Learning versus Unlearning:

An Experiment on Retractions

Duarte Gonçalves[*] Jonathan Libgober[**] Jack Willis[***]

[*]University College London

[**]University of Southern California

[***]Columbia University

June 8, 2022

Abstract. Widely discredited ideas nevertheless persist. Why do we fail to "unlearn"? We study one explanation: beliefs are resistant to retractions (the revoking of earlier information). Our experimental design allows us to identify updating from retractions—*unlearning*—and to compare it with updating from equivalent new information—*learning*. We find that subjects do not fully unlearn from retractions, irrespective of the initial beliefs or the updating direction, and that their beliefs update approximately one-third less from retractions than from equivalent new information, on average. While we document a number of well-known belief updating biases in our data, our results are inconsistent with any explanation that does not treat retractions as *inherently* different. Moreover, our analysis suggests retractions are harder to process, for instance, due to the intimate reliance on conditional reasoning.

Keywords. Belief Updating, Retractions, Misinformation, Learning.

JEL Codes. D83, D91, C91.

Contact. duarte.goncalves@ucl.ac.uk, libgober@usc.edu, jack.willis@columbia.edu. For helpful comments, we thank Andreas Aristidou, Vittorio Bassi, Dan Benjamin, Aislinn Bohren, Isabelle Brocas, Gabriele Camera, Juan Carrillo, Alessandra Casella, Thomas Chaney, David Cooper, Giorgio Coricelli, Mark Dean, Laura Doval, Martin Dufwenberg, Ben Enke, Dan Friedman, Mira Frick, Guillaume Fréchette, Jonas Hjort, John Horton, Judd Kessler, Brian Libgober, Rob Metcalfe, Muriel Niederle, Kirby Nielsen, Charles Noussair, Pietro Ortoleva, Jacopo Perego, Matthew Rabin, João Ramos, Martin Rotemberg, Joel Sobel, Charlie Sprenger, Dmitry Taubsinky, Michael Thaler, Séverine Toussaert, Sevgi Yuksel, and seminar audiences at NYU, Caltech, USC, and the SWEET Conference. Jeremy Ward and Malavika Mani provided outstanding research assistance. Funding from IEPR is gratefully acknowledged.

1. INTRODUCTION

Retracted information often continues to influence beliefs, even once widely discredited. Baseless rumors, mistaken earnings announcements, false claims of politicians; all tend to linger long after being revealed to be unfounded. Perhaps the best known example is the study claiming vaccines cause autism, whose publication in *The Lancet* in 1998 launched the anti-vaxxer movement. While the study has since been widely discredited, and indeed formally retracted in 2010 (National Consumer League, 2014), widespread vaccine hesitancy continues to cause social harm.[1]

Why is it so frequently easier to learn (incorrect) information than to subsequently unlearn it? Misinformation is inevitable and inevitably influences beliefs; even in science, information thought to be true is sometimes shown to be false. But any influence on beliefs should disappear once information is known to be incorrect. Understanding why it does not, and how people do unlearn, matters not only for the debate about misinformation—its harm and how to combat it—but also for designing information campaigns aimed at correcting beliefs. Is the failure to unlearn fundamental, or is it driven by context-specific factors? To correct beliefs, should we emphasize the error of the original information, or instead emphasize the correct alternative information?

Multiple potential explanations have been proposed for failures to unlearn. It could simply reflect it being harder to move beliefs away from what was deemed likely, as occurs with *confirmation bias* (Nickerson (1998)); or a failure to undo the emotional or motivated reaction arising from the initial piece of information, *motivated reasoning* (Mobius et al. (2013) and Zimmermann (2020)); or perhaps limitations of memory are responsible, an explanation advanced in the psychology literature (e.g. Ayers and Reder, 1998). However, as far as we are aware, existing explanations either apply to *all* forms of information processing, not just retractions, or are context or domain specific, and hence do not necessarily reflect fundamental limitations in Bayesian reasoning.

In this paper, we propose and analyze one hypothesis for this asymmetry between learning and unlearning: that beliefs display greater inertia to information in the form of a retraction—an amendment of earlier information—than to information which is *directly* informative about the state. As an implication, simply "deleting" information should generally not be expected to fully correct beliefs, and instead additional new information on top of the retraction may be necessary for unlearning.

[1] For example, vaccine hesitancy has contributed to the uptick in measles outbreaks in the United States during the 2010s (DeStefano and Shimabukuro, 2019).

To test this hypothesis, we present an experimental design which allows us to quantify (un)learning from retractions, and to compare it to learning from new signals about a state. We show that information still has residual impact even once retracted—retractions are not fully effective—and also that retractions are treated as less informative than equivalent new signals.[2] Our (pre-registered) analysis implies that "information about past information" is more difficult to interpret and internalize than evidence directly informative of the state, even if the informational content is otherwise identical. We therefore show retractions are *in themselves* different, in a setting that is free from context-specific confounds. We also uncover other determinants which do or do not influence the effectiveness of retractions in the course of our analysis.

Our design is deliberately abstract, for reasons described below, and is a variation on a classic bookbag-and-poker-chips (or urns-and-balls) experiment. We present subjects with draws of colored balls (blue or yellow) from a box with replacement, with one color being more likely depending on an underlying state. In particular, the box contains a "truth ball" which is either yellow or blue—the underlying state, over which we elicit subjects' beliefs—as well as four "noise balls," two yellow and two blue. After presenting subjects with a series of such draws, in which they are told the color but not the truth/noise status of each ball, we then either present another such draw, or inform subjects whether a randomly chosen earlier ball draw was the truth ball or a noise ball. This latter event—when an earlier draw is disclosed to be a noise ball and thus uninformative of the underlying state—is what we refer to as a *retraction*. After each event, we elicit beliefs on the underlying state (i.e., the color of the truth ball), allowing us to make two comparisons of particular interest: (a) beliefs following retractions versus beliefs without observing the retracted signal in the first place—testing whether retractions work; and (b) beliefs following retractions versus beliefs following new draws which yield identical Bayes updates (in our setup, a draw of the opposite color to that which is retracted)—testing retractions versus equivalent new signals.

Our first set of results show that subjects fail to fully unlearn from retractions. First, we compare beliefs after a signal is retracted, to beliefs after a history where the retracted signal was not observed to begin with. We find that beliefs consistently display a residual effect of the retracted signal—they assign greater probability to the state being of the same color as the retracted signal. Second, we show that subjects learn less from retractions than from direct information about the state. More precisely, beliefs update less in response to a retraction than to an equivalent new signal. Both results are robust across multiple variants of the experiment and

[2] By *equivalent* we mean yielding identical Bayesian posterior about the state given the same prior.

hold regardless of details of the retraction, for example, whether the information is confirmatory or not, or whether priors are more moderate or more extreme. The magnitude of this effect also appears economically meaningful; a quantification presented below suggests that beliefs move on average one-third less when information is a retraction (see Section 5.3.1).

Our second set of results leverages our design to study how this bias in updating from retractions interacts with (and is distinct from) other biases. We find that belief updating from retractions exhibits the opposite biases when compared to updating from new signals: when updating from new signals, subjects (slightly) overinfer and do more so when signals confirm the prior—indicating confirmation bias—whereas when updating from retractions they underinfer and exhibit anti-confirmation bias. This suggests these two types of information—direct information of new signals and indirect information of retractions—are treated in a fundamentally different manner, even though new signals and retractions are informationally equivalent in our setting.

Why are retractions less effective? Our third set of results speaks more directly to the mechanisms. Readers familiar with the experimental literature on belief formation could arrive at a variety of conjectures regarding whether and why retractions are distinct from otherwise equivalent information. We determine whether the diminished effectiveness of retractions could be explained by three plausible behavioral biases.

First, we consider whether acting on information makes it harder to unlearn it. If so, the residual effect of retracted signals could be due to subjects being unwilling to disregard information they *used* when stating their beliefs, prior to such information being retracted. To test this, in one treatment arm of the experiment, we randomly select subjects to only elicit their beliefs at the end of a sequence of signals and retractions, and not after each draw. We fail to reject the hypothesis that retractions have the same effect on belief updating as in our baseline treatment, where beliefs are elicited after each draw.

Second, we test if retractions are only less effective for earlier signals but not for more recent ones. When the most recent signal is retracted, one needs only to revert back to the belief held prior to observing the retracted signal. In contrast, there is an additional layer of complexity in reassessing one's beliefs when information obtained earlier is retracted. By comparing whether or not the retraction refers to the last observed signal, we do find that retractions are more effective for information which has been more recently received; however, retractions are still insufficient to induce subjects to unlearn the retracted signals and less effective than new direct information. This suggests that, while dispelling incorrect information is most effective immediately after its release, retractions remain unable to fully correct beliefs.

Third, past work has documented that subjects update less after having observed some signals already (see Benjamin, 2019); thus, one could conjecture our results reflect sensitivity to the underlying number of draws. We test this hypothesis by comparing across similar histories of draws and permuting the timing of the retractions, and we find no effect associated to retractions happening earlier rather than later. In short, we find that the continued effect of retracted signals is not primarily driven by either an endowment effect, recency, or sample-size-driven underinference.

If the preceding discussion highlights the specificity associated with retractions, what could then explain our experimental results? We posit that the conditional reasoning inherent to retractions makes it harder to interpret their informational content, reflected in their reduced effectiveness. Reasoning is conditional because retractions provide information about the state by providing information about past information, rather than directly. Our theoretical analysis illuminates this necessity; we consider a general class of non-Bayesian updating models (including, for instance, probability weighting), and show that the failure of retractions is independent from any "quasi-Bayesian" explanation that does not treat retractions as inherently different (even though, in principle, such models can explain a plethora of biases in belief updating). Thus, while our findings require retractions being treated differently, their reduced effectiveness can be due to subjects facing *higher* cognitive uncertainty about their informational content than that of direct information. Moreover, the data further supports this conjecture: we show that subjects take 10% longer on average updating from retractions than from new signals.

Finally, we examine updating after retractions. While this is not our primary focus, we believe that it speaks to policy-relevant questions: Do retractions foster a better understanding of new evidence? Or do individuals simply discount new information arriving after past evidence is retracted? Our results—consistent across all specifications—suggest that beliefs are more sensitive to new signals after a retraction, compared to both the case in which the retracted signal was never received and to having observed an equivalent new signal. However, we also find that decision times are slightly longer updating from new information after a retraction, indicating that retractions also render interpretation of ensuing new information harder.

While abstract, we believe that our design does represent the kinds of situations described in our introductory examples. Moreover, the design allows us to speak to many practically relevant questions which we would not be able to replicate in a less abstract environment. First, and most importantly, we wish to show that the diminished effectiveness of retractions is a general phenomenon, not tied to details of any particular domain (which typically motivates interest in similar designs). For instance, as we discuss further below, the closest precedent for our experiment comes from the literature on correcting political information. The fact that motivated reasoning is often at play in political domains might suggest it plays an important role in the limited effectiveness of retractions; in contrast, however, we find this effect even without motivation. Idiosyncratic features will arise naturally in any concrete application, making it impossible to conclude that retraction failure in those settings reflects pure limitations in Bayesian reasoning (as we conclude here). Second, we leverage the fact that we can quantify objectively correct beliefs, which is difficult or impossible to do directly in domains where beliefs are subjective or, perhaps more problematically, not concretely defined. Third, we can compare retractions to other pieces of equivalent information, and thus distinguish retraction failures from confirmation bias (which is also more difficult when information is subjective). Fourth, our design allows us to replicate and compare our findings with the existing literature on biases in belief updating, showing the failure of retractions to be a distinct phenomenon. Fifth, we are able to incentivize responses, which typically improves accuracy and reliability. Although some of these could be addressed in other creative designs, our experimental design succeeds in addressing all these issues while retaining the simplicity of the classic bookbag-and-poker-chips setup.

We believe that the observation that retractions are *fundamentally less effective* has significant practical value. Taken together, our results provide important guidelines regarding how individuals can be expected to update beliefs with information about information, and we hope these patterns will be helpful for those who are regularly involved in communicating information to the public. In particular, our results show that this finding is general, not tied to any particular domain. This last point has substantial practical relevance. A policymaker deciding whether to provide guidance that may need to be corrected later should understand that this may not be so easy, even if "this time seems different." This paper documents that it is in general unreasonable to expect a retraction to simply involve a "deletion" of a piece of information. Our suspicion is that in many real-world cases, appreciating the inability to correct retractions ex-post would have changed the calculus regarding decisions to disseminate information. By showing that it is harder to update from retractions relative to other kinds of information, our hope is that communicators will be better able to limit the channels through which incorrect beliefs propagate.

1.1. Literature Review of Related Experimental Evidence

Our work fits squarely within two literatures; one studying the impact of retractions, and one studying belief updating.

Retractions

In psychology, the idea that information may have a residual impact even after retracted is known as the *continued influence effect* (see e.g. Ecker et al., 2022). Johnson and Seifert (1994) articulated this bias in an important early study; in their experiment, subjects relied upon discredited information related to the cause of a fire, with the authors attributing this to the causal nature of the information provided. Lewandowsky et al. (2012) surveys this literature and highlights several possible reasons; we briefly mention that none appear capable of explaining our results.[3]

To the best of our knowledge, all past experiments on retractions involve information that is (at least partially) subjective. This leaves open that subjects are in fact interpreting them correctly within their subjective worldview. However, our theoretical framework highlights a more serious issue: unless one carefully implements retractions in particular way, it may be that actually subjects *should* rationally underreact to them. Finally, since equivalent new information is not presented in these experiments, they do not separately identify retraction failures from other well-known biases, such as confirmation bias; we discuss these biases at length below.

Our contribution is to demonstrate and quantify retraction ineffectiveness as distinct from biases in updating from direct information. By focusing on a "context-free" setting, our results suggest that the failure of retractions is a general phenomenon and not due to idiosyncratic features of each of the settings in which it had previously been documented. We briefly review some particular contexts were retractions and the continued influence effect have been studied.

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

[3] Two of the four explanations highlighted involve memory; our design explicitly shuts down the memory channel by reminding subjects of all information they have seen. One explanation relates to difficulties in dislodging mental models in settings with complex causal chains; the suggestion is that subjects cannot disregard information when a narrative is built around it. Our setting appears too stripped down for complex narratives to have significant role. The last explanation involves a distaste for acknowledging mistakes. While this factor does not *appear* relevant to our design, if anything we provide evidence against it as responsible for our results, since we find that it does *not* matter how often subjects are asked to state their beliefs.

in these studies, most show retractions have diminished effectiveness in political contexts.[4] For instance, in the context of the 2016 US Presidential election (Swire et al., 2017; Nyhan et al., 2019) 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. Many studies suggest motivated reasoning as the main explanation for the ineffectiveness of retractions in political contexts.[5] Although it may indeed play a significant role, our results indicate that retractions fail even in the absence of motivated reasoning.

Financial Information. Other work has focused on the effectiveness of retractions in financial settings, where designs tend to be 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.[6]

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.

Belief Updating Biases

Our paper builds on the experimental literature on errors in belief updating. Benjamin (2019) provides a comprehensive survey; of independent interest, we replicate many of its key findings.[7]

Our goal is to identify and distinguish the failure of unlearning from retractions from other

[4] In the context of highly politically charged topics, retractions may in rare cases *backfire*, leading subjects to believe more strongly in the retracted information. Nyhan and Reifler (2010) noted the occurrence of backfiring in an experiment where they provided subjects 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

[5] 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).

[6] Interestingly, Kogan et al. (2021) document an *overreaction* to retractions using a different measure, showing that revelation of fraud in an SEC investigation led individuals to discount all news, including legitimate sources.

[7] For recent papers studying these biases, see, for instance, Ambuehl and Li (2018), Coutts (2019), and Thaler (2021).

well-known biases. For instance, we document *base-rate neglect* (whereby agents underweight the prior when updating; see, e.g., Esponda et al. 2020), as well as *confirmation bias*, discussed above (see also Rabin and Schrag, 1999).[8] Several models have been proposed to explain these biases in belief updating, namely probability weighting (see e.g. Kahneman and Tversky, 1979, 1992) and cognitive imprecision (Woodford, 2020; Enke and Graeber, 2020; Thaler, 2021). As we show in our theoretical framework, 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 reasoning about "information about information" is harder to process than "direct information." Though our focus on *un*learning is new, there is precedent for the idea that contingent reasoning entails higher cognitive effort. One of the first documented difficulties of contingent reasoning was Charness and Levin (2005), in winner's curse settings.[9] Perhaps most related to our study is Enke (2020), which documents in a pure prediction setting that many subjects consistently fail to account for the informational content from the *absence* of a signal, suggesting a failure of contingent reasoning. One microfoundation for "information about information" being harder to process than "direct information" is that subjects face *higher cognitive imprecision* in their understanding of the informativeness of a retraction than of a signal.

A final connection worth highlighting is between our framework and the *principle of restricted choice* from Miller and Sanjurjo (2019). Miller and Sanjurjo (2019) argue that many famous mistakes in probabilistic reasoning emerge from failing to account for how signals provided by an information structure are *restricted*. Perhaps most relevant to our exercise is the *Monty Hall Problem*, where a subject is asked to select one of three doors, with one hiding a prize and two hiding goats. After making a choice, one of the *un*selected doors that hides a *goat* is revealed. The subject is then offered to switch their choice. The principle of restricted choice is relevant because only unselected doors *without a prize* can be revealed; thus, the unselected door *not* revealed to hide a goat is more likely to hide a prize. Despite this, Friedman (1998) shows that subjects err with striking consistency, choosing often to keep their choices.[10] Revealing the validity of a ball in this paper may strike some readers as analogous to revealing whether a door is hiding a goat.

[8] To avoid confounding factors, our design features exogenous information; Charness et al. (2020) study how biases may influence subjects' *choice* of sources of information.

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

[10] See also Borhani and Green (2018) for a theoretical treatment. To our knowledge, follow-on work to Friedman (1998) has not altered the underlying mathematical problem, instead varying other circumstances around it such as incentives (Palacios-Huerta, 2003) or how it is presented and explained to participants (James et al., 2018).

Two points on this connection are critical: First, since only incorrect information can be retracted, restrictions like those from Miller and Sanjurjo (2019) do naturally emerge for *certain* implementations of retractions. Second, *our* implementation of retractions is *unrestricted*, thus eliminating the relevance of what Miller and Sanjurjo (2019) identify as a source of Monty Hall mistakes. As this design detail is subtle, we defer further discussion to our theoretical framework.

2. FRAMEWORK

This section presents formal definitions, and includes our main framework and hypotheses.

2.1. Learning: Generating Information and Updating Beliefs

We first describe the "truth-or-noise" information arrival processes which we use in our experiment, and explain how many belief updating biases can be explained using *quasi-Bayesian models*, which we define below. In the next section, we articulate why our findings will not be explained by any such model alone, and instead requires an explanation specific to the nature of retractions.

We consider a decisionmaker who forms beliefs over a state θ, which takes one of two values with equal probability, say $\theta \in \{-1, 1\}$. The decisionmaker observes signals $s_t \in \{-1, 1\}$ about θ that are independent conditionally on the state; we use $P(\cdot)$ to denote *objective* probabilities associated with the data generating process, and $b(\theta \mid \cdot)$ to denote the decisionmaker's *subjective* beliefs about the state.

Each signal s_t can either be *true*, in which case $s_t = \theta$, or *noise*, in which case it is given by an independent ϵ_t. We denote the former event by $\{n_t = 0\}$ and the latter by $\{n_t = 1\}$. Formally,

$$s_t = (1 - n_t) \cdot \theta + n_t \cdot \epsilon_t, \tag{1}$$

where $n_t \in \{0, 1\}$ and n_t, ϵ_t and θ are independent. For simplicity, we write $S_t = \{s_1, \ldots, s_t\}$.

For a Bayesian decisionmaker, $b(\theta \mid S_t) = P(\theta \mid S_t)$. Past work has routinely found rejections of this hypothesis. One way to test for deviations from Bayesian updating (see Benjamin, 2019) is to note that log-odds updates are constant when signals are identically distributed; that is, if $K(s_{t+1}) = \log\left(P(s_{t+1}|\theta)/P(s_{t+1}| - \theta)\right)$, then for a Bayesian decisionmaker the following equation

$$\log\left(\frac{b(\theta \mid S_{t+1})}{b(-\theta \mid S_{t+1})}\right) = \alpha \log\left(\frac{b(\theta \mid S_t)}{b(-\theta \mid S_t)}\right) + \beta K(s_{t+1}), \tag{2}$$

should hold for $\alpha = 1$ and $\beta = 1$. Base rate neglect, for instance, corresponds to the hypothesis that $\alpha < 1$; underinference corresponds to the hypothesis that $\beta < 1$.

A common alternative is to instead assume a strictly increasing probability weighting function f exists such that:

$$b(\theta \mid S_t) = f(P(\theta \mid S_t)).$$

Even if $\alpha \neq 1$ or $\beta \neq 1$, as long as f is strictly increasing, it is invertible, so $f^{-1}(b(\theta \mid \cdot)) = P(\theta \mid \cdot)$. It then follows that $b(\theta \mid \cdot)$ is given by the following identity:

$$\log\left(\frac{f^{-1}(b(\theta \mid S_{t+1}))}{f^{-1}(b(-\theta \mid S_{t+1}))}\right) = \log\left(\frac{f^{-1}(b(\theta \mid S_t))}{f^{-1}(b(-\theta \mid S_t))}\right) + K(s_{t+1}), \tag{3}$$

As long as some f exists such that $b(\theta \mid \cdot) = f(P(\theta \mid \cdot))$, one could determine f by using (3) to "trace out" f.[11] Following Cripps (2021), we call such a decisionmaker "quasi-Bayesian:"[12]

Definition 1. *We say that a decisionmaker is a "quasi-Bayesian" if there exists a strictly increasing f such that $b(\theta \mid s)$ can be derived from $b(\theta)$ by (i) computing $f^{-1}(b(\theta))$, (ii) determining $f^{-1}(b(\theta \mid s))$ using (3), and (iii) composing the result with f to obtain $b(\theta \mid s)$.*

Updating rules satisfying this requirement are commonly used in experimental work (e.g. Angrisani et al., 2019). Since at least Kahneman and Tversky (1979), various forms of f have been proposed, criticized, and debated.[13] Among possible microfoundations for such distortion is the hypothesis that the agent faces some cognitive imprecision. Enke and Graeber (2020) and Thaler (2021) provide two different microfoundations for the agent's posterior belief to be characterized by a probability weighting function such that $b(\theta \mid \cdot) = f(P(\theta \mid \cdot))$, with underinference being intrinsically associated to the agent's cognitive uncertainty.

2.2. Unlearning: Testing that Retractions are Different

We now turn to updating *from retractions*, and highlight some subtleties that emerge in pursuit of our goal of showing that retractions are treated fundamentally differently from other information. In particular, we hope to highlight that this is rather subtle, and motivate our main hypotheses.

[11] For a general signal history, the decisionmaker's posterior belief is then

$$b(\theta \mid S_{t+1}) = f\left(\frac{f^{-1}(b(\theta \mid S_t))P(s_{t+1}|\theta)}{f^{-1}(b(\theta \mid S_t))P(s_{t+1}|\theta) + f^{-1}(b(-\theta \mid S_t))P(s_{t+1}| - \theta)}\right).$$

[12] Cripps (2021) axiomatizes quasi-Bayesian updating showing that a decisionmaker's belief updating will satisfy this property as long as their updating is "divisible"—roughly, that signals are treated as exchangeable.

