; following Khaw *et al.* (2021) and Enke and Graeber (2021), we adopt it as an indicator of the underlying complexity. The underlying intuition is that greater cognitive imprecision generates less precise choices. In our model, given the above, an increase in σ_ζ^2 increases the variance of log-odds posterior beliefs insofar as the posterior variance about $K(E)$ is at most half of the prior variance about $K(E)$, *i.e.* σ^2—see Supplementary Appendix G.1.

4.3. *Retraction complexity*

The preceding discussion motivates the following hypothesis, which we proceed to analyse:

Hypothesis 2 (Retractions are more complex). *Inference from retractions is more difficult than processing new observations, resulting in (a) lower belief accuracy, (b) longer decision time, and (c) higher belief variance.*

To test this hypothesis, we use an identification strategy similar to the one used to test the effects of retractions on belief updating (Section 3.3). Specifically, in Table 3, we estimate versions of the following:

$$y_{i,t} = \beta_1 \cdot r_{i,t} + \gamma_{i,t} + \varepsilon_{i,t}, \tag{3}$$

where $y_{i,t}$ is a dependent variable and $\gamma_{i,t}$ are the relevant fixed effects, as in Section 3.3 and under the same sample restrictions.

Specifically, to test if belief accuracy is lower and decision times longer when participants face a retraction, the dependent variable $y_{i,t}$ corresponds to participant i's absolute error in beliefs ($|\hat{p}_{i,t} - p_{i,t}|$), and to log decision time ($\ln(T_{i,t})$), respectively. We perform both comparisons outlined in Hypothesis 1: (a) retractions compared to histories where the draw was never

[27] Early evidence for this observation can be found in, for instance, Banks *et al.* (1976), Buckley and Gillman (1974), or Ratcliff (1978); see Gonçalves (2024) for a formal treatment.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

TABLE 3

Effect of retractions on complexity indicators (Hypothesis 2)

	No retracted draw			Equivalent new draw		
	(1)	(2)	(3)	(4)	(5)	(6)
Retraction versus	$\lvert\hat{p}_t - p_t\rvert$	$\ln(\mathrm{T}_t)$	$\mathrm{Var}(\hat{\ell}_t \mid h_t)$	$\lvert\hat{p}_t - p_t\rvert$	$\ln(\mathrm{T}_t)$	$\mathrm{Var}(\hat{\ell}_t \mid h_t)$
Retraction (r_t)	2.765^{***}	0.064^{***}	1.240^{***}	1.111^{***}	0.084^{***}	0.580^{***}
	(0.266)	(0.012)	(0.172)	(0.282)	(0.014)	(0.171)
Mean decision time		8.830			8.830	
Compressed history FEs	Yes	Yes	Yes	No	No	No
Sign history FEs	No	No	No	Yes	Yes	Yes
R^2	0.07	0.01	0.03	0.08	0.01	0.03
N	39,162	39,162	5,236	39,162	39,162	5,236

Notes: This table provides estimates of the effect of retractions on three indicators of complexity, following equation (3). There are two types of comparison: (a) updating from a retraction versus without the retracted observation (Columns (1)–(3)) and (b) updating from a retraction versus an equivalent new draw (Columns (4)–(6)). Columns (1) and (4) refer to the accuracy in belief updating as given by the absolute error in beliefs, defined as the absolute difference between beliefs and Bayesian posteriors. Columns (2) and (5) refer to the speed of response, defined as log decision time. Columns (3) and (6) refer to the variability of updating, defined as participant-level history-contingent log-odds belief variance. Decision time is measured in seconds. The sample includes all observations of participants in the baseline treatment, excluding periods in which the truth ball is disclosed or in which there was a retraction in an earlier period.
Clustered standard errors at the subject level in parentheses.
$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

observed, using compressed history fixed effects ($\gamma_{i,t} = \gamma_{C(\mathcal{H}_{i,t})}$) and (b) retractions versus an equivalent new draw, relying on sign history fixed effects ($\gamma_{i,t} = \gamma_{S(\mathcal{H}_{i,t})}$).

We test if retractions increase belief variance by taking the dependent variable to be the sample variance of beliefs computed at the participant level and conditional on (i) whether a retraction was observed and (ii) either the compressed history ($\mathrm{Var}(\hat{\ell}_{i,t} \mid C(\mathcal{H}_{i,t}), r_{i,t})$) or the sign history ($\mathrm{Var}(\hat{\ell}_{i,t} \mid S(\mathcal{H}_{i,t}), r_{i,t})$). Here, due to power considerations, we treat compressed/sign histories that are the same up to permutations as the same, and therefore, estimate within-participant belief variance at a given (permuted) compressed/sign history—for notational simplicity, we maintain the same notation.

Table 3 confirms Hypothesis 2. Retractions decrease accuracy in that the absolute error in beliefs increases both compared to not having seen the retracted draw (by almost 3 percentage points—Column (1)) and compared to an equivalent new observation (by over 1 point—Column (4)). Participants also take longer in reporting beliefs—approximately 6% compared without the retracted observation (Column (2)) and 10% longer when compared to an equivalent new draw (Column (5))—a conclusion that remains valid when controlling for experience and considering only later rounds.[28] Columns (3) and (6) provide an analogous comparison for the (log-odds) belief variance estimated at the participant level, where retractions increase significantly—by over one-third in either case. In both cases, we see that belief variance increases following a retraction. Figure 6 below provides a visualization of the results in Table 3.

Our results suggest that retractions are not only treated differently but also involve greater complexity. In line with the literature on cognitive imprecision, one interpretation consistent with

[28] See Section 6.2. While our results show participants take less time in later rounds, the increase in decision time caused by retractions remains consistent in later rounds, when participants have had more experience observing retractions. Note that participants are fully informed they may see a retraction prior to any round where they do, and the interface is as similar as possible for new draws and retractions; hence, it appears unsurprising that we do not detect a difference depending on whether participants have seen more retractions in the past.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

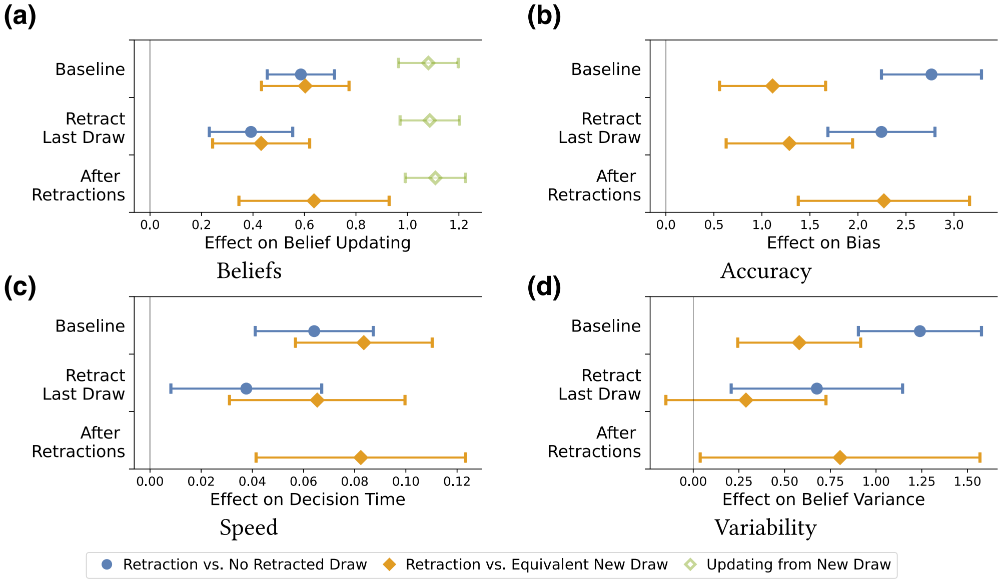


FIGURE 6

Retracting recent evidence and evidence after retractions (Hypotheses 3 and 4) (a) Beliefs. (b) Accuracy. (c) Speed and (d) variability

Notes: This figure provides estimates for the effect of retractions on belief updating and on three complexity indicators, across settings in which we expect complexity to change. "Retract Last Draw" restricts the sample of retractions to retractions in which the most recent draw is retracted, corresponding to Hypothesis 3. "After Retractions" considers updating from new draws contingent on whether or not a retraction occurred in the past, corresponding to Hypothesis 4. (a) The effect of retractions on belief updating, $\hat{\ell}_t$, under the same specifications as for Figure 3 is displayed. (b–d) Effects on our three complexity indicators—accuracy ($|\hat{p}_t - p_t|$), speed ($\ln(\mathrm{T}_t)$), and variability ($\mathrm{Var}(\hat{\ell}_t \mid h_t)$)—under the same specifications as for Table 3 are displayed. Plot whiskers represent 95% confidence intervals.

our results is that such increased complexity is reflected in a noisier perception of a retraction's informativeness relative to direct information about the state of the world.

4.4. *Validating and varying complexity*

We now show that variation in the strength of belief updating moves together with predictions that would emerge from a complexity-based mechanism. In particular, we complement our analysis by assessing whether, in situations that we would expect to be more complex, our proxies for complexity are aligned, and if beliefs are correspondingly less responsive to more complex information.

We first exploit the natural variation in our experimental design to consider cases in which retractions should be less complex. If, at time t, the observation received at $t-1$ is retracted, participants need only to revert to the belief they held at $t-2$, that is, before receiving that observation. In contrast, inferring from a retraction of previous evidence involves forming beliefs about a dataset not previously observed, thus involving counterfactual reasoning. Hence, we expect retractions of more recent observations to be easier to process than retractions of less recent observations and, consequently, more effective in moving beliefs:

Hypothesis 3 (Retracting recent observations is easier). *Retractions of recent observations are (a) more effective and (b) less complex, compared to those in the overall sample.*

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

To assess Hypothesis 3, we use the same regression specifications and contrast the estimates of the effect of retractions on belief updating (Table 2), belief accuracy, decision time, and belief variance (Table 3) in our baseline treatments to the estimates one obtains when considering only retractions of the more recent observation.