[13] Two particularly popular functional forms are $f(p) = \frac{p^\gamma}{(p^\gamma+(1-p)^\gamma)^{1/\gamma}}$ (Kahneman and Tversky, 1992) and $f(p) = \exp(-(-\log p)^\gamma)$ (Prelec, 1998). See McGranaghan et al. (2022) for a recent contribution and a discussion.

We begin with a formal definition:

Definition 2. *Consider the data generating process described in (1). A retraction consists in informing the decisionmaker that $n_\tau = 1$, thus implying the observed signal s_τ was noise.*

In order to update beliefs as a Bayesian following a retraction, the decisionmaker must know how the retraction is generated, that is, how τ was chosen. We consider the following:

Definition 3. *A* verifying retraction *is a retraction in which $\{\tau = t\}$ is independent from observed signals.*

In our experiment, this is implemented by:

- selecting τ uniformly at random from $\{1, ..., T\}$ and
- subsequently revealing n_τ to the decisionmaker; that is, whether this signal is noise or not.[14]

Note that a Bayesian decisionmaker should be able to follow Bayes rule and update beliefs following retractions without any ambiguity.[15] We have the following result:

Proposition 1. *Consider any quasi-Bayesian updating rule, as described in Definition 1. For verifying retractions, the following are identical:*

(1) $b(\theta|S_t, n_\tau = 1)$*, the decisionmaker's belief after observing the retraction* $n_\tau = 1$*;*

(2) $b(\theta|S_t \setminus s_\tau)$*, the decisionmaker's belief had the retracted signal* s_τ *never been observed.*

Moreover, the above are equivalent to

(3) $b(\theta|S_t \cup s_{t+1})$*, the decisionmaker's belief after observing a new signal realization* s_{t+1} *instead of the retraction*

if and only if its loglikelihood is negative of the retracted signal, $K(s_{t+1}) = -K(s_\tau)$*.*

As a result, quasi-Bayesian models alone cannot explain any differences between (1) and (2) or (3)—unless retractions are treated as intrinsically different. The proof of this proposition essentially follows from a careful application of Bayes rule and observing that quasi-Bayesian updating rules still satisfy this identity under the transformation f^{-1}.

[14] Note that this implies that when $n_\tau = 0$, the decisionmaker learns their past information was actually true and, in the current setting, this would result in degenerate Bayesian posterior beliefs.

[15] This lack of ambiguity distinguishes our experiment from Liang (2020), Shishkin and Ortoleva (2021), and Epstein and Halevy (2020).

We emphasize, however, that the assumption that the retractions are verifying is important and the result is not generally true without it: unless retractions are explicitly verifying, a Bayesian decisionmaker typically *should* update from them differently, even when signals are independent and identically distributed conditional on the state. For instance, if information about past evidence is disclosed only when the evidence is found to be uninformative of the state—as occurs in the retraction of academic papers or with fact-checkers targeting misinformation—then the retraction of a piece of evidence would give more credence to *non-retracted* evidence.[16]

On this point, note that the principle of restricted choice (see Section 1.1) clarifies why Monty Hall issues (a) could be relevant for non-verifying retractions, but (b) not for verifying retractions. With verifying retractions, any signal is targeted for "retraction" with equal probability, even if it is actually true (in which case it is verified). Thus, a retraction faces no additional restriction, in contrast to non-verifying retractions.[17] And indeed, Proposition 1 is no longer true if retractions provide information related to how other evidence was generated—if, for instance, only noise balls are targeted. While these issues are certainly relevant in a number of circumstances, we deliberately preclude this phenomenon to make updating from retractions not only as simple as possible, but especially to make it equivalent to deleting retracted evidence and nothing more. Additionally, this discussion suggests that verifying retractions should be the simplest case for subjects, and thus appears to be the natural starting point.

To summarize, updating from retractions in our setup is made as simple as possible: over a broad class of belief updating rules—including Bayesian updating and generalizations common in the literature—it is equivalent to deleting the retracted signal *and* it is equivalent to receiving an opposite new signal, it does not depend on which other signals were observed, nor does it require any information on past data. Furthermore, as we detail in Section 4, our experimental design emphasizes this simplicity by also removing other sources of complications:

1. The prior about the state and the noise are both symmetric ($P(\theta = 1) = P(\epsilon_t = 1) = 1/2$).
2. Signals are independent and identically distributed conditional on the state and the log-likelihood of their realizations is symmetric around zero ($K(s_t) = -K(-s_t)$) and therefore

[16] In ongoing research we also 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, as in this case it is not in general true that $P(\theta \mid S_t, n_\tau = 1) = P(\theta \mid S_t \setminus s_\tau) = P(\theta \mid S_t \cup -s_\tau)$. These results are available from the authors upon request.

[17] Or, for that matter, the information provided in the Monty Hall problem, since only doors with goats can be revealed.

retracting s_τ is equivalent to observing an additional signal $s_{t+1} = -s_\tau$, a necessary and sufficient condition for such equivalence as per Proposition 1.

3. The details of the data-generating process are graphically described in an intuitive manner and both these and full history of signals always visible to subjects.
4. The decisionmaker observes a small number of signals (up to four).

3. HYPOTHESES

The purpose of this paper, simply put, is to understand patterns in updating from retractions. Insofar as our view, informed by the anecdotal evidence mentioned in the introduction, is that retractions are indeed more difficult to process, this suggests our first and indeed main hypothesis:

Hypothesis 1 (Retractions are ineffective)**.** *(a) Subjects fail to fully internalize retractions, and (b) subjects treat retractions as less informative than an otherwise equivalent piece of new information.*

We emphasize that the use of the term "retractions" in this hypothesis reflects the meaning in Definition 2, with "otherwise equivalent" reflecting the last case of Proposition 1. Thus, relative to the work on retractions surveyed above, this hypothesis conjectures that retraction failures can emerge solely as a (specific) departure from Bayesian updating.

Note that while (a) and (b) both reflect retractions being less effective, and that one conclusion may be *suggestive* of the other, they are ultimately distinct. In principle, both new information and retractions could be treated as equivalent and less informative than an earlier signal, leading to (a) without (b). Conversely, new information and retractions could be treated as different, but with retractions being internalized fully and a distinct departure from Bayesian updating yielding overreactions to new information, leading to (b) without (a).

The complexity and subtleties associated with belief updating often motivate the emergence of other commonly studied biases. Insofar as retractions might be harder to intepret, this suggests our second hypothesis:

Hypothesis 2 (Retractions accentuate biases)**.** *Updating from retractions accentuates biases present in updating from signals.*

As part of testing this hypothesis, it is important to show that we do in fact find the same kinds of biases in updating from signals as those reported in existing literature, which is indeed the case.

If indeed retractions are less effective, what can be driving this phenomenon? Our next hypotheses relate to four mechanisms that can underlie the effectiveness of retractions.

One natural conjecture is that retractions may fail because it is more difficult to disregard evidence that has been acted upon. In other words, we hypothesize that it is the fact that subjects previously used the retracted information (when stating their beliefs) that causes the failure to unlearn it. In contrast, were subjects not to have used the information, one could expect subjects to correctly unlearn the retracted signals.

Hypothesis 3 (Retracting internalized signals)**.** *Retractions are ineffective only when agents have acted upon the observed signals.*

Relatedly, since our experiment involves dynamic information arrival, the effectiveness of retractions may depend on the timing of the signal that is retracted (as well as the timing of retractions themselves). When retractions refer to signals observed earlier, there is a layer of added complexity in belief updating: ‘unlearning’ signals acquired earlier entails forming beliefs about a dataset not previously observed. In contrast, retracting the most recently observed signal requires returning to the belief held prior to observing the retracted signal—$b(\theta \mid S_t, \text{Retraction of } s_t) = b(\theta \mid S_{t-1})$. Thus, by removing a significant layer of complexity in updating from retractions, a conjecture is that retractions lead to (more effective) unlearning of past information when the information is immediately retracted. This motivates the following hypothesis:

Hypothesis 4 (Timing of retracted signals)**.** *The effect of retractions on belief updating is stronger when it refers to the signals observed more recently.*

A third potential mechanism rendering retractions ineffective is that retractions may become less effective as more information is acquired. While existing experimental evidence is ambiguous on this point, some studies on bookbag-and-poker-chips experiments have found that, in contrast to Bayesian updating, beliefs become less sensitive to new signals as more signals are observed (see e.g. Benjamin, 2019). Our design allows us to test whether retractions occurring after more signals are observed have a lower impact on belief updating:

Hypothesis 5 (Timing of retractions)**.** *Experiencing a retraction later leads to a lower impact on beliefs compared to when the agent experiences it earlier, fixing the same history of signals.*

We then test for the above-mentioned explanation as to why retractions may work differently from new signals: retractions are harder to process. More specifically, we examine whether more time is spent in updating from retractions relative to new signals, considering decision time as a proxy for difficulty in processing the information. That is, we conjecture that an extra layer of complexity is introduced for retractions, as subjects must consider what a retraction implies about past information; since new draws (without retractions) are exchangeable, this step is not required when learning from new draws. Our conjecture is that the complexity inherent to this kind of conditional reasoning would imply the following:

Hypothesis 6 (Retractions are harder to process)**.** *Updating from retractions takes longer.*

Lastly, we take an exploratory approach to updating after retractions. To our knowledge, this is the first time that data of this kind is collected and analyzed, and, therefore, existing literature provides little guidance on what to expect. While our setup precludes any considerations on drawing inferences regarding the credibility of the source following a retraction, one can conjecture that, if a retraction is more difficult to process, it may be more difficult to update following a retraction. This observation, however, does not point toward any particular direction regarding how updating from signals ensuing a retraction compares to updating in absence of a retraction. Our hypothesis retains this agnostic view:

Hypothesis 7 (Signals after retractions)**.** *Subjects update from signals differently, depending on whether or not a signal has been retracted.*

Taken together, our hypotheses posit that a diminished effectiveness of retractions due to the conditional reasoning they necessitate and consider different conditions which may enhance or further curtail their efficacy. In particular, we examine whether the effectiveness of retractions depend on prior beliefs, on how recent the retracted evidence was, or on whether the decisionmaker previously acted upon the retracted information. While these elements are (to varying degrees) known to influence belief updating and can be captured by models of quasi-Bayesian updating, as Proposition 1 makes clear, this is orthogonal to the question of whether they influence the effectiveness of retractions.

4. EXPERIMENTAL DESIGN

In this section, we describe the overall experimental design—which is summarized visually in Figure 1—and then we provide details on the experimental interface and protocols. The basic data generating process matches the theoretical framework in Section 2 and subjects were provided full information regarding how observations would be drawn and how performance-based compensation would be provided.

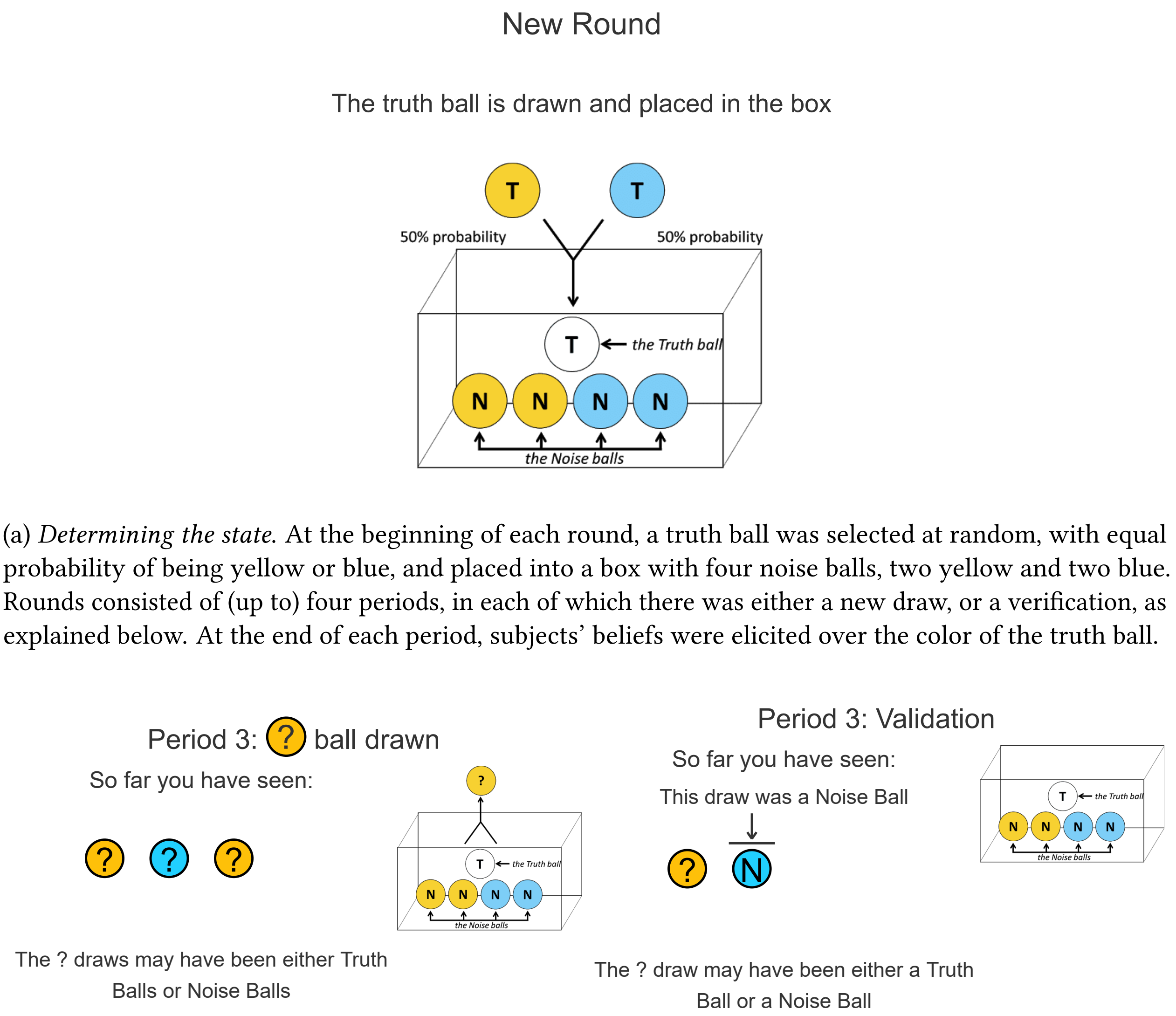
4.1. Basic Design

We first describe one round of the basic experimental design. Each *round* of the experiment has up to four *periods*, with beliefs elicited at the end of each period. Each subject plays a total of 32 rounds, and no feedback on performance is provided until the end of the experiment, when performance-based payouts are made. In each round, the sequence of events is as follows:

1. At the start of the round, a *truth ball* (referring to the state θ) is chosen at random to be either yellow or blue, with equal probability. The truth ball is then placed into the box with four *noise balls*, two yellow and two blue (corresponding to $P(n_t = 1) = 1/5$ and $P(\epsilon_t = 1) = 1/2$ in the information arrival process described in Section 2).
2. In periods one and two, subjects observe a *new signal*: a draw from the box, with replacement. They are told the color of the ball but not whether it is the truth ball or a noise ball.
3. In periods three and four, and independently across periods, subjects either observe a new draw (as above), with probability 1/2, or they observe a verification of an earlier signal from the same round, with complementary probability. Under a verification, one of the prior draws is chosen at random and it is revealed whether it was a noise ball—a *retraction*—or the truth ball. If the draw is revealed to have been the truth ball, the round ends, as at that point the state (the color of the truth ball) is fully revealed.

Additionally, at the end of each period—that is, after each new piece of information—subjects report their belief regarding the probability that the truth ball is blue vs. yellow. These reports are incentivized, as detailed below.

As explained in detail in Section 5.1, this basic design suggests several comparisons of beliefs. Comparing beliefs after a given ball draw is retracted to beliefs in histories in which the retracted draw was not made to begin with allows us to test whether retractions are effective (Hypothesis 1a). Comparing changes in beliefs in response to retractions with changes in beliefs in response to equivalent new ball draws allows us to test whether learning from retractions is different from learning from new signals (Hypothesis 1b). Indeed, a key aspect of our design is that a new draw


New Round
The truth ball is drawn and placed in the box
T
T
50% probability
50% probability
T
the Truth ball
N N N N
the Noise balls
Period 3: ? ball drawn
So far you have seen:
? ? ?
The ? draws may have been either Truth Balls or Noise Balls
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

(a) *Determining the state.* At the beginning of each round, a truth ball was selected at random, with equal probability of being yellow or blue, and placed into a box with four noise balls, two yellow and two blue. Rounds consisted of (up to) four periods, in each of which there was either a new draw, or a verification, as explained below. At the end of each period, subjects' beliefs were elicited over the color of the truth ball.

(b) *Ball draws and retractions.* In periods where there was a new draw (left), a ball was drawn from the box (with replacement), and its color was disclosed, but whether it was the truth ball or a noise ball was not. In periods where there was a verification (right), an earlier draw was chosen at random, and it was disclosed whether that ball was a noise ball (a retraction) or the truth ball. If it was the truth ball, the round ended. The history of the round was displayed throughout.

Figure 1: Summary of Experimental Visuals

of one color is informationally equivalent, for a Bayesian (and under a general model allowing for deviations from Bayesian updating), to a retraction of the opposite color—what matters for updating from the prior is the difference in the number of balls of each color. In turn, making these comparisons across different histories tests *when* retractions are ineffective (Hypotheses 2, 4, 5), and comparing updating after a retraction to after a new signal tests whether retractions affect subsequent updating (Hypothesis 7).

4.2. Single-Elicitation Treatment

Our experiment features a between-subject treatment. At the start of the experiment, each subject is randomly allocated to one of two treatments. With 1/2 probability they are allocated to the *baseline treatment*, as described above. With 1/2 probability, they are allocated to the *single-elicitation treatment*, whose purpose is to test whether requiring subjects to report their beliefs in *every* period—and hence to act on draws before they are retracted—affects the efficacy of retractions (Hypothesis 3).

In the single-elicitation treatment, the sequence of events is the same as in 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 are nevertheless able to form estimates for beliefs after two draws, as well as beliefs after three draws when the third draw is either a retraction or a new signal.

4.3. Implementation Details

Experimental Interface. A summary of the explanatory visuals shown to subjects is given in Figure 1 and the full instructions of the experiment can be found in Online Appendix C. Beliefs were reported using a slider, which displayed both the probability they assign to the truth ball being yellow, as well as the probability they assign to the truth ball being blue. Immediately after the instructions, subjects were given two rounds of unincentivized "practice" to familiarize themselves with the interface.

Subject Pool and Comprehension Checks. The experiment was run on Amazon Mechanical Turk (henceforth MTurk) on June 16-18, 2020. In order to ensure adequate statistical power, we targeted 200 subjects per treatment group. We recruited a total of 415 subjects, 211 subjects for our baseline setup and 204 for the single elicitation treatment.

We took four main steps in order to ensure that our subject pool was of high quality. First, we included Captchas throughout the experiment in order to filter out bots. Second, we included comprehension questions in the instructions which subjects needed to answer correctly in order to proceed with the experiment.[18] The questions summarized the key points the subjects needed to understand, and would have been very difficult to answer correctly without having understood the instructions. While unincentivized, the majority of the subjects answered all questions correctly on the first try (55%), and 90% answered correctly by the second try—with uniform random guesses, the probability of answering all correctly on first try would be lower than 1%. Third, as detailed below, we used a payment scheme which involved a high baseline and reward pay. Fourth, we restricted our study to be held only during business hours (Eastern Standard Time), and we restricted eligibility to US adults and precluded the possibility of repeating the experiment.

These quality checks were important for us to be able to meaningfully test our hypotheses. Excessively noisy answers would have attenuated our results: while a subject answering 50-50 to everything would not be a Bayesian, they would also demonstrate no differential updating from retractions. It was also important that subjects understood retraction should *not* be treated as evidence for the opposite state. Misinterpreting the instructions in this way would suggest retractions should be treated as *more* informative than new information, again working against us finding evidence for our hypothesis.

Consistent with our quality controls being largely effective, our results are robust to multiple sample restrictions. For example, our results do no meaningfully change if we exclude those who appear to be answering randomly or inconsistently, or those who did not answer the comprehension test questions correctly on their first (or second) attempts — see Section 5.6 for details.

Payments. We incentivized subjects to report their beliefs truthfully using a binarized scoring rule (see Hossain and Okui (2013) and Mobius et al. (2013)). By reporting $b \in [0, 100]$, a subject would receive \$12 with probability $(1 - (\mathbf{1}\{\theta = 1\} - b/100)^2)$ and \$6 with complementary probability, where θ equals 1 (-1) when the truth ball is yellow (blue). In the instructions—but not in the main interface—we provided information on the elicitation procedure, phrased as eliciting the probability the truth ball was either yellow or blue, and explained that the procedure was meant to ensure they were incentivized to answer truthfully. To determine payments, we used a report from a single randomly selected period of a randomly selected round. We also asked additional questions on mathematical ability, which were incentivized by providing a \$0.50 reward if they answered correctly a randomly chosen question.

[18] See Online Appendix C for the instructions as presented to the subjects in the experiment.

The average compensation was of $20.02/hour, with subjects spending on average 29 minutes in the experiment. For comparison, this rate is similar to the MTurk experiment of Enke and Graeber (2020), and four times the MTurk average of $5.

4.4. Preregistration

Our experiment was registered using the AEA RCT Registry under RCT ID AEARCTR-0003820. The registration lists Hypotheses 1-4 as our Primary Hypotheses and Hypothesis 7 as a secondary hypothesis. Hypotheses 5 and 6 were introduced subsequently, as feedback we received convinced us they helped interpret our results. The registration lists one other secondary hypothesis—that we can replicate findings from the belief updating literature. While we speak to this in our analysis (Section 5.2), as it did not directly pertain to retractions we did not list it above.

5. RESULTS

In this section, we begin by explaining our empirical strategy in Section 5.1, then we turn to four sets of results. In Section 5.2 we validate our experimental setting by showing that updating from new draws—*learning*—is similar to that found in the existing literature. Then, in Section 5.3, we turn to the main topic of the paper, updating from retractions—*unlearning*. We ask: do retractions work?; do people update differently from retractions versus new draws?; and how do retractions interact with previously documented deviations from Bayesian updating? (Hypotheses 1-2). Next, in Section 5.4, we examine possible mechanisms for why retractions fail (Hypotheses 3-6). Finally, in Section 5.5, we test whether retractions affect *subsequent* updating (Hypothesis 7).

5.1. Empirical Strategy

There are two distinct empirical tasks: identifying the effectiveness of unlearning (vs. learning) for a given history, and aggregating the results across different histories. For both, we lean on the simplicity of our experimental design to make the analysis non-parametric when possible.

5.1.1. Identifying Learning versus Unlearning

To test the effectiveness of unlearning and to compare it to learning, we perform two distinct comparisons throughout our analysis, corresponding to parts (a) and (b) of Hypothesis 1 and explained visually in Figure 2:

(a) *Testing unlearning*: Are subjects' beliefs after seeing a retraction the same as if the retracted signal had never been observed in the first place?

$$b(\theta \mid \text{Signals, Retraction of Signal } s_\tau) = b(\theta \mid \text{Signals} \setminus \text{Signal } s_\tau)$$

(b) *Comparing unlearning to learning*: Do subjects update equally from retractions as from equivalent (in terms of Bayesian belief updates) new information?

$$b(\theta \mid \text{Signals, Retraction of Signal } s_\tau) = b(\theta \mid \text{Signals} \cup \text{New Signal } - s_\tau)$$

To outline the regressions for these basic tests, we introduce some notation. Denote by b the subject's beliefs—the probability they assign to the truth ball being yellow—and by s the signal in question. We treat signals as $+1$ if they favor the belief that the truth ball is yellow—new draws of a yellow ball or retractions of a blue ball—and -1 if they favor it being blue.[19] Finally, denote by r a dummy variable indicating whether the signal is a retraction ($r = 1$) or a new draw ($r = 0$).

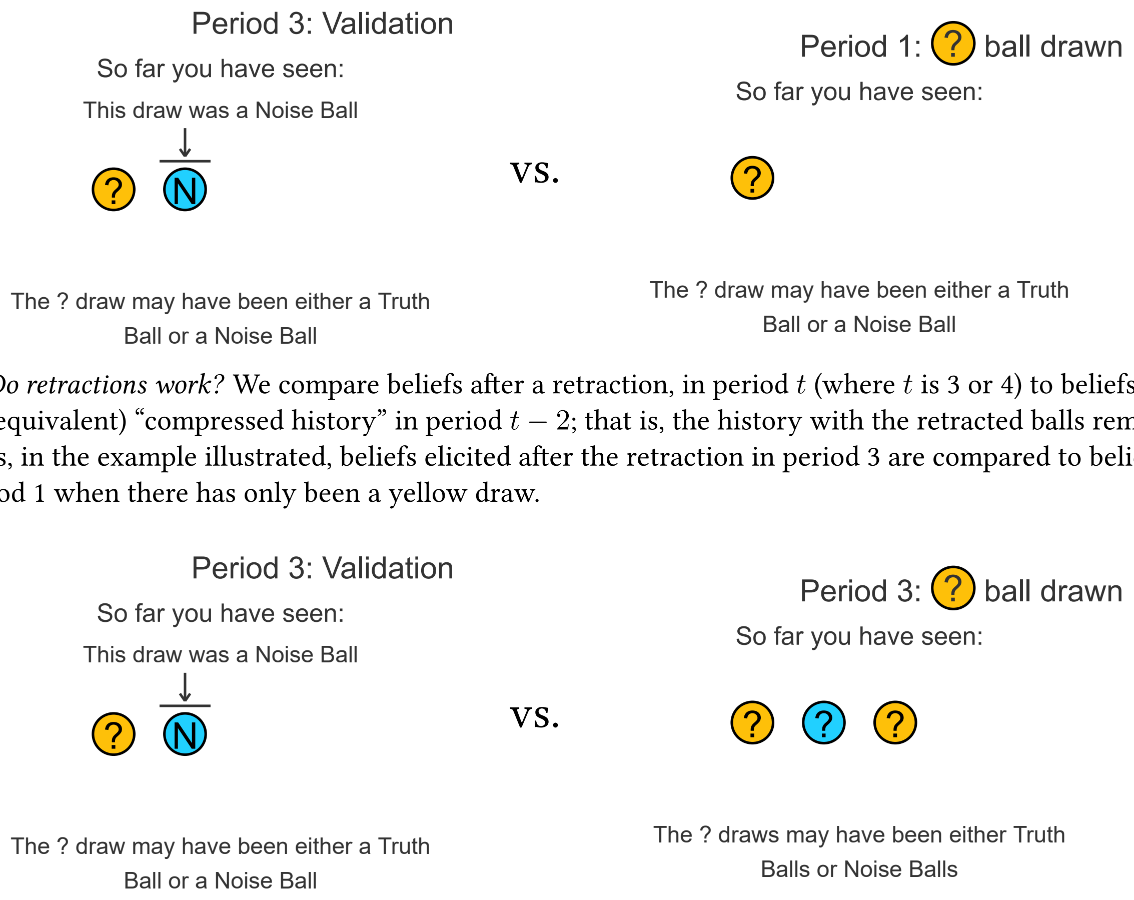
With this notation in hand, for a specific history, we can perform both tests (a) and (b) with the regression:

$$b = \beta_0 + \beta_1 \cdot r \cdot s, \tag{4}$$

where the sample for test (a) comprises beliefs after the retraction as well as when the retracted signal had not been observed to begin with, while for test (b) it comprises beliefs after the retraction and beliefs after a new signal of the opposite sign. The coefficient of interest for both tests is β_1. Under test (a), if β_1 is zero, retractions work: beliefs are as if the retracted signal was never seen; if it is negative, retracted signals continue to influence beliefs. Under test (b), if β_1 is negative, beliefs move less in response to retractions than to equivalent new signals. To give concrete examples, as illustrated in Figure 2, test (a) would compare beliefs in period 3 having observed *(yellow, blue, retraction of the blue)*, to those in period 1 having just observed *(yellow)*; while test (b) would compare beliefs in period 3 having observed *(yellow, blue, retraction of the blue)*, to those in period 3 having observed *(yellow, blue, yellow)*.

When analyzing tests of unlearning, test (a), we compare belief reports in levels, while for test (b) we compare effects on beliefs in both levels and changes (first differences), since the test is specifically about how beliefs *change* in response to retractions. We use beliefs as reported by subjects, on a linear scale (0 to 100), except when we analyze biases in belief updating in Sections

[19] To be precise, if an earlier signal of value v is retracted, then $s = -v$


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
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

(a) *Do retractions work?* We compare beliefs after a retraction, in period t (where t is 3 or 4) to beliefs after an (equivalent) "compressed history" in period $t-2$; that is, the history with the retracted balls removed. Thus, in the example illustrated, beliefs elicited after the retraction in period 3 are compared to beliefs in period 1 when there has only been a yellow draw.

(b) *Are retractions treated differently from equivalent new signals?* We compare beliefs after a retraction, in period t (where t is 3 or 4) to beliefs after an equivalent new draw (of opposite color to the draw which was retracted), also in period t. Thus, in the example illustrated, beliefs elicited after the retraction of the blue ball in period 3 are compared to beliefs elicited in period 3 when the history through period 2 is the same, but then a draw of a yellow ball occurs in period 3.

Figure 2: Illustrative examples to explain the empirical strategy

5.2 and 5.3.2, where we use the log odds scale to be consistent with existing literature. Levels has the advantage that extreme beliefs, near 0 or 100, are not overly inflated; log odds has the advantage that the experimental signals should lead to a constant change in the log-odds belief, independent of the prior. As we show in Online Appendix B.4, our conclusions are robust to relying exclusively on log odds.

5.1.2. Aggregating Results Across Histories, Using Fixed Effects

While we report results disaggregated by case in Figures 4 and 5, showing that they are qualitatively consistent across histories, the results are simpler to digest when aggregated. We do so by pooling the sample across histories in suitably modified versions of the above regressions. The basic identification concern in pooling across histories is that heterogeneity in updating from new signals across histories has been well documented. As such so we do not want to use identifying variation which compares updating from retractions in one history to updating from new signals in a different history.

We ensure that we are only identifying off within-history variation by using appropriately defined fixed effects. To explain them, denote by H_t the history up to and including period t, that is, the set of all the draws observed as well as the retractions, fixing the order. For the tests of unlearning, (a), we use fixed effects for what we refer to as a *compressed history*, $C(H_t)$: the history, removing any retracted ball draws as if they had never occurred to begin with, keeping the order fixed. For instance, a history of *(yellow, blue, retraction of the blue)* would be equivalent to *(yellow).*[20] For the comparisons to new information, test (b), we include fixed effects at the level of the *sign history*, $S(H_t)$, which is the history without distinguishing whether signals were new draws or retractions. For example, *(blue, yellow, retraction of the blue)* is equivalent to *(blue, yellow, yellow).* Once we include these fixed effects in the pooled regression, if there have not been retractions in previous periods, then we compare the a retraction of the ball of one color to the informationally equivalent new draw of the opposite color, conditional on what happened in all previous periods of the round.

5.2. Learning: Preliminary Analysis of Subject Responses and Belief Paths

As a first step in our analysis, in part as a test of validity of experimental setting, we examine the belief paths of subjects when they are not shown retractions using the same empirical approach as previous papers in the literature. One concern about our design is that the complexity would make it difficult for subjects to understand the instructions. Figure 3 dispels this concern: reported beliefs track Bayesian posteriors, implying that, on average, subjects are able to interpret correctly

[20] Note that while the signal order does matter, we do *not* include a timestamp on each signal draw for these fixed effects.

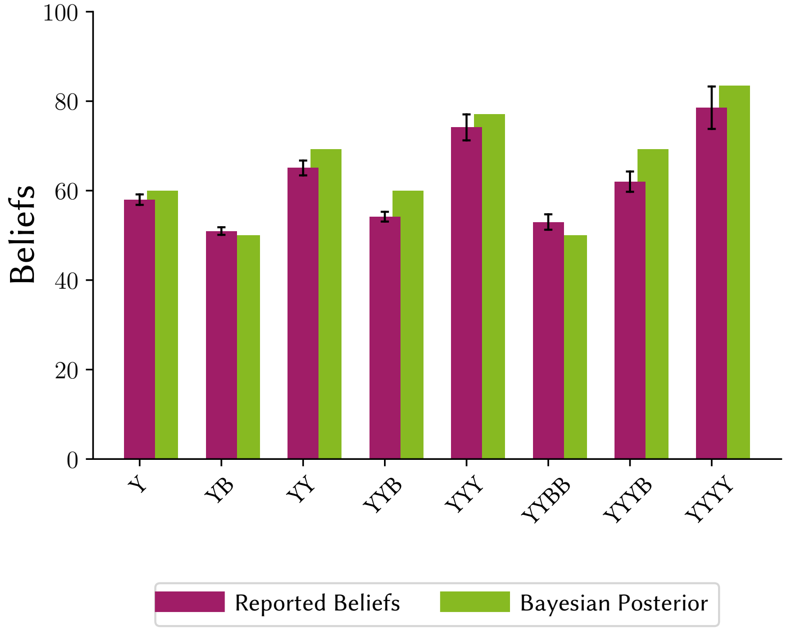


Figure 3: Reported Beliefs and Bayesian Posteriors

Notes: The figure compares mean reported beliefs with Bayesian posteriors. Belief reports are symmetrized around 50, e.g. signals BY are treated as YB. The sample is restricted to the baseline treatment and to sequences of observations in which no signal is retracted. The whiskers denote 95% confidence intervals using standard errors clustered at the subject level.

information from new draws.[21]

In the absence of a retraction, the design is very similar to many others surveyed by Benjamin (2019). We show that the results are largely consistent with the main findings from the literature, suggesting that any differences in our subsequent analysis can indeed be attributed to distinct features of retractions. In Table 1, we present Grether-style (Grether, 1980) log-odds regressions—a workhorse model of analysis in this literature—enabling a direct comparison to existing experimental results on belief updating. Specifically, Table 1 shows the following specification, restricted to the cases where there has not been a retraction (so only new draws):

$$l_t = \beta_0 + \beta_1 \cdot l_{t-1} + \beta_2 \cdot s_t \cdot K \tag{5}$$

and $$l_t = \beta_0 + \beta_1 \cdot l_{t-1} + \beta_2 \cdot s_t \cdot K + \beta_3 \cdot s_t \cdot K \cdot c_t \tag{6}$$

[21] Figures 6 and 7 in the Online Appendix B.1 present the distance the belief reports are from the truth in real and absolute terms: in both cases, we see that reports tend to be fairly close to the truth on average.

where t is the period, l_t is the log-odds of the beliefs reported at t,[22] s_t is the signal in round t (+1 or -1), $c_t := \mathbf{1}\{\text{sign}(l_{t-1}) = \text{sign}(s_t)\}$ is an indicator function that equals 1 when the signal at t confirms the prior at $t-1$, and $K > 0$ is a constant factor of Bayesian updating.

The usefulness of using a log odds framework is that a perfect Bayesian updater would move log odds by a constant amount, which depends only on the likelihood of each signal. Hence the above regression, regressing log-odds of belief reports on this log odds ratio would yield a coefficient $\beta_2 = 1$ for a Bayesian updater. Benjamin (2019) notes that this tends not to be the case: subjects tend to under-react to new information. For the two incentivized studies he reviews with sequential observations, the estimate on this coefficient is $.528$ times the likelihood. Thaler (2021) provides evidence that subjects 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.

In the most parsimonious of our specifications, we find this coefficient estimate to be $\hat{\beta}_2 = 1.219$, indicating mild over-updating from new information. Once we include the effect of confirmatory information, we uncover an interesting finding: the estimated coefficient on the likelihood becomes 0.998 (not significantly different from 1; p-value = .929), while $\beta_3 > 0$ (p-value$< .001$). We note that strict overinference resulting from confirmatory information—that is, $\beta_2 + \beta_3 > 1$ (p-value $< .001$)— has been previously documented (e.g. Charness and Dave, 2017). Together, this suggests that our subjects slightly over-react to new information but that this is mostly driven by confirmation bias: they update more from a signal when the belief movement is in the direction of their prior. We also verify another deviation from Bayesian updating identified in the literature: subjects exhibit base-rate neglect. In other words, they underweight the prior, as evidenced by $\beta_1 < 1$.

It is helpful to keep these general patterns in mind below when interpreting our results; we emphasize that while subjects depart from Bayesian updating, they do so in a way consistent with what one would expect from the literature. It also suggests that, since we find these biases in the "new information" treatment, any additional departure due to retractions cannot be attributed to explanations that are not specific to the nature of the information source.

To summarize, in our analysis of this data, we do not see any consistent departure from the

[22] All tables involving log-odds of beliefs treat $b_t = 100$ and $b_t = 0$ respectively as $b_t = 100 - \delta$ and $b_t = \delta$. We chose $\delta = 0.1$ so as to avoid biasing the regression with extreme outliers. The results are robust to varying δ and to dropping subjects that answer $b_t \in \{0, 100\}$.

	(1)	(2)
	l_t	l_t
Prior (l_{t-1})	0.875***	0.800***
	(0.026)	(0.023)
Signal (s_t)	1.219***	0.998***
	(0.034)	(0.025)
Signal Confirms Prior ($s_t \cdot c_t$)	–	0.417***
		(0.062)
R-Squared	0.48	0.41
Observations	18491	18491

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1: Updating from New Draws

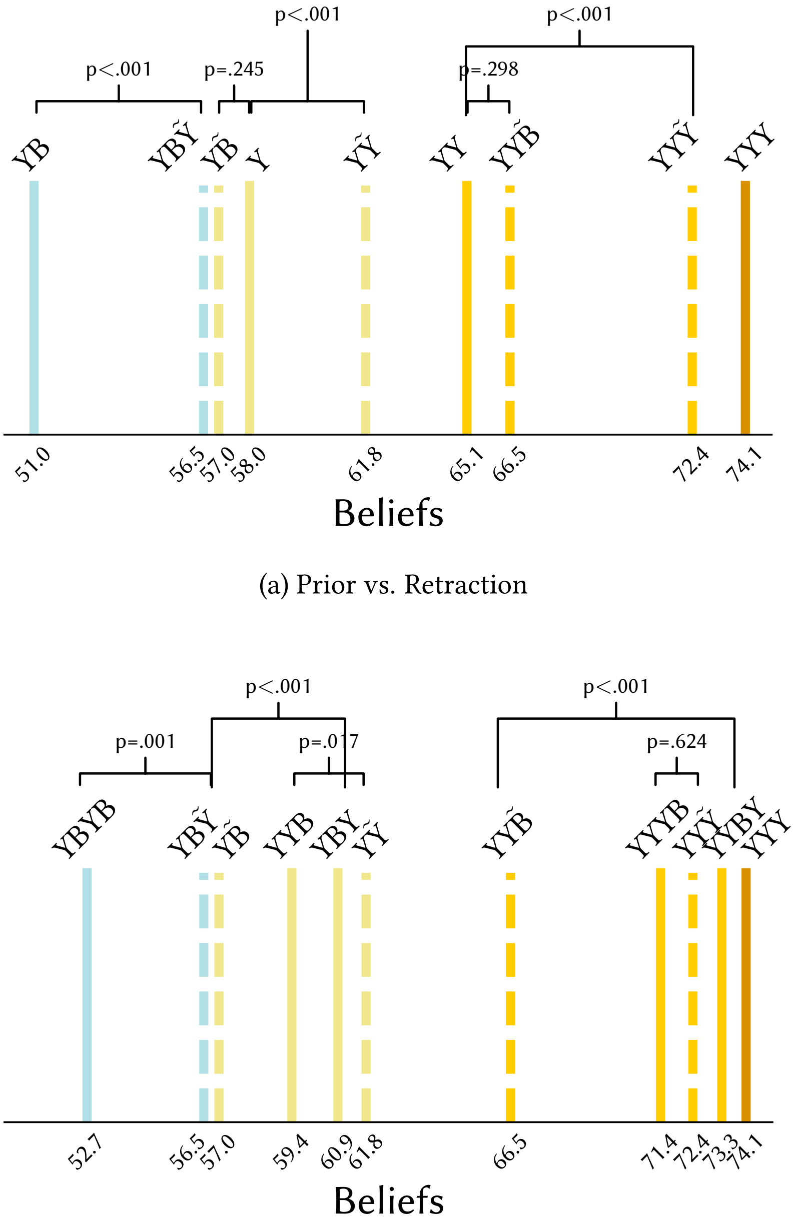
Notes: This table represents updating from standard new ball draws. The sample consists of subjects in the baseline treatment (beliefs are elicited each period). We include beliefs of all periods (1-4) but, within a given round, we exclude any beliefs which are elicited after a verification (regardless of whether a truth or noise ball is revealed). The regressions correspond to equations (5) and (6). The outcome is the log odds of beliefs in period t, l_t. s_t is the signal in round t (+1 or -1, multiplied by K, a constant factor of Bayesian updating, such that the coefficient on s_t would be 1 under Bayesian updating), $c_t := \mathbf{1}\{\text{sign}(l_{t-1}) = \text{sign}(s_t)\}$ is an indicator function that equals 1 when the signal at t confirms the prior at $t-1$.

prior literature on belief updating.[23] We do not find any significant departures from existing literature and therefore do not have strong reasons to suspect our results are driven by, for instance, the choice of venue.

5.3. Unlearning: Updating from Retractions

This section presents our first main findings, on the failure to fully "unlearn" from retractions and on the differences in belief updating from retractions as opposed to new signals. While we begin with the aggregate results, the richness of the design also allows us to break down the comparison of retractions to new signals across various belief paths, and hence to study how retractions interact with existing deviations from Bayesian updating.

[23] In Online Appendix B.3, we reestimate the specifications in Table 1 using probability weights so as to render different histories equally likely. Not only do the conclusions remain unchanged, the estimates are extremely similar.

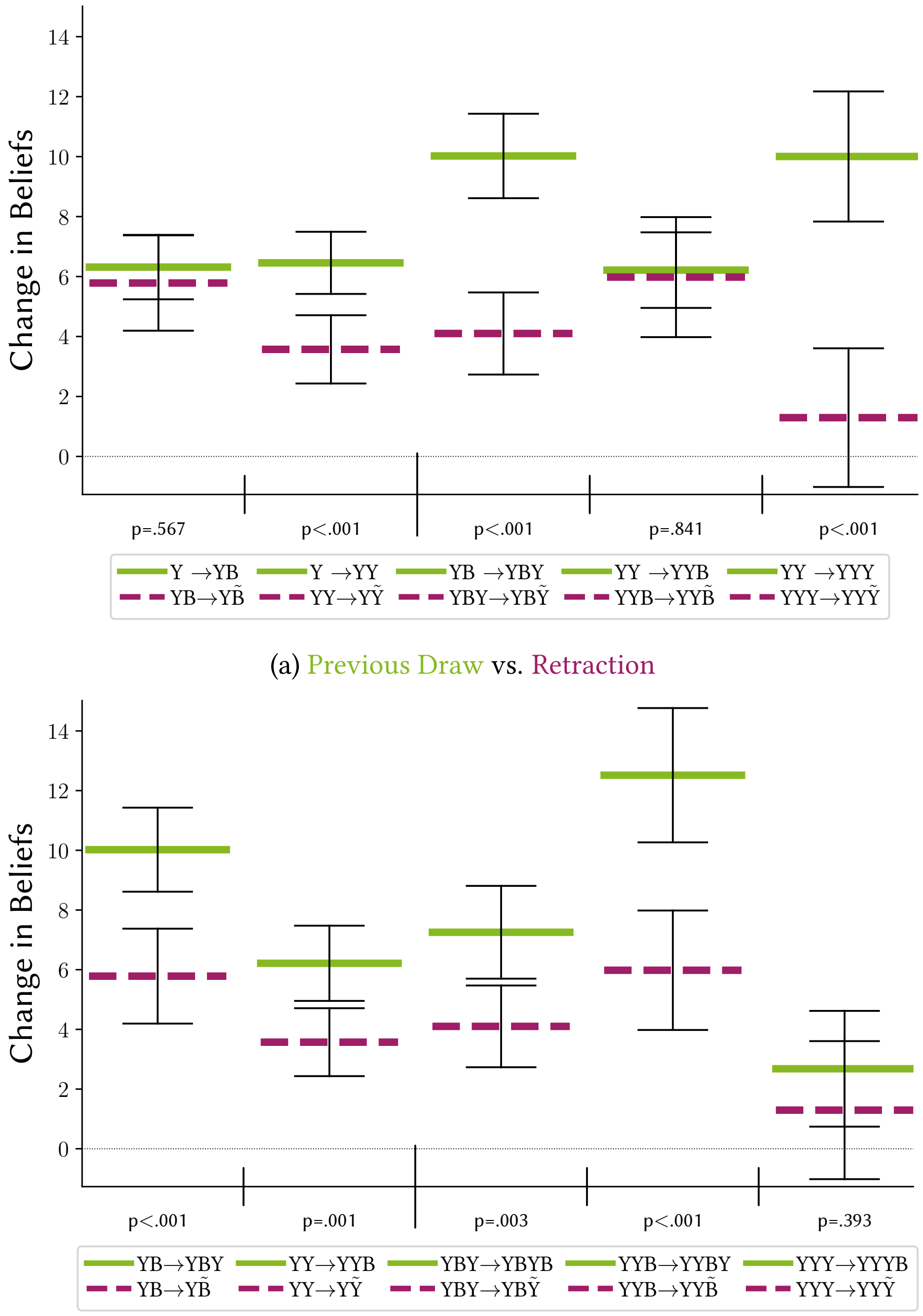


(a) Prior vs. Retraction

(b) New Draw vs. Retraction

Figure 4: Updating from Retractions: Beliefs (Hypothesis 1)

Notes: The figure displays the mean (thick vertical lines) of the reported beliefs disaggregated by history, as indicated above the line. Dashed lines indicate beliefs at histories in which a retraction occurred; solid lines those in which no retraction occurred. Lines of the same color correspond to histories inducing the same Bayesian posterior. Belief reports are symmetrized around 50, i.e. $100 - b(B\tilde{Y})$ is treated as $b(Y\tilde{B})$, where a tilde denotes a retracted signal. The sample paths are not conditioning on the sequence order: e.g. $Y\tilde{B}$ and $\tilde{B}Y$ are bundled together. The sample consists of subjects in the baseline treatment (beliefs are elicited each period). p-values were obtained by a regression similar to columns (1) and (2) of Table 2, using standard errors clustered at the subject level, but restricting to the disaggregated histories.