We also examine how retractions affect inference from subsequent new evidence. Our posited mechanism suggests that if a retraction is harder to process, then it may be more difficult to update following a retraction. To see why, we note that a signal history S_t will generally influence how a participant should respond to s_{t+1} via its implications on θ; the added complexity of retractions would then imply spillovers as participants would correspondingly face greater difficulty understanding what this implication should be. This idea underlies another expression of our proposed mechanism, which we articulate as a related hypothesis:

Hypothesis 4 (Updating after retractions). *Following a retraction, (a) participants update less from new observations, and (b) inference is more difficult.*

Since participants update differently from a retraction than from an equivalent new draw, a difference in beliefs $\hat{\ell}_t$ following a retraction in period $t-1$ may just be an expression of the difference in the history at $t-1$. In order to test if participants update less after a retraction, one needs to now explicitly consider how the *change* in log-odds beliefs at a particular sign history is affected by having observed a retraction in the previous period. For this reason, we use the change in log-odds beliefs, $\Delta\hat{\ell}_{i,t} := \hat{\ell}_{i,t} - \hat{\ell}_{i,t-1}$, as our dependent variable when testing this hypothesis. Thus, we estimate the following: $\Delta\hat{\ell}_{i,t} = \beta_0 + \beta_1 \cdot r_{i,t-1} + \gamma_{S(\mathcal{H}_{i,t})} + \varepsilon_{i,t}$, where $\gamma_{S(\mathcal{H}_{i,t})}$ denotes sign-history fixed effects. To test Hypothesis 4b, we consider an analogous version of equation (3): $y_{i,t} = \beta_1 \cdot r_{i,t-1} + \gamma_{S(\mathcal{H}_{i,t})} + \varepsilon_{i,t}$, where $y_{i,t}$ is a dependent variable. We exclude periods in which the truth ball was revealed for obvious reasons.

We find support for both Hypotheses 3 and 4. As shown in Figure 6(a), retractions of more recent observations are significantly more effective. Specifically, participants update about 35–40% less from retractions of recent draws than from equivalent new draws, in contrast to approximately 50–55% in our baseline. In line with greater effectiveness, we find that belief reporting is starkly faster when the retraction refers not to an earlier but to the last draw (Figure 6(c)) and also that retractions of more recent observations induce lower belief variances (Figure 6(d)). We further observe that belief accuracy is attenuated (Figure 6(b)), although not significantly different from our baseline in one case. Regarding Hypothesis 4, we find that participants update less after retractions than after equivalent new draws (a), are less accurate (b), take longer (c), and exhibit higher variability in their reports (d).

To summarize, consistent with our posited mechanism, the data suggest retractions of more recent observations are less cognitively demanding and that inference from new draws is more complex if they follow a retraction. In both cases, the intensity of belief updates aligns with our complexity indicators.

5. BELIEF UPDATING PATTERNS UNDER RETRACTIONS

So far, we have provided evidence that complexity considerations can explain the diminished effectiveness of retractions. Here, we discuss how retractions entail significantly different belief-updating patterns compared to updating from new direct evidence.

While our results imply that retractions—indirect information—are treated differently from direct information, one possibility is that retractions simply magnify known updating biases. To examine this, we rely on Grether (1980) log-odds regressions, the main workhorse in the existing literature (cf. Benjamin, 2019). Starting from the observation that, with Bayesian

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

TABLE 4
Belief updating patterns under retractions: grether regressions

	(1) $\hat{\ell}_t$	(2) $\hat{\ell}_t$
Signal (K_t)	1.102^{***}	0.907^{***}
	(0.060)	(0.060)
Prior (l_{t-1})	0.801^{***}	0.747^{***}
	(0.032)	(0.032)
Confirmatory signal ($K_t \cdot c_t$)	–	0.651^{***}
		(0.097)
Retraction (r_t) × Signal (K_t)	-0.768^{***}	-0.516^{***}
	(0.071)	(0.074)
Retraction (r_t) × Prior (l_{t-1})	0.042	0.106^{***}
	(0.037)	(0.039)
Retraction (r_t) × Confirmatory signal ($K_t \cdot c_t$)	–	-0.807^{***}
		(0.130)
R^2	0.42	0.42
N	39,162	39,162

Notes: This table shows that patterns in belief updating from retractions do not simply reflect a strengthening of known updating biases. It reports estimates of equation (5) interacting the independent variables with whether or not the signal was a retraction (r_t). The sample includes all observations of participants in the baseline treatment, excluding periods in which the truth ball is disclosed or in which there was a retraction in an earlier period.
Clustered standard errors at the subject level in parentheses.
$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

updating, the log-odds posterior probability equals the prior log-odds plus the log-likelihood ($\ell_{i,t} = \ell_{i,t-1} + K_{i,t}$), a Grether regression relaxes the weight on the prior log-odds and the log-likelihood, allowing them to be different from one, *i.e.* $\hat{\ell}_{i,t} = \beta_0 + \beta_1 \cdot \hat{\ell}_{i,t-1} + \beta_2 \cdot K_{i,t}$. Following Benjamin (2019), we estimate variants of the following:

$$\hat{\ell}_{i,t} = \beta_0 + \beta_1 \cdot \hat{\ell}_{i,t-1} + \beta_2 \cdot K_{i,t} + \beta_3 \cdot K_{i,t} \cdot c_{i,t} + \varepsilon_{i,t}, \tag{4}$$

where $\hat{\ell}_{i,t}$ denotes i's log-odds belief at period t, $K_{i,t}$ the log-likelihood of the signal—that is, $K(s_{i,t})$ in the case of a new draw $s_{i,t}$, and $-K(s_{i,\rho_{i,t}})$ for retractions—and $c_{i,t}$ an indicator variable that equals 1 whenever the signal observed confirms the prior belief ($\text{sign}(\hat{\ell}_{i,t-1}) = \text{sign}(K_{i,t})$) and 0 if otherwise. Bayesian updating implies that $\beta_1 = 1$, $\beta_2 = 1$, and $\beta_3 = 0$. Base rate neglect, for instance, corresponds to $\beta_1 < 1$; under- and overinference are expressed by $\beta_2 < 1$ and > 1, respectively; and confirmation bias, to updating relatively more from signals when these confirm one's prior belief, that is, $\beta_3 > 0$.

In examining how patterns in updating from retractions differ from updating from direct evidence, we fully interact the specification given above with the dummy variable r_t indicating whether or not the signal corresponds to a retraction or a new draw:

$$\begin{aligned}\hat{\ell}_{i,t} = \beta_0 + \beta_1 \cdot \hat{\ell}_{i,t-1} + \beta_2 \cdot K_{i,t} + \beta_3 \cdot K_{i,t} \cdot c_{i,t} + r_{i,t} \\ \times [\gamma_0 + \gamma_1 \cdot \hat{\ell}_{i,t-1} + \gamma_2 \cdot K_{i,t} + \gamma_3 \cdot K_{i,t} \cdot c_{i,t}] + \varepsilon_{i,t}.\end{aligned} \tag{5}$$

The interaction terms allow us to examine how previously documented deviations from Bayesian updating vary depending on whether or not the signal is a retraction. Table 4 presents these results.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

As foreshadowed in Section 3.4, we replicate known updating patterns. In line with results by Augenblick *et al.* (2023),[29] we find $\hat{\beta}_2 = 1.102$, indicating weak overinference from new observations, although not statistically different from 1. Once we consider whether the signal is confirmatory, we then obtain underinference from new observations, with $\hat{\beta}_2 = 0.907$ and not statistically different from 1, while $\hat{\beta}_3 = 0.651 > 0$ indicates confirmation bias, resulting in over-inference from confirmatory information ($\beta_2 + \beta_3 > 1$)—a phenomenon previously documented by, for example, Charness and Dave (2017). Together, this finding suggests that our participants slightly overreact to new observations. However, this conclusion is primarily driven by confirmation bias: participants update more from a signal when it corroborates their prior belief. We also verify another deviation from Bayesian updating identified in the literature: participants exhibit base-rate neglect. In other words, they underweight the prior, as evidenced by $\beta_1 < 1$.

A striking difference emerges: while updating from new draws exhibits slight overinference ($\beta_2 \geq 1$) driven by confirmation bias ($\beta_3 > 0$), updating from retractions leads to marked *under*inference ($0 < \beta_2 + \gamma_2 < 1$) and *anti*confirmation bias ($\beta_3 + \gamma_3 < 0$). In sum, belief updating from retractions exhibits biases opposite those that emerge when updating from new draws, a conclusion which is robust across specifications. This nuance strengthens our finding that retractions are treated differently from new signals, as the behavioural responses to retractions are not simply accentuating pre-existing biases. In fact, retractions induce opposite biases in belief-reporting behaviour.

These results suggest a specific form of heterogeneity in the diminished effect of retractions across different histories. We examine this heterogeneity using our baseline identification strategy (Section 3.3). Figure 7 shows the difference between beliefs updated from retractions and equivalent new draws for each sign history. In line with the documented expression of anti-confirmation bias in Table 4, participants update less from confirmatory retractions than from confirming new draws at extreme histories, following which they hold more extreme beliefs. Table 4 documents (1) a general diminished updating from retractions relative to new draws, (2) confirmatory bias from new draws, and (3) anticonfirmatory bias from retractions. Figure 7 illustrates this finding: it is exactly at more extreme histories, entailing more extreme beliefs, when observing a confirmatory signal induces participants to update less from retractions relative to new draws, as (2) and (3) there enhance (1). In contrast, (2) and (3) counter (1) for disconfirmatory signals, explaining why the difference between beliefs following retractions and new draws is small in this case, even if with the anticipated sign. We emphasize that this result does not speak to which beliefs are more difficult to update from[30] —rather, it speaks to the differential impact of retractions.

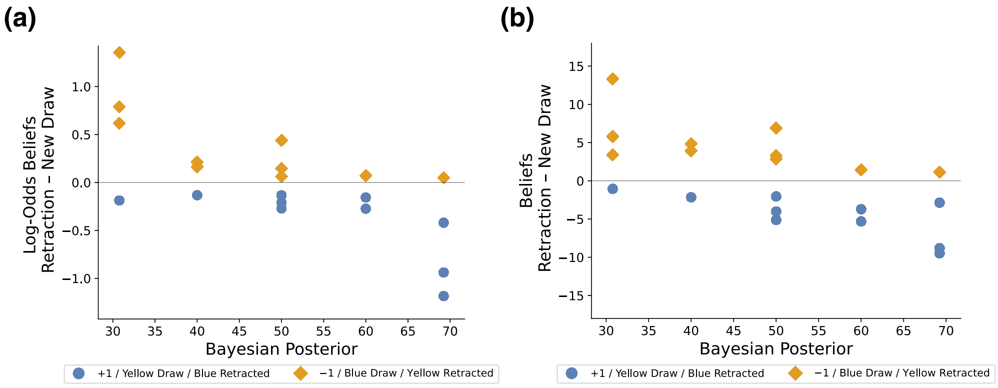
6. ROBUSTNESS OF THE FINDINGS

We performed extensive robustness checks to assess the validity of our results. In this section, we examine the extent to which our results (1) are driven by participant understanding, (2) reflect

[29] Augenblick *et al.* (2023) provide evidence that participants overinfer (resp. underinfer) from signals in similar symmetric environments whenever $P(s_t = \theta \mid \theta) \geq 1/2$ is below (resp. above) approximately 3/5, coinciding with our parameters in the experimental design.