(a) Previous Draw vs. Retraction

(b) New Draw vs. Retraction

Figure 5: Updating from Retractions: Changes in Beliefs (Hypothesis 1)

Notes: The figure compares the change in beliefs following a retraction (e.g. $\Delta b(Y\tilde{B}) = b(Y\tilde{B}) - b(YB)$) to (a) the change in beliefs induced by the retracted signal when first drawn ($\Delta b(YB) = b(YB) - b(Y)$), and (b) to the change in beliefs ensuing an equivalent new draw ($\Delta b(YBY) = b(YBY) - b(YB)$). A tilde denotes a retracted signal. Belief reports are symmetrized around 50, i.e. $-(b(B\tilde{Y}) - b(BY))$ is treated as $(b(Y\tilde{B}) - b(YB))$; equivalently, we normalize the direction in which the updating should occur by considering $\Delta b_t \cdot s_t$. The sample paths are not conditioning on the sequence order: e.g. $Y\tilde{B}$ and $\tilde{B}Y$ are bundled together. The sample consists of subjects in the baseline treatment (beliefs are elicited each period). The whiskers denote 95% confidence intervals using standard errors clustered at the subject level; p-values were obtained by auxiliary regressions similar to column (3) of Table 2 (using standard errors clustered at the subject level), but restricting to the disaggregated histories.

	Prior vs. Retraction	Retraction vs. New Draw		
	(1)	(2)	(3)	(4)
	b$_t$	b$_t$	Δb$_t$	Δb$_t$
Retraction (r_t)	0.201	-0.167	-0.351	-0.242
	(0.275)	(0.369)	(0.355)	(0.363)
Retracted Signal ($r_t \cdot s_t$)	-3.134***	-3.628***	-3.701***	-3.316***
	(0.601)	(0.727)	(0.670)	(0.675)
Signal (s_t)	–	–	–	8.658***
				(0.510)
Compressed History FEs	Yes	No	No	No
Sign History FEs	No	Yes	Yes	No
Lagged Sign History FEs	No	No	No	Yes
R-Squared	0.29	0.39	0.16	0.15
Observations	17591	9074	9074	9074

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Updating from Retractions: do they work and how do they compare to equivalent new signals (Hypothesis 1)

Notes: This table provides the main specification of interest in this paper. It tests whether retractions are effective in inducing 'un-learning' and compares their effectiveness relative to new direct information. The sample consists of subjects in the baseline treatment (beliefs are elicited each period). Column (1) tests whether retractions work, by comparing beliefs after a retraction to beliefs after the equivalent compressed history. We include beliefs of all periods (1-4) but, within a given round, we exclude any beliefs which are elicited after the truth ball is disclosed and we exclude period 4 beliefs if there was a retraction in period 3. In periods 3 and 4 we only include beliefs when there was a retraction in that period. The outcome is the beliefs in period t, $b_t \in [0, 100]$. $r_t \cdot s_t$ is the opposite sign of the retracted signal in round t (+1 if a -1 signal is retracted, -1 if a +1 signal is retracted). The regression includes fixed effects for the compressed history of draws. Columns (2) to (4) test whether people update more or less from retractions compared to equivalent new signals. The sample is restricted to beliefs in periods 3 and 4, once again dropping beliefs after the truth ball is disclosed or in period 4 if there is a retraction in period 3. The specifications include fixed effects for the sign history. In column (2), the outcome is the beliefs in period t, b_t. In columns (3) and (4), the outcome is the first difference in beliefs. Column (4) uses lagged sign history fixed effects to enable us to compare the magnitude of $r_t \cdot s_t$ to s_t, which is otherwise absorbed by the fixed effects.

5.3.1. Failure to "Unlearn" and Retractions Versus New Signals (Hypothesis 1)

Our first result, and the key finding of the paper, is the empirical support of Hypothesis 1: retractions are ineffective, in that (a) retracted signals are not fully disregarded (Prior vs. Retraction), and (b) beliefs are less responsive to retractions than to equivalent new signals (Retraction vs. New Draw). Figure 4 depicts mean beliefs across different histories and demonstrates both parts of the hypothesis. Figure 5 confirms this when looking at changes in beliefs: the change in beliefs following a retraction is smaller than the changes in beliefs both when (a) the subsequently-retracted signal was originally observed, and (b) when, instead of a retraction, subjects observe an equivalent new draw.

We pool these results across different histories in Table 2. Column (1) is a test of (a) and corresponds to the following regression:[24]

$$b_t = \beta_0 + \beta_1 \cdot r_t \cdot s_t + \beta_2 \cdot r_t + F_{C(H_t)}, \tag{7}$$

where the sample is all beliefs in periods 1 and 2, and beliefs in periods 3 and 4 if there was a retraction in that period (except in period 4 if there was a retraction in period 3). As explained in Section 5.1.2, controlling for compressed history fixed effects $F_{C(H_t)}$ compares, for example, the beliefs after observing $(s_1, s_2, n_2 = 1)$ to those reported when only signal s_1 was seen.

Columns (2)-(4) test (b) and correspond to variants of the following regression:

$$b_t = \beta_0 + \beta_1 \cdot r_t \cdot s_t + \beta_2 \cdot r_t + \beta_3 \cdot s_t + F_{S(H_t)}, \tag{8}$$

where the sample is all beliefs in periods 3 and 4, except those in period 4 if there was a retraction in period 3. As explained in Section 5.1.2, controlling for sign history fixed effects, $F_{S(H_t)}$, means we compare for example beliefs reported after $(s_1, s_2, n_2 = 1)$ to those reported after $(s_1, s_2, s_3 = -s_2)$. In Column (2) the dependent variable is belief levels, whereas in column (3) it is *change* in beliefs. Column (4) uses less stringent fixed effects—those for *lagged* signed history $F_{S(H_{t-1})}$—so that the signal term s_t is not absorbed by the fixed effects. This enables us to benchmark the differential effect of retractions, β_1, by comparing it with the effect of new signals, β_3.

The key finding for both tests is that the differential effect of retractions on beliefs, β_1 the coefficient on $r_t \cdot s_t$, is negative and consistent in magnitude across all of the specifications we study. Retractions are treated *differently*, and in particular as if they were less informative than

[24] The coefficient on r, which is added when we aggregate across histories, is not of primary interest. For test (a), here, it identifies whether beliefs in period 3 are on average shifted towards yellow, compared to period 1, across all retractions. Since the probability of a blue versus yellow retraction may not be balanced, and depends on the history, we do not have a simple prediction for this term. For test (b), below, it reflects whether beliefs in period 3 are on average shifted towards yellow under retractions versus new signals. Again, since retractions may be more likely in one direction than another at a given history, we do not have a simple prediction for this term.

equivalent new signals. To quantify this effect, a simple comparison shows that beliefs move approximately one-third less when information is in the form of a retraction. This can be seen from column (4) of Table 2, by comparing the coefficient on the retracted signal—the interaction term between the signal and the retraction variables—to the coefficient on the signal variable itself. Performing this back-of-the-envelope calculation in other ways, for example by dividing the coefficient on $r_t \cdot s_t$ in column (3) by the average update from a new signal in the corresponding sample, consistently finds that beliefs update around $1/3$ from retractions relative to new signals.

5.3.2. Retractions Accentuate Biases in Updating (Hypothesis 2)

Having illustrated that retractions are treated differently, with beliefs reacting less on average, we next seek to determine *when* this effect is relatively more or less pronounced, and how variation across belief paths in updating from retractions compares to well-documented variation in updating from new signals. Specifically, we return to the regression specifications which were the focus of Section 5.2, equations (5) and (6), which are the conventional specifications in the literature analysing other deviations from Bayesian updating in similar experiments. We then fully interact these specifications with the retraction variable, r_t, corresponding to whether the signal was in the form of a retraction:

$$\begin{aligned} l_t = \beta_0 + \beta_1 \cdot l_{t-1} + \beta_2 \cdot s_t \cdot K + \beta_3 \cdot s_t \cdot K \cdot c_t + \\ + r_t \cdot [\gamma_0 + \gamma_1 \cdot l_{t-1} + \gamma_2 \cdot s_t \cdot K + \gamma_3 \cdot s_t \cdot K \cdot c_t] \end{aligned} \tag{9}$$

The inclusion of the interactions allows us to detect how previously documented deviations from Bayesian updating vary, depending on whether or not the signal is a retraction. In other words, they provide a flexible functional form in order to capture the effect of retractions as discussed in Section 2.

The results can be found in Table 3. A striking pattern emerges: when updating from new draws subjects (slightly) overinfer from signals ($\beta_2 \geq 1$) and do more so when signals confirm the prior ($\beta_3 > 0$); in contrast, when updating from retractions they *under*-infer ($0 < \beta_2 + \gamma_2 < 1$) and exhibit *anti*-confirmation bias ($\beta_3 + \gamma_3 < 0$).[25] In sum, belief updating from retractions

[25] We conduct F-tests to analyze the statistical significance of such observations: β_2 is not significantly different from 1 in column (2) (p-values$=.929$), and significantly larger than 1 in the remaining columns (p-value< 0.01); $\beta_2 + \gamma_2$ is always significantly smaller than 1 and larger than 0 (p-value$< .001$ in all cases); $\beta_3 + \gamma_3$ is always negative and significantly different from zero (p-value$= .023$ for column (3) and $.018$ for column (4)).

	All Periods		Period 3	
	(1)	(2)	(3)	(4)
	l$_t$	l$_t$	l$_t$	l$_t$
Prior (l_{t-1})	0.834***	0.800***	0.904***	0.839***
	(0.021)	(0.023)	(0.039)	(0.048)
Signal (s_t)	1.126***	0.998***	1.647***	1.314***
	(0.023)	(0.025)	(0.068)	(0.092)
Signal Confirms Prior ($s_t \cdot c_t$)	–	0.417***	–	0.705***
		(0.062)		(0.178)
Retraction (r_t)	-0.033*	-0.030	-0.034	-0.026
	(0.019)	(0.019)	(0.034)	(0.034)
Retraction x Prior ($r_t \cdot l_{t-1}$)	0.019	0.070*	-0.093*	-0.002
	(0.033)	(0.038)	(0.055)	(0.067)
Retracted Signal ($r_t \cdot s_t$)	-0.768***	-0.541***	-1.286***	-0.825***
	(0.050)	(0.068)	(0.085)	(0.119)
Retraction x Signal Confirms Prior ($r_t \cdot s_t \cdot c_t$)	–	-0.675***	–	-1.051***
		(0.131)		(0.231)
R-Squared	0.44	0.44	0.43	0.43
Observations	22578	22578	6081	6081

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3: How Retractions Interact with Other Biases (Hypothesis 2)

Notes: This table runs standard tests for biases in belief updating, interacting them with whether the signal was a retraction. The sample consists of subjects in the baseline treatment (beliefs are elicited each period). The regressions correspond to equation (9), using data from period 3. The sample excludes the cases where the truth ball is disclosed. The outcome is the log odds of beliefs in period t, l_t. s_t is the signal in round t (+1 or -1, multiplied by K, a constant factor of Bayesian updating, such that the coefficient on s_t would be 1 under Bayesian updating), r_t is an indicator variable for whether the signal in period t can from a retraction, $c_t := \mathbf{1}\{\text{sign}(l_{t-1}) = \text{sign}(s_t)\}$ is an indicator function that equals 1 when the signal at t confirms the prior at $t-1$.

exhibits the opposite biases when compared to updating from new draws, a conclusion which is robust across all specifications. This strengthens the conclusion that retractions are treated differently from new signals, inasmuch as the behavioral responses to retractions are not simply accentuating pre-existing biases; in fact, retractions induce opposite biases in belief reporting behavior.[26]

5.4. Why Do Retractions Fail?

We now turn to the question of why retractions are not effective, and especially why they affect beliefs less than equivalent new draws. We present three sets of results. First, we consider the question of whether retractions are less effective once signals have been acted upon, and hence internalized. We test this by comparing the effect of retractions when beliefs have already been elicited versus when they have not, holding constant the history of signals. Second, we analyze the heterogeneous effect of retractions based upon which signal was retracted. Third, we discuss whether the timing of retractions themselves affects their effectiveness by considering whether beliefs are updated differently when retractions occur after the second or the third signals. Taken together, the results suggest that the failure of retractions is not driven by one particular instance, but instead reflect retractions being treated as uniformly less informative. We then examine decision time data to understand whether this reflects the increased difficulty of the contingent reasoning required to interpret retractions.

5.4.1. Retracting Internalized Signals (Hypothesis 3)

Assuming that our results on the ineffectiveness of retractions is a cognitive effect, a reasonable hypothesis is that this residual effect is stronger if the earlier draw has been acted upon and hence potentially internalized. In other words, retraction failure could be due to an informational version of 'endowment' effect, with individuals resisting to 'delete' past information that was acted upon, in absence of which retractions would successfully induce 'unlearning.' We test this hypothesis by comparing updating from retractions when beliefs have already been elicited versus when they have not, by comparing beliefs across intermediate (baseline) versus final (single) elicitation treatments.

The results from this comparison are documented in Table 4. The specifications correspond to equations (7) and (8) which we described in Section 5.3.1, with the addition of the final elicitation treatment as an interaction term.[27] The result is a well-identified null result: having acted upon a

[26] As for Table 1, we reestimate the specifications in Table 3 using probability weights so as to render different histories equally likely. Again, the conclusions are the same and the estimates are very similar. The results can be found in Online Appendix B.3.

[27] When beliefs are elicited only at the end of each round, it is not possible to obtain the first difference in beliefs and there is therefore no way to estimate columns (3) and (4) of Table 2.

	Prior vs. Retraction	Retraction vs. New Draw
	(1)	(2)
	b_t	b_t
Final (Fin_t)	0.175	0.143
	(0.938)	(0.997)
Retraction (r_t)	-0.048	-0.230
	(0.326)	(0.400)
Retracted Signal ($r_t \cdot s_t$)	-2.404***	-3.658***
	(0.622)	(0.712)
Final x Retraction ($\text{Fin}_t \cdot r_t$)	–	0.049
		(0.754)
Final x Retracted Signal ($\text{Fin}_t \cdot r_t \cdot s_t$)	0.132	0.154
	(0.999)	(0.995)
Compressed History FEs	Yes	No
Sign History FEs	No	Yes
R-Squared	0.21	0.31
Observations	11213	9920

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Intermediate versus Final Belief Elicitation (Hypothesis 3)

Notes: This table tests whether updating from retractions is different if beliefs have previously been elicited before a signal is retracted. The sample includes all subjects, both those in the baseline treatment (beliefs are elicited each period) as well as those in the final elicitation treatment (beliefs are elicited only at the end of the round). Column (1) restricts to period 1 and to period 3 when there is a retraction, interacting the specification from column (1) in Table 2—are retractions effective—with a dummy for being in the final period only elicitation group ($\text{Fin}_t \cdot r_t$ and $\text{Fin}_t \cdot s_t$ are spanned by the other controls and hence omitted, since period 3 is only in the sample when it is a retraction, making Fin_t = $\text{Fin}_t \cdot r_t$ within the sample). Column (2) restricts to period 3, interacting the specification from column (2) in Table 2—is updating from retractions different from updating from new signals—with a dummy for being in the final period only elicitation group.

piece of information does not change the effect of it being retracted. As before, beliefs move in the directions of signals, but less so for retractions relative to new signals, and none of the other lessons we have described so far are changed in this treatment. While this does not imply that retractions are as (in)effective when individuals acted upon past information in other settings, it does suggest that the intermediate elicitation of beliefs is underlying the effect in our setting.

	Prior vs. Retraction	Retraction vs. New Draw		
	(1)	(2)	(3)	(4)
	b$_t$	b$_t$	Δb$_t$	Δb$_t$
Retraction (r_t)	0.630	-0.158	-0.745**	-0.403
	(0.412)	(0.438)	(0.358)	(0.381)
Retracted Signal ($r_t \cdot s_t$)	-4.341***	-4.258***	-3.893***	-3.534***
	(0.685)	(0.800)	(0.763)	(0.760)
Last Draw Retracted (rl_t)	-0.981	-0.029	0.896	0.362
	(0.662)	(0.697)	(0.552)	(0.475)
Retracted Signal x Last Draw Retracted ($rl_t \cdot s_t$)	2.737***	1.436**	0.432	0.500
	(0.678)	(0.641)	(0.649)	(0.639)
Signal (s_t)	–	–	–	8.657***
				(0.510)
Compressed History FEs	Yes	No	No	No
Sign History FEs	No	Yes	Yes	No
Lagged Sign History FEs	No	No	No	Yes
R-Squared	0.29	0.39	0.16	0.15
Observations	17591	9074	9074	9074

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Timing of Retracted Signals (Hypothesis 4)

Notes: This table tests whether there is a difference in responding to retractions depending on whether the last signal was retracted or an earlier signal was retracted. The sample consists of subjects in the baseline treatment (beliefs are elicited each period). In column (1), we include beliefs of all periods (1-4) but, within a given round, we exclude any beliefs which are elicited after the truth ball is disclosed. We also exclude any beliefs if there was no retraction in period 3 or 4. In columns (2) to (4), we further restrict the beliefs to periods 3 and 4. In column (2), the outcome is the beliefs in period t, b_t. In columns (3) and (4), the outcome is the first difference in beliefs. Column (4) uses lagged sign history fixed effects to enable us to compare the magnitude of $r_t \cdot s_t$ to s_t.

5.4.2. The Timing of Retracted Signals (Hypothesis 4)

One feature of our design is that, while the signals subjects receive are exchangeable, they are observed in sequence. If updating also happens in sequence, then retractions of signals received earlier may induce a more complex reevaluation of previously observed signals, relative to retractions of signals received later, since they require subjects to reverse their belief updating until before the retracted signal was received, and then add back the subsequent signals as if the retracted signal had not been observed. In this case, a natural conjecture is that retractions are effective in inducing 'unlearning' when they to refer to information that was just received (Hypothesis 4), in which case subjects only need report their beliefs from the previous period. This section assesses the relevance of the timing of retracted signals in accounting for the failure of retractions.

The importance of the timing of retracted signals is presented in Table 5. It reports the same basic specifications described in Section 5.3.1, tests of (a) retractions and (b) retractions versus new information, but with the addition of an indicator variable for whether the last signal observed was retracted (rl_t), as well as its interaction with the signal itself. We find that the effectiveness of retractions is slightly increased when they correspond to the most recently received ball draw; it is easier to disregard a piece of information if it arrived more recently. This can be seen in column (1), by the positive and statistically significant estimated coefficient on $rl_t \cdot s_t$. However, subjects still fail to fully disregard retracted signals, even when they are of the most recent draw, as reflected by the sum of the coefficients on $r_t \cdot s_t$ and $rl_t \cdot s_t$ being negative (p-value$= .021$) and over 1/3 of the size of $r_t \cdot s_t$. In the comparison of retractions versus new information, the difference with respect to the timing of the retracted signal is significantly muted (Column (2)) or not statistically significant (Columns (3) and (4)).

To summarize, our results suggest that while the timing of the retracted signal may have an effect on the effectiveness of retractions—more recently observed signals appear slightly easier to retract—retractions are still ineffective even when they apply to the most recently observed draw. We caveat that the limited effect of timing may have been driven in part by the experimental design, since signals are observed in close succession to one another. We are unable to say whether these effects may or may not be present when information arrives over a longer timescale, and leave this to future work.

5.4.3. The Timing of Retractions (Hypothesis 5)

Existing literature has documented that beliefs are less sensitive to new signals the more signals have been observed in the past. Motivated by this observation, we ask whether the timing of a retraction itself has any bearing on its effectiveness. We are able to do so because our experimental design involves the possibility of observing either a new draw or a retraction in both periods 3 and

	Prior vs. Retraction	Retraction vs. New Draw
	(1)	(2)
	b_t	b_t
Retraction ($r_3 + r_4$)	1.077*	0.747
	(0.596)	(0.815)
Retracted Signal ($r_3 \cdot s_3 + r_4 \cdot s_4$)	-4.514***	-4.208***
	(0.739)	(0.906)
Retraction in Period 4 (r_4)	-0.371	-0.554
	(0.735)	(0.745)
Retracted Signal x Retraction in Period 4 ($r_4 \cdot s_4$)	-0.162	0.710
	(0.765)	(0.917)
Compressed History FEs	Yes	No
Sign History FEs	No	Yes
R-Squared	0.33	0.40
Observations	9432	4350

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Timing of Retractions (Hypothesis 5)

Notes: We test whether retractions have a different effect if they come in period 3 or period 4, based on beliefs in period 4 and holding the history otherwise fixed. The sample consists of subjects in the baseline treatment (beliefs are elicited each period). In column (1) period 4 is compared to period 2. Beliefs in period 4 are included if there was a retraction in period 3 or 4, but not both. In column (2) only beliefs in period 4 are considered, and they are dropped if there is retractions in both periods 3 and 4. The coefficient of interest, $r_4 \cdot s_4$ is a difference in differences. Based on beliefs in period 4, it is the effect of a retraction in period 4 versus an equivalent new signal in period 4, compared to the effect of a retraction in period 3 versus an equivalent new signal in period 3.

4. We consider similar specifications to Section 5.3.1, and introduce interaction terms to compare the effect of retractions in period 3 to those in period 4.

In order to provide a clear identification of the effect of the timing of the retraction itself, we consider only situations where retractions occurred only in either period 3 or in period 4. Furthermore, in order to compare the effectiveness of retractions relative to new draws, we use fixed effects to restrict identifying comparisons to cases where the compressed history at period 4 is the same, but in one case the retraction took place in period 3 and in the other case in period 4, with both retracting the same signal. That is, in Column (1), we run test (a), comparing beliefs after histories $(s_1, s_2, n_\tau = 1, s_4)$ with those following $(s_1, s_2, s_3 = s_4, n_\tau = 1)$, where $\tau \in \{1, 2\}$, both compared in turn to those after history (s_1, s_2). In Column (2), we run test (b), again comparing $(s_1, s_2, n_\tau = 1, s_4)$ with $(s_1, s_2, s_3 = s_4, n_\tau = 1)$, but this time both compared with $(s_1, s_2, s_3, s_4 = -s_\tau)$.

As can be seen in Table 6, the timing of retractions has no noticeable nor significant impact on the effectiveness of retractions in leading subjects to disregard particular pieces of information, nor does it impact the relative (in)efficiency of retractions vis-à-vis equivalent new signals.

5.4.4. Retractions are Harder to Process (Hypothesis 6)

We now consider a mechanism which could explain why retractions are less effective regardless of when they occur: retractions are simply harder to process. This is a natural hypothesis, when we consider that retractions are information *about* information, while new draws are simply information. To test this it, we rely on a natural proxy for processing cognitive difficulty is the subjects' decision time dt_t, under the assumption that the greater the difficulty the longer the time taken to interpret the information provided. Specifically, we regress decision time on a dummy variable indicating whether or not a retraction occurs in that period. As before, we control for the sign history such that we compare decision times for retraction to those of informationally-equivalent new signals. We estimate

$$dt_t = \beta_0 + \beta_1 \cdot r_t + F_{S(H_t)}, \tag{10}$$

where decision time is measured in seconds in our baseline specification, but we also report results with log decision time.

The results, in Table 7, confirm our hypothesis: subjects take longer to report their beliefs when updating from retractions, suggesting that retractions are not only treated differently, they are also harder to process. We conjecture that the kind of contingent reasoning which is inherent to interpreting information about information renders retractions inherently more complex—as reflected by the decision times. In line with the literature on cognitive imprecision, one interpretation consistent with our results is that such increased complexity is reflected in a noisier perception of the informativeness of a retraction relative to direct information about the state of the world.

	Retraction vs. New Draw	
	(1)	(2)
	dt_t	$\log(\text{dt}_t)$
Retraction (r_t)	0.501***	0.101***
	(0.090)	(0.014)
Mean Decision Time (secs)	5.557	1.547
Sign History FEs	Yes	Yes
R-Squared	0.01	0.02
Observations	8983	8983

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Decision Times in Updating from Retractions (Hypothesis 6)

Notes: This table tests whether the time taken to report beliefs is different following a retraction compared to following an equivalent new signal (Hypothesis 6). The sample consists of subjects in the baseline treatment (beliefs are elicited each period). The specifications compare updating from retractions versus from an equivalent new signal, in periods 3 and 4 (we drop period 4 if there was a retraction in period 3). We drop the top 1% of decision times from the sample.

5.5. Updating After Retractions (Hypothesis 7)

Our design also identifies how retractions affect updating from *subsequent* information, a question on which we are unaware of other work. Were individuals to discount further evidence following a retraction, then retractions would not only be less effective than direct evidence, they would also hamper the interpretation of future evidence, further dampening the absorption of new information. We test whether observing a retraction affects *subsequent* updating, both in the beliefs themselves and in decision times, our proxy for cognitive difficulty.

In Table 8, we study the effect on subsequent belief updating by estimating two variants of the following equation:

$$\Delta b_t = \beta_0 + \beta_1 s_t + \beta_2 r_{t-1} + \beta_3 s_t \cdot r_{t-1} + \beta_4 s_{t-1} \cdot r_{t-1} + \beta_5 s_t \cdot s_{t-1} \cdot r_{t-1} + F. \qquad (11)$$

In columns (1) and (2), the sample is restricted to periods 2 and 4 and fixed effects F correspond to the lagged compressed history, $F_{C(H_{t-1})}$. This enables us to compare changes beliefs after observing, e.g. (s_1, s_2) and $(s_1, s_2, n_2 = 1, s_4 = s_2)$. In columns (3) and (4), we restrict the sample to period 4 and we have fixed effects F corresponding to the lagged sign history, i.e.

	Prior vs. Retraction		Retraction vs. New Draw	
	(1)	(2)	(3)	(4)
	$\Delta \mathrm{b}_t$	$\Delta \mathrm{b}_t$	$\Delta \mathrm{b}_t$	$\Delta \mathrm{b}_t$
Signal (s_t)	6.690***	6.690***	7.964***	7.960***
	(0.397)	(0.397)	(0.648)	(0.647)
Retraction in Previous Period (r_{t-1})	0.655	0.669	0.786	0.777
	(0.542)	(0.536)	(0.810)	(0.811)
Retraction in Previous Period (r_{t-1}) x Signal (s_t)	1.281**	1.296**	0.008	0.018
	(0.516)	(0.517)	(0.697)	(0.698)
Retracted Signal in Previous Period ($r_{t-1} \cdot s_{t-1}$)	–	-1.874***	–	-1.201*
		(0.515)		(0.722)
Retracted Signal in Previous Period ($r_{t-1} \cdot s_{t-1}$) x Signal (s_t)	–	-0.141	–	-0.157
		(0.497)		(0.499)
Lagged Compressed History FEs	Yes	Yes	No	No
Lagged Sign History FEs	No	No	Yes	Yes
R-Squared	0.16	0.16	0.17	0.17
Observations	9779	9779	3027	3027

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Belief Updating After Retractions (Hypothesis 7)

Notes: This table tests updating from new signals *after* retractions, compared to after the equivalent compressed history, and also compared to *after* the equivalent new signal. The sample consists of subjects in the baseline treatment (beliefs are elicited each period). In columns (1) and (3), we restrict the sample to period 2 and 4, comparing updating in period 4, after a retraction in period 3, to updating in period 2, after the equivalent compressed history in period 1. In columns (2) and (4), we restrict the sample to period 4, comparing updating after a retraction in period 3 to updating after the equivalent new signal in period 3.

$F_{S(H_{t-1})}$. As such, we can compare the change in belief reports at histories $(s_1, s_2, n_2 = 1, s_4)$ and $(s_1, s_2, s_3 = -s_2, s_4)$. Our results—consistent across all specifications—suggest that beliefs are *more* sensitive to new signals after a retraction.

Turning to the effect on cognitive difficulty, we test whether experiencing a retraction in the past results in subjects taking longer in subsequent updating from new signals, by estimating:

$$dt_t = \beta_0 + \beta_1 \cdot r_{t-1} + F_{S(H_t)}, \qquad (12)$$

where we control for the sign history. Analogously to Section 5.4.4, we also report on the effect on

	Retraction vs. New Draw	
	(1)	(2)
	dt_t	$\log(\text{dt}_t)$
Retraction in Previous Period (r_{t-1})	0.320^{*}	0.071^{***}
	(0.165)	(0.023)
Mean Decision Time (secs)	5.722	1.576
Lagged Sign History FEs	Yes	Yes
R-Squared	0.00	0.01
Observations	2996	2996

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Decision Times in Updating After Retractions

Notes: This table tests whether the time taken to report beliefs is different after a retraction in the previous period compared to after an equivalent new signal in the previous period. The sample consists of subjects in the baseline treatment (beliefs are elicited each period). The specifications consider updating from new signals in period 4, and compare the decision time based on whether or not there was a retraction in period 3. That is, we compare updating from a given new signal, based upon whether in the previous period there was a new signal or an equivalent retraction.

log decision time. The estimated coefficients—reported in Table 9—suggest that a retraction in the previous period does lead to an increase in decision time, albeit a mild one which is statistically significant only in logs.

5.6. Robustness Checks and Individual Heterogeneity

We strove to ensure that our results were not driven by inattentive subjects. While behavior of participants in our choice of subject pool (Amazon Mechanical Turk) has been shown to approximate well a representative population sample, it is also the case that behavior is 'noisy' relative to a traditional laboratory subject pool (Snowberg and Yariv, 2021; Gupta et al., 2021). We went to lengths to filter out bots and overly inattentive subjects at the start of the experiment, as described in Section 4.3. We also check robustness to excluding subjects based on different measures of inattentiveness.

The results are robust, and indeed slightly stronger, when restricting the sample to those subjects who appear attentive, as defined in three different ways. First, using the (unincentivized) comprehension questionnaire that followed the presentation of the instructions, we restrict our sample to subjects who answered all questions correctly on their first try, who account for a majority. Second, we further restrict the sample to subjects who, when the state is revealed, correctly report that they know the state. Third, we remove subjects whose belief reports are excessively noisy, which we define as updating in the opposite direction to the signal more than 10% of the time.[28] The robustness and indeed slight strengthening of the results (see Online Appendix B.5) is consistent with noisy subjects if anything attenuating the effect, and shows that inattention is not driving our results.

Underinference from retractions appears to be a robust feature within our subject pool, with our results being driven by a substantial fraction of subjects, not just a small minority. To test this, we estimate the specifications in Table 2 at the *subject*-level. We report summary statics on the subject-level estimates of the coefficient of interest (Retracted signal) in Online Appendix B.7. 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 subject. However, both the mean and the median of these coefficient estimates are similar to our baseline estimates, and the estimates are strictly negative for a substantial majority of the subjects (approx. 70%).

We examine whether retractions are more effective for subjects with higher quantitative ability, proxied for by their scores on incentivized quantitative multiple-choice questions which were asked at the end of the experiment. Expanding our main specifications to account for heterogeneity with respect to quantitative ability, we fail to find any significant effects, as reported in Online Appendix B.6.

6. CONCLUSION

This paper has shown that people continue to be influenced by information even once told that it is meaningless. We find that there is a residual impact of information after it is retracted, and that this is a consistent phenomenon across a variety of different kinds of beliefs (extreme vs. moderate) and of retractions (confirming vs. contradicting). We demonstrated this in an abstract setting, where this comparison can be made cleanly and precisely, in an incentivized manner.

In the process of illustrating that retractions are in themselves treated as less informative, we formulated a number of hypotheses relating retraction failure to a number of plausible biases.

[28] As expected, these checks are correlated. For example the first two samples contain a substantially smaller fraction of subjects with excessively noisy reports.

Hypothesis	Documented (✓) or not detected (✕)
1 (a): Subjects fail to fully internalize retractions	✓
(b): Subjects treat retractions as less informative than equivalent new information	✓
2: Updating from retractions accentuates biases in updating	✓[†]
3: Retractions are ineffective only when subjects have acted upon observed signals	✕[‡]
4: Retractions are more effective when referring to more recent signals	✓
5: Later retractions have a different impact on beliefs compared to earlier retractions	✕
6: Updating from retractions takes longer	✓
7: Subjects update differently after retractions	✓[§]

Table 10: Assessment of Main Hypotheses

Notes: See Section 3 for a more complete description of each, as well as the reasoning involved with formulating each one. [†] Retractions reverse the direction of the biases, resulting in under-inference and anti-confirmation bias. [‡] Precise null. [§] Subjects update more from signals after retractions.

Table 10 revisits each of these hypotheses, and assesses our findings. Our analysis suggests that retraction failure is due to difficulties in contingent reasoning particular to information about information—rendering retractions harder to interpret—and not just an expression of well-known biases. We also argued that our design is the simplest possible which still enables the desired comparisons in our two main empirical tests illustrated in Figure 2.

While our main goal in this paper was to document that retractions had a differential impact, and to determine any significant sources of variation, our results point to a number of interesting potential directions for future work. We see two as being particularly natural.

First, exploring this phenomenon in particular contexts, where it may interact with other behavioral biases, is likely to generate further insights. By using an abstract design, we provided, to the best of our knowledge, the first identification of retraction failure (or the continued influence effect, as in the psychology literature) as distinct from other biases, such as confirmation bias. But just as substantial work has explored confirmation bias or base-rate neglect in particular domains, we believe it would be valuable for future applied work to determine particular real-world settings where retractions appear relatively more difficult to process. As mentioned in our review of the literature, a significant body of work on political behavior suggests that this class of settings features several factors which interfere with Bayesian reasoning. We mentioned that difficulties in constructing neutral comparisons with subjective information makes it less clear how one would show retractions are discounted more than they "should" be in these settings. Nevertheless, it may still be possible to explore whether different ways of framing retractions influence their relative effectiveness. Any such nuance, we believe, could yield insights which help elucidate the failure or success of retractions and fact-checking in practice.

Second, we have not fully explored the possible heterogeneity in the reactions to retractions. For instance, in ongoing work we examine whether beliefs are affected by changes in which information is checked, and how. If only information of a specific kind gets checked and retracted—e.g., only articles that challenge the scientific consensus get checked, only political statements supporting specific agendas—would retractions be less effective? Additionally, if only corrections are announced—as is the case in many circumstances—would people correctly infer when retractions render unretracted evidence more reliable?

On this note, our design is limited in how strongly it can address memory, or in the impact of the timing of retractions. The uniformity of our results is somewhat striking, but we also suspect that more targeted designs addressed on these questions may yield interesting and useful results. Following Bordalo et al. (2021), a natural question is whether relying on memory increases or decreases retraction effectiveness, as both retractions and information akin to the retracted one become more salient. Indeed, the psychology literature on the continued influence effect has devoted significant attention to explanations based on the processes behind memorization and memory retrieval. While it is interesting we obtain our results despite shutting this channel down, we do not claim it is unimportant. Understanding this is important insofar as it provides lessons for how to more effectively retract information, a question with a high degree of policy relevance.

References

Ambuehl, S. and S. Li (2018): "Belief Updating and the Demand for Information," *Games and Economic Behavior*, 109, 21–39. 7

Angelucci, C. and A. Prat (2020): "Measuring Voters' Knowledge of Political News," *Working Paper*. 7

Angrisani, M., A. Guarino, P. Jehiel, and T. Kitagawa (2019): "Information Redundancy Neglect versus Overconfodence: A Social Learning Experiment," *AEJ: Microeconomics*, Forthcoming. 10

Ayers, M. S. and L. M. Reder (1998): "A Theoretical Review of the Misinformation Effect: Predictions from an Activation-Based Memory model," *Psychological Bulletin & Review*, 5, 1–21.

1

Barrera, O., S. Guriev, E. Henry, and E. Zhuravskaya (2020): “Fake news, fact-checking and information in times of post-truth politics,” *Journal of Public Economics*, 182. 7

Benjamin, D. (2019): “Errors in Probabilistic Reasoning and Judgment Biases,” in *Handbook of Behavioral Economics*, ed. by B. D. Bernheim, S. DellaVigna, and D. Laibson, Elsevier Press. 4, 7, 9, 14, 24, 25

Bordalo, P., J. J. Conlon, N. Gennaioli, S. Y. Kwon, and A. Shleifer (2021): “Memory and Probability,” Working Paper 29273, National Bureau of Economic Research. 44

Borhani, F. and E. Green (2018): “Identifying the Occurrence or Non-Occurrence of Cognitive Bias in Situations Resembling the Monty Hall Problem,” *Working Paper.* 8

Charness, G. and C. Dave (2017): “Confirmation bias with motivated beliefs,” *Games and Economic Behavior*, 104, 1–23. 25

Charness, G. and D. Levin (2005): “When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect,” *American Economic Review*, 95, 1300–1309. 8

Charness, G., R. Oprea, and S. Yuksel (2020): “How Do People Choose Between Biased Information Sources? Evidence from a Laboratory Experiment.” *Journal of the European Economic Association*, Forthcoming. 8

Coutts, A. (2019): “Good news and bad news are still news: experimental evidence on belief updating,” *Experimental Economics*, 22, 369–395. 7

Cripps, M. (2021): “Divisible Updating,” *Working Paper.* 10

DeStefano, F. and T. Shimabukuro (2019): “The MMR Vaccine and Autism,” *Annual Review of Virology*, 6, 585–600. 1

Ecker, U. K. H., S. Lewandowsky, J. Cook, P. Schmid, L. K. Fazio, N. Brashier, P. Kendeou, E. K. Vraga, and M. A. Amazeen (2022): “The psychological drivers of misinformation belief and its resistance to correction,” *Nature Reviews Psychology*, 1, 13–29. 6

Enke, B. (2020): “What You See is All There Is,” *Quarterly Journal of Economics*, 135, 1363–1398. 8

Enke, B. and T. Graeber (2020): “Cognitive Uncertainty,” *Working Paper.* 8, 10, 20

Epstein, L. and Y. Halevy (2020): “Hard-to-Interpret Signals,” *Working Paper.* 11

Esponda, I. and E. Vespa (2014): “Hypothetical Thinking and Information Extraction in the Laboratory,” *AEJ: Microeconomics*, 6, 180–202. 8

Esponda, I., E. Vespa, and S. Yuksel (2020): “Mental Models and Learning: The Case of Base-Rate Neglect,” *Working Paper.* 8

Fein, S., A. L. McCloskey, and T. M. Tomlinson (1997): "Can the Jury Disregard that Information? The Use of Suspicion to Reduce the Prejudicial Effects of Pretrial Publicity and Inadmissible Testimony," *Personality and Social Psychology Bulletin*, 23, 1215–1226. 7

Friedman, D. (1998): "Monty Hall's Three Doors: Construction and Deconstruction of a Choice Anomaly," *American Economic Review*, 88, 933–946. 8

Grant, S., F. Hodge, and S. Seto (2021): "Can Prompting Investors to be in a Deliberative Mindset Reduce Their Reliance on Fake News?" *Working Paper.* 7

Grether, D. M. (1980): "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *The Quarterly Journal of Economics*, 95, 537–557. 24

Gupta, N., L. Rigotti, and A. Wilson (2021): "The Experimenters' Dilemma: Inferential Preferences over Populations," *Working Paper.* 41

Hossain, T. and R. Okui (2013): "The Binarized Scoring Rule," *Review of Economic Studies*, 80, 984–1001. 19

James, D., D. Friedman, C. Louie, and T. O'Meara (2018): "Dissecting the Monty Hall Anomaly," *Economic Inquiry*, 56, 1817–1826. 8

Johnson, H. M. and C. M. Seifert (1994): "Sources of the Continued Influence Effect: When Misinformation in Memory Affects later Influences," *Journal of Experimental Psychology*, 20, 1420–1436. 6

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47, 263–292. 8, 10

——— (1992): "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty*, 5, 297–323. 8, 10

Kassin, S. M. and S. R. Sommers (1997): "Inadmissible Testimony, Instructions to Disregard, and the Jury: Substantive Versus Procedural Considerations," *Personality and Social Psychology Bulletin*, 23, 1046–1054. 7

Kogan, S., T. J. Moskowitz, and M. Niessner (2021): "Social Media and Financial News Manipulation," *Working Paper.* 7

Lewandowsky, S., U. K. H. Ecker, C. M. Seifert, N. Schwarz, and J. Cook (2012): "Misinformation and Its Correction: Continued Influence and Successful Debiasing," *Psychological Science in the Public Interest*, 13, 106–131. 6

Liang, Y. (2020): "Learning from unknown information sources," *Working Paper.* 11

Martínez-Marquina, A., M. Niederle, and E. Vespa (2019): "Failures in Contingent Reasoning:

The Role of Uncertainty," *American Economic Review*, 109, 3437–3474. 8

McGranaghan, C., K. Nielsen, T. O'Donoghue, J. Somerville, and C. Sprenger (2022): "Distinguishing Common-Ratio Preferences from Common-Ratio Effects Using Paired Valuation Tasks," *Working Paper*. 10

Miller, J. and A. Sanjurjo (2019): "A Bridge from Monty Hall to the Hot Hand: The Principle of Restricted Choice," *Journal of Economic Perspectives*, 33, 144–162. 8, 9

Mobius, M. M., M. Niederle, P. Niehaus, and T. Rosenblat (2013): "Managing Self-Confidence: Theory and Experimental Evidence," *Working Paper*. 1, 19

National Consumer League (2014): "Survey: One third of American parents mistakenly link vaccines to autism," `https://nclnet.org/surveyonethirdofamericanparentsmistakenlylinkvaccinestoautism/`, accessed: 2021-06-16. 1

Nickerson, R. S. (1998): "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology*, 2, 175–220. 1

Nyhan, B. (2021): "Why the backfire effect does not explain the durability of political misperceptions," *Proceedings of the National Academy of Sciences*, 118. 7

Nyhan, B., E. Porter, J. Reifler, and T. Wood (2019): "Taking Fact-checks Literally But Not Seriously? The Effects of Journalistic Fact-checking on Factual Beliefs and Candidate Favorability," *Political Behavior*, forthcoming. 7

Nyhan, B. and J. Reifler (2010): "When corrections fail: The persistence of political misperceptions," *Political Behavior*, 32, 303–330. 7

Palacios-Huerta, I. (2003): "Learning to Open Monty Hall's Doors," *Experimental Economics*, 6, 235–251. 8

Prelec, D. (1998): "The Probability Weighting Function," *Econometrica*, 66, 497–527. 10

Rabin, M. and J. Schrag (1999): "First Impressions Matter: A Model of Confirmatory Bias," *Quarterly Journal of Economics*, 144, 37–82. 8

Shishkin, D. and P. Ortoleva (2021): "Ambiguous Information and Dilation: An Experiment," *Working Paper*. 11

Snowberg, E. and L. Yariv (2021): "Testing the Waters: Behavior across Participant Pools," *American Economic Review*, 111, 687–719. 41

Swire, B., A. J. Berinsky, S. Lewandowsky, and U. K. H. Ecker (2017): "Processing political misinformation: comprehending the Trump phenomenon," *Royal Society of Open Science*, 4. 7

Taber, C. S. and M. Lodge (2006): "Motivated Skepticism in the Evaluation of Political Beliefs," *American Journal of Political Science*, 50, 755–769. 7

Tan, H.-T. and S.-K. Tan (2009): "Investors' reactions to management disclosure corrections: Does presentation format matter?" *Contemporary Accounting Research*, 26, 605–626. 7

Tan, S.-K. and L. Koonce (2011): "Investors' reactions to retractions and corrections of management earnings forecasts," *Accounting, Organizations and Society*, 36, 382–397. 7

Thaler, M. (2020): "The "Fake News" Effect: Experimentally Identifying Motivated Reasoning Using Trust in News," *Working Paper.* 7

——— (2021): "Overinference from Weak Signals, Underinference from Strong Signals, and the Psychophysics of Interpreting Information," *Working Paper.* 7, 8, 10, 25

Thompson, W. C., G. T. Fong, and D. L. Rosenhan (1981): "Inadmissible evidence and juror verdicts," *Journal of Personality and Social Psychology*, 40, 453–463. 7

Woodford, M. (2020): "Modeling Imprecision in Perception, Valuation, and Choice," *Annual Review of Economics*, 579–601. 8

Zimmermann, F. (2020): "The Dynamics of Motivated Beliefs," *American Economic Review*, 110, 337–361. 1

Online Appendix for Learning versus Unlearning

A. OMITTED PROOFS

Proof of Proposition 1

Let $f : [0,1] \to [0,1]$ be a strictly increasing function and ℓ denote the logit function $\ell(p) = \log(p/(1-p))$. From Definition 1, $(\ell \circ f^{-1})(b(\theta \mid S_t)) = (\ell \circ f^{-1})(b(\theta)) + \sum_{j=1}^{t} K(s_j)$. Now, let $\alpha(\tau \mid S_t) = \frac{P(\text{Retraction of } s_\tau \mid S_t, \theta=1)}{P(\text{Retraction of } s_\tau \mid S_t, \theta=-1)}$. With symmetric noise, if signal s_τ is retracted, the Bayesian update should be

$$P(\theta \mid S_t, n_\tau = 1) = \frac{P(\theta) K(s_\tau)^{\eta_t - s_\tau} \alpha(\tau \mid S_t)}{P(-\theta) + P(\theta) K(s_\tau)^{\eta_t - s_\tau} \alpha(\tau \mid S_t)},$$

where $\eta_t := \sum_{\ell=1}^{t} s_\ell$. For a retraction, the log odds update of a Bayesian decisionmaker is therefore:

$$\ell\left(P(\theta \mid S_t, n_\tau = 1)\right) = \ell\left(P(\theta \mid S_t)\right) - K(s_\tau)\mathbf{1}[s_\tau \text{ retracted}] + \log(\alpha(\tau \mid S_t)). \tag{13}$$

Notice that $\alpha(\tau \mid S_t) = 1$, and hence $\log(\alpha(\tau \mid S_t)) = 0$, for all verifying retractions. Therefore, for any $\tau \in \{1, \ldots, t\}$,

$$\begin{aligned}
(\ell \circ f^{-1})(b(\theta \mid S_t, n_\tau = 1)) &= (\ell \circ f^{-1})(b(\theta \mid S_t)) - K(s_\tau) \\
&= (\ell \circ f^{-1})(b(\theta)) + \sum_{j \in \{1,\ldots,t\}} K(s_j) - K(s_\tau) \\
&= (\ell \circ f^{-1})(b(\theta)) + \sum_{j \in \{1,\ldots,t\} \setminus \tau} K(s_j) \\
&= (\ell \circ f^{-1})(b(\theta \mid S_t \setminus s_\tau)).
\end{aligned}$$

As $(\ell \circ f^{-1})$ is injective, then $b(\theta \mid S_t, n_\tau = 1) = b(\theta \mid S_t \setminus s_\tau)$.