[30] Indeed, evidence for our complexity indicators is mixed, suggesting one should not infer that the heterogeneity across histories is motivated by varying degree complexity in updating. While we do find that decision time patterns by sign history are strongly related to those in Figure 7, the difference in the absolute error in beliefs when updating from retractions and new draws, however, is greater for histories inducing more moderate posteriors, and a similar phenomenon seems to occur with belief variability—see Supplementary Appendix H.2.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025



Log-Odds

Levels

FIGURE 7

Updating from retractions: heterogeneity by history (a) log-odds and (b) levels

Notes: This figure displays the difference between beliefs following retractions versus equivalent new draws disaggregated by sign history. Blue circles represent sign histories in which the last signal was either the retraction of a blue draw or a new yellow draw. Orange diamonds represent sign histories in which the last signal was either the retraction of a yellow draw or a new blue draw. (a) Results in log-odds are presented, while (b) results in levels are presented. In both cases, the x-axis is the Bayesian posterior of the sign history. The sample includes all observations of participants in the baseline treatment, excluding periods in which the truth ball is disclosed or in which there was a retraction in an earlier period.

a general feature of behaviour or rather depend on specific individual characteristics, and (3) are affected by design choices.

6.1. *Robustness 1: participant screening and understanding*

Participant screening. We strove to ensure that our results were not driven by inattentive participants. While the behaviour of participants on Amazon Mechanical Turk and Prolific has been shown to approximate well representative population samples, it can sometimes be "noisy" relative to traditional laboratory participants (Gupta *et al.*, 2021; Snowberg and Yariv, 2021). To ensure our data were of high quality, we restricted participation to U.S. residents with high approval rates (over 95%) and held our study during business hours (Eastern Standard Time), added captchas throughout the experiment, employed an incentivization scheme involving a high baseline and reward pay (see Section 2.3), and precluded the possibility of repeating the experiment. Additionally, we included comprehension questions in the instructions, which participants had to answer correctly to proceed. These quality checks were important for us to be able to meaningfully test our hypotheses. If participants were simply answering randomly, they would be biased relative to Bayesian updating but would exhibit no difference between updating from retractions relative to direct evidence.

Participant understanding. We further examined the robustness of our results to excluding participants based on different measures of inattentiveness. The results are robust, and if anything slightly stronger, when restricting the sample to those participants who appear attentive, as defined in four different ways. First, using the comprehension questionnaire, we restrict our sample to participants who answered all questions correctly on their first try ("Comprehension Correct"). While unincentivized, the majority of the participants demonstrated clear understanding: approximately 60% and 90% answered all questions correctly on the first and second try, respectively; when answering randomly, the probability of answering all correctly on the first

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

try would be 0.2% (see Appendix D). Second, we further restrict the sample to participants who, when the state is revealed, correctly report that they know the state ("Understands Disclosure"). Third, we remove participants whose belief reports are excessively noisy, which we define as updating in the opposite direction to the signal more than 10% of the time ("Fewer Mistakes").[31] Fourth, we exclude participants who could be mistaking sampling with and without replacement ("Understands Replacement").[32]

In Figure 8, we exhibit the estimates of the coefficient of interest corresponding to our baseline tables (Tables 2 and 3); the supporting regression tables can be found in Supplementary Appendix I.1. The robustness of the results is consistent with noisy participants if anything attenuating the effect and shows that inattention is not driving our results.

Participant confidence. We examine the possibility that retractions are associated with lower confidence, which would express greater cognitive uncertainty. For this, we included a question regarding participant confidence in all treatments in experiment C: similar to Enke and Graeber (2022), following the input of a belief report of $\hat{p} \in [0, 100]$, we ask participants "Out of 100, how certain are you that the optimal estimate of the Truth Ball being yellow lies between $\hat{p} - 1$ and $\hat{p} + 1$?" Participants then report a value between 0, labelled "completely uncertain," and 100, "completely certain."

In line with Enke and Graeber (2022), higher confidence is associated with participants inferring more from new draws. However, this effect seems to be driven by greater confidence being associated with greater reliance on *confirmatory* signals.[33] Furthermore, confidence increases from approximately 60 out of 100, on average, to about 90 out of 100 when the truth ball is disclosed—a figure that is even closer to 100 for any of the sample restrictions discussed above.

While we do not find a significant difference in updating from retractions and direct evidence (see Supplementary Appendix J.1.) depending on whether participants are more or less confident, we do observe an effect of updating from retractions (relative to new draws) on participant confidence, albeit a small one: about 2 "confidence points" on average, and about 0.1 standard deviations in confidence, normalized within-participant (see Supplementary Appendix I.2). This suggests that participants are aware of, but ultimately underestimate, the greater complexity associated with updating from retractions, indicating—in the terminology of Enke *et al.* (2023b) and Enke and Shubatt (2023)—that objective complexity (as revealed by behaviour) is more severe than participants' subjective perception, as given by reported confidence or cognitive certainty. Still, this finding provides reassurances that our results are not driven by participants being uncomfortable with retractions or considering their interpretation insufficiently clear.

[31] We considered various degrees of mistake-propensity: 1, 5, 10, and 20%; our conclusions remain the same. We also note that these checks are correlated. For example, the first two samples contain a substantially smaller fraction of participants with excessively noisy reports.

[32] If sampling were without replacement, observing three draws of the same colour would reveal the colour of the truth ball. Less than 10% of all participants hold extreme beliefs (close to 1 or 0) in these cases. Removing these participants from the sample leaves results virtually unchanged.

[33] Specifically, we find that participants that are, on average, more confident than the median infer slightly more, especially from confirmatory new draws. Perhaps more interesting is that when *within*-participant confidence is higher—that is, using measures of confidence normalized for each participant—participants do infer significantly more from signals, even though, again, this effect seems to be driven by inference from confirmatory signals. However, we find no significant correlation between confidence and absolute error in beliefs—if anything, there is a weak positive correlation.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

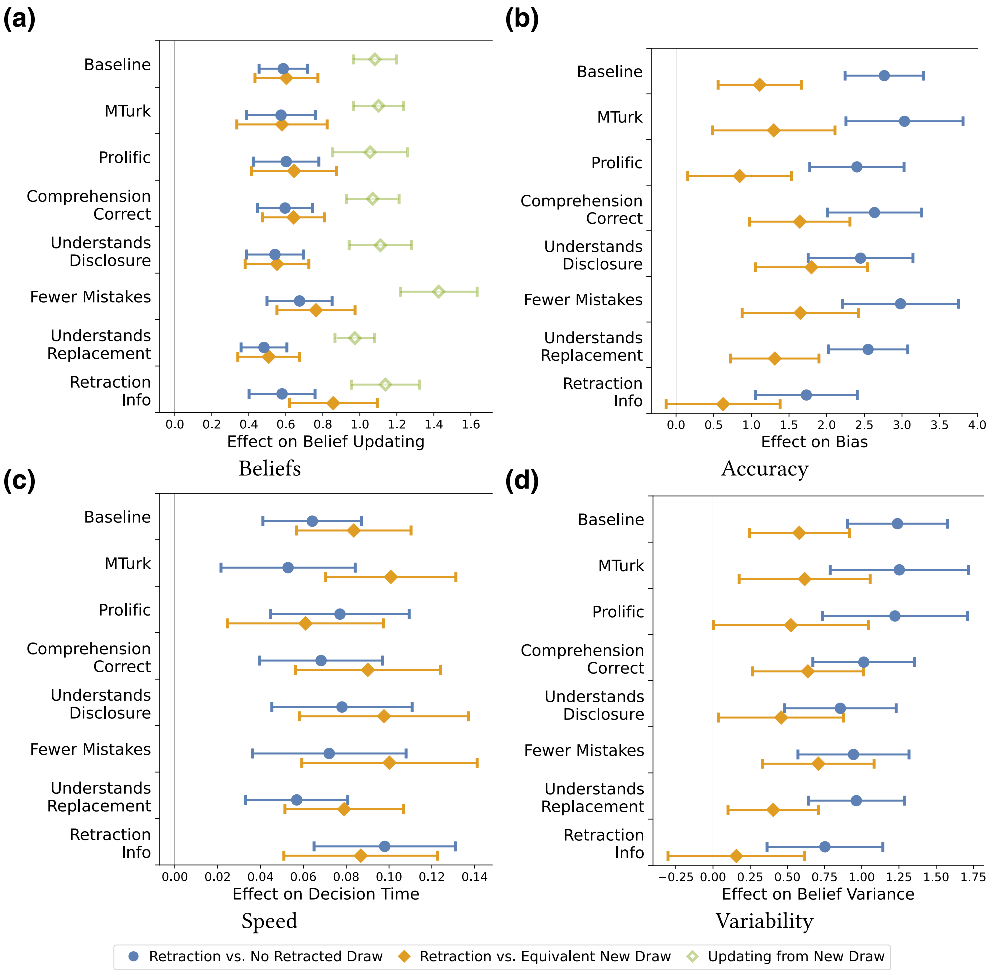


FIGURE 8

Robustness 1: participant screening and understanding (a) Beliefs. (b) Accuracy. (c) Speed and (d) variability

Notes: This figure provides estimates of the effect of retractions on belief updating and on our three complexity measures across sample restrictions and experimental variants designed to test robustness to participant understanding. "Baseline" is the pooled sample of our baseline treatments; "MTurk" and "Prolific" split the sample by those platforms. We restrict the baseline sample to participants who appear attentive in four ways: "Comprehension Correct" restricts to participants who answered all experimental comprehension questions correctly on their first try; "Understand Disclosure" restricts to participants who, when the state is revealed, correctly report that they know the state; "Fewer Mistakes" removes participants who update in the opposite direction to the signal more than 10% of the time; "Understands Replacement" excludes participants who could be mistaking sampling with and without replacement. "Retraction Info" reports results from experiment C, where participants are told retracted observations should be ignored. Panel (a) displays the effect of retractions on belief updating, $\hat{\ell}_t$, under the same specification as for Figure 3. Panels (b)–(d) display effects on our three complexity indicators—accuracy ($|\hat{p}_t - p_t|$), ($\ln(\mathrm{T}_t)$), and variability ($\mathrm{Var}(\hat{\ell}_t \mid h_t)$)—under the same specifications as for Table 3. Plot whiskers represent 95% confidence intervals.