If, moreover, $K(s_{t+1}) = -K(s_\tau)$, then it is immediate that $(\ell \circ f^{-1})(b(\theta \mid S_t, n_\tau = 1)) = (\ell \circ f^{-1})(b(\theta \mid S_t \cup s_{t+1}))$ and therefore $b(\theta \mid S_t, n_\tau = 1) = b(\theta \mid S_t \cup s_{t+1})$. □

B. TABLES AND FIGURES

B.1. Additional Comparisons of Belief Reports to Bayesian Posteriors

Figure 6: Distribution of reported beliefs

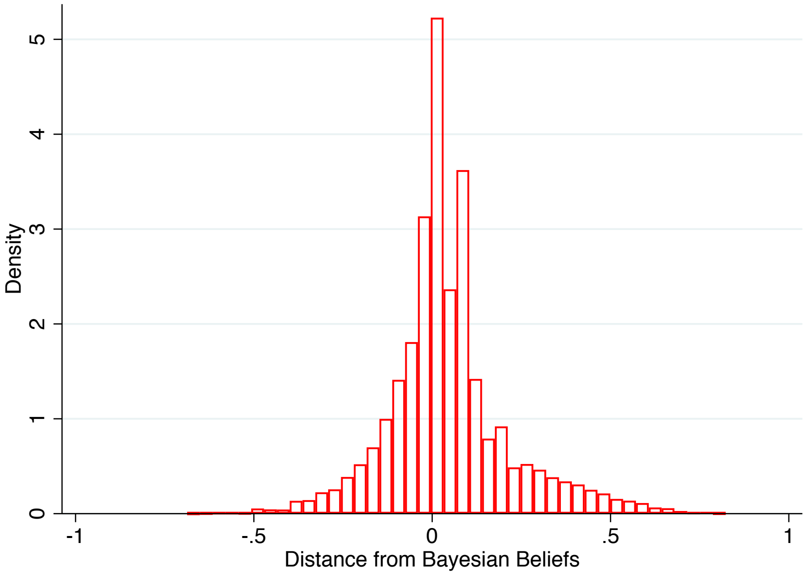
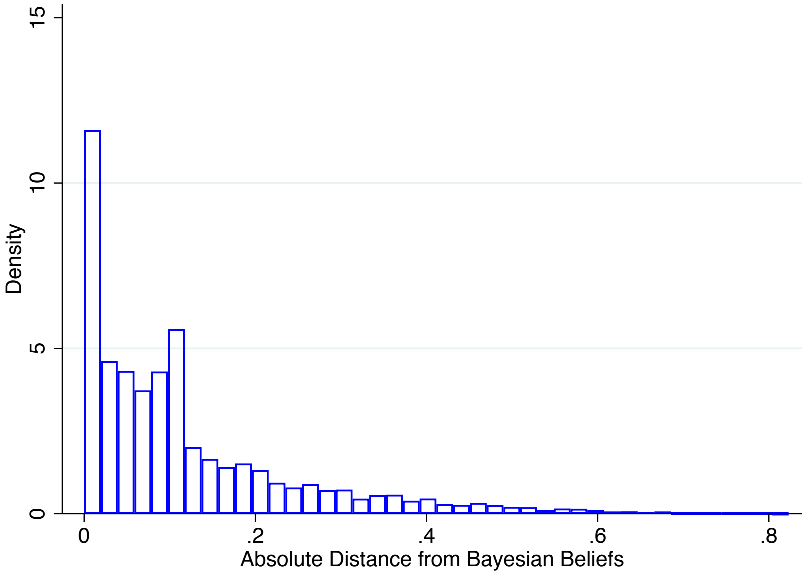


Figure 7: Distribution of reported beliefs



B.2. Sample Characteristics

	(1)	(2)	(3)
	All	Final Treatment	Baseline Treatment
Age	38.58	39.52	37.66
	(15.66)	(19.84)	(10.06)
Female	0.40	0.40	0.40
High School	0.11	0.11	0.10
College	0.21	0.23	0.19
Bachelor's or equivalent	0.45	0.41	0.49
Postgrad or equivalent	0.18	0.19	0.18
Other Education level	0.05	0.05	0.04
Answered all numeracy questions correctly	0.57	0.55	0.59
Total score on numeracy measures	1.74	1.79	1.69
	(1.01)	(1.00)	(1.02)

Table 11: Sample Characteristics

Notes: This table provides a comparison of the socio-demographic characteristics of the subjects in our sample. Column (1) considers all subjects and columns (2) and (3) provide summary statics by treatment.

B.3. Tables 1 and 3 with Probability Weights

In contrast to Tables 1 and 3, Tables 12 and 13 weight observations so as to make histories equally likely. We note that our conclusions remain the same.

Our weighting strategy is as follows: First, we create fine-grained groups of histories, excluding cases in which the state (the truth ball) is revealed. Two histories H_t and $H'_{t'}$ are treated as the same history $W(H_t)$ if (1) they refer to the same period, $t = t'$; (2) prior to the current period's signal, the absolute sum of signals (treating retractions as signals of opposing sign) was the same; (3) given the current period's signal, the absolute sum of signals is also the same; and (4) if a retraction occurs in period t in history H_t, then the same is true for history H'_t—and vice-versa. Conditions (2) and (3) induce Bayesian posteriors that are symmetric around .5 before and after the current period's signal. That is, $YBBY$, $BYYB$, and $BYBY$ are treated as the same history; this is so that we do not overweight such histories relative to, for example, $YYYY$. Our weights correspond to the inverse probability of history $W(H_t)$ being observed in period t. This procedure also ensures that each period is made equally likely.

	(1)	(2)
	l_t	l_t
Prior (l_{t-1})	0.875^{***}	0.829^{***}
	(0.026)	(0.030)
Signal (s_t)	1.219^{***}	0.963^{***}
	(0.034)	(0.043)
Signal Confirms Prior ($s_t \cdot c_t$)	–	0.653^{***}
		(0.091)
R-Squared	0.48	0.48
Observations	18491	18491

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 12: Updating from New Draws

Notes: This table represents updating from standard new ball draws. The sample consists of subjects in the baseline treatment (beliefs are elicited each period). We include beliefs of all periods (1-4) but, within a given round, we exclude any beliefs which are elicited after the truth ball is disclosed. Thus, for example, if there is a retraction in period 3, we exclude beliefs in both period 3 and 4. The regressions correspond to equations (5) and (6). Inverse probability weights are used to make each history equally likely. The outcome is the log odds of beliefs in period t, l_t. s_t is the signal in round t (+1 or -1, multiplied by K, a constant factor of Bayesian updating, such that the coefficient on s_t would be 1 under Bayesian updating), $c_t := \mathbf{1}\{\text{sign}(l_{t-1}) = \text{sign}(s_t)\}$ is an indicator function that equals 1 when the signal at t confirms the prior at $t-1$.

	All Periods		Period 3	
	(1)	(2)	(3)	(4)
	l_t	l_t	l_t	l_t
Prior (l_{t-1})	0.875***	0.829***	0.924***	0.824***
	(0.026)	(0.030)	(0.043)	(0.054)
Signal (s_t)	1.219***	0.963***	1.713***	1.195***
	(0.034)	(0.043)	(0.077)	(0.111)
Signal Confirms Prior ($s_t \cdot c_t$)	–	0.653***	–	1.100***
		(0.091)		(0.218)
Retraction (r_t)	-0.050**	-0.044*	-0.039	-0.027
	(0.024)	(0.024)	(0.036)	(0.036)
Retraction x Prior ($r_t \cdot l_{t-1}$)	-0.002	0.056	-0.114*	0.013
	(0.036)	(0.042)	(0.058)	(0.072)
Retracted Signal ($r_t \cdot s_t$)	-0.833***	-0.497***	-1.346***	-0.700***
	(0.060)	(0.080)	(0.092)	(0.134)
Retraction x Signal Confirms Prior ($r_t \cdot s_t \cdot c_t$)	–	-0.853***	–	-1.444***
		(0.155)		(0.262)
R-Squared	0.51	0.51	0.42	0.43
Observations	22578	22578	6081	6081

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 13: How Retractions Interact with Other Biases (Hypothesis 2)

Notes: The sample consists of subjects in the baseline treatment (beliefs are elicited each period). The regressions correspond to equation (9), using data from period 3. The sample excludes the cases where the state of the world is fully disclosed. Inverse probability weights are used to make each history equally likely. The outcome is the log odds of beliefs in period t, l_t. s_t is the signal in round t (+1 or -1, multiplied by K, a constant factor of Bayesian updating, such that the coefficient on s_t would be 1 under Bayesian updating), r_t is an indicator variable for whether the signal in period t can from a retraction, $c_t := \mathbf{1}\{\text{sign}(l_{t-1}) = \text{sign}(s_t)\}$ is an indicator function that equals 1 when the signal at t confirms the prior at $t-1$.

B.4. Log Odds

In this section, we re-estimate the main specifications in the paper, but using log-odds beliefs as the dependent variable instead.

	Prior vs. Retraction	Retraction vs. New Draw		
	(1)	(2)	(3)	(4)
	l_t	l_t	Δl_t	Δl_t
Retraction (r_t)	0.005	-0.015	-0.044*	-0.043
	(0.023)	(0.034)	(0.025)	(0.029)
Retracted Signal ($r_t \cdot s_t$)	-0.233***	-0.235***	-0.249***	-0.276***
	(0.039)	(0.051)	(0.046)	(0.048)
Signal (s_t)	–	–	–	0.668***
				(0.055)
Compressed History FEs	Yes	No	No	No
Sign History FEs	No	Yes	Yes	No
Lagged Sign History FEs	No	No	No	Yes
R-Squared	0.18	0.30	0.15	0.14
Observations	17591	9074	9074	9074

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 14: Updating from Retractions: do they work and how they compare to equivalent new signals (Hypothesis 1); Log-odds

Notes: This table provides the main specification of interest in this paper. It tests whether retractions are effective in inducing ‘un-learning’ and compares their effectiveness relative to new direct information. The sample consists of subjects in the baseline treatment (beliefs are elicited each period). Column (1) tests whether retractions work, by comparing beliefs after a retraction to beliefs after the equivalent compressed history. We include beliefs of all periods (1-4) but, within a given round, we exclude any beliefs which are elicited after the truth ball is disclosed and we exclude period 4 beliefs if there was a retraction in period 3. In periods 3 and 4 we only include beliefs when there was a retraction in that period. The outcome is the log-odds beliefs in period t, $l_t \in \mathbb{R}$. $r_t \cdot s_t$ is the opposite sign of the retracted signal in round t (+1 if a -1 signal is retracted, -1 if a +1 signal is retracted). The regression includes fixed effects for the compressed history of draws. Columns (2) to (4) test whether people update more or less from retractions compared to equivalent new signals. The sample is restricted to beliefs in periods 3 and 4, once again dropping beliefs after the truth ball is disclosed or in period 4 if there is a retraction in period 3. The specifications include fixed effects for the sign history. In column (2), the outcome is the beliefs in period t, b_t. In columns (3) and (4), the outcome is the first difference in beliefs. Column (4) uses lagged sign history fixed effects to enable us to compare the magnitude of $r_t \cdot s_t$ to s_t.

B.5. Robustness Checks

B.5.1. Comprehension Questions Correct at First Try

In this section, we re-estimate the main specifications in the paper, but restricting to subjects who correctly answered all the comprehension questions at first try.

Figure 8 shows the proportion of subjects who successfully answered the comprehension questionnaire by the n-th try and compares this with the case in which they would be choosing uniformly at random. In particular, we take the case of a sophisticated randomizer that understands which questions were incorrect and only randomizes among the ones that were not revealed incorrect.

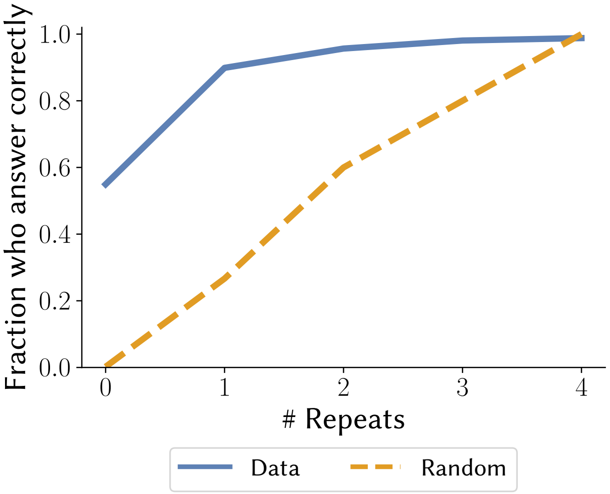


Figure 8: Comprehension Questions

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

	Prior vs. Retraction	Retraction vs. New Draw		
	(1)	(2)	(3)	(4)
	b$_t$	b$_t$	Δb$_t$	Δb$_t$
Retraction (r_t)	0.276	-0.003	0.071	0.162
	(0.305)	(0.331)	(0.349)	(0.362)
Retracted Signal ($r_t \cdot s_t$)	-3.522***	-4.129***	-3.779***	-3.549***
	(0.724)	(0.779)	(0.763)	(0.760)
Signal (s_t)	–	–	–	9.712***
				(0.502)
Compressed History FEs	Yes	No	No	No
Sign History FEs	No	Yes	Yes	No
Lagged Sign History FEs	No	No	No	Yes
R-Squared	0.39	0.53	0.26	0.26
Observations	10592	5446	5446	5446

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 15: Updating from Retractions: do they work and how they compare to equivalent new signals (Hypothesis 1); Robustness Check 1

Notes: This table provides the main specification of interest in this paper. It tests whether retractions are effective in inducing 'un-learning' and compares their effectiveness relative to new direct information. The sample consists of subjects in the baseline treatment (beliefs are elicited each period) who correctly answered all comprehension questions at first try. Column (1) tests whether retractions work, by comparing beliefs after a retraction to beliefs after the equivalent compressed history. We include beliefs of all periods (1-4) but, within a given round, we exclude any beliefs which are elicited after the truth ball is disclosed and we exclude period 4 beliefs if there was a retraction in period 3. In periods 3 and 4 we only include beliefs when there was a retraction in that period. The outcome is the beliefs in period t, $b_t \in [0, 100]$. $r_t \cdot s_t$ is the opposite sign of the retracted signal in round t (+1 if a -1 signal is retracted, -1 if a +1 signal is retracted). The regression includes fixed effects for the compressed history of draws. Columns (2) to (4) test whether people update more or less from retractions compared to equivalent new signals. The sample is restricted to beliefs in periods 3 and 4, once again dropping beliefs after the truth ball is disclosed or in period 4 if there is a retraction in period 3. The specifications include fixed effects for the sign history. In column (2), the outcome is the beliefs in period t, b_t. In columns (3) and (4), the outcome is the first difference in beliefs. Column (4) uses lagged sign history fixed effects to enable us to compare the magnitude of $r_t \cdot s_t$ to s_t.

	Retraction vs. New Draw	
	(1)	(2)
	dt_t	$\log(\text{dt}_t)$
Retraction (r_t)	0.620^{***}	0.114^{***}
	(0.103)	(0.016)
Mean Decision Time (secs)	5.3	1.507
Sign History FEs	Yes	Yes
R-Squared	0.02	0.02
Observations	5400	5400

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 16: Decision Times in Updating from Retractions (Hypothesis 6); Robustness Check 1

Notes: This table tests whether the time taken to report beliefs is different after retractions compared to equivalent new signals (Hypothesis 6). The sample consists of subjects in the baseline treatment (beliefs are elicited each period) who correctly answered all comprehension questions at first try. The specifications compare updating from retractions versus from an equivalent new signal in periods 3 and 4 (we drop period 4 if there was a retraction in period 3). We drop the top 1% of decision times from the sample.

B.5.2. Correct Belief Reports when Truth Ball is Revealed

In this section, we re-estimate the main specifications in the paper, not only restricting to subjects who correctly answered all the comprehension questions at first try, but further remove from the sample subjects who failed to correctly report beliefs close to 0 or 100 when the truth ball was revealed. In particular, we remove from the sample any subject who, when state θ is revealed, failed to report beliefs $|b_t - \theta| \leq \epsilon$. We present the results for $\epsilon = .05$, but the results are robust to the choice of small ϵ.

	Prior vs. Retraction	Retraction vs. New Draw		
	(1)	(2)	(3)	(4)
	b_t	b_t	Δb_t	Δb_t
Retraction (r_t)	0.174	-0.399	-0.069	-0.010
	(0.327)	(0.352)	(0.320)	(0.306)
Retracted Signal ($r_t \cdot s_t$)	-3.647***	-3.897***	-3.671***	-3.526***
	(0.762)	(0.816)	(0.784)	(0.791)
Signal (s_t)	–	–	–	10.434***
				(0.476)
Compressed History FEs	Yes	No	No	No
Sign History FEs	No	Yes	Yes	No
Lagged Sign History FEs	No	No	No	Yes
R-Squared	0.52	0.67	0.45	0.45
Observations	8001	4119	4119	4119

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 17: Updating from Retractions: do they work and how they compare to equivalent new signals (Hypothesis 1); Robustness Check 2

Notes: This table provides the main specification of interest in this paper. It tests whether retractions are effective in inducing 'un-learning' and compares their effectiveness relative to new direct information. The sample consists of subjects in the baseline treatment (beliefs are elicited each period); we removed from the sample all subjects who did not correctly answer all comprehension questions at first try or who did not correctly report beliefs when the state was revealed. Column (1) tests whether retractions work, by comparing beliefs after a retraction to beliefs after the equivalent compressed history. We include beliefs of all periods (1-4) but, within a given round, we exclude any beliefs which are elicited after the truth ball is disclosed and we exclude period 4 beliefs if there was a retraction in period 3. In periods 3 and 4 we only include beliefs when there was a retraction in that period. The outcome is the beliefs in period t, $b_t \in [0, 100]$. $r_t \cdot s_t$ is the opposite sign of the retracted signal in round t (+1 if a -1 signal is retracted, -1 if a +1 signal is retracted). The regression includes fixed effects for the compressed history of draws. Columns (2) to (4) test whether people update more or less from retractions compared to equivalent new signals. The sample is restricted to beliefs in periods 3 and 4, once again dropping beliefs after the truth ball is disclosed or in period 4 if there is a retraction in period 3. The specifications include fixed effects for the sign history. In column (2), the outcome is the beliefs in period t, b_t. In columns (3) and (4), the outcome is the first difference in beliefs. Column (4) uses lagged sign history fixed effects to enable us to compare the magnitude of $r_t \cdot s_t$ to s_t.

	Retraction vs. New Draw	
	(1)	(2)
	dt_t	$\log(\text{dt}_t)$
Retraction (r_t)	0.587***	0.116***
	(0.109)	(0.018)
Mean Decision Time (secs)	5.09	1.481
Sign History FEs	Yes	Yes
R-Squared	0.02	0.03
Observations	4080	4080

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 18: Decision Times in Updating from Retractions (Hypothesis 6); Robustness Check 2

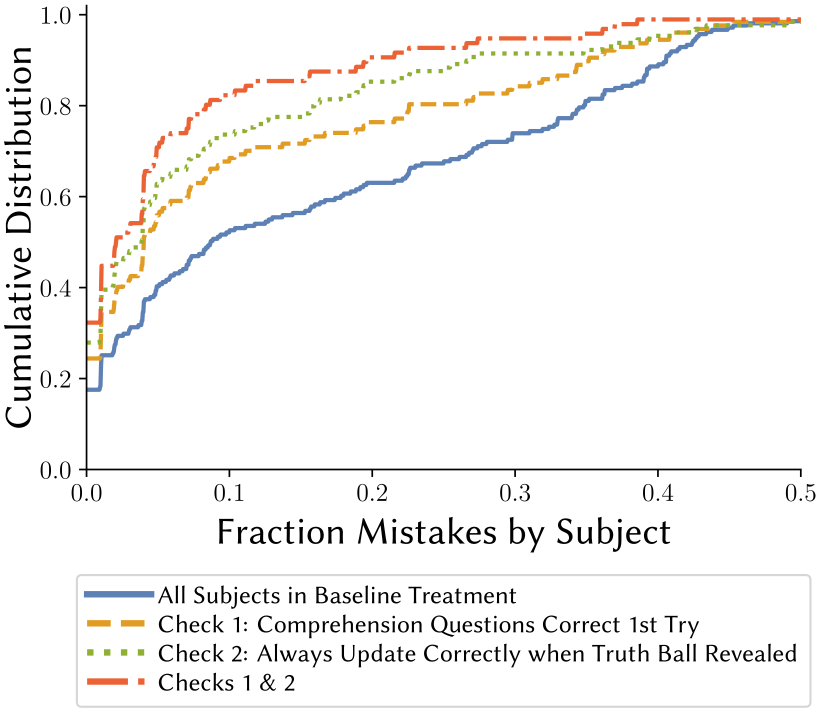
Notes: This table tests whether the time taken to report beliefs is different after retractions compared to equivalent new signals (Hypothesis 6). The sample consists of subjects in the baseline treatment (beliefs are elicited each period); we removed from the sample all subjects who did not correctly answer all comprehension questions at first try or who did not correctly report beliefs when the state was revealed. The specifications compare updating from retractions versus from an equivalent new signal in periods 3 and 4 (we drop period 4 if there was a retraction in period 3). We drop the top 1% of decision times from the sample.

B.5.3. Noisy Belief Reports

In this section, we re-estimate the main specifications in the paper, removing subjects who seem to be answering randomly.

For this purpose, we consider that the subject makes a mistake when they update their beliefs in the opposite direction to the signal, i.e. $(b_t - b_{t-1}) \cdot s_t < 0$. For each subject, we compute updating mistakes as a fraction of the total number of belief elicitations and, in Figure 9, we show the distribution of the individual-level mistakes by robustness check. It is immediate that the previous robustness checks do reduce the fraction of subjects who seem to be answering randomly.

Figure 9: Subject-Level Mistakes



In the following tables, we remove from the sample any subject who updates their beliefs in the opposite direction to the signal more than $x\%$ of the time. We present the results for $x = 10\%$; the coefficients are virtually unchanged when considering 5% and 25% instead.

	Prior vs. Retraction	Retraction vs. New Draw		
	(1)	(2)	(3)	(4)
	b_t	b_t	$\Delta\mathrm{b}_t$	$\Delta\mathrm{b}_t$
Retraction (r_t)	-0.586**	-0.730**	-0.692**	-0.457
	(0.286)	(0.309)	(0.284)	(0.279)
Retracted Signal ($r_t \cdot s_t$)	-5.105***	-5.912***	-5.748***	-5.360***
	(0.725)	(0.803)	(0.761)	(0.766)
Signal (s_t)	–	–	–	11.896***
				(0.404)
Compressed History FEs	Yes	No	No	No
Sign History FEs	No	Yes	Yes	No
Lagged Sign History FEs	No	No	No	Yes
R-Squared	0.60	0.74	0.47	0.46
Observations	9179	4732	4732	4732

Clustered standard errors at the subject level in parentheses

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

Table 19: Updating from Retractions: do they work and how they compare to equivalent new signals (Hypothesis 1); Robustness Check 3

Notes: This table provides the main specification of interest in this paper. It tests whether retractions are effective in inducing ‘un-learning’ and compares their effectiveness relative to new direct information. The sample consists of subjects in the baseline treatment (beliefs are elicited each period); we removed from the sample all subjects who made mistakes in more than 10% of the periods. Column (1) tests whether retractions work, by comparing beliefs after a retraction to beliefs after the equivalent compressed history. We include beliefs of all periods (1-4) but, within a given round, we exclude any beliefs which are elicited after the truth ball is disclosed and we exclude period 4 beliefs if there was a retraction in period 3. In periods 3 and 4 we only include beliefs when there was a retraction in that period. The outcome is the beliefs in period t, $b_t \in [0, 100]$. $r_t \cdot s_t$ is the opposite sign of the retracted signal in round t (+1 if a -1 signal is retracted, -1 if a +1 signal is retracted). The regression includes fixed effects for the compressed history of draws. Columns (2) to (4) test whether people update more or less from retractions compared to equivalent new signals. The sample is restricted to beliefs in periods 3 and 4, once again dropping beliefs after the truth ball is disclosed or in period 4 if there is a retraction in period 3. The specifications include fixed effects for the sign history. In column (2), the outcome is the beliefs in period t, b_t. In columns (3) and (4), the outcome is the first difference in beliefs. Column (4) uses lagged sign history fixed effects to enable us to compare the magnitude of $r_t \cdot s_t$ to s_t.