Additional retraction information. We examine if providing additional information about retractions improves outcomes significantly. In experiment C, we included a treatment ("Retraction Info") in which, when presented with a retraction, participants are not only informed that a particular earlier draw was a noise ball but also told that "A noise ball is not informative about the colour of the Truth Ball and you should ignore that you have seen it." The treatment is identical to our baseline but for this extra information. While some outcomes seem to improve (*e.g.* accuracy increases, as does belief variance, and confidence in updating from retractions increases),

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

the differences with respect to our baseline are not statistically significant—see Supplementary Appendix I.3. This suggests participants err in interpreting the indirect evidence provided by retractions.[34]

6.2. *Robustness 2: consistency across heterogeneity*

Heterogeneous treatment effects. We explore heterogeneity in updating from retractions across multiple dimensions. We consider heterogeneity by whether participants (1) have higher quantitative ability, as proxied for by their scores on incentivized quantitative multiple-choice questions which were asked at the end of the experiment ("High Quant Ability"); (2) are more confident on average than the median participant ("High Confidence"); and (3) are on average closer to the Bayesian posterior when updating from new draws than the median participant ("More Bayesian"). We reestimate our main specifications on these groups (Figure 9) and expand our main specifications with interaction terms to account for heterogeneity (Supplementary Appendix J.1), failing to find any relevant deviations from our baseline.

We also examine whether experience with the task affects our results. For this, we perform a similar heterogeneity analysis considering the second half of the experiment (Rounds 17–32), at which point almost all participants will have encountered a retraction. Again, we find no significant difference.

Finally, attesting to the robustness of our findings, we highlight that we replicated results using our baseline treatment in two different recruitment platforms, Amazon Mechanical Turk and Prolific, two years apart (see Appendix C).

Individual heterogeneity. Underinference from retractions appears to be a robust feature within our sample, reflecting the overwhelming majority of participants' behaviour rather than a small minority. To show this, we estimate the specifications in Table 2 at the *participant* level. We report summary statistics on the participant-level estimates of the coefficient of interest in Supplementary Appendix J.2. It is difficult to fully decompose the heterogeneity in these estimates into underlying population heterogeneity versus sampling noise, given the small number of belief reports per participant.

That said, the following observations are notable: First, the estimates are strictly negative for most participants (approx. 70%). Second, the mean estimate is higher than the median; thus, while most participants infer less from retractions (with the median participant's absolute error still substantial), the distribution is skewed. Bootstrapped standard errors for both mean and median coefficients of interest show that these estimates are several standard deviations above 0, implying that these estimates are sufficiently precise to conclude that the diminished effectiveness of retractions is the rule, not the exception, among our participant pool. Finally, the individual-level estimates are single-peaked around the mean, pointing to a continuous spectrum of intensity of diminished inference from retractions rather than clearly distinguishable heterogeneous types.

[34] This pattern is also reminiscent of findings documented in experimental tests of the Monty Hall problem, where individuals often fail to recognize the error even when told the correct way to reason through it (*e.g.* Friedman, 1998). Pearl and Mackenzie (2018) discuss famous anecdotal instances of sophisticated individuals unwilling to admit errors in paradoxes involving reasoning with colliders.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

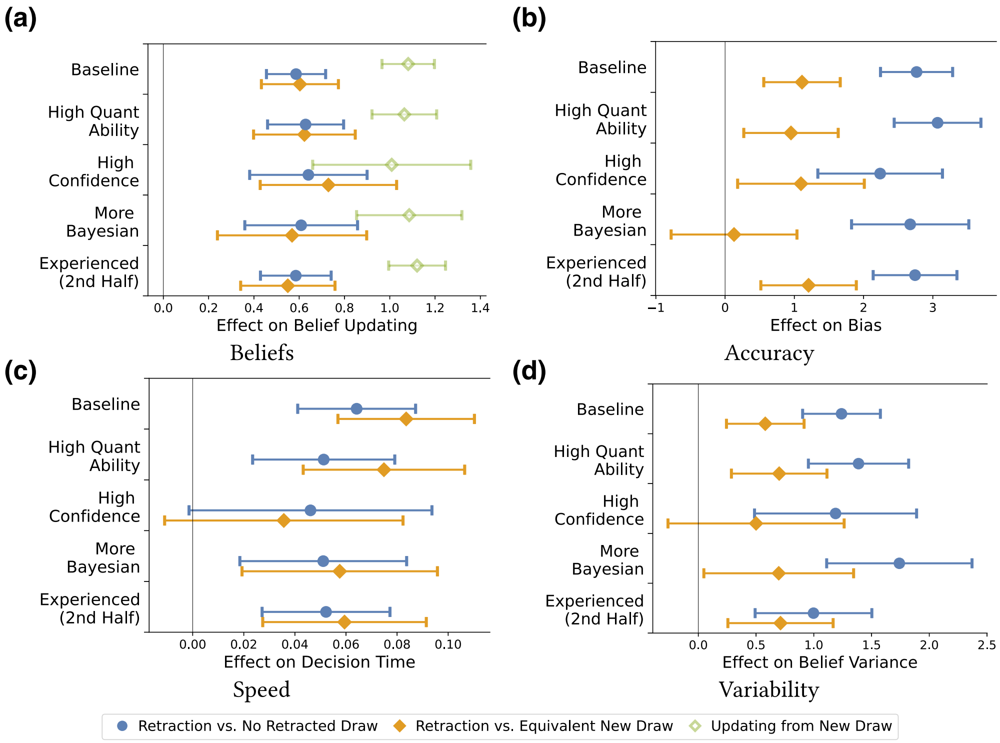


FIGURE 9

Robustness 2: consistency across heterogeneity (a) Beliefs. (b) Accuracy. (c) Speed and (d) Variability

Notes: This figure provides estimates of the effect of retractions on belief updating and on our three complexity measures across sample restrictions designed to test for heterogeneity. “Baseline” is the pooled sample of our baseline treatment. We restrict the baseline sample to test for heterogeneity in four ways: “High Quant Ability” restricts to participants with above median score on a quantitative test in the experiment; “High Confidence” restricts to participants with above median confidence in their beliefs; “More Bayesian” restricts to those who are more Bayesian than the median participant when updating from new draws; and “Experienced” restricts to the second half of rounds for each participant. (a) The effect of retractions on belief updating, $\hat{\ell}_t$, under the same specification as for Figure 3 is displayed. (b)–(d) Effects on our three complexity indicators—accuracy ($|\hat{p}_t - p_t|$), speed ($\ln(\mathrm{T}_t)$), and variability ($\mathrm{Var}(\hat{\ell}_t \mid h_t)$)—under the same specifications as for Table 3 are displayed. Plot whiskers represent 95% confidence intervals.

6.3. *Robustness 3: variations on the design*

We now discuss our experimental design. We begin by revisiting how it contributes to our identification of mechanisms and subsequently examining the robustness of our results with respect to various design features.

6.3.1. Alternative explanations ruled out by design. We first take stock of alternative explanations for retraction failure that we rule out based on the design itself.

First, our use of a balls-and-urns design was motivated by our desire to tie the limited effectiveness of retractions to belief updating itself, minimizing the role of explanations related to particular domains (*e.g.* scientific understanding or political preferences). The fact that motivated reasoning is often at play in political domains might suggest it plays a crucial role in the limited effectiveness of retractions. While it could magnify it, we find this effect even without motivated reasoning. Additionally, even if we recognize memory is bound to play an important role in many settings, our baseline design also precludes memory-based explanations for retraction’s limited effectiveness, as all information remained on the screen making the recollection of

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

past signals simple.[35] Furthermore, issues of whether retractions lead to questioning the source's reliability, while interesting in their own right, are also precluded in our setting: a Bayesian decision-maker should be able to update beliefs from retractions without any ambiguity.[36]

Second, as Proposition 1 demonstrates, only explanations specific to retractions can rationalize retraction's diminished effectiveness. Indeed, we designed the experiment to compare retractions to informationally equivalent direct evidence. The paradigm we build on allows us to quantify objectively correct beliefs, which is difficult or impossible in domains where beliefs are subjective or, perhaps more problematically, not concretely defined. We can thus distinguish retraction failures from any explanation that applies to all forms of information processing and belief updating, such as confirmation bias. Our results studying such biases further show that they are also *qualitatively* different for retractions as compared to new observations, as shown above in Section 5: biases in updating from retractions are not simply accentuated versions of known biases.

6.3.2. Variations on the design. We ran several variations of our baseline design as different treatments in our four experiments. We discuss each of them, referring to our summary Figure 10 and additional analysis in Supplementary Appendix K.

Elicit at the end. An alternative explanation for the diminished effect of retractions is that it is difficult to disregard evidence that has been actively used, as might be suggested by explanations based on cognitive dissonance. We test whether this hypothesis could drive our results by contrasting updating from retractions when beliefs have already been elicited to when they have not. In order to do so, we compare beliefs across our baseline—in which beliefs are elicited every period within a round—and the "Elicit at End" treatment in experiment A—in which beliefs are elicited only at the end of each round.[37] The difference is null: having acted upon a piece of information or not does affect how much less one updates from retractions relative to equivalent new draws. Interestingly, accuracy in updating is lower, but a heterogeneity analysis reveals it to be only marginally significant (Supplementary Appendix K.1.1). While this does not imply that retractions are as (in)effective when individuals act upon past information in other contexts, it does strengthen our conviction that our results are not due to design details.

No history of past draws. It is often the case that, in real-world settings, past evidence remains available even if invalid, and retractions (*e.g.* of academic papers by journals or of news reports by media outlets) do not simply remove incorrect information but also describe what was corrected. Nevertheless, in many cases, the full history of past evidence may not be readily available either; it will necessarily be less salient and require being recalled. It is, therefore, natural to ask how omitting the history of past draws affects our baseline results. To speak to this, in our treatment "No History," the interface was kept exactly the same as in our baseline, except that the screen only showed the ball that had just been drawn and no other draws. When presenting retractions, we showed the retracted ball with the noise label, as in the original design, without

[35] Ratcliff's (1978) seminal paper already provided evidence that recall is imperfect even when referring to very short periods of time—and the more so, the greater the elapsed time.

[36] This lack of ambiguity distinguishes our experiment from Liang (2020), Shishkin and Ortoleva (2023), and Epstein and Halevy (2020).