B.6. Heterogeneous Effects

In this section, we test for the existence of heterogeneous treatment effects relative to the subjects quantitative ability. In the last part of our experiment, we posed three multiple-choice quantitative questions. The median number of correct answers per subject was two. We expand our specifications by interacting all the regressors with a dummy variable that equals 1 if the subject answered all questions correctly and 0 if otherwise.

	Prior vs. Retraction	Retraction vs. New Draw		
	(1)	(2)	(3)	(4)
	b_t	b_t	Δb_t	Δb_t
Retraction (r_t)	0.414	-0.166	-0.434	-0.309
	(0.349)	(0.484)	(0.467)	(0.484)
Retracted Signal ($r_t \cdot s_t$)	-2.958***	-3.483***	-4.294***	-3.038***
	(0.779)	(0.887)	(0.796)	(0.890)
Signal (s_t)	–	–	–	7.821***
				(0.659)
All correct	-1.051**	-1.895***	-0.590*	-0.586*
	(0.415)	(0.600)	(0.340)	(0.340)
Retraction (r_t) x All correct	-0.753	-0.026	0.306	0.257
	(0.504)	(0.629)	(0.603)	(0.600)
Retracted signal ($r_t \cdot s_t$) x All correct	-0.644	-0.535	2.129**	-0.975
	(1.098)	(1.096)	(0.984)	(1.224)
Signal (s_t) x All correct	–	–	–	3.033***
				(0.868)
Compressed History FEs	Yes	No	No	No
Sign History FEs	No	Yes	Yes	No
Lagged Sign History FEs	No	No	No	Yes
R-Squared	0.29	0.39	0.16	0.16
Observations	17591	9074	9074	9074

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 20: Updating from Retractions: do they work and how they compare to equivalent new signals (Hypothesis 1); Heterogeneous Effects

Notes: This table provides the main specification of interest in this paper. It tests whether retractions are effective in inducing 'un-learning' and compares their effectiveness relative to new direct information. 'All correct' denotes a dummy variable that equals 1 when the subject answered all quantitative questions correctly, and 0 if otherwise. We investigate the existence of heterogeneous effects with respect to quantitative ability by expanding our baseline specifications from Table 2 with interaction terms. The sample consists of subjects in the baseline treatment (beliefs are elicited each period). Column (1) tests whether retractions work, by comparing beliefs after a retraction to beliefs after the equivalent compressed history. We include beliefs of all periods (1-4) but, within a given round, we exclude any beliefs which are elicited after the truth ball is disclosed and we exclude period 4 beliefs if there was a retraction in period 3. In periods 3 and 4 we only include beliefs when there was a retraction in that period. The outcome is the beliefs in period t, $b_t \in [0, 100]$. $r_t \cdot s_t$ is the opposite sign of the retracted signal in round t (+1 if a -1 signal is retracted, -1 if a +1 signal is retracted). The regression includes fixed effects for the compressed history of draws. Columns (2) to (4) test whether people update more or less from retractions compared to equivalent new signals. The sample is restricted to beliefs in periods 3 and 4, once again dropping beliefs after the truth ball is disclosed or in period 4 if there is a retraction in period 3. The specifications include fixed effects for the sign history. In column (2), the outcome is the beliefs in period t, b_t. In columns (3) and (4), the outcome is the first difference in beliefs. Column (4) uses lagged sign history fixed effects to enable us to compare the magnitude of $r_t \cdot s_t$ to s_t.

	Retraction vs. New Draw	
	(1)	(2)
	dt_t	$\log(\text{dt}_t)$
Retraction (r_t)	0.423***	0.090***
	(0.101)	(0.017)
All correct	0.119	0.053
	(0.331)	(0.059)
Retraction (r_t) x All correct	0.280	0.039
	(0.192)	(0.028)
Mean Decision Time (secs)	5.557	1.547
Sign History FEs	Yes	Yes
R-Squared	0.01	0.02
Observations	8983	8983

Clustered standard errors at the subject level in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 21: Decision Times in Updating from Retractions (Hypothesis 6); Heterogeneous Effects

Notes: This table tests whether the time taken to report beliefs is different after retractions compared to equivalent new signals (Hypothesis 6). ‘All correct’ denotes a dummy variable that equals 1 when the subject answered all quantitative questions correctly, and 0 if otherwise. We investigate the existence of heterogeneous effects with respect to quantitative ability by expanding our baseline specifications from Table 7 with interaction terms. The sample consists of subjects in the baseline treatment (beliefs are elicited each period). The specifications compare updating from retractions versus from an equivalent new signal in periods 3 and 4 (we drop period 4 if there was a retraction in period 3). We drop the top 1% of decision times from the sample.

B.7. Subject-Level Estimates

	Prior vs. Retraction	Retraction vs. New Draw		
	(1)	(2)	(3)	(4)
	b_t	b_t	$\Delta \mathrm{b}_t$	$\Delta \mathrm{b}_t$
Mean	-3.285	-3.740	-3.892	-3.448
Median	-2.103	-2.602	-2.797	-3.236
Fraction Coeff $<$ 0	0.69	0.72	0.72	0.72
Mean std error	2.696	4.446	5.044	4.777
Compressed History FEs	Yes	No	No	No
Sign History FEs	No	Yes	Yes	No
Lagged Sign History FEs	No	No	No	Yes

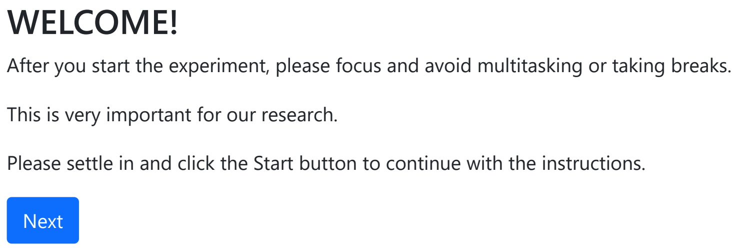
Table 22: Updating from Retractions: do they work and how they compare to equivalent new signals (Hypothesis 1); Subject-Level Estimates

Notes: This table provides summary statistics on distribution of subject-level estimates of the coefficient of interest in the main specification of interest in this paper. We investigate the existence of individual-level heterogeneity by estimating the specifications in Table 2 for each subject. The sample consists of subjects in the baseline treatment (beliefs are elicited each period) who observed at least 8 retractions in the 32 rounds. Column (1) tests whether retractions work, by comparing beliefs after a retraction to beliefs after the equivalent compressed history. We include beliefs of all periods (1-4) but, within a given round, we exclude any beliefs which are elicited after the truth ball is disclosed and we exclude period 4 beliefs if there was a retraction in period 3. In periods 3 and 4 we only include beliefs when there was a retraction in that period. The outcome is the beliefs in period t, $b_t \in [0, 100]$. $r_t \cdot s_t$ is the opposite sign of the retracted signal in round t (+1 if a -1 signal is retracted, -1 if a +1 signal is retracted). The regression includes fixed effects for the compressed history of draws. Columns (2) to (4) test whether people update more or less from retractions compared to equivalent new signals. The sample is restricted to beliefs in periods 3 and 4, once again dropping beliefs after the truth ball is disclosed or in period 4 if there is a retraction in period 3. The specifications include fixed effects for the sign history. In column (2), the outcome is the beliefs in period t, b_t. In columns (3) and (4), the outcome is the first difference in beliefs. Column (4) uses lagged sign history fixed effects to enable us to compare the magnitude of $r_t \cdot s_t$ to s_t.

C. INSTRUCTIONS AND SCREENSHOTS

C.1. Start Screen and Instructions

Below are screenshots of the start screen and the instructions as presented to the subjects.

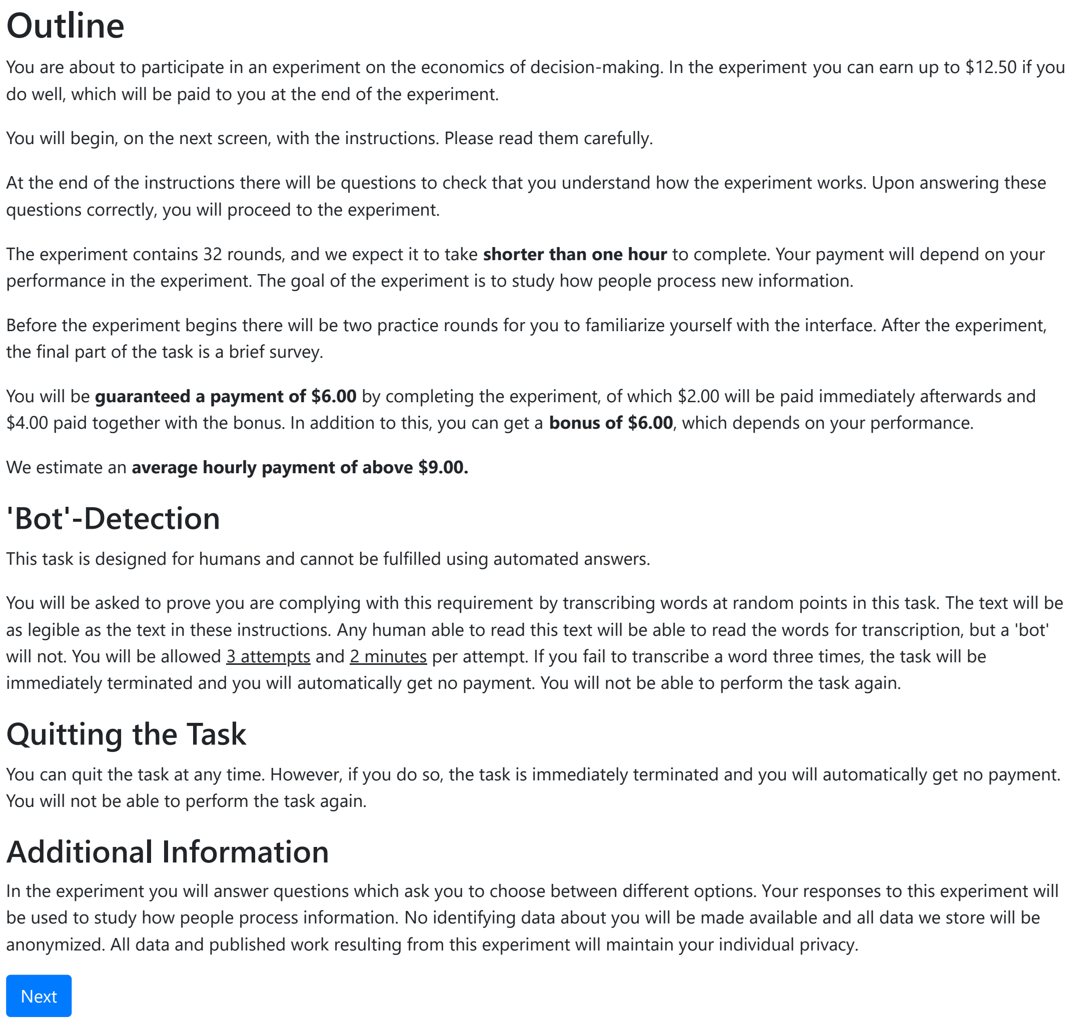


WELCOME!

After you start the experiment, please focus and avoid multitasking or taking breaks.

This is very important for our research.

Please settle in and click the Start button to continue with the instructions.



Outline

You are about to participate in an experiment on the economics of decision-making. In the experiment you can earn up to $12.50 if you do well, which will be paid to you at the end of the experiment.

You will begin, on the next screen, with the instructions. Please read them carefully.

At the end of the instructions there will be questions to check that you understand how the experiment works. Upon answering these questions correctly, you will proceed to the experiment.

The experiment contains 32 rounds, and we expect it to take **shorter than one hour** to complete. Your payment will depend on your performance in the experiment. The goal of the experiment is to study how people process new information.

Before the experiment begins there will be two practice rounds for you to familiarize yourself with the interface. After the experiment, the final part of the task is a brief survey.

You will be **guaranteed a payment of $6.00** by completing the experiment, of which $2.00 will be paid immediately afterwards and $4.00 paid together with the bonus. In addition to this, you can get a **bonus of $6.00**, which depends on your performance.

We estimate an **average hourly payment of above $9.00.**

'Bot'-Detection

This task is designed for humans and cannot be fulfilled using automated answers.

You will be asked to prove you are complying with this requirement by transcribing words at random points in this task. The text will be as legible as the text in these instructions. Any human able to read this text will be able to read the words for transcription, but a 'bot' will not. You will be allowed 3 attempts and 2 minutes per attempt. If you fail to transcribe a word three times, the task will be immediately terminated and you will automatically get no payment. You will not be able to perform the task again.

Quitting the Task

You can quit the task at any time. However, if you do so, the task is immediately terminated and you will automatically get no payment. You will not be able to perform the task again.

Additional Information

In the experiment you will answer questions which ask you to choose between different options. Your responses to this experiment will be used to study how people process information. No identifying data about you will be made available and all data we store will be anonymized. All data and published work resulting from this experiment will maintain your individual privacy.

Instructions

Welcome!

In the experiment you will be asked to estimate the probability that a given ball in a box is blue or yellow.
The experiment is divided into 32 rounds, each round with up to 4 periods, plus two practice rounds before you start for you to get familiar with the interface.
We expect the overall experiment to last for less than 1 hour, although you are free to move at your own pace.
We also expect that, with an adequate amount of effort, participants get on average $9.00, of which $6.00 depends only on completing the task.

Truth Balls and Noise Balls

At the beginning of each round, 5 balls are put inside a box.
The balls in that box are of two kinds:

- 4 Noise Balls Ⓝ, of which 2 are yellow Ⓝ and 2 are blue Ⓝ; and
- 1 Truth Ball Ⓣ, which can be either yellow Ⓣ or blue Ⓣ.

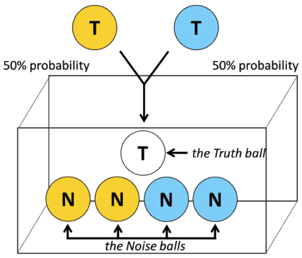
Your task is to estimate the probability that the Truth Ball Ⓣ is yellow Ⓣ or blue Ⓣ, upon observing random draws from the selected box in each round.

Your Task

A Round

At the beginning of each round, the Truth Ball Ⓣ is chosen to be either Ⓣ or Ⓣ with equal probability.
The Truth Ball Ⓣ is then put inside the box with all 4 Noise Balls, 2 Ⓝ and 2 Ⓝ
.
All balls remain inside the box throughout the round.

The round lasts for 4 periods, each of which may help you to guess the color of the Truth Ball Ⓣ.



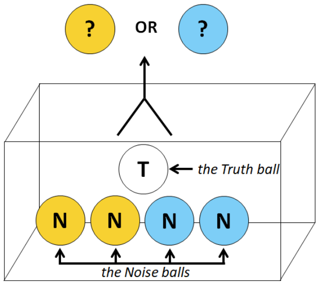
Note that the Truth Ball remains the same throughout the round but changes across different rounds.

This means that the draws you observe from a particular round are not helpful to estimate the color of a Truth Ball in another round and every round you need to start afresh.

Periods 1 and 2

In periods 1 and 2, a ball is drawn from the box at random and you are told its color, 🟡 or 🔵.
The ball is then placed back into the box.
You will not be told whether it is a Noise Ball Ⓝ or the Truth Ball Ⓣ. Because of this, the ball will be labelled with a question mark ?.
Since the balls are drawn at random, the drawn ball ?:

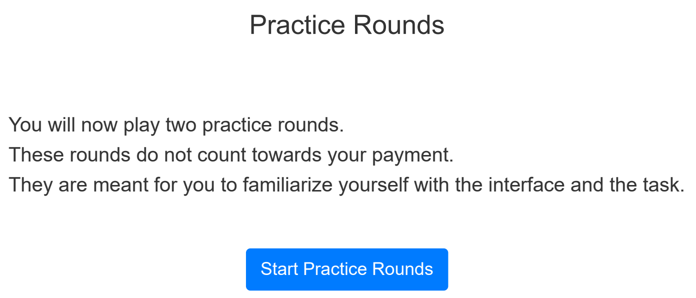
- is the Truth Ball Ⓣ with 20% probability;
- is a Noise Ball Ⓝ with 80% probability.



Naturally, the more draws you observe, the more likely that one of them is the Truth Ball, and the more balls of one color you observe, the more likely it is that the Truth Ball is of that color. However, because in each period the ball you are shown is placed back into the box, it can be that you are shown the Truth Ball multiple times or even that you are only shown Noise Balls.

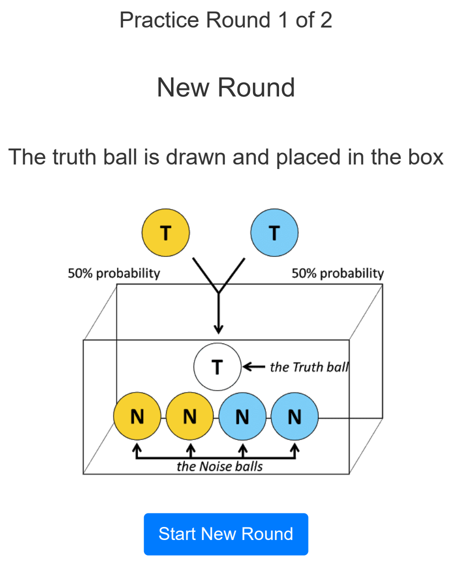
C.2. Practice Round

Subjects played had two practice rounds before starting the task. It was explicitly mentioned that these would not count toward their payment.



Practice Rounds

You will now play two practice rounds.
These rounds do not count towards your payment.
They are meant for you to familiarize yourself with the interface and the task.

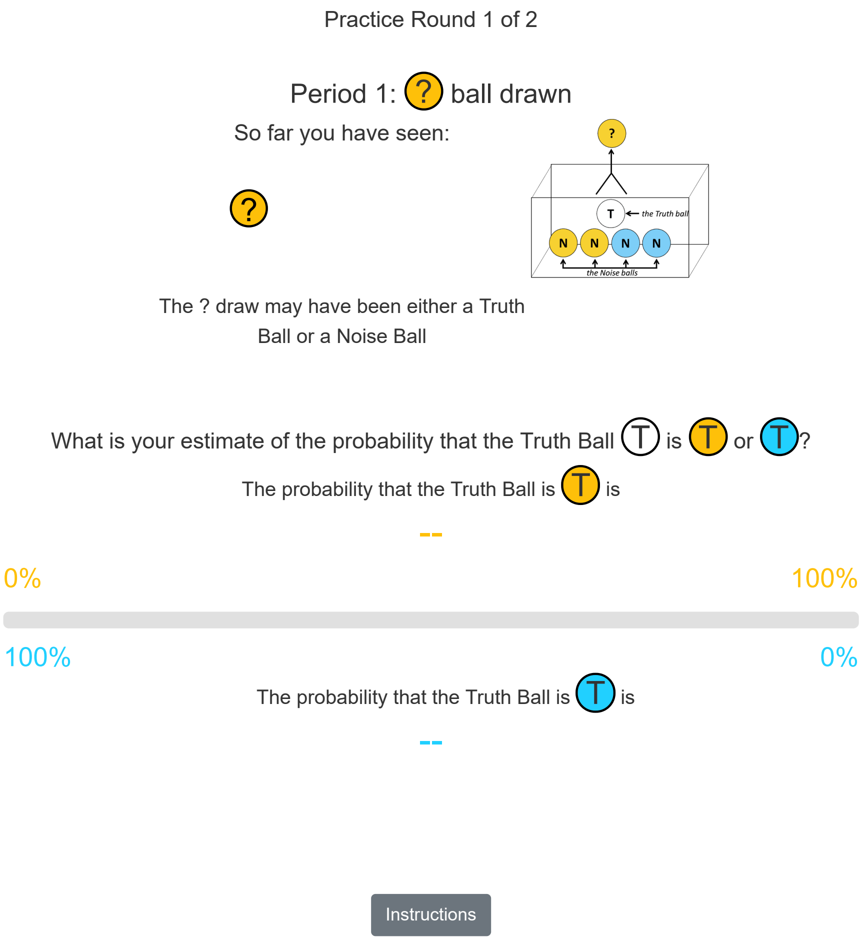


Practice Round 1 of 2

New Round

The truth ball is drawn and placed in the box

One the page loaded, the slider was blank and only activated once the subjects clicked on it.



Practice Round 1 of 2

Period 1: ? ball drawn

So far you have seen:

?

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

What is your estimate of the probability that the Truth Ball T is T or T?

The probability that the Truth Ball is T is

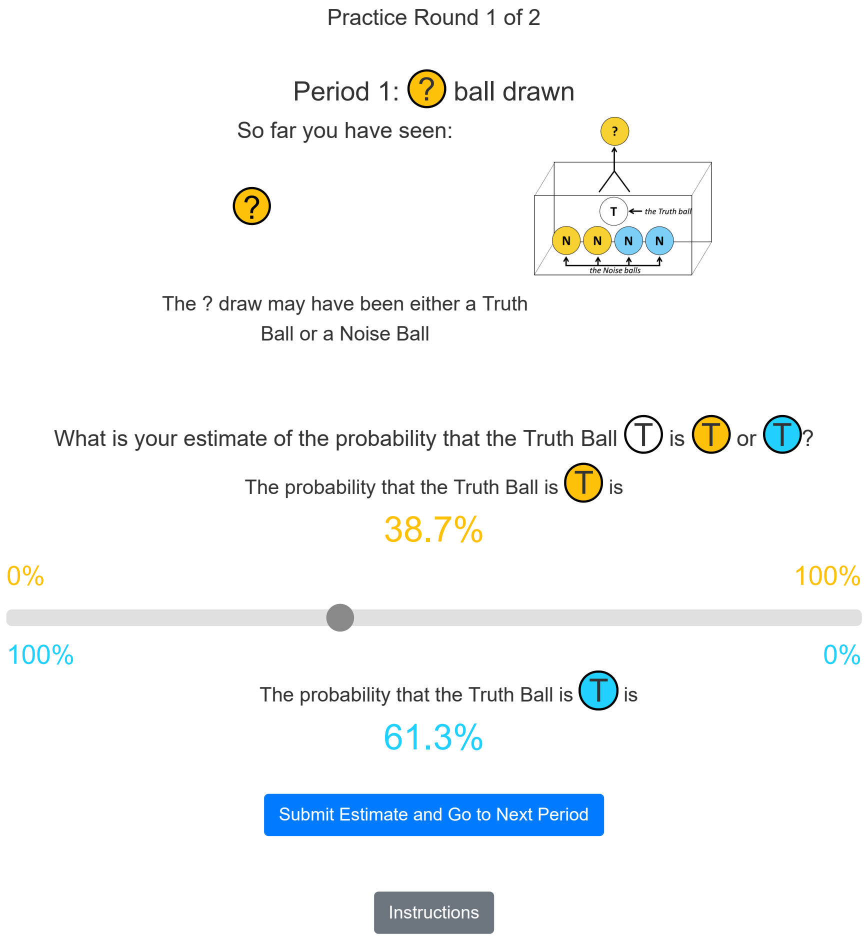
--

0% 100%

100% 0%

The probability that the Truth Ball is T is

--



Practice Round 1 of 2

Period 1: ? ball drawn

So far you have seen:

?