[37] Specifically, the "Elicit at End" treatment consisted of a sequence of events identical to the baseline treatment, except for two differences: (1) beliefs are only elicited at the end of each round, rather than each period; (2) with probability 1/3, the round ends in period two; with probability 2/3, the round ends in period three. The design ensures that, while we do not observe the *entire* belief path, we can nevertheless observe beliefs after two draws, as well as in Period 3, whether there is a third draw or a retraction.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

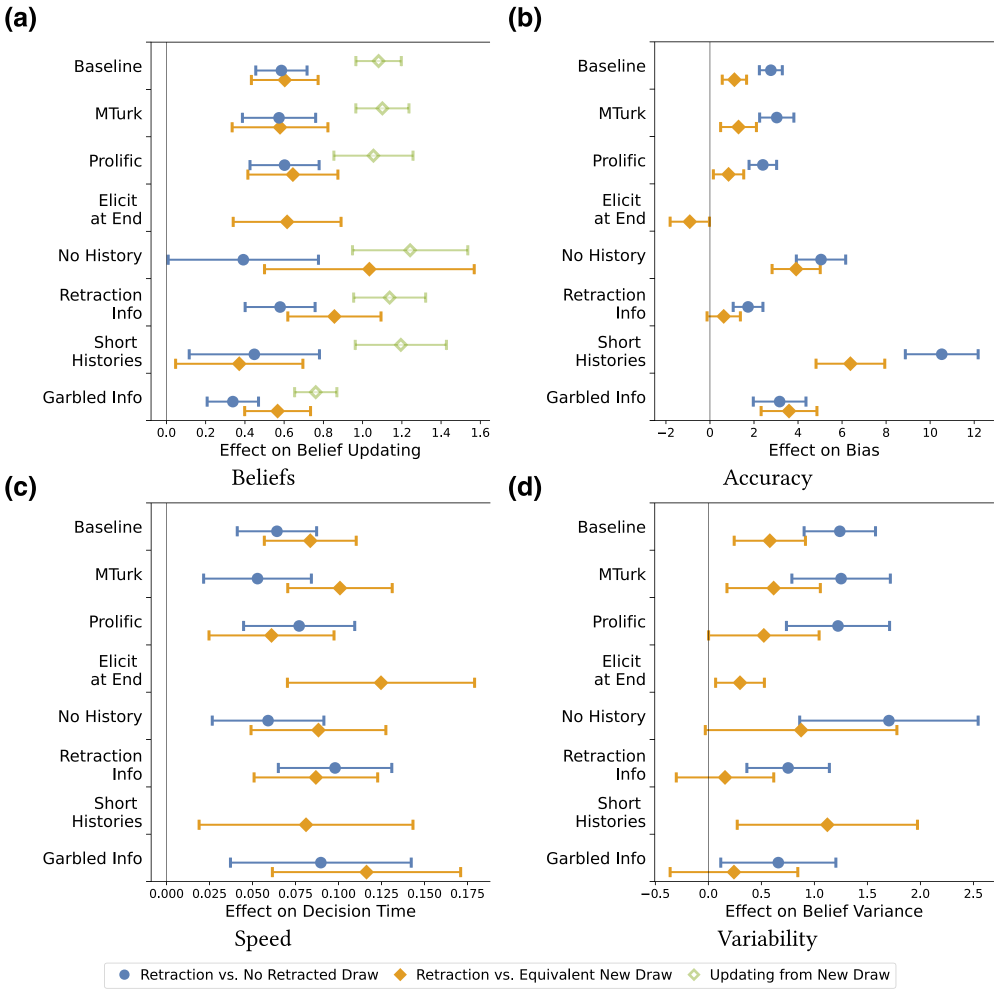


FIGURE 10

Robustness 3: variations on the design (a) Beliefs. (b) Accuracy. (c) Speed and (d) variability

Notes: This figure provides estimates of the effect of retractions on belief updating and on our three complexity measures across multiple variations on the baseline design. "Baseline" is the pooled sample of our baseline treatment; "MTurk" and "Prolific" split the sample by those platforms. In the "Elicit at End" variant, beliefs are elicited only at the end of each round. In "No History," participants were only shown the current observation, not the history of all observations in the current round. In "Short Histories," there were only two periods per round, rather than four. In "Garbled Information," truth balls were not fully informative. (a) The effect of retractions on belief updating, $\hat{\ell}_t$, under the same specification as for Figure 3 is displayed. "Retraction Info" is described in Figure 8 and included to facilitate comparison to other treatments. (b–d) Effects on our three complexity indicators—accuracy ($|\hat{p}_t - p_t|$), speed ($\ln(\mathrm{T}_t)$), and variability ($\mathrm{Var}(\hat{\ell}_t \mid h_t)$)—under the same specifications as for Table 3 are displayed. Plot whiskers represent 95% confidence intervals.

any other draws. It was unclear if this would prompt participants to misinterpret retractions as evidence for the opposite state and therefore lead to treating retractions as *more* informative than new draws and thus to updating more, not less, from retractions.

While removing the history does not result in statistically significant differences from our baseline in terms of how participants update from retractions relative to new draws (Supplementary Appendix K.1.2), the data suggest that retractions become harder to interpret and that participants update even less from retractions relative to new draws. Interestingly, removing the history of draws leads to notably higher variability in beliefs and lower accuracy,

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

resulting in less precise estimates for retraction effectiveness. Note that we would not expect to find this effect if participants only paid attention to the last draw observed—the only piece of evidence necessary to update beliefs—suggesting that theoretically redundant past evidence plays a role in belief formation.

Short histories. In order to assess whether and how much our main finding that retractions entail diminished belief updating is due to the limited understanding of a complicated setup, we made the setup as simple as possible: In a follow-up experiment, D, we presented participants with an updating task identical to our baseline, except that in this treatment—labelled "Short Histories"—histories were shorter and ran for two periods only. Specifically, participants were provided one new draw in the first period, with the second signal being either a retraction or new draw. Our findings are robust even for short histories: participants infer less from retractions, take longer, and are more biased when updating from retractions than from equivalent new draws, and the variability of beliefs is also higher. Although direct comparisons to our baseline are not well-founded, as these would partly reflect the documented heterogeneity of effects across histories (Section 5), we feel compelled to comment on the similarities and differences. The effect of retractions on diminished belief updating and decision time is similar to that in our baseline. In contrast, the "Short Histories" treatment features lower belief accuracy and greater belief variability, in line with suggestive evidence that these tend to be greater at histories leading to more moderate beliefs.

Garbled information. Our last design variation (experiment B, "Garbled Info" treatment) considered the case in which participants never perfectly learn θ. Our goal was to allow participants to form nondegenerate beliefs about θ *even* following an observation of a truth ball, thereby assessing robustness of our main results to an alternative specification of the information structure. As before, when a draw s_t is labelled as noise ($n_t = 1$), it is an independently drawn uniform ϵ_t. Unlike our baseline, however, even when labelled as a truth ball ($n_t = 0$), s_t matches θ with 80% probability and is uniform noise with complementary probability.

The specific implementation of this design was as follows: At the start of each round, a *truth box* (instead of a truth ball) is chosen at random to be either "mostly yellow" or "mostly blue," each with equal probability. A "mostly yellow" box has 9 yellow balls and 1 blue ball, and vice versa for a "mostly blue" box. Participants could observe draws from the truth box or from a *noise box* consisting of 5 yellow and 5 blue balls. For periods 1 and 2, a ball is drawn (with replacement) and shown to the participant; with probability 1/2, the ball is from the noise box, and with probability 1/2 the ball is from the truth box. In period 3, there is either a new draw or a "fact-check" (a slight variation in terminology relative to "validation" from the baseline design). In a fact-check, one of the prior draws is chosen uniformly at random, and the participant is told which box the ball is drawn from. In short, we simultaneously vary (1) the likelihood of new draws (from 3/2 to 7/3), (2) the probability of drawing a noise ball, and (3) the fact that now observing a ball from the truth box does not fully reveal the urn composition.

Despite the changes to the design, our results stand. We again here find that participants update less from retractions than from direct evidence and behave as if it is more complex as per our indicators: they take longer and exhibit lower belief accuracy and greater variability.[38]

[38] Interestingly, they also update less from new draws—something in line with existing evidence that underinference from evidence is higher the greater its likelihood (see Augenblick *et al.*, 2023).

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

7. CONCLUSION

This article identifies and quantifies diminished updating from retractions and shows updating from retractions is revealed more complex. Our analyses distinguish diminished updating from retractions and other information-processing patterns that may not have been previously recognized as relevant to retraction effectiveness. These findings provide insights into the design of interventions to address erroneous information. Specifically, we find that presenting direct evidence is more effective in correcting beliefs than retractions or corrections. Furthermore, corrections of erroneous evidence are more effective when they occur swiftly.

The minimality of our design facilitated a clear link between empirical results and their theoretical interpretation. But it certainly overlooks significant dimensions of real-world scenarios, where outcomes (*e.g.* citations) reflect factors other than probabilistic likelihood assessments, and domain-specific factors (*e.g.* memory frictions, motivated reasoning about health outcomes, etc.) may influence how individuals respond to retractions. However, the information structure in our study does approximate certain aspects of retractions in scientific articles, fact-checking, or other mechanisms of information correction. Furthermore, we interpret the consistency of our results across variations of our baseline design as evidence for the external validity of our mechanism. As such, our contribution is to propose that the additional layers of complexity in updating from retractions are generically an important factor in explaining the diminished updating from retractions. To the extent that other factors are significant, our work suggests their impacts should be separately identified from—and interacted with—the effects of retractions on information processing analysed here.

Our results point to several interesting potential directions for future work. Two strike us as particularly natural.

First, studying what makes indirect information more complex. Our experiment was designed to highlight how errors in information processing contribute to retraction failures. The richness afforded to us by variation in the design spoke to our proposed mechanism without altering the fundamental nature of the task at hand. Our findings suggest scope to further elucidate patterns in cognitive noise in indirect information. In particular, our results point toward the need for theoretical models of costly information processing to distinguish direct from indirect information. Additional research is necessary to document how belief updating depends on the degree of contingent reasoning involved. This agenda is not only of theoretical interest but also practical importance, as it aims to clarify how to correct misinformation and improve information transmission.

Second, exploring the implications of these patterns on optimal information design policies. In many settings—for example, interactions between politicians and the media, or firms and financial auditors—information receivers obtain results from strategic interplay between senders and third-party verification (*e.g.* Levkun, 2021). While our results suggest receivers may be susceptible to err following certain kinds of information, we do not speak to how endogenous changes in information may influence belief-updating patterns. Furthermore, a broader implication of our work is that the way in which information is generated can influence its perception beyond the objective informational content. While work in information design commonly reduces information to posterior beliefs, such reductions may omit important economic forces that seem worth exploring. For instance, knowing that retractions are not fully effective in correcting beliefs, to what extent could an information designer (*e.g.* a partisan media outlet or a political campaign strategist) exploit under-reaction to retractions? How would our findings shape their information policy, and how should a third party design a verification or fact-checking policy to counter it? If corrections but not validations are announced, will people correctly treat unretracted evidence as more reliable? When policies target evidence favouring

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

a particular view, are the resulting corrections or fact-checks perceived as less informative? We believe answering these and related questions has substantial practical value.

APPENDIX

Appendix A. Additional discussion of the related literature

In this appendix, we discuss existing domain-specific evidence of retraction ineffectiveness. We emphasize that this discussion focuses on the relationship between our design and those from past work, and is not meant to be a systematic survey or meta-analysis. As such, we mention broad themes that have been productively explored across a variety of research lines, but do not formally assess these papers or the state of these literatures.

Political information

Perhaps the largest number of experiments in this literature have studied the correction of information in political settings. While interpreting magnitudes is sometimes difficult in these studies, most show retractions have diminished effectiveness in political contexts.[39] For instance, in the context of the 2016 U.S. Presidential election (Swire *et al.*, 2017; Nyhan *et al.*, 2020) and the 2017 French Presidential election (Barrera *et al.*, 2020), fact-checking did improve factual knowledge, but was less effective than the original corrected information. Guriev *et al.* (2023), however, document relatively small impacts of fact-checking on perceived veracity in the context of the 2022 U.S. Midterm elections. Many studies suggest motivated reasoning as the main explanation for the ineffectiveness of retractions in political contexts.[40] Although it may indeed play a significant role, our results indicate that retractions fail even in the absence of motivated reasoning.

Fake news

Prior literature on fake news across psychology, political science and economics has studied the effectiveness of fact checking in combating misinformation; Pennycook and Rand (2021) discuss several reasons for this apparent diminished effectiveness. It is worth emphasizing that many papers in this literature vary the *nature of the fact-check itself*, with the pattern of interest being whether some presentations of fact-checks are viewed as subjectively more informative; see Ecker *et al.* (2020) for both an insightful discussion and an example.

Financial information

Other work has focused on the effectiveness of retractions in financial settings, where designs tend to involve presentations of earnings reports or related financial statements and then instructions to disregard. The focus is typically less on beliefs themselves, but rather how the information is *used* in assessments or investments. Grant *et al.* (2021), Tan and Tan (2009), and Tan and Koonce (2011) run experiments using such designs, finding that retractions have diminished effectiveness in these domains, and discuss ways this can be combated.

Jury trials

Jury trials often feature information which jurors are instructed to disregard. Experiments on this question tend to focus on whether the reason evidence should be disregarded matters. Kassin and Sommers (1997), Thompson *et al.* (1981), and Fein *et al.* (1997) conduct experiments documenting that juries do not always simply disregard information if instructed to do so. While these studies do show retracted information is not so easily disregarded, it is less clear that this reflects a departure from Bayesian rationality, since the retracted information is often meaningful.

[39] In the context of highly politically charged topics, retractions may in rare cases *backfire*, leading participants to believe more strongly in the retracted information. Nyhan and Reifler (2010) noted the occurrence of backfiring in an experiment where they provided participants with information about the presence of weapons of mass destruction in Iraq during the early 2000s, and subsequently provided them with corrections. This extreme form of retraction failure, for the most part, has not been replicated. See Nyhan (2021) for an authoritative discussion.

[40] Various studies have articulated how motivated reasoning influences belief processing in political domains; for instance, see Angelucci and Prat (2020), Thaler (2020), and Taber and Lodge (2006).

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

Academic papers

In addition to work studying society's beliefs in the association between vaccines and autism discussed in the introduction, other existing literature on retractions of scientific articles typically focuses on documenting the reasons why papers are retracted, as well as assessing the consequences for researchers. While fraud and academic misconduct are the main reasons behind retractions, error and failure to replicate constitute a significant fraction of the retraction notices (Fang *et al.*, 2012; Brainard and You, 2018)—and it is important to note that many papers that do not replicate are not retracted (Serra-Garcia and Gneezy, 2021). Among the academic community, there seems to be a significant penalty for researchers associated to retractions: a decrease in citations not only of the authors' prior work, but also of their collaborators', and, more generally, of work in related topics (Lu *et al.*, 2013; Azoulay *et al.*, 2015; Hussinger and Pellens, 2019). Of course, there are many reasons citations may be an imperfect proxy for retraction effectiveness (strategic citation motives, information about retractions not reaching the target audience, among others). Existing experimental evidence focusing on this setting suggests that retractions induce insufficient belief updating, even when the cited reason is fabricated data, and points to availability of a causal narrative as a possible reason (see, *e.g.* Greitemeyer, 2014). While these studies do show that retracted information is not so easily disregarded, relying on observational data is challenging. Indeed at least in some cases, the diminished updating from retractions may not reflect a misperception of its informational content and instead be consistent with Bayesian inference; for instance, a scientific article's retraction may involve a dispute with unclear implications, or follow-on work may find the retracted article made certain contributions which were accepted as valid (see Fang *et al.*, 2012 for examples).

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

Appendix B. Sample characteristics

TABLE 5
Sample characteristics

	MTurk	Elicit at End	Garbled Info	Prolific	No history	Retraction Info	Short Histories
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Total subjects	211	204	164	155	164	164	150
Age	37.7	39.5	38.9	38.5	37.8	37.6	35.5
Female	0.398	0.402	0.433	0.477	0.439	0.500	0.473
High school	0.900	0.887	0.927	0.865	0.860	0.884	0.880
College degree	0.673	0.613	0.683	0.587	0.506	0.561	0.560
Postgraduate	0.180	0.201	0.189	0.148	0.177	0.177	0.153
High comprehension	0.602	0.495	0.561	0.574	0.634	0.555	0.660
High quant	0.275	0.294	0.171	0.290	0.317	0.317	0.307
Experiment	A	A	B	C	C	C	D
Date	2020-06	2020-06	2021-05	2024-01	2024-01	2024-01	2024-02
Platform	MTurk	MTurk	MTurk	Prolific	Prolific	Prolific	Prolific

Notes: The table shows sample characteristics for each of our treatments in each of our experiments. “Age” is measured in years; “Female” denotes the fraction of the sample that identifies as a woman; “High school,” “College degree,” and “Postgraduate studies” denote the fraction of the sample that has completed the respective level of education. “Comprehension correct” shows the fraction of the sample that answered all comprehension questions correctly at first try; “High quant” shows to the fraction of participants who answer all the quantitative questions correctly at on their first try. Finally, “Date” denotes when the data was collected, and “Platform” the venue used to recruit participants.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

Appendix C. Comparison of recruitment platforms

TABLE 6

Treatment effects across recruitment platform

	No retracted draw				Equivalent new draw			
Retraction versus	(1) $\hat{\ell}_t$	(2) $\lvert\hat{p}_t - p_t\rvert$	(3) $\ln(\mathrm{T}_t)$	(4) $\mathrm{Var}(\hat{\ell}_t \mid h_t)$	(5) $\hat{\ell}_t$	(6) $\lvert\hat{p}_t - p_t\rvert$	(7) $\ln(\mathrm{T}_t)$	(8) $\mathrm{Var}(\hat{\ell}_t \mid h_t)$
Retraction (r_t)	0.009	2.454^{***}	0.061^{***}	1.146^{***}	−0.031	0.797^{**}	0.080^{***}	0.486^{*}
	(0.027)	(0.306)	(0.016)	(0.232)	(0.030)	(0.331)	(0.016)	(0.251)
Retracted draw ($r_t \cdot K(s_{\rho_t})$)	0.600^{***}	–	–	–	0.608^{***}	–	–	–
	(0.090)				(0.104)			
Retraction (r_t) × MTurk	0.004	0.544	0.005	0.163	0.021	0.544	0.007	0.164
	(0.035)	(0.464)	(0.021)	(0.275)	(0.034)	(0.464)	(0.020)	(0.276)
Retracted draw × MTurk	−0.024	–	–	–	−0.007	–	–	–
	(0.131)				(0.131)			
Mean decision time			8.830				8.830	
Compressed history FEs	Yes	Yes	Yes	Yes	No	No	No	No
Sign history FEs	No	No	No	No	Yes	Yes	Yes	Yes
R^2	0.27	0.08	0.18	0.03	0.27	0.08	0.18	0.03
N	39,162	39,162	39,162	5,236	39,162	39,162	39,162	5,236

Notes: This table compares average treatment effects in our baseline treatment across experiments A (MTurk) and C (Prolific). MTurk corresponds to an indicator variable that equals 1 when the observation is from our baseline treatment in Experiment A. There are two types of comparison: (a) updating from a retraction versus without the retracted observation (Columns (1)–(4)) and (b) versus an equivalent new draw (Columns (5)–(8)). Columns (1) and (5) show effects on log-odds beliefs; (2) and (6) on the accuracy of belief updating; (3) and (7) on the speed of updating; (4) and (8) on the variability of updating. The sample includes all observations of participants in the baseline treatments, excluding periods in which the truth ball is disclosed or in which there was a retraction in an earlier period.

Clustered standard errors at the subject level in parentheses.

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Appendix D. Comprehension questionnaire

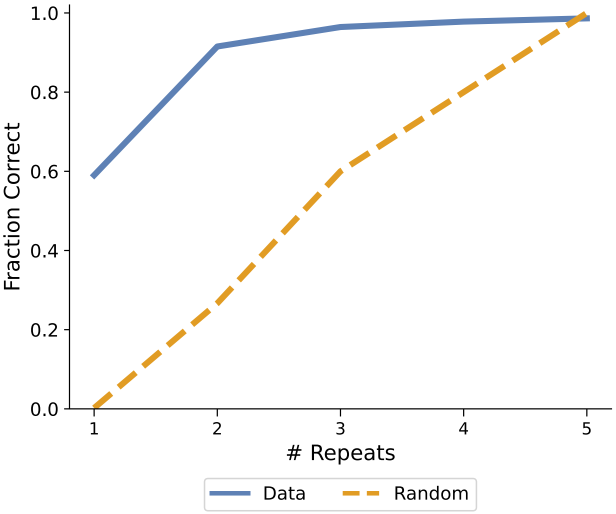


FIGURE 11

Comprehension questions

Notes: The comparison is to the case in which participants randomize uniformly over answers that were not previously tried and only in questions that were marked wrong.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

Acknowledgments. An earlier version of this article was circulated in June 2021 and then as an NBER working paper in November 2021 under the title “Learning versus Unlearning: An Experiment on Retractions.” We benefited from helpful comments from many individuals, but we especially thank Dan Benjamin, Thomas Chaney, Giorgio Coricelli, Cary Frydman, Terri Kneeland, Rani Spiegler, Michael Thaler, Séverine Toussaert, Sevgi Yuksel, and seminar audiences at Bonn, Nottingham, NYU, Caltech, Ohio State, USC, UCL, and the SWEET, and ESA conferences. Jeremy Ward, Malavika Mani, and Julen Zarate-Pina provided excellent research assistance. We are also indebted to the editor, Andrea Galeotti, and four anonymous referees whose excellent comments and suggestions contributed significantly to improving this article. Funding from IEPR is gratefully acknowledged.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

Data Availability

The data and code underlying this research is available on Zenodo at https://dx.doi.org/10.5281/zenodo.15044955.