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

What is your estimate of the probability that the Truth Ball T is T or T?

The probability that the Truth Ball is T is

38.7%

0% 100%

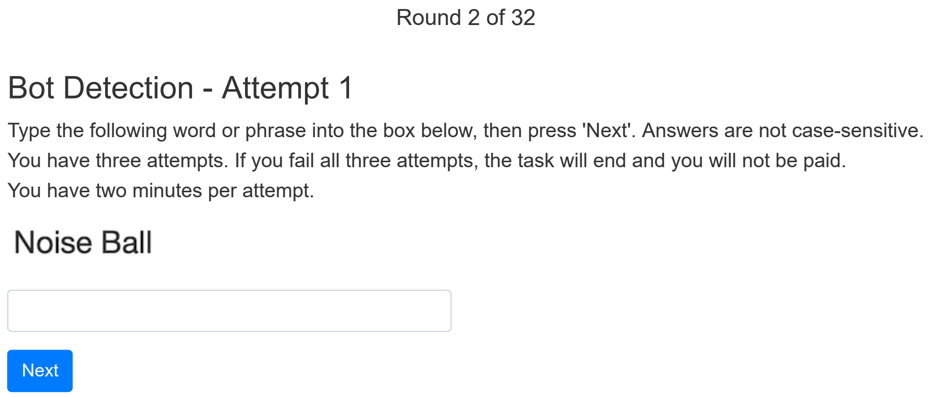
100% 0%

The probability that the Truth Ball is T is

61.3%

C.3. Captchas

Subjects face five different captchas at different rounds. They had 3 tries and one minute to submit for each try. Were they to fail the 3 tries, the task ended and they would not receive any bonus.



Round 2 of 32

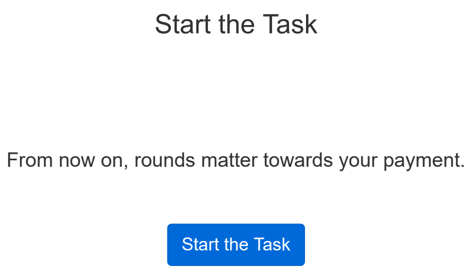
Bot Detection - Attempt 1

Type the following word or phrase into the box below, then press 'Next'. Answers are not case-sensitive.
You have three attempts. If you fail all three attempts, the task will end and you will not be paid.
You have two minutes per attempt.

Noise Ball

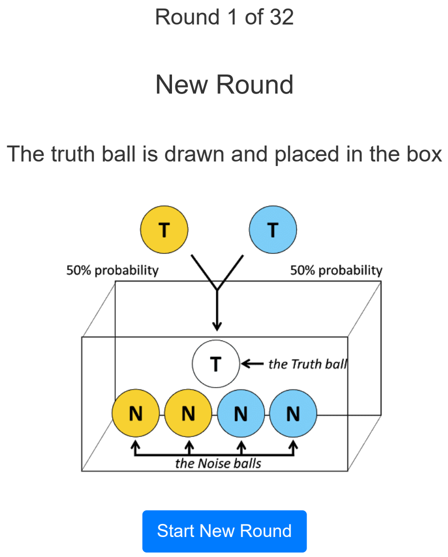
C.4. Rounds

The rounds were described in Section 4.



Start the Task

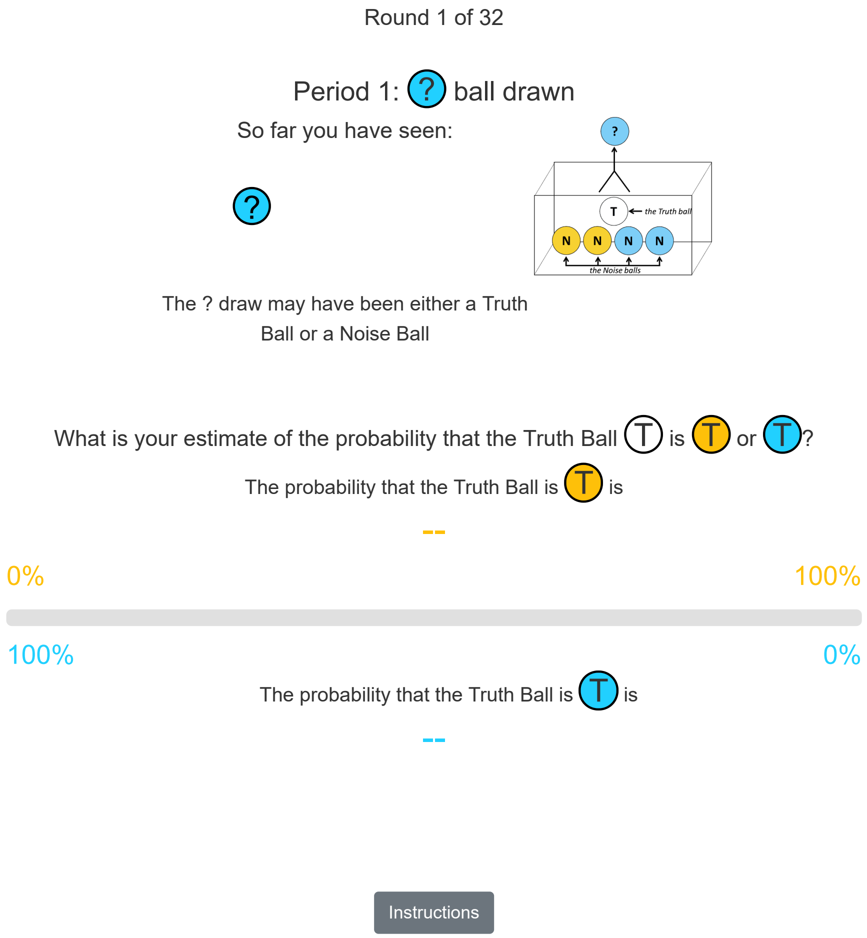
From now on, rounds matter towards your payment.



Round 1 of 32

New Round

The truth ball is drawn and placed in the box



Round 1 of 32

Period 1: ? ball drawn

So far you have seen:

?

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

What is your estimate of the probability that the Truth Ball T is T or T?

The probability that the Truth Ball is T is

--

0% 100%

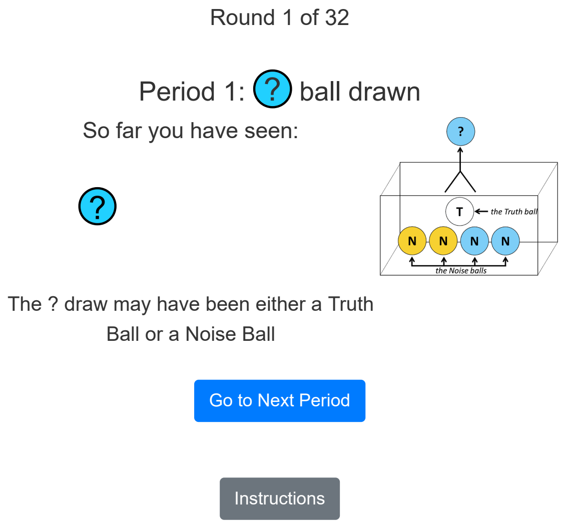
100% 0%

The probability that the Truth Ball is T is

--

C.5. Final Period Elicitation Only

Were the subjects to be in the treatment arm in which beliefs were elicited only at the last period of each round, the last period would be just as before. In periods in which there was no belief elicitation, they would observe just the ball draw:



Round 1 of 32

Period 1: ? ball drawn

So far you have seen:

?

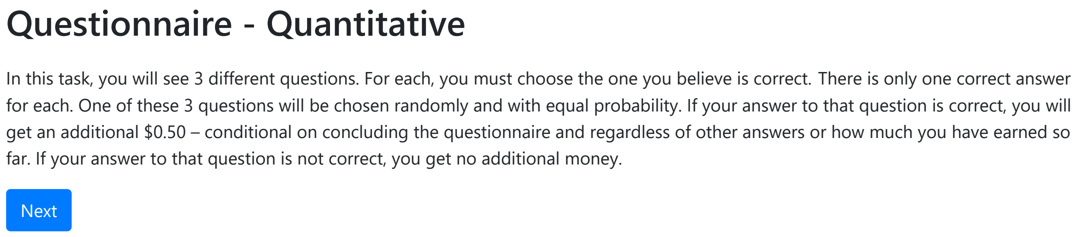
The ? draw may have been either a Truth Ball or a Noise Ball

Go to Next Period

Instructions

C.6. Quantitative Questions

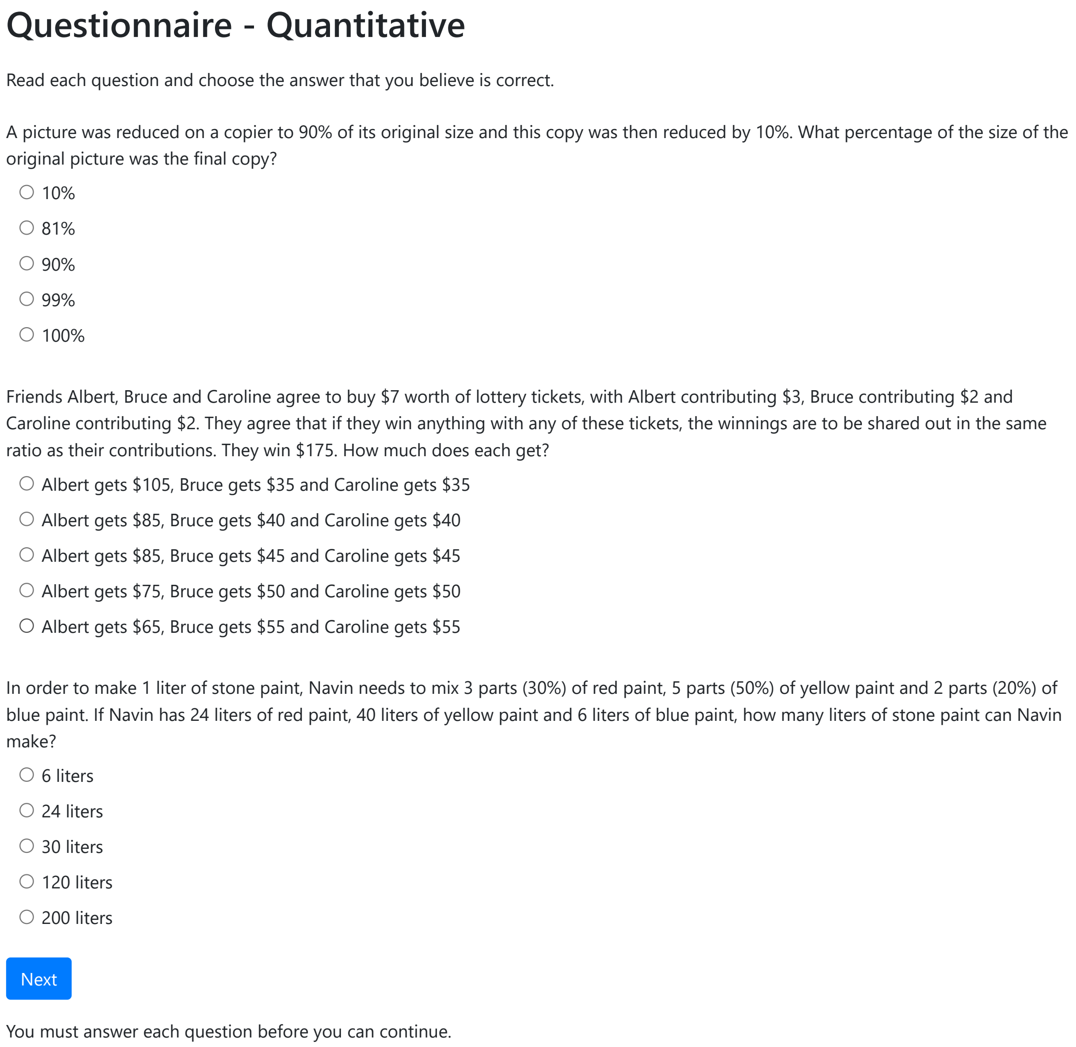
After the main task, the subjects had to answer three questions meant to assess their quantitative ability; these were incentivized.



Questionnaire - Quantitative

In this task, you will see 3 different questions. For each, you must choose the one you believe is correct. There is only one correct answer for each. One of these 3 questions will be chosen randomly and with equal probability. If your answer to that question is correct, you will get an additional $0.50 – conditional on concluding the questionnaire and regardless of other answers or how much you have earned so far. If your answer to that question is not correct, you get no additional money.

Next



Questionnaire - Quantitative

Read each question and choose the answer that you believe is correct.

A picture was reduced on a copier to 90% of its original size and this copy was then reduced by 10%. What percentage of the size of the original picture was the final copy?

○ 10%

○ 81%

○ 90%

○ 99%

○ 100%

Friends Albert, Bruce and Caroline agree to buy $7 worth of lottery tickets, with Albert contributing $3, Bruce contributing $2 and Caroline contributing $2. They agree that if they win anything with any of these tickets, the winnings are to be shared out in the same ratio as their contributions. They win $175. How much does each get?

○ Albert gets $105, Bruce gets $35 and Caroline gets $35

○ Albert gets $85, Bruce gets $40 and Caroline gets $40

○ Albert gets $85, Bruce gets $45 and Caroline gets $45

○ Albert gets $75, Bruce gets $50 and Caroline gets $50

○ Albert gets $65, Bruce gets $55 and Caroline gets $55

In order to make 1 liter of stone paint, Navin needs to mix 3 parts (30%) of red paint, 5 parts (50%) of yellow paint and 2 parts (20%) of blue paint. If Navin has 24 liters of red paint, 40 liters of yellow paint and 6 liters of blue paint, how many liters of stone paint can Navin make?

○ 6 liters

○ 24 liters

○ 30 liters

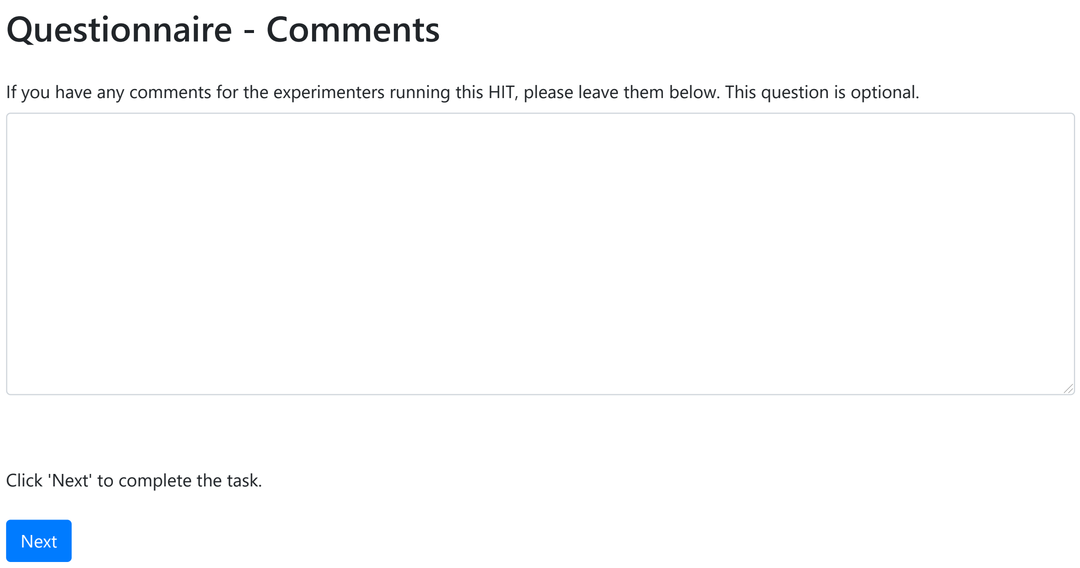
○ 120 liters

○ 200 liters

You must answer each question before you can continue.

C.7. Debrief and Payments

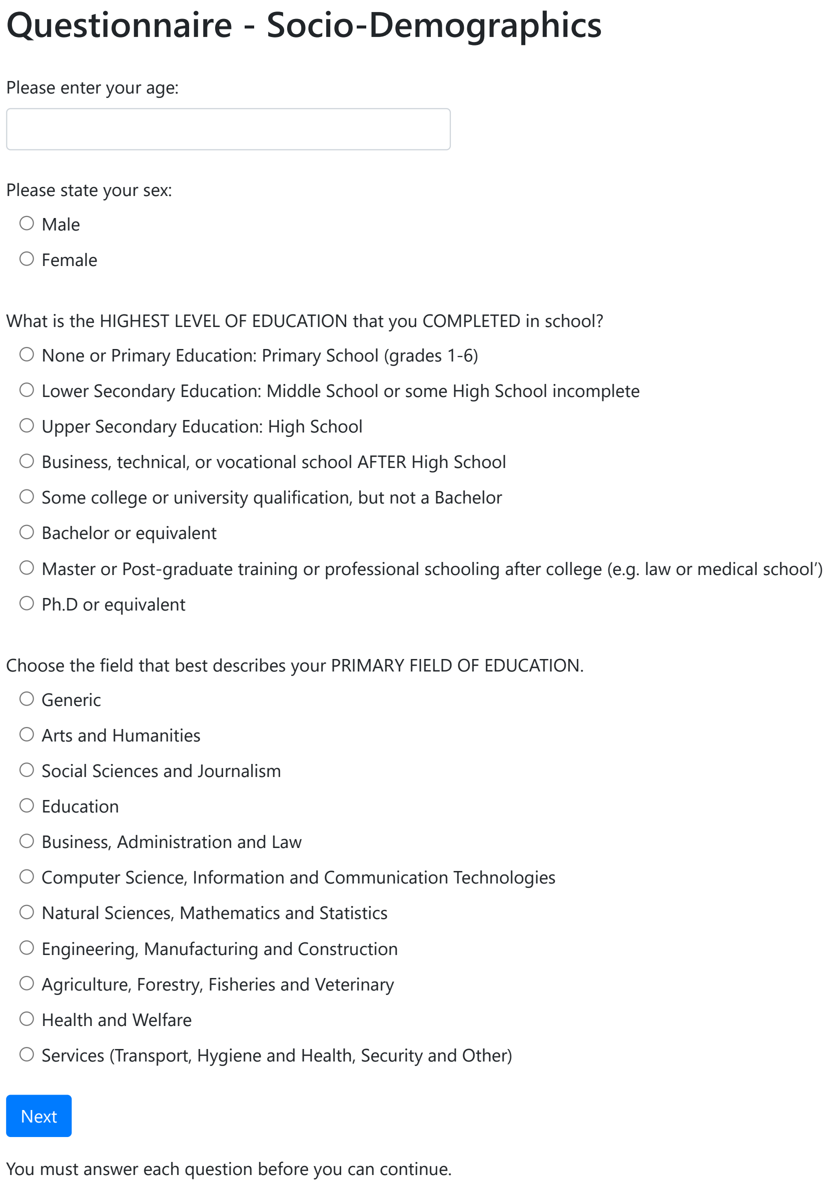
Following the task, we gathered subjects comments, socio-demographic information, and informed them of the payment they would receive.



Questionnaire - Comments

If you have any comments for the experimenters running this HIT, please leave them below. This question is optional.

Click 'Next' to complete the task.



Questionnaire - Socio-Demographics

Please enter your age:

Please state your sex:

- ○ Male
- ○ Female

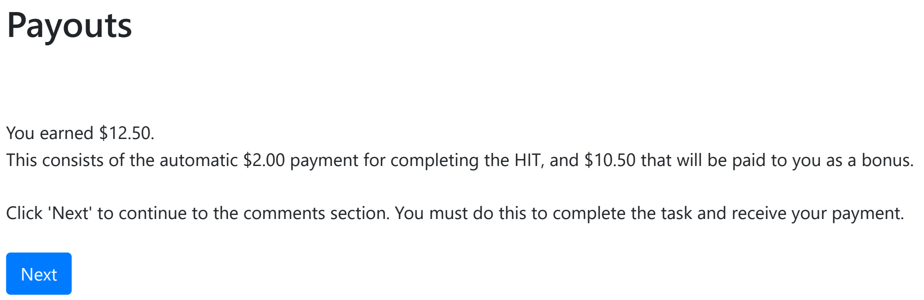
What is the HIGHEST LEVEL OF EDUCATION that you COMPLETED in school?

- ○ None or Primary Education: Primary School (grades 1-6)
- ○ Lower Secondary Education: Middle School or some High School incomplete
- ○ Upper Secondary Education: High School
- ○ Business, technical, or vocational school AFTER High School
- ○ Some college or university qualification, but not a Bachelor
- ○ Bachelor or equivalent
- ○ Master or Post-graduate training or professional schooling after college (e.g. law or medical school')
- ○ Ph.D or equivalent

Choose the field that best describes your PRIMARY FIELD OF EDUCATION.

- ○ Generic
- ○ Arts and Humanities
- ○ Social Sciences and Journalism
- ○ Education
- ○ Business, Administration and Law
- ○ Computer Science, Information and Communication Technologies
- ○ Natural Sciences, Mathematics and Statistics
- ○ Engineering, Manufacturing and Construction
- ○ Agriculture, Forestry, Fisheries and Veterinary
- ○ Health and Welfare
- ○ Services (Transport, Hygiene and Health, Security and Other)

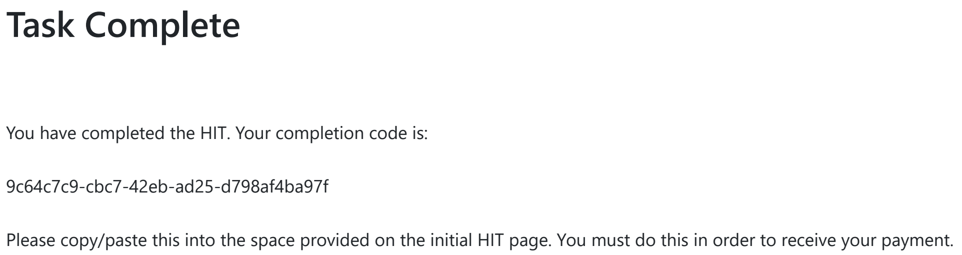
You must answer each question before you can continue.



Payouts

You earned $12.50.
This consists of the automatic $2.00 payment for completing the HIT, and $10.50 that will be paid to you as a bonus.

Click 'Next' to continue to the comments section. You must do this to complete the task and receive your payment.



Task Complete

You have completed the HIT. Your completion code is:

9c64c7c9-cbc7-42eb-ad25-d798af4ba97f

Please copy/paste this into the space provided on the initial HIT page. You must do this in order to receive your payment.