REFERENCES

ABDELLAOUI, M., KLIBANOFF, P. and PLACIDO, L. (2015), “Experiments on Compound Risk in Relation to Simple Risk and to Ambiguity”, *Management Science*, **61**, 1306–1322.

AGRANOV, M., LOPEZ-MOCTEZUMA, G., STRACK, P., *et al.* (2022), “Learning Through Imitation: An Experiment” (Working Paper, NBER).

ALAOUI, L., JANEZIC, K. and PENTA, A. (2020), “Reasoning about Others’ Reasoning”, *Journal of Economic Theory*, **189**, 105091.

ALAOUI, L. and PENTA, A. (2016), “Endogenous Depth of Reasoning”, *Review of Economic Studies*, **83**, 1297–1333.

ALI, S. N., MIHM, M., SIGA, L., *et al.* (2021), “Adverse and Advantageous Selection in the Laboratory”, *American Economic Review*, **111**, 2152–2178.

ALÓS-FERRER, C., FEHR, E. and NETZER, N. (2021), “Time Will Tell: Recovering Preferences When Choices Are Noisy”, *Journal of Political Economy*, **129**, 1828–1877.

AMBUEHL, S. and LI, S. (2018), “Belief Updating and the Demand for Information”, *Games and Economic Behavior*, **109**, 21–39.

ANDERSON, L. R. and HOLT, C. A. (1997), “Information Cascades in the Laboratory”, *American Economic Review*, **87**, 847–862.

ANGELUCCI, C. and PRAT, A. (2020), “Measuring Voters’ Knowledge of Political News” (Working Paper).

ANGRISANI, M., GUARINO, A., JEHIEL, P., *et al.* (2021), “Information Redundancy Neglect versus Overconfidence: A Social Learning Experiment”, *American Economic Journal: Microeconomics*, **13**, 163–197.

AUGENBLICK, N., LAZARUS, E. and THALER, M. (2023), “Overinference from Weak Signals and Underinference from Strong Signals” (Working Paper).

AZOULAY, P., FURMAN, J. L., KRIEGER, J. L., *et al.* (2015), “Retractions”, *Review of Economics and Statistics*, **97**, 1118–1136.

AZRIELI, Y., CHAMBERS, C. P. and HEALY, P. J. (2018), “Incentives in Experiments: A Theoretical Analysis”, *Journal of Political Economy*, **126**, 1472–1503.

BA, C., BOHREN, J. A. and IMAS, A. (2022), “Over and Underreaction to Information” (Working Paper).

BANKS, W. P., FUJII, M. and KAYRA-STEWART, F. (1976), “Semantic Congruity Effects in Comparative Judgments of Magnitudes of Digits”, *Journal of Experimental Psychology: Human Perception and Performance*, **2**, 435–447.

BARRERA, O., GURIEV, S., HENRY, E., *et al.* (2020), “Fake News, Fact-Checking and Information in Times of Post-Truth Politics”, *Journal of Public Economics*, **182**, 104123.

BENJAMIN, D. (2019), “Errors in Probabilistic Reasoning and Judgment Biases”, in Douglas Bernheim, B., DellaVigna, S. and Laibson, D. (eds) *Handbook of Behavioral Economics* (Amsterdam: Elsevier Press) 69–186.

BHUI, R. and GERSHMAN, S. J. (2018), “Decision by Sampling Implements Efficient Coding of Psychoeconomic Functions”, *Psychological Review*, **125**, 985–1001.

BRAINARD, J. and YOU, J. (2018), “What a Massive Database of Retracted Papers Reveals About Science Publishing’s ‘Death Penalty”’ Accessed: 2022-04-14.

BROOCKMAN, D. and KALLA, J. (2016), “Durably Reducing Transphobia: A Field Experiment on Door-To-door Canvassing”, *Science*, **352**, 220–224.

BUCKLEY, P. B. and GILLMAN, C. B. (1974), “Comparison of Digits and Dot Patterns”, *Journal of Experimental Psychology*, **103**, 1131–1136.

CAPLIN, A., CSABA, D., LEAHY, J., *et al.* (2020), “Rational Inattention, Competitive Supply, and Psychometrics”, *Quarterly Journal of Economics*, **135**, 1681–1724.

CHAN, M.-P. S., JONES, C. R., JAMIESON, K. H., *et al.* (2017), “Debunking: A Meta-analysis of the Psychological Efficacy of Messages Countering Misinformation”, *Psychological Science*, **28**, 1531–1546.

CHARNESS, G. and DAVE, C. (2017), “Confirmation Bias with Motivated Beliefs”, *Games and Economic Behavior*, **104**, 1–23.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

CHARNESS, G. and LEVIN, D. (2005), "When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect", *American Economic Review*, **95**, 1300–1309.

CHARNESS, G., OPREA, R. and YUKSEL, S. (2021), "How Do People Choose Between Biased Information Sources? Evidence from a Laboratory Experiment", *Journal of the European Economic Association*, **19**, 1656–1691.

COUTTS, A. (2019), "Good News and bad News are Still News: Experimental Evidence on Belief Updating", *Experimental Economics*, **22**, 369–395.

CRAWFORD, V. P., COSTA-GOMES, M. A. and IRIBERRI, N. (2013), "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications", *Journal of Economic Literature*, **51**, 5–62.

CRIPPS, M. (2021), "Divisible Updating" (Working Paper).

DANZ, D., VESTERLUND, L. and WILSON, A. J. (2022), "Belief Elicitation and Behavioral Incentive Compatibility", *American Economic Review*, **112**, 2851–83.

DEAN, M. and ORTOLEVA, P. (2019), "The Empirical Relationship between Nonstandard Economic Behaviors", *Proceedings of the National Academy of Sciences of the United States of America*, **116**, 16262–16267.

ECKER, U. K. H., LEWANDOWSKY, S., COOK, J., *et al.* (2022), "The Psychological Drivers of Misinformation Belief and its Resistance to Correction", *Nature Reviews Psychology*, **1**, 13–29.

ECKER, U. K. H., O'REILLY, Z., REID, J. S., *et al.* (2020), "The Effectiveness of Short-Format Refutational Fact-Checks", *British Journal of Psychology*, **111**, 36–54.

ENKE, B. (2020), "What You See is All There Is", *Quarterly Journal of Economics*, **135**, 1363–1398.

ENKE, B. and GRAEBER, T. (2021), "Cognitive Uncertainty in Intertemporal Choice" (Working Paper).

—— —— (2022), "Cognitive Uncertainty", *Quarterly Journal of Economics*, **138**, 2021–2067.

ENKE, B., GRAEBER, T. and OPREA, R. (2023a), "Complexity and Time" (Working Paper).

—— —— —— (2023b), "Confidence, Self-Selection, and Bias in the Aggregate", *American Economic Review*, **113**, 1933–1966.

ENKE, B. and SHUBATT, C. (2023), "Quantifying Lottery Choice Complexity" (Working Paper).

ENKE, B. and ZIMMERMANN, F. (2019), "Correlation Neglect in Belief Formation", *Review of Economic Studies*, **86**, 313–332.

EPSTEIN, L. and HALEVY, Y. (2020), "Hard-to-Interpret Signals" (Working Paper).

ESPONDA, I., OPREA, R. and YUKSEL, S. (2022), "Contrast-Biased Evaluation" (Working Paper).

ESPONDA, I. and VESPA, E. (2014), "Hypothetical Thinking and Information Extraction in the Laboratory", *American Economic Journal: Microeconomics*, **6**, 180–202.

—— —— (2021), "Contingent Thinking and the Sure-Thing Principle: Revisiting Classic Anomalies in the Laboratory" (Working Paper).

ESPONDA, I., VESPA, E. and YUKSEL, S. (2024), "Mental Models and Learning: The Case of Base-Rate Neglect", *American Economic Review*, **114**, 752–782.

EYSTER, E., RABIN, M. and VAYANOS, D. (2019), "Financial Markets Where Traders Neglect the Informational Content of Prices", *Journal of Finance*, **74**, 371–399.

FANG, F. C., STEEN, R. G. and CASADEVALL, A. (2012), "Misconduct Accounts for the Majority of Retracted Scientific Publications", *Proceedings of the National Academy of Sciences*, **109**, 17028–17033.

FEIN, S., MCCLOSKEY, A. L. and TOMLINSON, T. M. (1997), "Can the Jury Disregard That Information? The Use of Suspicion to Reduce the Prejudicial Effects of Pretrial Publicity and Inadmissible Testimony", *Personality & Social Psychology Bulletin*, **23**, 1215–1226.

FRIEDMAN, D. (1998), "Monty Hall's Three Doors: Construction and Deconstruction of a Choice Anomaly", *American Economic Review*, **88**, 933–946.

FRYDMAN, C. and JIN, L. J. (2022), "Efficient Coding and Risky Choice", *Quarterly Journal of Economics*, **137**, 161–213.

FUDENBERG, D., STRACK, P. and STRZALECKI, T. (2018), "Speed, Accuracy, and the Optimal Timing of Choices", *American Economic Review*, **108**, 3651–3684.

GABIS, L. V., ATTIA, O. L., GOLDMAN, M., *et al.* (2022), "The Myth of Vaccination and Autism Spectrum", *European Journal of Paediatric Neurology*, **36**, 151–158.

GONÇALVES, D. (2023), "Sequential Sampling Equilibrium" (Working Paper).

—— (2024), "Speed, Accuracy, and Complexity" (Working Paper).

GRANT, S., HODGE, F. and SETO, S. (2021), "Can Prompting Investors to be in a Deliberative Mindset Reduce Their Reliance on Fake News?" (Working Paper).

GREITEMEYER, T. (2014), "Article Retracted, but the Message Lives on", *Psychonomic Bulletin & Review*, **21**, 557–561.

GRETHER, D. M. (1980), "Bayes Rule as a Descriptive Model: The Representativeness Heuristic", *The Quarterly Journal of Economics*, **95**, 537–557.

GUAY, B., BERINSKY, A. J., PENNYCOOK, G., *et al.* (2023), "How to Think about Whether Misinformation Interventions Work", *Nature Human Behavior*, **7**, 1231–1233.

GUL, F., NATENZON, P. and PESENDORFER, W. (2021), "Random Evolving Lotteries and Intrinsic Preference for Information", *Econometrica: Journal of the Econometric Society*, **89**, 2225–2259.

GUPTA, N., RIGOTTI, L. and WILSON, A. (2021), "The Experimenters' Dilemma: Inferential Preferences over Populations" (Working Paper).

GURIEV, S., HENRY, E., MARQUIS, T., *et al.* (2023), "Curtailing False News, Amplifying Truth" (Working Paper).

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

HALIM, E., RIYANTO, Y. E. and ROY, N. (2019), "Costly Information Acquisition, Social Networks, and Asset Prices: Experimental Evidence", *Journal of Finance*, **74**, 1975–2010.

HARMON-JONES, E. and MILLS, J. (2019), "An Introduction to Cognitive Dissonance Theory and an Overview of Current Perspectives on the Theory", in Harmon-Jones, E. (ed) *Cognitive Dissonance: Reexamining a Pivotal Theory in Psychology* (Washington, DC: American Psychological Association) 3–24.

HARTZMARK, S. M., HIRSHMAN, S. D. and IMAS, A. (2021), "Ownership, Learning, and Beliefs", *Quarterly Journal of Economics*, **136**, 1665–1717.

HEALY, P. J. and KAGEL, J. (2023), "Testing Elicitation Mechanisms via Team Chat" (Working Paper).

HOSSAIN, T. and OKUI, R. (2013), "The Binarized Scoring Rule", *Review of Economic Studies*, **80**, 984–1001.

HUSSINGER, K. and PELLENS, M. (2019), "Guilt by Association: How Scientific Misconduct Harms Prior Collaborators", *Research Policy*, **48**, 516–530.

JACOBY, L., KELLEY, C., BROWN, J., *et al.* (1989), "Becoming Famous Overnight: Limits on the Ability to Avoid Unconscious Influences of the Past", *Journal of Personality and Social Psychology*, **56**, 326–338.

JOHNSON, H. M. and SEIFERT, C. M. (1994), "Sources of the Continued Influence Effect: When Misinformation in Memory Affects Later Influences", *Journal of Experimental Psychology*, **20**, 1420–1436.

KASSIN, S. M. and SOMMERS, S. R. (1997), "Inadmissible Testimony, Instructions to Disregard, and the Jury: Substantive Versus Procedural Considerations", *Personality & Social Psychology Bulletin*, **23**, 1046–1054.

KHAW, M. W., LI, Z. and WOODFORD, M. (2021), "Cognitive Imprecision and Small-Stakes Risk Aversion", *Review of Economic Studies*, **88**, 1979–2013.

KNEELAND, T. (2015), "Identifying Higher-Order Rationality", *Econometrica: Journal of the Econometric Society*, **83**, 2065–2079.

KRAJBICH, I., ARMEL, C. and RANGEL, A. (2010), "Visual Fixations and the Computation and Comparison of Value in Simple Choice", *Nature Neuroscience*, **13**, 1292–1298.

KRAJBICH, I., LU, D., CAMERER, C., *et al.* (2012), "The Attentional Drift-Diffusion Model Extends to Simple Purchasing Decisions", *Frontiers in Psychology*, **3**, 235–251.

LEVKUN, A. (2021), "Communication with Strategic Fact-checking" (Working Paper).

LEWANDOWSKY, S., ECKER, U. K. H., SEIFERT, C. M., *et al.* (2012), "Misinformation and Its Correction: Continued Influence and Successful Debiasing", *Psychological Science in the Public Interest*, **13**, 106–131.

LIANG, Y. (2020), "Learning from Unknown information sources" (Working Paper).

LU, S. F., JIN, G. Z., UZZI, B., *et al.* (2013), "The Retraction Penalty: Evidence from the Web of Science", *Scientific Reports*, **3146**, 1–5.

MARTÍNEZ-MARQUINA, A., NIEDERLE, M. and VESPA, E. (2019), "Failures in Contingent Reasoning: The Role of Uncertainty", *American Economic Review*, **109**, 3437–3474.

MASATLIOGLU, Y., ORHUN, A. Y. S. and RAYMOND, C. (2021), "Intrinsic Information Preferences and Skewness" (Working Paper).

MILLER, J. and SANJURJO, A. (2019), "A Bridge from Monty Hall to the Hot Hand: The Principle of Restricted Choice", *Journal of Economic Perspectives*, **33**, 144–162.

MOBIUS, M. M., NIEDERLE, M., NIEHAUS, P., *et al.* (2022), "Managing Self-Confidence: Theory and Experimental Evidence", *Management Science*, **68**, 7793–7817.

MOTTA, M. and STECULA, D. (2021), "Quantifying the Effect of Wakefield *et al.* (1998) on Skepticism about MMR Vaccine Safety in the U.S", *PLoS One*, **16**, e0256395.

NYHAN, B. (2021), "Why the Backfire Effect Does not Explain the Durability of Political Misperceptions", *Proceedings of the National Academy of Sciences*, **118**, e1912440117.

NYHAN, B., PORTER, E., REIFLER, J., *et al.* (2020), "Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability", *Political Behavior*, **42**, 939–960.

NYHAN, B. and REIFLER, J. (2010), "When Corrections Fail: The Persistence of Political Misperceptions", *Political Behavior*, **32**, 303–330.

OPREA, R. (2020), "What Makes a Rule Complex?", *American Economic Review*, **110**, 3913–3951.

—— (2022), "Simplicity Equivalents" (Working Paper).

OPREA, R. and YUKSEL, S. (2022), "Social Exchange of Motivated Beliefs", *Journal of the European Economic Association*, **20**, 667–699.

PEARL, J. (2009), *Causality: Models, Reasoning and Inference* (New York: Cambridge University Press).

PEARL, J. and MACKENZIE, D. (2018), *The Book of Why* (New York: Basic Books).

PENNYCOOK, G., BINNENDYK, J., NEWTON, C., *et al.* (2021), "A Practical Guide to Doing Behavioral Research on Fake News and Misinformation", *Collabra: Psychology*, **7**, 25293.

PENNYCOOK, G. and RAND, D. G. (2021), "The Psychology of Fake News", *Trends in Cognitive Sciences*, **25**, 388–402.

PLUVIANO, S., WATT, C. and SALA, S. D. (2017), "Misinformation Lingers in Memory: Failure of Three Pro-Vaccination Strategies", *PLoS One*, **12**, e0181640.

PULLAN, S. and DEY, M. (2021), "Vaccine Hesitancy and Anti-Vaccination in the Time of COVID-19: A Google Trends Analysis", *Vaccine*, **39**, 1877–1881.

RABIN, M. and SCHRAG, J. (1999), "First Impressions Matter: A Model of Confirmatory Bias", *Quarterly Journal of Economics*, **144**, 37–82.

RATCLIFF, R. (1978), "A Theory of Memory Retrieval", *Psychological Review*, **85**, 59.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025

RATCLIFF, R., SMITH, P. L., BROWN, S. D., *et al.* (2016), "Diffusion Decision Model: Current Issues and History", *Trends in Cognitive Sciences*, **20**, 260–281.

Retraction Watch (2023), "Top 10 Most Highly Cited Retracted Papers" https://retractionwatch.com/the-retraction-watch-leaderboard/top-10-most-highly-cited-retracted-papers/, Accessed: Saturday 12th April, 2025.

SERRA-GARCIA, M. and GNEEZY, U. (2021), "Nonreplicable Publications are Cited More Than Replicable Ones", *Science Advances*, **7**, 7.

SHISHKIN, D. and ORTOLEVA, P. (2023), "Ambiguous Information and Dilation: An Experiment", *Journal of Economic Theory*, **208**, 105610.

SNOWBERG, E. and YARIV, L. (2021), "Testing the Waters: Behavior across Participant Pools", *American Economic Review*, **111**, 687–719.

SUSMANN, M. W. and WEGENER, D. T. (2022), "The Role of Discomfort in the Continued Influence Effect of Misinformation", *Memory & Cognition*, **50**, 435–448.

SWIRE, B., BERINSKY, A. J., LEWANDOWSKY, S., *et al.* (2017), "Processing Political Misinformation: Comprehending the Trump Phenomenon", *Royal Society of Open Science*, **4**, 160802.

TABER, C. S. and LODGE, M. (2006), "Motivated Skepticism in the Evaluation of Political Beliefs", *American Journal of Political Science*, **50**, 755–769.

TAN, H.-T. and TAN, S.-K. (2009), "Investors' Reactions to Management Disclosure Corrections: Does Presentation Format Matter?", *Contemporary Accounting Research*, **26**, 605–626.

TAN, S.-K. and KOONCE, L. (2011), "Investors' Reactions to Retractions and Corrections of Management Earnings Forecasts", *Accounting, Organizations and Society*, **36**, 382–397.

THALER, M. (2020), "The "Fake News" Effect: Experimentally Identifying Motivated Reasoning Using Trust in News" (Working Paper).

THOMPSON, W. C., FONG, G. T. and ROSENHAN, D. L. (1981), "Inadmissible Evidence and Juror Verdicts", *Journal of Personality and Social Psychology*, **40**, 453–463.

WALTER, N. and TUKACHINSKY, R. (2020), "A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It?", *Communication Research*, **47**, 155–177.

WEI, X.-X. and STOCKER, A. A. (2015), "A Bayesian Observer Model Constrained by Efficient Coding Can Explain 'Anti-Bayesian' Percepts", *Nature Neuroscience*, **18**, 1509–1517.

WEIZSÄCKER, G. (2010), "Do We Follow Others When We Should? A Simple Test of Rational Expectations", *American Economic Review*, **100**, 2340–2360.

WOODFORD, M. (2020), "Modeling Imprecision in Perception, Valuation, and Choice", *Annual Review of Economics*, **12**, 579–601.

WRIGHT, G. and AYTON, P. (1988), "Decision Time, Subjective Probability, and Task Difficulty", *Memory & Cognition*, **16**, 176–185.

Downloaded from https://academic.oup.com/restud/advance-article/doi/10.1093/restud/rdaf032/8159675 by University College London user on 04 August 2025